Device-Algorithm Co-Optimization of TiO_x-based Resistive Switching Devices

by

Yu Shi

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2025

© Yu Shi 2025

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:
Fabien Alibart
Associate Professor, Dept. of Electrical Engineering and Computer Science,
Université de Sherbrooke

Supervisor(s):
Guo-Xing Miao
Associate Professor, Dept. of Electrical and Computer Engineering,
University of Waterloo

Manoj Sachdev
Professor, Dept. of Electrical and Computer Engineering,
University of Waterloo

Committee Members:
Lan Wei
Associate Professor, Dept. of Electrical and Computer Engineering,
University of Waterloo

Youngki Yoon
Associate Professor, Dept. of Electrical and Computer Engineering,
University of Waterloo

Adam Wei Tsen
Associate Professor, Dept. of Chemistry,
University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The rapid advancement of artificial intelligence (AI) and growing demand for high-performance computing have exposed key limitations in von Neumann architectures, particularly energy inefficiency and data movement bottlenecks. Resistive Random Access Memory (RRAM) has emerged as a promising candidate for next-generation memory technologies, offering non-volatility, high integration density, and the potential for energy-efficient Compute-In-Memory (CIM) architectures. This thesis focuses on the development and optimization of TiO_x-based RRAM devices, with a particular emphasis on their application in CIM systems.

The work begins with a comprehensive exploration of the device characteristics of TiO_x-based RRAM, including the influence of electrode materials, oxygen stoichiometry, and physical dimensions on device performance. Through interface engineering and material stack optimization, we achieve significant improvements in forming voltage, endurance, and retention, enabling low-power operation and high reliability. The optimized devices exhibit stable bipolar switching behavior with forming voltages below 1.5 V and operation currents under 100 µA, making them suitable for integration with advanced CMOS technologies.

Building on the optimized device performance, we propose a state-aware multi-bit programming algorithm that significantly reduces the number of programming steps and improves the efficiency of multi-bit operations. The algorithm leverages the state-dependent conductance modulation of RRAM devices, enabling precise control over resistance states and mitigating the effects of fast relaxation and retention loss. Additionally, we introduce an electrical annealing method to further enhance the long-term stability of multi-bit programming, demonstrating its effectiveness in extending the refresh period for CIM applications.

To bridge the gap between device-level optimization and system-level implementation, we present a back-end-of-line (BEOL) integration process for TiO_x-based RRAM devices on CMOS chips. The integration process is validated through the successful fabrication and characterization of 1T1R arrays, demonstrating reliable resistive switching behavior and compatibility with existing CMOS technologies. This integration paves the way for the development of RRAM-based CIM macros, which combine RRAM arrays with peripheral circuits for high-performance AI computing.

Finally, we discuss future directions for device optimization, hardware-aware CIM design, and system-level integration. Key challenges include reducing programming current, improving retention stability, and developing reconfigurable CIM architectures for emerging AI workloads. The insights and methodologies developed in this thesis provide a

foundation for the continued advancement of RRAM technologies and their integration into next-generation computing systems.

In summary, this thesis contributes to the field of RRAM-based CIM by addressing critical challenges in device optimization, multi-bit programming, and CMOS integration. The proposed solutions not only enhance the performance and reliability of RRAM devices but also provide a pathway for their practical implementation in energy-efficient AI hardware.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Guo-Xing Miao, for his invaluable guidance and support throughout my PhD journey. His expertise and encouragement have been instrumental in shaping my research and academic development. I am also grateful to my co-supervisor, Professor Manoj Sachdev, for his insightful feedback and mentorship, which have greatly expanded my understanding of the field.

I would like to extend my appreciation to the members of my examining committee for their constructive comments and suggestions that have improved the quality of this thesis.

I would like to thank my lab mates and colleagues for their collaboration and support. I am also grateful to the staffs of the Quantum Nano Fabrication and Characterization Facility (QNFCF) at the University of Waterloo for their assistance with device fabrication and characterization.

I am profoundly grateful to my wife, Tianqi Li, for her endless love, support, and encouragement throughout my PhD journey. Her unwavering belief in me has been a source of strength during challenging times. I also want to thank my parents and family members in China, despite the distance, their love and support have always been with me.

Finally, I would like to thank everyone again who has contributed to my research and personal growth during my time at the University of Waterloo.

Dedication

This is dedicated to my wife, Tianqi Li, whose love and support have been my greatest motivation.

To my parents, who have always believed in me and supported my dreams.

To my family and friends, who have been there for me through ups and downs.

Table of Contents

Examining Committee ii

Author's Declaration iii

Abstract iv

Acknowledgements vi

Dedication vii

List of Figures xi

List of Tables xv

List of Abbreviations xvi

1 Introduction 1

1.1 Background 1

1.2 Device Characteristics of RRAM 3

1.2.1 Forming 4

1.2.2 SET/RESET 8

1.2.3 Endurance and Retention 9

1.2.4 Multi-Level Programming 10

1.3 RRAM based Compute-In-Memory . . . 13
1.4 Challenges and Limitations of RRAM Devices for CIM . . . 15
1.5 Contributions of This Thesis . . . 17

2 Bipolar Resistive Switching of TiO_x based Devices 19
2.1 Device Fabrication and Characterization . . . 20
2.2 Influence of Electrode Material . . . 25
2.3 Influence of Oxygen Ratio . . . 27
2.4 Influence of Physical Dimensions . . . 30
2.5 Retention and Endurance . . . 32
2.6 Summary . . . 34

3 Device Optimization Through Interface Engineering 38
3.1 Scaling of Forming Voltage . . . 40
3.2 Influence of TiN Bottom Electrode . . . 43
3.3 Influence of Al_2O_3 Insertion . . . 46
3.4 Summary . . . 49

4 State-Aware Multi-bit Write Algorithm 52
4.1 Write Direction Dependent Fast Relaxation . . . 53
4.2 State Dependent Conductance Modulation of RESET . . . 55
4.3 Design of State Aware Multi-bit Write Algorithm . . . 56
4.4 Benchmark of Multi-bit Writing . . . 58
4.5 Electrical Annealing to Mitigate Retention Loss . . . 63
4.6 Benchmark of CIM Performance . . . 64
4.7 Summary . . . 68

5 BEOL Integration of TiO_x based RRAM Devices on CMOS Chips **71**

5.1 Circuit Design of the Chip 71

5.2 BEOL Integration Fabrication Process Flow 73

5.3 Packaging and Characterization 75

5.4 Summary 76

6 Conclusions and Future work **77**

6.1 Device Optimization 78

6.2 Hardware Aware CIM Design 80

6.3 System Level Integration 81

References **83**

List of Figures

1.1 (a) SRAM cell size[1, 2] and (L2 + L3) capacity of CPU versus technology node. (b) Comparison of LLM model size and compute capacity of GPU versus Year. 2

1.2 (a) Schematic of the device structure of RRAM devices. I-V characteristics of unipolar (b) and bipolar (c) resistive switching devices[3]. I-V characteristics of filamentary (d)[4] and interfacial (e) switching devices[5]. 5

1.3 (a) Schematic of two-stage forming process[6]. (b) Schematic of SET and RESET process[7]. 7

1.4 Benchmark of retention of RRAM devices at different ambient temperature[8]. 10

1.5 Schematic of fixed pulse time and voltage method (a) and incremental pulse voltage (b) and time (c) method. Flowchart of the write-verify method (d). 12

1.6 (a) Schematic of vector-matrix multiplication using CIM. Circuit diagram of crossbar array (b)[9] and 1T1R array (c)[10]. 14

1.7 Schematic of current mode sensing (a) and voltage mode sensing (b)[11]. . 16

2.1 (a) Crossbar structure of the device. (b) Pillar structure of the device. (c) Via structure of the device. 21

2.2 Fabrication process flow of the pillar structure. 23

2.3 Fabrication process flow of the via structure. 24

2.4 (a) Abnormal forming I-V characteristic of Pt/TiO_x/Pt devices. (b) Switching I-V characteristic of Pt/TiO_x/Pt devices. 26

2.5 Interface analysis of (a) Pt/TiO_x/TiN, (b) TiN/TiO_x/Pt and (c) TiN/TiO_x/TiN devices. 27

2.6 Switching I-V characteristic of (a) Pt/TiO_x/TiN, (b) TiN/TiO_x/Pt and (c) TiN/TiO_x/TiN devices (positive forming voltage). 28

2.7 Switching I-V characteristic of TiN/TiO_x/TiN devices before and after interface engineering. 29

2.8 Influence of thickness and area on the forming voltage, HRS resistance and HRS breakdown voltage. 30

2.9 Transient of RESET process: (a) Competition between RESET oxidation and breakdown. (b) Breakdown of the device in HRS. 31

2.10 Retention of the devices at room temperature. 33

2.11 (a)Switching I-V of the device with different stop voltage of RESET. (b) Retention of the varied resistance states. 34

2.12 RTN behavior of the devices in HRS under different read voltage. 35

2.13 I_t vs I_{t+1} of RTN behavior under different read voltage. 36

2.14 Endurance of the devices under (a) DC sweeping and (b) pulse based measurement. 37

3.1 (a) Schematic of the fabricated TiN/TiO_x/TiN devices. (b) Cross-section TEM image of the device. (c) Fabrication process flow of the devices. (d) Endurance test of the devices. (e) Forming I-V (green) and switching I-V (red) of the formed devices. (f)Retention test of LRS and HRS. 39

3.2 (a) Forming and breakdown I-V of fabrication devices. Electrical characterization of devices with varying oxide thicknesses: Dependence of LRS and HRS resistance on forming voltage across different thicknesses (2-4 nm). (c) Cumulative distribution function (CDF) of forming voltages. (d) Two-dimensional boxplot analysis correlating HRS resistance with forming voltage. 41

3.3 (a) Nucleation step of forming. (b) Growth step of forming. Defect profile of High HRS (c) and Low HRS (d) devices. 42

3.4 X-ray diffraction patterns of TiN films deposited with varying N_2 flow rates. 43

3.5 (a) I-V of fresh devices with varied TiN bottom electrode in log scale. (b) Forming voltage with varied TiN bottom electrode. (c) Comparison of I-V of fresh device between $Q_{N_2} = 1.5sccm$ and $Q_{N_2} = 5sccm$ in linear scale. . 44

3.6 Electrical characteristics of devices with 1.5 sccm N_2 TiN bottom electrodes: (a) Resistance of LRS and HRS versus forming voltage . (b) Two-dimensional boxplot analysis correlating HRS resistance with forming voltage. (c) Cumulative distribution function (CDF) of forming voltages. (d) Comparison of resistance of HRS versus forming voltage for varied interface. 45

3.7 (a) Forming voltage and breakdown voltage of device with varied thickness of switching layer (averaged over 50 devices). (b) Forming voltage and breakdown voltage of varied interface. (averaged over 50 devices) 47

3.8 (a) Voltage drop diagram of devices under forming bias. (b) Defect profile of devices and possible hazards during material growth and fabrication. . . 48

4.1 (a) Schematic of pulse train used for characterization of fast relaxation. (b) Average conductance change after programming with varied G_{Target}. Top figure is programming terminated with SET while the bottom one is terminated with RESET. 54

4.2 3D mapping of average conductance after RESET versus pulse amplitude and pulse time, with varying initial conductance states. 56

4.3 (a) Conductance change (ΔG) versus pulse time, initial conductance state before RESET is 30 μS. (c) The conductance change ratio versus the pulse amplitude. (d) Conductance change versus the initial conductance with the inset figure being the fitting slope versus the voltage amplitude (average over 30 devices). 57

4.4 (a) Flowchart of the proposed multi-bit writing algorithm. (b) Flowchart of status flag update module. (c) Flowchart of RESET voltage and time calculation module. (d) Example of programming process. The bar height of SET/RESET (RESET time) represents the voltage amplitude (pulse time) applied. The triggering of monotone and oscillation status flags are illustrated in the conductance change plot. 58

4.5 (a) After programming state distribution of ISPP versus proposed algorithm. (b) Retention test of 16 conductance states programmed using the proposed algorithm. 60

4.6 Benchmark of number of programming steps and total RESET time between ISPP, proposed algorithm and quantized proposed algorithm 61

4.7 (a) Fast relaxation results and (b) retention results of conductance states after programming. 62

4.8 Customized voltage waveform for electrical annealing. 63
4.9 4-bit Retention results (a) without and (b) with electrical annealing. 65
4.10 (a) Schematic of the CIM simulation setup. (b) Schematic of the weight mapping. 67
4.11 CIM simulation results of classification accuracy versus retention time. . . 68
4.12 CIM simulation results of classification accuracy with ADC precision bits being (a)5 and (b)8. 70

5.1 (a) Schematic of Integrated RRAM device on CMOS chip (b) Layout of the 1T1R array. 72
5.2 (a) Layout of the two-finger transistor and the position of the RRAM device; (b) Layout of the whole chip; (c) Layout of the 1T1R array. 73
5.3 BEOL integration process flow of TiO_x based RRAM devices on CMOS chips. 74
5.4 (a) Wire bonding of the chip; (b) PCB board for the chip; (c) Microscope image of the probed chip. 75
5.5 I-V characteristics of the integrated RRAM devices. (a) Forming process of the device; (b) RESET process of the device; (c) SET process of the device. 76

List of Tables

1.1 Comparison of Memory Technologies[12] 3

2.1 Growth methods and parameters of TiO_x, TiN and Pt used in this thesis. . 20
2.2 Forming voltage of devices with varied electrode combinations. 27
2.3 Effects of oxygen ratio on the forming voltage and HRS resistance (The thickness of TiO_x is 5 nm). 28
2.4 Comparison of the forming voltage and HRS resistance of devices with different heterostructure. 29

3.1 Effects of N_2 plasma treatment on top interface. 49
3.2 Effect of N_2 plasma treatment on bottom electrode. 49
3.3 Comparison of optimized RRAM devices with other works. 51

4.1 Shared parameters between ISPP and proposed algorithm. 59
4.2 VGG8 Architecture Summary 69

List of Abbreviations

BEOL Back-End-Of-Line 17, 18, 74, 76, 77, 81

CBRAM Conductive Bridge Random Access Memory 4, 8

CIM Compute-In-Memory xi, 3, 13–18, 77, 78, 80–82

CPU Central Processing Unit xi, 1–3

DRAM Dynamic Random Access Memory 1, 3, 13, 16

FeRAM Ferroelectric Random Access Memory 1, 3, 13, 16

GPU Graphics Processing Unit xi, 2, 3

HBM High Bandwidth Memory 3

HRS High Resistance State 3, 8, 17, 18, 25, 28–34, 39–42, 46, 48–50, 78–80

LLM Large Language Model xi, 2, 13, 81

LRS Low Resistance State 3, 6, 8, 25, 32, 33, 39–41, 46, 49, 50, 78, 80

MRAM Magnetic Random Access Memory 1, 3

MVM Matrix-Vector Multiplication 2, 3, 13, 15, 17

PCM Phase Change Memory 1, 13, 16

RRAM Resistive Random Access Memory xi, 1, 3–6, 8–18, 28, 34, 38, 50, 52, 55, 65–69, 71–73

SRAM Static Random Access Memory xi, 1–3, 13, 16

TPU Tensor Processing Unit 3

Chapter 1

Introduction

1.1 Background

The slowdown of Moore's Law has become increasingly evident, as the area density of static random access memory (SRAM)—used as cache memory in central processing units (CPUs)—has failed to scale with advancing technology nodes. This trend indirectly signals the end of the FinFET era. Despite innovations such as gate-all-around transistors and backside power delivery, SRAM density has not improved as anticipated. Modern CPUs rely heavily on the a hierarchical cache structure to reduce the latency of data access, including the L1, L2 and L3 cache, the area of which occupies more than 50% of the total area of the CPU. The large area of the cache memory not only increases the cost of the CPU but also limits the scalability of the CPU, which is not desirable for the development of the high-performance CPUs. Emerging applications such as cloud computing and artificial intelligence demand increasingly aggressive on-chip cache integration—a trend that is unsustainable under current technology projections. The slowdown in memory density scaling affects not only SRAM but also dynamic random access memory (DRAM) and NAND Flash, which serve as main memory and storage, respectively, in modern computer systems.

To extend the scaling of memory technologies, emerging memory technologies including magnetic random access memory (MRAM)[13], resistive random access memory (RRAM)[3], phase change memory (PCM)[14] and ferroelectric random access memory (FeRAM)[15] have been proposed. All the emerging memory technologies are based on two-terminal structure, while the data are stored in the form of magnetization, resistance, phase or polarization state of the devices. The simple device structure of the emerging memory

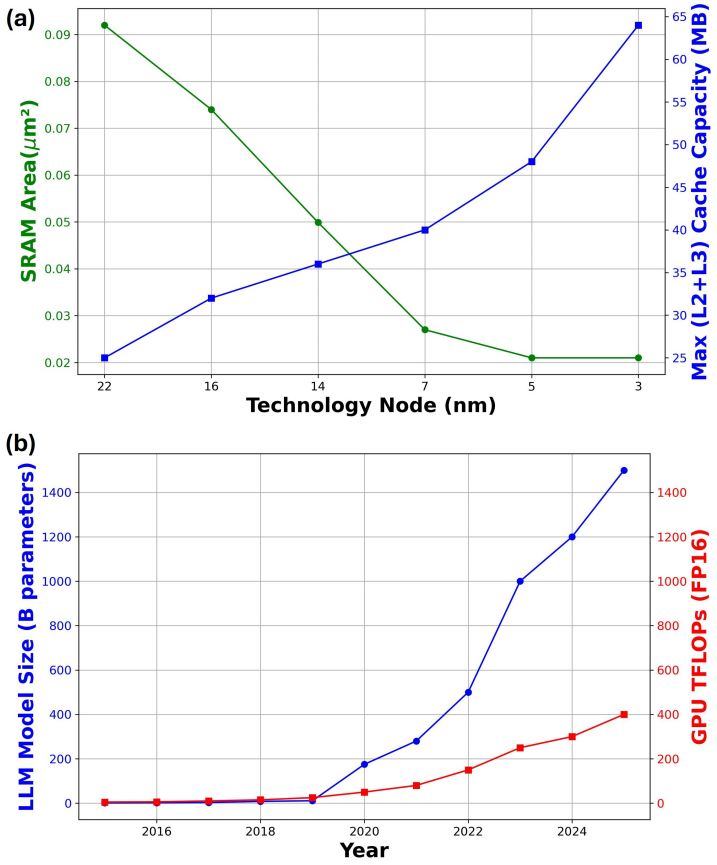


Figure 1.1: (a) SRAM cell size[1, 2] and (L2 + L3) capacity of CPU versus technology node. (b) Comparison of LLM model size and compute capacity of GPU versus Year.

technologies enables the high integration density of the devices. Moreover, the non-volatile nature of the emerging memory technologies fills the gap between the SRAM and the NAND Flash, which is critical for the development of the memory hierarchy of the computer system.

Apart from the scaling issue of memory technologies, another challenge of the modern computer system is the energy efficiency of the system. With the development of artificial intelligence including deep learning and large language model (LLM), the demand of computation has transitioned from the high precision scalar computation to the low precision high dimensional tensor computation, with the matrix-vector multiplication (MVM) as the core operation. The von Neumann architecture, which separates the memory and the com-

Technology	Cell Size (F^2)	Density (Gb/mm^2)	Endurance	Retention
SRAM	120	~0.1-0.5	unlimited	Volatile
DRAM	6-8	~1-2	$\sim 10^{16}$	Volatile
NAND Flash	4.5-5.5	~1-3	$\sim 10^3$ - 10^5	>10 years
STT-MRAM	20-50	~0.5-1	$>10^{12}$	>10 years
RRAM	4-10	~1-2	$>10^6$	>10 years
PcRAM	5-10	~1-2	$>10^6$	>10 years
FeRAM	6-8	~0.5-1	$>10^{10}$	>10 years

Table 1.1: Comparison of Memory Technologies[12]

puting unit, has been found to be inefficient for the large-scale MVM computation, which requires frequent data movement between the memory and the computing unit. The data shuffling between the memory and the computing unit increases the computation delay as well as the energy consumption, which is not desirable for the energy efficiency of the system. Architecture innovations including systolic array, near-memory computing and compute-in-memory (CIM) have been proposed to solve this problem. The systolic array (Google's TPUs)[16, 17] and the near-memory computing (HBM used in GPUs) bring the memory unit (SRAM or DRAM) physically closer to the computing unit (CPU or GPU), which mainly reduces the memory access time and increases the computation throughput. The CIM architecture integrates the memory and the computing unit together, and perform the computation inside the memory array, which eliminates the data movement between the memory and the computing unit, and reduces the energy consumption of the system.

1.2 Device Characteristics of RRAM

Resistive switching random access memory (RRAM) is a non-volatile memory technology that stores data using the resistance state of the devices. The resistance state of RRAM devices can be switched between high resistance state (HRS) and low resistance state (LRS) by applying a voltage bias to the device, which is normally referred to as SET and RESET operation[3]. Based on the polarity of the switching voltage bias, RRAM devices can be classified into bipolar and unipolar resistive switching devices. Bipolar resistive switching devices require a voltage bias with opposite polarity to switch the device back to HRS, while unipolar resistive switching devices require a voltage bias with the same polarity to switch the device back to HRS.

The device structure of RRAM devices is a simple MIM (metal-insulator-metal) stack, which includes a top electrode, a switching layer and a bottom electrode. The switching layer is normally a dielectric layer, which includes transition metal oxide (HfO_2[4], Ta_2O_5[18], TiO_2[19]), transition metal nitride (SiN[20]), 2D layered dielectric(hBN[21], MoS_2) or organic materials. The choice of the switching layer is critical for the performance of RRAM devices, which includes the forming voltage, the resistance state stability, the endurance and the retention of the devices. The electrodes of RRAM devices are normally made of noble metals (Pt, Au)[22] or active metals (Ag, Ti, Hf, Ta)[23, 24], the choices of the electrodes determines the switching type of the RRAM devices (filamentary or interface) as well as the switching performance of the devices. The filamentary switching devices normally have higher ON/OFF ratio and lower forming voltage compared to the interface switching devices, but it also has higher operation current and larger device-to-device variation. The conducting filament of the filamentary switching devices is composed of oxygen vacancy (transition metal based RRAM) or metal atoms (such as Ag, Cu, etc) in conductive bridge random access memory (CBRAM) devices, which is formed during the forming step and ruptured during the RESET step. The interface switching devices are normally based on the interface between the electrode and the switching layer, the defect concentration of the interface determines the resistance state of the device[25, 26].

Compared to other types of switching layer, transition oxide based RRAM stands out due to its process compatibility with CMOS technology, which is critical for the integration of RRAM devices with CMOS chips. Instead of using the conventional metal layer, metal nitrides (TiN, TaN)[27, 28] are normally used as the electrode of transition oxide based RRAM devices not only because of its process compatibility but the tunability of the electrode properties, which includes the work function and thermal diffusion barrier. Transition oxide based RRAM is chosen as the main focus of this thesis, the details of the device characteristics of transition oxide based RRAM devices will be discussed in the following sections.

1.2.1 Forming

Transition oxide based RRAM devices requires a forming step to form conducting filaments inside the switching layer, the formation and rupture of conducting filaments controls the resistance state of the device. During forming, a voltage bias is applied to the device, oxygen vacancies (or metal atoms in CBRAM) would migrate along the electric field direction and form a conduction channel between the electrodes. The amplitude of the forming voltage is around $1 \sim 3$ V, which is normally higher than the operation voltage of the device. The amplitude of forming voltage is determined by the physical dimensions of the

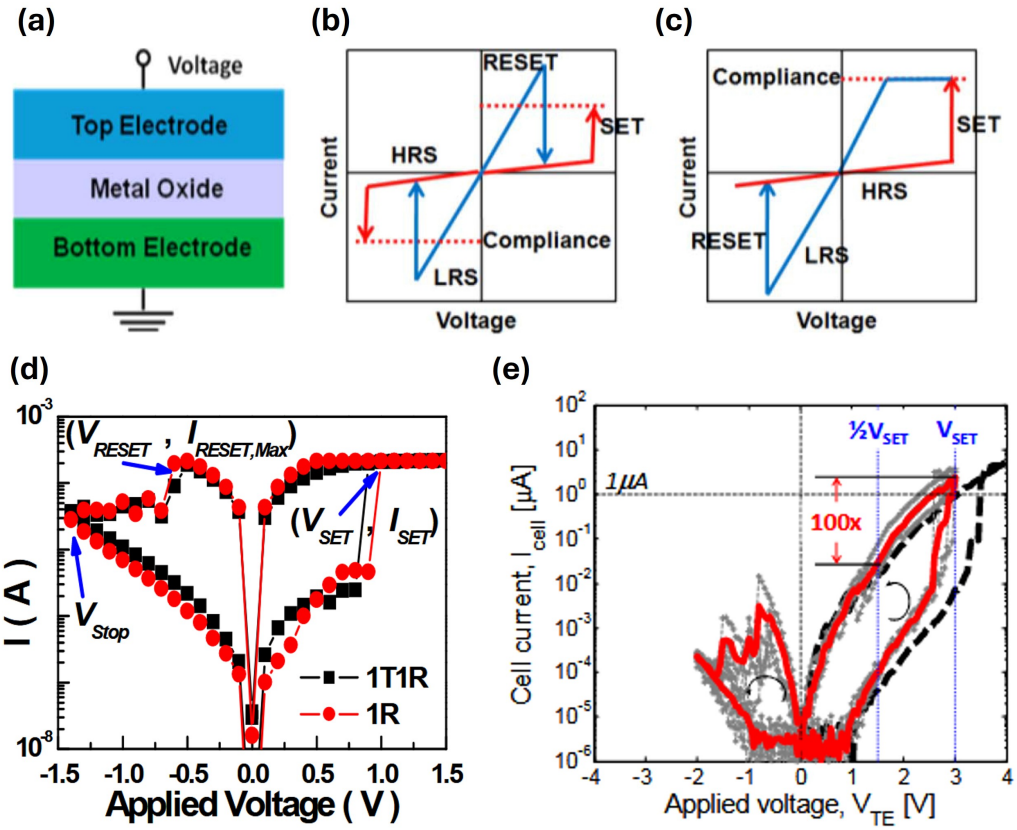


Figure 1.2: (a) Schematic of the device structure of RRAM devices. I-V characteristics of unipolar (b) and bipolar (c) resistive switching devices[3]. I-V characteristics of filamentary (d)[4] and interfacial (e) switching devices[5].

device (thickness of the switching layer, device area, material stack structure, etc)[29, 30], the properties of the switching layer (oxygen ratio, defect density, etc) as well as the electrodes.

Due to the dielectric breakdown nature of the forming step, the forming voltage is anti-correlated to the device area and correlated to the thickness of the switching layer[31, 32, 33]. Scaling down the thickness of the switching layer would lower down the forming voltage, but it also leads to the increase of the operation current, which is not desirable for the energy efficiency. To integrate with the CMOS technology, RRAM devices are normally placed between metal layers (between M-1 layer and M layer), requiring the device area to be scaled with the scaling of the CMOS technology. However, smaller device area requires

higher forming voltage, making the integration of RRAM devices with CMOS technology simply based on physical dimensions challenging[34].

The scaling of the forming voltage can also be achieved by introducing more defects to the switching layer, which includes lowering down the oxygen ratio of transition oxide dielectric, using active metal layer (Ti, Hf. etc)[35] as one of the electrodes or plasma treatment using Argon or Hydrogen gas[15, 36]. Further researches reveal that asymmetric defect profile (top active metal layer, oxide heterostructure) of the switching layer has better device performance (lower forming voltage, low power operation) compared to homogeneous defect distribution (tuning oxygen ratio of the oxide layer)[37, 38]. The work function of the electrodes also plays an important role in the forming voltage due to the Schottky barrier formed between the electrode and the switching layer, earlier researches prefer high-work function metal like Pt or Au[19] while later researches prefer TiN, TaN or W due to its lower forming voltage required and better CMOS process compatibility. It is also observed that the forming voltage is dependent on the temperature, higher temperature would lower down the forming voltage[39], which is attributed to the thermally activated migration of oxygen vacancies.

By varying the compliance current, atmosphere temperature as well as the forming voltage amplitude. It is proposed that the forming is a two-step convoluted process[40], which includes the filament nucleation process and growth process. At low compliance current (< 0.1 μA), the oxygen vacancy propagate through the grain boundaries of the oxide layer, forming a conduction channel with high resistance LRS but with higher device-to-device variation due to stochastic nature of this process; at high compliance current (> 10 μA), the formed conduction channel is further expanded by the oxygen vacancy, forming a conduction channel with low resistance LRS and less variability. This model aligns with observation of current overshoot during forming, where larger forming voltage leads to permanent breakdown of the device[29]: the moment the nucleation process is triggered, the device resistance drops rapidly, leading to a large current overshoot, which expands the conduction channel and triggers the hard breakdown of the device (hard to RESET).

Transient analysis of the forming process reveals that forming, as a dielectric breakdown process, is a stochastic process, the average forming time needed is dependent on the amplitude of the forming voltage. The stochastic nature of forming adds up the complexity of write circuit design, as the forming time is not deterministic, the write circuit should be able to detect the forming completion and tune the forming parameters accordingly to shorten the forming time. Different strategies have been proposed, which includes incremental pulse time, incremental voltage, two-step forming (page forming)[41] and self-terminated forming. The incremental pulse time and incremental voltage method[42, 43]

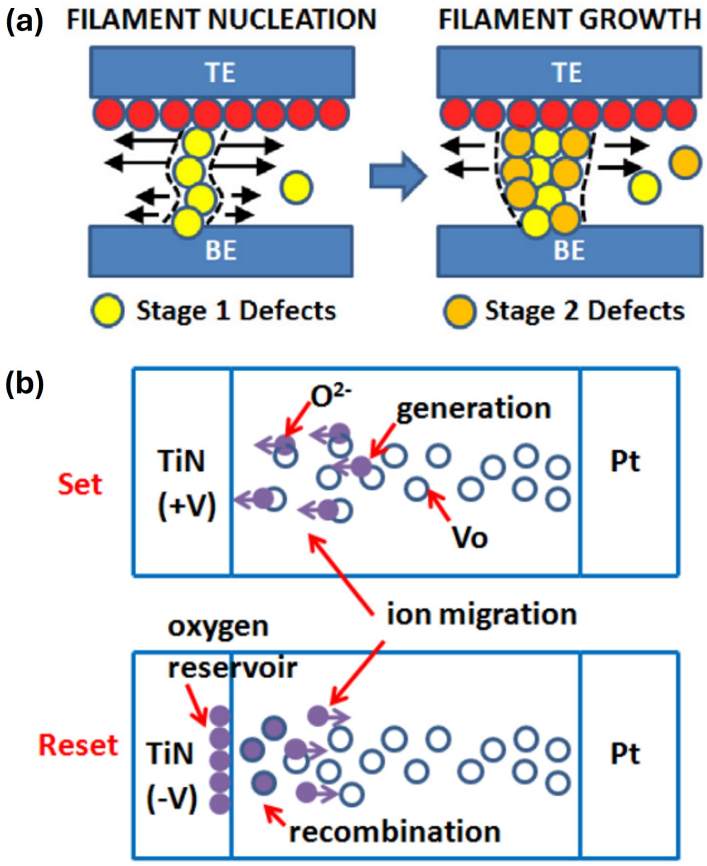


Figure 1.3: (a) Schematic of two-stage forming process[6]. (b) Schematic of SET and RESET process[7].

depends on the verify step to detect the forming completion, and increase the pulse time or voltage amplitude accordingly. Both methods are energy inefficient and time consuming, aggressive parameter choices (larger step time or voltage) would lead to over-forming (hard-breakdown) and device failure, which is not desirable for the device reliability. The two-step forming method[41] requires a pre-forming step (high voltage, low compliance current and longer bias time) to form a partial conduction channel, longer bias time helps to alleviate the device-to-device variation, and a second forming step (low voltage, high compliance current and short bias time) to form the rest of the conduction channel. The self-terminated forming[44] methods uses negative feedback (current) to flag the forming completion, which is more energy efficient and faster compared to the verify method, but it requires a more

complex circuit design.

1.2.2 SET/RESET

After the forming step, the RRAM devices are in low resistance state (LRS), a RESET operation is applied to switch the device back to high resistance state (HRS). For most transition metal oxide based RRAM devices, which has bipolar resistive switching behavior, a bias that has opposite polarity versus forming bias is applied to RESET the devices to HRS. During RESET, the oxygen vacancies (or metal atoms in CBRAM) would migrate back, breaking the conduction channel (filamentary switching) or re-oxidize the switching layer (interface switching)[45]. The amplitude of the RESET voltage is around 1 $\sim$ 2 V, which is normally lower than the forming voltage. No compliance current is needed during the RESET process, but for RESET applied using voltage sweeping, a maximum stop voltage needs to be applied to avoid HRS breakdown (negative forming)[46, 47, 48, 49], which would lead to conduction filament formed from another electrode and permanently degrade the resistance state of the device. The amplitude of the RESET voltage is found to be independent on the physical dimensions of the devices as the diameter of the conducting filament ($\sim$ 10 nm)[50] is much smaller compared to the device size (100 $\sim$ 1000 nm), the amplitude of the RESET voltage is mostly determined by the properties of the switching layer (oxygen ratio, defect density, etc) as well as the electrodes. The RESET process is generally proposed as a thermal dominated process[51], joule heating generated by the current flowing through the conduction channel would promote the oxygen ion diffusion, the mobility of which is strongly dependent on the temperature[52], and the oxygen ion would recombine with the oxygen vacancies, leading to the rupture of the conducting filaments. By varying the stop voltage of RESET as well as the biasing time[53], it is observed that the devices can be partially RESET, showing the multi-bit capability of the RRAM devices, the details of multi-bit operation will be discussed in the following sections.

The SET operation is similar to the forming process, a bias with the same polarity as the forming bias is applied to the device to switch the device from HRS to LRS. During the SET process, oxygen vacancy will migrate along the initial conduction channel direction until the conducting filament is reconstructed[54]. Due to the corona discharge effect of the conducting filament tip, the voltage amplitude of SET is normally smaller than that of forming, but the compliance current is still needed to avoid hard breakdown of the device. The amplitude of the SET voltage is also found to be independent on the physical dimensions of devices as SET is a localized process (dependent on the conducting filament). Different from RESET, the SET process is a electric field driven process, critical field is needed to trigger the migration of the oxygen vacancy. The devices can be partially SET

by alternating the compliance current applied, which can be easily implemented by varying the gate voltage in 1T1R structure.

1.2.3 Endurance and Retention

The switching speed of SET and RESET show exponential dependence on the amplitude of the bias, the higher the amplitude of the bias, the faster the switching speed[55, 56, 57]. The switching speed of RESET is normally slower than that of SET due to different switching mechanisms behind: thermal induced diffusion VS electric field driven migration. The endurance of RRAM is defined as the number of switching cycles the device can sustain before the resistance state degrades. State-of-the-art RRAM devices normally has endurance of around 10^6 cycles[58, 59, 60], which is comparable to the endurance of phase change memory and NAND Flash. The degradation of endurance of RRAM devices can be classified into three scenarios[61]: 1. Stuck-at-SET; 2. Stuck-at-RESET; 3. Resistance window collapse. Stuck-at-SET is because of the excessive migration of oxygen vacancy or hard breakdown of RRAM devices, stuck-at-SET is not recoverable in most cases, thermal recover of SiN based RRAM is reported in[20]; stuck-at-RESET is mainly due to annihilation of oxygen vacancy, which can be recovered by another forming step, but the endurance after recovery is limited due to the aging of switching dielectric (defects generated). Resistance window collapse is mainly due to anode oxidation induced interface reaction, excessive oxygen ions generated during the switching process would oxidize the anode, leading to the resistance state degradation of the device[61]. The degradation of endurance illustrates the importance of balancing the switching parameters of SET and RESET, which includes the switching voltage, compliance current and biasing time. A well balanced SET/RESET combinations can lead to a long endurance of RRAM devices[62, 63], which is critical for the application of RRAM in memory and computing.

Retention of RRAM devices shows the stability of the resistance state of the device. The retention of RRAM devices is measured by applying a small read voltage and monitor the resistance change, the retention time is defined as the time duration the resistance state degrades to a certain level. The degradation of the retention is mainly due to oxidation of the conducting filament by surrounding oxygen ions, which is a thermally activated process[64]. To shorten the characterization time, retention at elevated temperatures is measured to speed up the aging and be used to extract the activation energy of retention degradation, which can be used to project the retention time. The retention of RRAM devices (Figure. 1.4) can reach 10 year @ 150 °C[65], making it a promising candidate for non-volatile memory applications in consumer electronics and automotive industry. It is also reported that the retention performance of RRAM devices is also dependent on

the amplitude and polarity of the read voltage[66, 67], read voltage induced degradation (read-disturb) is critical for the design of peripheral circuits of RRAM devices. Larger read voltage would lead to faster read speed (larger sensing current) but also faster retention degradation, which is a trade-off between read speed and retention time. Based on this, a sweeping voltage based biasing scheme is proposed to speed up the characterization of read-disturb, which eliminates the need of long retention time measurement[68, 69].

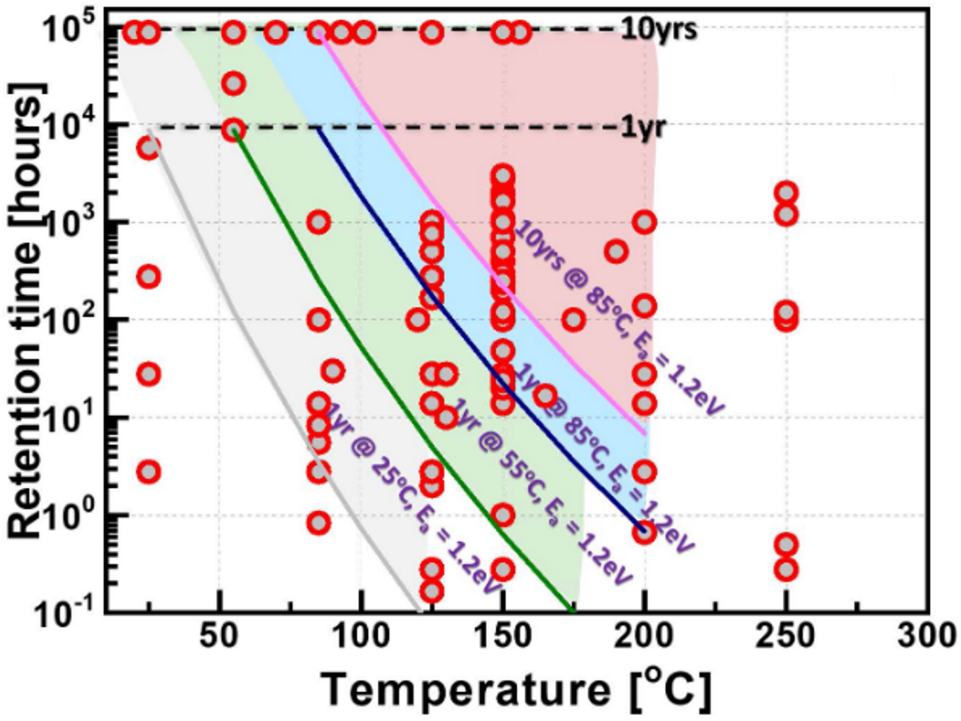


Figure 1.4: Benchmark of retention of RRAM devices at different ambient temperature[8].

1.2.4 Multi-Level Programming

As discussed in the previous section, RRAM devices can be partially SET or RESET by varying the compliance current, voltage amplitude and the biasing time. Due to the stochastic nature of SET/RESET, effective multi-level programming of RRAM devices is challenging, as the resistance state of the device is not deterministic by simply applying SET or RESET operation. To help determine the resistance state of the device, a verify step is needed to detect the resistance state of the device, which is normally done by

applying a small read voltage ($0.1 \sim 0.3$ V) and monitor the resistance difference (ΔR) between the measured one (R_i) and the target value R_{target}: $\Delta R = R_i - R_{target}$. The writing direction of the next programming step is determined by the resistance difference (ΔR), if $\Delta R > 0$, a SET operation is applied, if $\Delta R < 0$, a RESET operation is applied[70, 71]. Different strategies have been proposed to tune the writing parameters of SET/RESET for the next programming step, which can be categorized as: fixed pulse time and voltage and incremental pulse time/voltage.

- The fixed pulse time and voltage method is the most straightforward method, which applies a fixed pulse time and voltage of SET/RESET for each programming step until the target value (with accepted noise margin) is reached. Such method is energy inefficient and time consuming, and the programming efficiency is dependent on the choice of the fixed writing parameters, which is hard to determine with the existence of device-to-device variation and cycle-to-cycle variation. Despite that, researches on fixed pulse time and voltage method leads to an interesting area of RRAM research, which is the analog switching of RRAM devices and the neuromorphic computing (such as spike timing dependent plasticity (STDP)[72, 73]) based on RRAM devices.

- The incremental pulse time and voltage method[70, 74] is proposed to speed up the programming as it will keep increasing the voltage/time if the target value has not been reached. A fixed incremental step of voltage/time is normally applied, which is determined by the device characteristics and the design of the strategies. Voltage is preferred as the tuning parameter to minimize the programming time, but it can lead to degradation of the device (SET induced hard breakdown or RESET breakdown) if the voltage is too high, which is not desirable for the device reliability. To solve this problem, for RRAM devices in 1T1R configuration, the tuning of SET has gate voltage as the tuning parameter while keeping a fixed SET voltage and pulse time; the tuning of RESET has the gate full open and the RESET voltage as the tuning parameter, a maximum stop voltage is set to avoid RESET breakdown.

Apart from the write-verify method discussed above, self-termination writing[75, 76, 77] is also proposed to avoid repeated verify steps. The self-termination writing method uses the negative feedback (sensing current) to detect the termination point of the programming, which is more energy efficient and faster compared to the write-verify method. However, it requires more complex design of circuit and less immune to the device-to-device variation, which is not suitable for large-scale integration.

Another challenge of multi-level programming of RRAM is the state stability after programming, which includes short term relaxation[78] and long term retention[79]. To

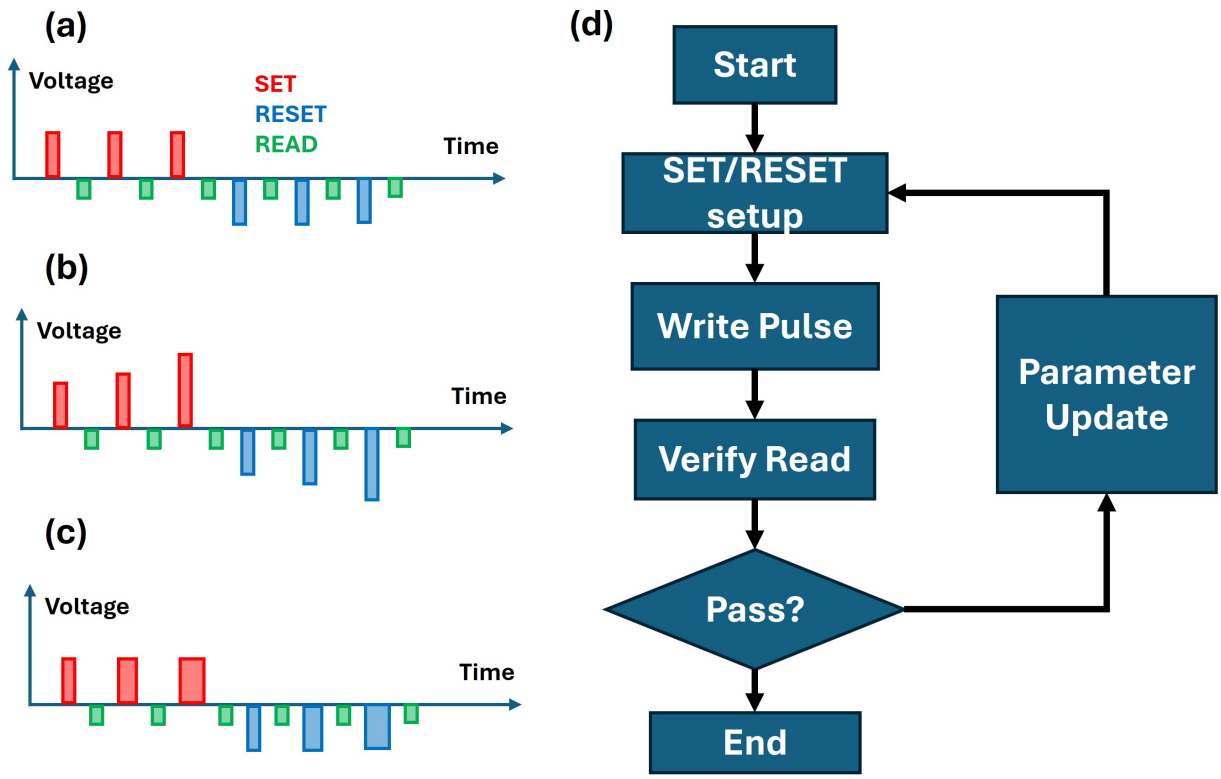


Figure 1.5: Schematic of fixed pulse time and voltage method (a) and incremental pulse voltage (b) and time (c) method. Flowchart of the write-verify method (d).

have more states stored in one single RRAM device, the noise margin between states for multi-bit RRAM devices is more stringent compared to the binary RRAM devices. The short term relaxation is defined as the resistance change of the device after programming in a short time duration (< 1 ms), which is normally ignored in previous characterization of RRAM devices due to loose timing setup of write-verify method. The origin of the short term relaxation is still not clear, published results suggest that the short term relaxation is dependent of the resistance state as well as the writing direction[80, 81], which may suggest that it is correlated with the spontaneous oxygen vacancy diffusion and recombination at room temperature. Thermal capping layer on top of switching dielectric layer have been found to be effective to suppress the short term relaxation[82], different programming strategies have also been reported to mitigate the influence. The long term retention of multi-level RRAM devices shares the same degradation mechanism with the binary RRAM devices, however, the shrunk noise margin makes it hard to achieve 10 year retention time,

especially for devices with bits more than 2. Different methods including the material stack engineering as well as the programming strategies[83, 81, 84, 85] have been proposed to improve the retention time of multi-level RRAM devices.

1.3 RRAM based Compute-In-Memory

Recent rapid development of artificial intelligence including deep learning and large language model (LLM) has called for more efficient computing architecture to handle the large amount of data and computation. The von Neumann architecture, which separates the memory and the computing unit, has been found to be inefficient for the large-scale matrix-vector multiplication (MVM) computation, which is the core operation of deep learning and LLM. The data movement between the memory and the computing unit consumes most of the energy and time, which is not desirable for the energy efficiency of the system. To solve this problem, compute-in-memory (CIM) architecture has been proposed, which integrates the memory and the computing unit together, and perform the computation inside the memory array. CIM based on different memory technologies including SRAM[86], DRAM[87], PCM[88], FeRAM[15] and RRAM[89, 90] have been proposed, among which RRAM based CIM shows its advantages in energy efficiency and scalability due to its non-volatile nature and higher integration density. The working principle of RRAM based CIM is based on Kirchoff's current law, where $I = \Sigma V_i \cdot G_i$, the current flowing through the memory array is the sum of the product of the voltage applied to each cell and the conductance of each cell[9]. By mapping the matrix element to the conductance value of each cell in memory array, the matrix-vector multiplication (MVM) can be performed in parallel by sampling the summing current of each bitline of the memory array, which is the output of the MVM operation. The sampled current vector can be converted to digital signal for further processing.

The initial proposal of RRAM based CIM was based on pure crossbar array structure, however, the leakage current due to partial biasing of un-selected cells leads to the degradation of the sensing margin. Several solutions including "1/2" (or "1/3")sensing scheme[91, 92] and two step sensing[93] have been proposed to mitigate the influence of the leakage current, which is effective for small-scale array ($<$ 1k bit)[94]. To suppress the leakage flowing through the un-selected cells, one way is increase the non-linearity of I-V of RRAM devices, which utilizes passive selector devices (such as diode or threshold switching devices)[95, 96, 97, 98]; another way is to use active switching devices (such as transistor) to block the leakage current sneak path[99]. Due to immaturity of RRAM compatible selector technology, transistor based approach is normally used for large-scale array.

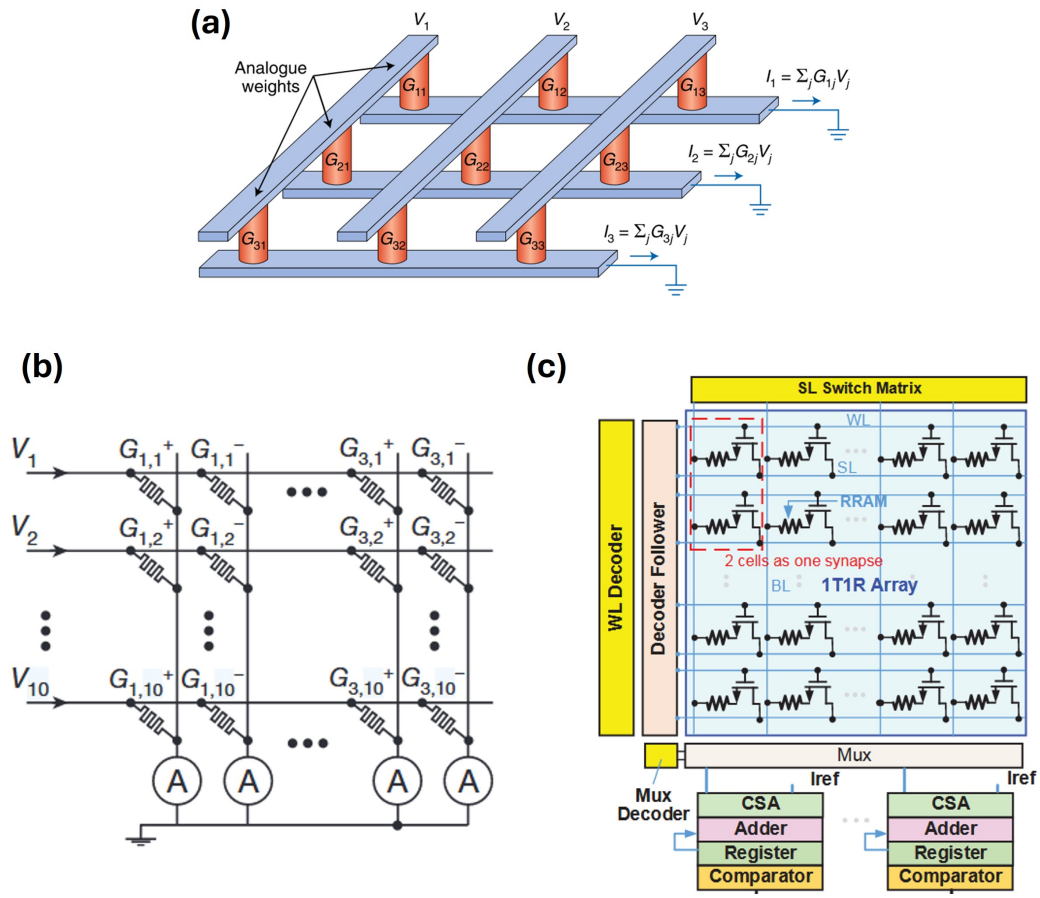


Figure 1.6: (a) Schematic of vector-matrix multiplication using CIM. Circuit diagram of crossbar array (b)[9] and 1T1R array (c)[10].

The transistor based approach is normally implemented in 1T1R configuration, where the transistor is connected in series with the RRAM device, the access of the RRAM device is controlled by the gate voltage of the transistor, which is normally connected to the wordline of the memory array. The adoption of 1T1R configuration also provides other advantages, which includes the tuning of SET/RESET of the RRAM device by the gate voltage, which is critical for multi-level programming of RRAM devices. Variants of 1T1R configuration including 2T1R[100] and 1txR ($x \geq 2$)[101, 102, 103] have been proposed to further improve the computing efficiency and memory density of RRAM based CIM.

The sensing of the summing current of the memory array is critical for the efficiency

and accuracy of the MVM operation, different approaches have been proposed, including the current mode sensing and voltage mode sensing.

- Current mode sensing is the most straightforward method, which converts the summing current to voltage signal using current sense amplifier[75]. However, with the increasing size of the memory array, especially for multi-bit RRAM devices, the current mode sensing is not efficient due to the large current flowing through the bitline as well as the longer sampling time needed, which increases the energy consumption of the system and leads to non-ideal effects such as IR-drop[104]. Apart from that, to accommodate the large sensing current, the area of the sensing circuit also needs to be increased, leading to multi bitlines sharing one sensing circuit. The shared sensing circuit needs to be time multiplexed to sample the current of each bitline, which limits the parallelism of the MVM operation.
- Compared to current mode sensing, voltage mode sensing normally utilizes a sampling capacitor to convert the summing current to voltage signal ($V = \frac{\int Idt}{C}$)[105, 11]. The voltage stored in the sampling capacitor is further sampled by ADCs and pass to next stage for further processing. The voltage mode sensing only require fixed time sampling of the summing current, which is more energy efficient compared to the current mode sensing. Moreover, more bitlines can be switched on in parallel, which increases the computational parallelism.

Both current mode and voltage mode requires ADCs for analog/digital conversion, which are area and power hungry in real implementation. ADC-free approach is proposed to eliminate the need of ADCs[106, 107], which samples the summing current using the sample capacitor and compare the sampled the voltage to a ramp voltage to generated the encoded PMW signal, which is passed to next RRAM CIM array for next stage computation. The ADC-free approach is more energy efficient compared to the ADC based approach, but it requires more complex design of the circuit and the scalability of the system remains a challenge as the signal to noise ratio of the ADC-free approach is lower compared to the ADC based approach.

1.4 Challenges and Limitations of RRAM Devices for CIM

As discussed in the previous sections, RRAM based CIM shows its advantages in energy efficiency and scalability, but there are still several challenges and limitations of RRAM

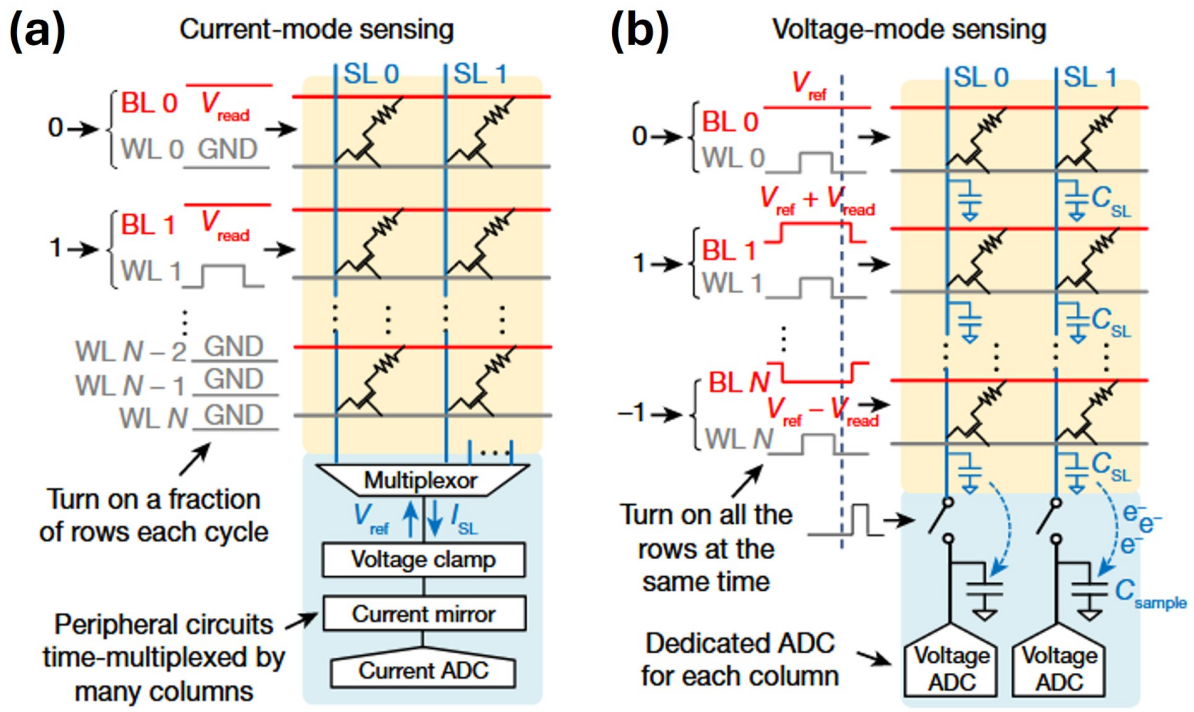


Figure 1.7: Schematic of current mode sensing (a) and voltage mode sensing (b)[11].

devices that prevents further adoption of RRAM based CIM. The challenges and limitations includes:

- Endurance: Compared to other memory technologies like SRAM or DRAM, the limited endurance of RRAM devices limits the application of RRAM based CIM to inference-only tasks, which requires less frequent programming of the memory devices compared to training tasks. Although it has been demonstrated that RRAM devices can achieve up to 10^{12} endurance cycles, RRAM devices from top foundries including TSMC and Intel only have endurance of around 10^6 cycles mostly because of the large device-to-device variation. Material stack innovations including doping of the switching layer[108], interface engineering[37] and having 2D layered dielectric[21] as switching layer have been proposed to improve the endurance of RRAM devices, but the endurance of RRAM devices is still not comparable to that of SRAM or DRAM.
- ON/OFF resistance: Compared to other non-volatile memory technologies like PCM or FeRAM, the endurance performance of RRAM is comparable. However, the low

ON resistance (typically $< 10\ K\Omega$) of RRAM devices is not desired in the application of CIM, as the summing current of the CIM array is determined by the ON resistance of the RRAM devices, the low ON resistance of RRAM devices results in high summing current, reducing energy efficiency. Large bitline current also leads to non-ideal effects like IR-drop, causing the voltage applied to the devices at far-end smaller than the desired value, which would lead to deviation between the actual and the target value of the MVM operation. ON resistance of RRAM can be controlled by the compliance current of the SET operation, but it would lead to the degradation of the sense margin (ON/OFF ratio), which requires higher OFF resistance (HRS) to compensate the degradation. Higher HRS, however, cannot simply be achieved by increasing the RESET voltage due to the RESET induced breakdown, which remains a challenge in the optimization of RRAM devices.

- State stability: The requirement of resistance state stability in CIM is higher than that in memory storage, as the resistance state of the RRAM devices are used to store the weight of the neural network, drift of the resistance state would influence the computational results, which in turn leads to the degradation of the accuracy of the neural network. Resistance drift of RRAM devices, including short term fast relaxation and long term retention loss, is primarily due to the spontaneous oxygen vacancy diffusion and recombination at room temperature. Both short term relaxation and long term retention loss can be mitigated using extra writing steps, but simply increasing the number of writing steps would lead to the degradation of the endurance of RRAM devices as well as programming speed to the system.

1.5 Contributions of This Thesis

In this thesis, we proposed several strategies to mitigate the challenges and limitations of RRAM devices for CIM, which includes the optimization of the device characteristics of RRAM devices, the development of multi-bit programming algorithm, the programming methods to mitigate the retention loss of RRAM devices and the back-end-of-line (BEOL) integration of RRAM devices on CMOS chips. The contributions of this thesis are summarized as follows:

- Bipolar resistive switching of TiO_x devices (Chapter 2): Details of the device fabrication and characterization of TiO_x RRAM devices are discussed, which includes the device structure, electrode dependence, physcial scaling and retention/endurance performance of the fabricated devices. The observed electrode dependence serves as

the foundation of the device optimization of TiO_x RRAM devices, which is discussed in the next chapter.

- Device optimization (Chapter 3): We observed a dependence between the resistance of resistance states and the forming voltage of our fabricated TiO_x RRAM devices. We later found a way to lower down the forming voltage by tuning the nitrogen concentration of the bottom TiN electrode, which further promotes us to improve the overall performance of RRAM devices using interface engineering. The interface engineering includes the tuning of nitrogen concentration of bottom TiN electrode, insertion of Al_2O_3 layer and N_2 plasma treatment of top and bottom interfaces. The optimized devices show lower forming and operation voltage and higher HRS, which is critical for the application of RRAM based CIM.

- Multi-bit programming algorithm (Chapter 4): We proposed a state aware multi-bit programming algorithm for RRAM devices, which utilizes the dependence between conductance modulation and conductance state to deterministically calculate the writing parameters of SET/RESET for the next programming step. The proposed algorithm was validated using the experimental data of RRAM devices, demonstrating the effectiveness of the algorithm in improving the programming efficiency of RRAM devices. We also observed a electrical annealing method to mitigate the retention loss of RRAM devices after programming. The method utilizes sub-write voltage to heat up the RRAM devices and promote the oxygen vacancy diffusion and recombination, which in turn is proved to effective in mitigating the long terms retention loss, especially for high conductance states. The influence of retention loss of CIM is evaluated using simulation, which shows the effectiveness of the proposed method in improving the accuracy of the neural network.

- BEOL integration of RRAM devices on CMOS chips (Chapter 5): We proposed a BEOL integration method of RRAM devices on foundry CMOS chips, which is a pioneer part of the long term development project of RRAM based CIM. The integration method includes the design of the CMOS chip, the fabrication process flow of RRAM devices and the integration of RRAM devices on CMOS chips. The packaging and testing of the integrated chip is also discussed.

Chapter 2

Bipolar Resistive Switching of TiO_x based Devices

Titanium oxide (TiO_x) is selected as the switching layer in this study. Compared to other well-studied transition metal oxides such as hafnium oxide (HfO_2), TiO_x exhibits a more complex phase diagram and a broader stoichiometric range. These characteristics contribute to its diverse material properties, including superconductivity, diluted ferromagnetism, and resistive switching behavior—making it a compelling candidate for memory applications.

Unipolar resistive switching[109] and bipolar resistive switching[110] behavior have both been reported on TiO_x based devices, it has also been found that TiO_x based devices with bipolar resistive switching behavior can be transitioned to unipolar resistive switching behavior by increase the compliance current during operation. While TiO_x based devices with unipolar resistive switching behavior has the advantage of large ON/OFF ratio, the requirement of higher switching current and voltage makes it not suitable for integration with CMOS chips as larger transistors are needed to provide the programming current.

Depending on the switching mechanism, TiO_x based resistive switching devices can also be categorized into interfacial switching and filamentary switching. For interfacial switching devices, the concentration of oxygen ions around the metal/oxide interfaces is modulated by applied electric filed, which would in turn change the interface barrier, leading to the resistance change of the devices[19, 111]. Compared to filamentary switching devices, interfacial switching normally has higher resistance of resistance states and larger non-linearity of its I-V behavior as it normally requires no forming operation. However, interfacial switching TiO_x based devices normally requires higher operation voltage and

the switching speed as well as the state stability remains a challenge. In this work, the research focus is on the bipolar resistive switching behavior of TiO_x based devices and the switching mechanism is filamentary switching.

2.1 Device Fabrication and Characterization

The TiO_x thin films are deposited by reactive sputtering of a 2 inch Ti target in an Ar/O_2 atmosphere. The deposition is carried out at room temperature with a base pressure of 3×10^{-8} torr, and the deposition pressure is 3 mTorr. The flow rate of Ar is fixed at 30 sccm, and the flow rate of O_2 is varied to control the oxygen content in the TiO_x film. The sputtering power is fixed at 200 W, and the deposition time is varied to control the thickness of the TiO_x film. To minimize the batch-to-batch variation, a 2 inch Ti target is specified for the growth of TiO_x and the oxidation layer on the Ti target after each TiO_x deposition need to be removed by sputtering the target with pure Ar plasma. The vacuum level of the sputtering chamber need to be smaller than $3e^{-8}$ torr before each deposition to avoid unexpected oxidation during growth.

Material	Growth Methods	Growth parameters
TiO_x	Reactive sputtering	Deposition power: 200 W (2 inch Ti target) $Ar:O_2 = 30 : X$ $(2 \leq X \leq 12)$ Gas pressure: 3 mTorr Pre-sputtering time: 10 min
TiN	Reactive sputtering	Deposition power: 100 W (1.5 inch Ti target) $Ar:N_2 = 30 : X$ $(1.5 \leq X \leq 10)$ Gas pressure: 3 mTorr Back sputtering power(optional): 10 W Pre-sputtering time: 10 min
Pt	DC sputtering	Deposition power: 30 W (1.5 inch Pt target) Gas pressure: 3 mTorr

Table 2.1: Growth methods and parameters of TiO_x, TiN and Pt used in this thesis.

The initial choice of electrode material is Pt, which is a common electrode material for resistive switching devices. The Pt electrode is deposited by sputtering a Pt target in an Ar atmosphere (base pressure being $< 3 \times 10^{-8}$ torr, deposition pressure being 3 mTorr, flow rate of Ar is 30 sccm, sputtering power of 30 W). The thickness of the Pt electrode is fixed at 30 nm. However, the poor adhesion of Pt to the TiO_x layer leads to the partial

delamination of the Pt electrode during the device fabrication process. Apart from that, the Pt electrode provides less flexibility in tuning the device performance. Therefore, TiN is chosen as the electrode material due to its process compatibility and its tunability of the device performance. The TiN electrode is deposited by reactive sputtering of a Ti target in an Ar/N_2 atmosphere (base pressure being $< 3 \times 10^{-8}$ torr, deposition pressure being 3 mTorr, flow rate of Ar is 30 sccm, flow rate of N_2 is varied to control the nitrogen content in the TiN film, sputtering power being 100 W).

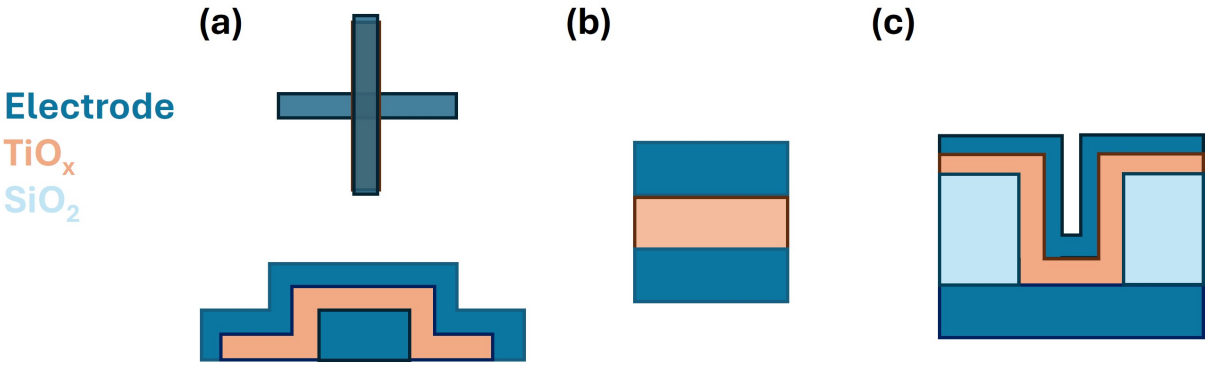


Figure 2.1: (a) Crossbar structure of the device. (b) Pillar structure of the device. (c) Via structure of the device.

There are three ways to fabricate the sandwich structure of the device:

- Crossbar (Fig. 2.1(a)): The top and bottom electrodes are perpendicular to each other, and the switching layer is sandwiched between them. The device area is determined by the dimensions of the top and the bottom electrode. The bottom electrode is patterned first, and then the switching layer is deposited on top of it. The top electrode is patterned on top of the switching layer. The advantage of this structure is that it can be achieved using a shadow mask process, making it suitable for fast prototyping, but it is hard to scale down the dimensions of the device below 10 μm with shadow mask. Further scaling down the device dimensions requires lithography steps, however, the sharp edges of the lithographically patterned electrodes can cause the breakdown of the switching layer.

- Pillar (Fig. 2.1(b)): The top and bottom electrodes are parallel to each other, and the switching layer is sandwiched between them. The material stack is deposited all together and then patterned using a lithography and etching process. The advantage of this structure is that it can be scaled down to sub-100 nm dimensions and the

process flow can be easily integrated into the CMOS process flow. However, such whole stack patterning requires a high aspect ratio etching process, which would lead to fabrication hazards such as side wall re-deposition and etch profile distortion. The process temperature of the etching process can also affect the properties of the switching layer.

- Via (Fig. 2.1(c)): Similar to the fabrication process of the crossbar structure, the bottom electrode is patterned first for electrode connection, and a passivation layer (such as SiO_2 and SiN) is deposited on top of it. A via is etched in the passivation layer to expose the bottom electrode, the dimensions of the via determines the device area. The switching layer is deposited on top of the passivation layer, and the top electrode is patterned on top of the switching layer. Such method overcomes the structure limitation of the crossbar and the fabrication hazards of the pillar structure. The device area can be scaled down to sub-100 nm dimensions. Pre-fabrication of the bottom electrode also allows the use of a shadow mask process or lift-off process for the top electrode patterning, which is suitable for fast prototyping.

Both pillar (Fig. 2.2) and Via (Fig. 2.3) structure have been tested in this study, and the via structure is chosen for the current and future device development due to its fast prototyping capability and the ease of scaling down the device dimensions. The via structure is fabricated as follows: a 30 nm thick TiN (or Pt) bottom electrode is deposited on a 3-inch Si/SiO_2 wafer substrate and patterned using a lithography and etching process. A 50 nm SiO_2 passivation layer is deposited on top of the bottom electrode, and a via is patterned using optical lithography (and e-beam lithography for sub-1 μm patterns) and reactive-ion-etching (RIE) process. Due to the selectivity of the RIE process of SiO_2 to TiN, the bottom electrode is not affected during the via etching process. After RIE etching, the wafer is patterned again using optical lithography to define the shape of the top electrode. The fabricated 3-inch wafer is diced into small chips ($\sim 1\ cm \times 1\ cm$) using a dicing saw, small chips are loaded into the sputtering chamber for the deposition of the TiO_x switching layer and the top electrode (Pt or TiN) using the same process as described above. Followed by a lift-off process to remove the excess material on the top electrode, the chip is ready for electrical characterization using the probe station.

The fabricated chips are probed using a probe station with a semiconductor parameter analyzer (Keithley 4200A-SCS) to measure the electrical characteristics (DC and pulse) of the devices. The parameter analyzer is equipped with three mid-voltage source measurement unit (SMU), one high-voltage SMU, one two-channel pulse measurement unit (PMU), one capacitance measurement unit (CVU) and two remote preamplifier/switcher (RPMs). Two of the mid-voltage SMUs (SMU1 and SMU2) are equipped with pre-amplifiers to

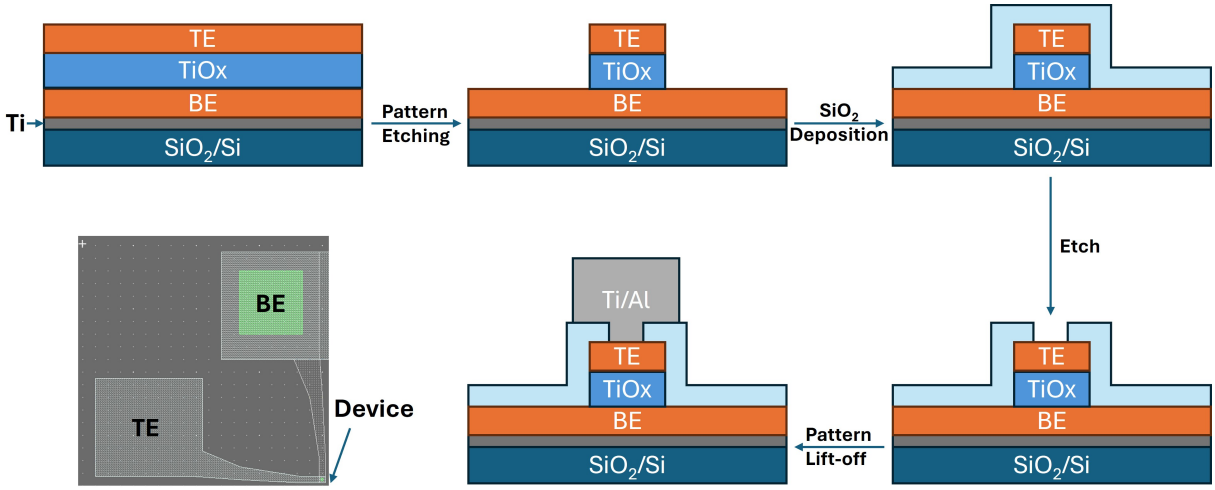


Figure 2.2: Fabrication process flow of the pillar structure.

measure the current in the sub-pA range. SMU1, SMU2, PMU and CVU are connected to the RPMs (RPM1-1 and PRM1-2), which can be configured to switch the measurement units. RPM1-1 is connected to the top electrode of the device under test (DUT), and RPM1-2 is connected to the bottom electrode of the DUT.

For the DC measurement of DUT, custom code is developed to achieve bipolar voltage sweeping of the SMUs. The developed DC measurement module enables setting different compliance current at positive and negative polarity of voltage applied, repeated voltage sweeping is also enabled to evaluate the endurance of devices using DC sweeping. For pulse measurement, a series of test modules are developed, including transient analysis of switching current, custom voltage waveform for memory endurance test and custom multi-bit writing algorithm.

The standard electrical characterization of the devices include:

- Dielectric leakage current test: A voltage sweeping from $0V \rightarrow 1V \rightarrow 0V \rightarrow -1V \rightarrow 0V$ is applied to the devices to measure the initial resistance of the devices, compliance current being $1e-6$ or $1e-7A$ is applied during the measurement.
- Forming test: A voltage sweeping from $0V \rightarrow V_{forming} \rightarrow 0V$ with compliance current being I_{comp} is applied to FORM the devices. The amplitude of $V_{forming}$ and I_{comp} are tuned accordingly (start with low I_{comp}, increase I_{comp} first then $V_{forming}$ until the devices is formed).

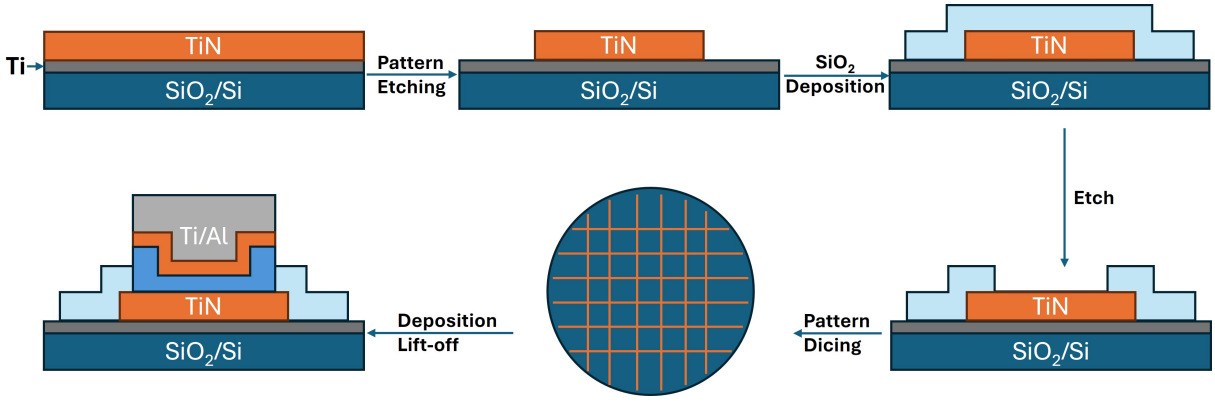


Figure 2.3: Fabrication process flow of the via structure.

- RESET test. A voltage sweeping from $0V \rightarrow V_{RESET} \rightarrow 0V$ with no compliance current applied is applied to RESET the devices. For the first few measured devices (~ 20) of each batch, the amplitude of V_{RESET} need to be increased gradually until the RESET induced breakdown is observed. The average RESET breakdown voltage can be used as the maximum applied RESET voltage applied to the remaining devices of the same batch.

- Sweeping test: A voltage sweeping from $0V - V_{SET} - 0V - V_{RESET} - 0V$ is applied to the devices. Compliance current is applied during SET, and the amplitude of the compliance current is equal to or smaller than the compliance current used during forming.

- Retention test: The devices can be SET(or RESET) to LRS(or HRS) by DC sweeping or voltage pulses. A constant small read voltage (-0.2 V) is applied to read the resistance of the devices. The sampled interval is varied depending on the test requirement.

- Endurance test: DC sweeping or pulse based method can be applied to conduct the endurance test. The resistance of LRS and HRS are sampled at each programming cycle to monitor the cycle-to-cycle variation and state degradation.

For each batch of devices, a yield test which includes Forming test, RESET test and Sweeping test (10 cycles) is applied to at least 50 devices to estimate the yield rate. A devices is classified as PASS if HRS (read at -0.2 V) is larger than the setting value (100 KΩ for most cases) and LRS is smaller than HRS during the Sweeping test.

2.2 Influence of Electrode Material

Devices with pillar structure (Pt/TiO_x/Pt) are initially fabricated to explore the resistive switching behavior of TiO_x based devices. The devices are firstly fabricated based on the pillar structure and the smallest device area is 4 μm $\times$ 4 μm. The I-V characteristic of the fresh devices show strong diode behavior and can only be formed by applying a negative voltage bias (> -5 V) to the top electrode. Contrary to the common resistive switching behavior, where devices are in LRS after forming, such devices are in HRS after the forming (Figure. 2.4), a subsequent SET operation is needed to switching the device from HRS to LRS. The abnormal resistive switching behavior is mainly due to the Schottky barrier between the Pt electrode and the TiO_x layer, the forming step breaks the Schottky barrier and allows the voltage to be applied to the TiO_x layer in the following operation. After the forming, a SET operation with positive voltage applied to the top electrode is applied to switch the device to LRS. The devices can be further RESET to HRS by applying a negative voltage, showing bipolar resistive behavior. The devices with this structure is not further explored due to the high forming voltage and high operation current despite the thickness scaling of the switching layer.

The experimental results on Pt/TiO_x/Pt inspires us to think about the role of work function of the electrode material in the resistive switching behavior of the devices. The high work function of Pt leads to the formation of a Schottky barrier at the Pt/TiO_x interface, which hinders the formation of the conducting filament. To overcome the high Schottky barrier, TiN is chosen as the electrode material due to its lower work function (4.5 eV) compared to Pt (5.65 eV). Apart from that, high temperature process (PECVD SiO_2 deposition requiring 300 °C ambient temperature) during fabrication leads to uncertainties in conditioning of the TiO_x layer, which could also lead to the abnormal resistive switching behavior.

Via structure is developed to overcome such issue. Both Pt and TiN are used to fabricate devices using the via structure, the smallest device area is 4 μm $\times$ 4 μm. Devices with material stack TiN/TiO_x/Pt, Pt/TiO_x/TiN, Pt/TiO_x/Pt and TiN/TiO_x/TiN are fabricated to explore the influence of the electrode materials. Pt/TiO_x/Pt devices suffer from partial delamination of the electrodes due to poor adhesion between Pt and TiO_x layer, however, no such issue is observed for devices with the same material stack using the pillar structure, which further supports our assumption that process temperature during fabrication modulates the device properties. All other three material stacks show bipolar resistive switching behavior, the polarity of the forming voltage is the same as the SET voltage. Moreover, the forming process is found be dependent on the dielectric/metal interfaces: TiN/TiO_x/Pt and Pt/TiO_x/TiN can only be formed by applying positive volt-

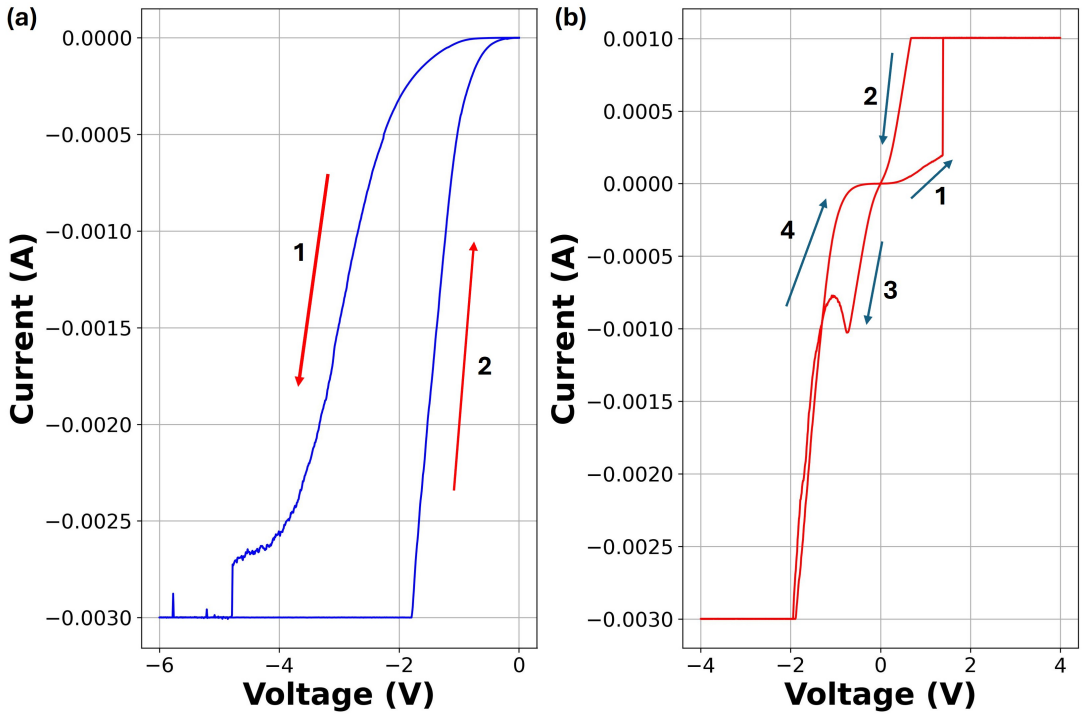


Figure 2.4: (a) Abnormal forming I-V characteristic of Pt/TiO_x/Pt devices. (b) Switching I-V characteristic of Pt/TiO_x/Pt devices.

age to the TiN electrode, while TiN/TiO_x/TiN can be formed with both polarities. The forming voltage of the devices is summarized in Table 2.2. It can also be noticed that the amplitude of forming voltage is dependent on the interface position (Figure. 2.5). The switching I-V of the three material stacks are shown in Figure. 2.6. For TiN/TiO_x/TiN, which can be formed with both polarities, forming with positive voltage is preferred as the forming voltage is smaller and the operation voltage/current is also lower. Compared to devices with Pt as top or bottom electrode, TiN/TiO_x/TiN is chosen for the following studies due to smaller operation voltage.

Further optimizations show that the deposition condition of the TiN bottom electrode also plays a crucial role in the performance of the devices. DC sputtering is found to be better than RF sputtering, and back bias during deposition is found to have greatly improved the device performance (Figure. 2.7). These observations suggest that the TiN/TiO_x interfaces are the key to the resistive switching behavior of the devices, the details of the interface engineering will be further discussed in the next chapter.

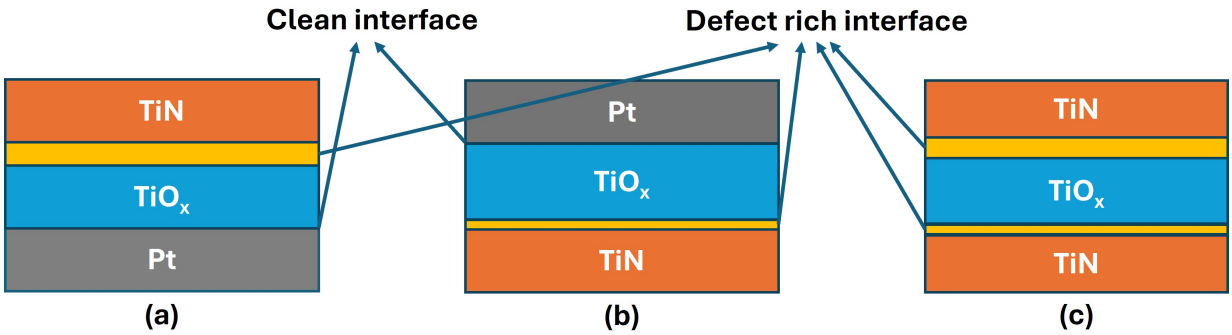


Figure 2.5: Interface analysis of (a) Pt/TiO_x/TiN, (b) TiN/TiO_x/Pt and (c) TiN/TiO_x/TiN devices.

Material Stack	Forming voltage (V)
TiN(BE)/TiO_x/Pt(TE)	-2
Pt(BE)/TiO_x/TiN(TE)	1.3
TiN(BE)/TiO_x/TiN(TE)	1.5
	-2

Table 2.2: Forming voltage of devices with varied electrode combinations.

2.3 Influence of Oxygen Ratio

The oxygen vacancy inside the TiO_x layer plays an dominant role in the formation of conducting filaments and the performance of the devices. There are many ways to introduce oxygen vacancies in the TiO_x layer, including from the interface (active metal electrode or H_2 annealing) or from the stoichiometry of the bulk layer. The role of the oxygen vacancy in bulk layer is firstly investigated by varying the oxygen ratio during the deposition of TiO_x layer. The oxygen ratio is varied by changing the flow rate of O_2 during the deposition of TiO_x layer, the flow rate of Ar is fixed at 30 sccm, and the flow rate of O_2 is varied from 1.5 sccm to 12 sccm. The deposition power is fixed at 200 W, the deposition rate of the sub-stoichiometric TiO_x layer is found to be dependent on the oxygen ratio, the deposition rate decreases with the increase of oxygen ratio and saturates when the flow rate of oxygen ratio is layer than 8 sccm.

Devices with O_2 flow rate of 1.5 sccm, 4 sccm and 6 sccm are fabricated using the via structure, the thickness of the TiO_x layer is fixed at 5 nm and the bottom electrode is the optimized TiN (with back sputtering enabled during deposition). For each oxygen

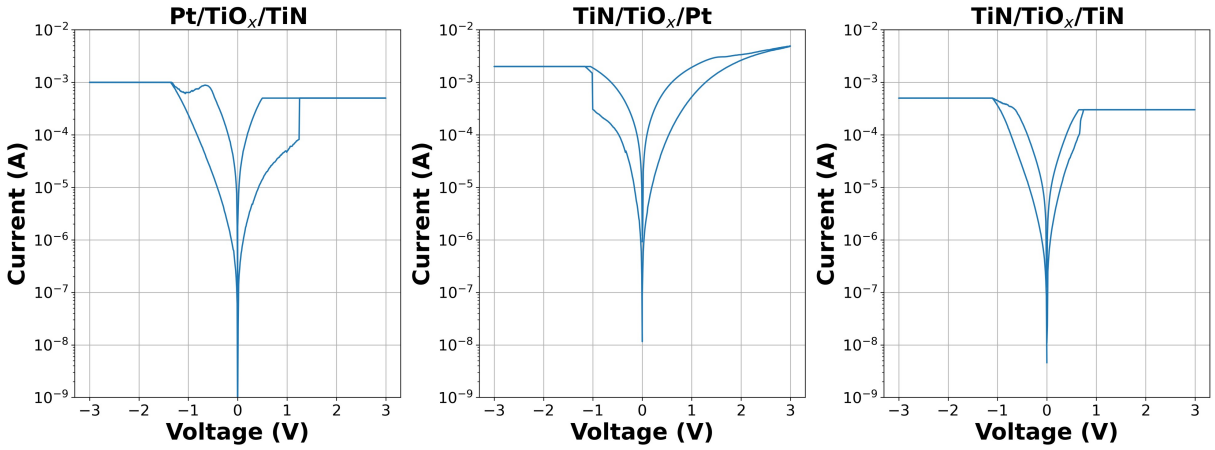


Figure 2.6: Switching I-V characteristic of (a) Pt/TiO$_x$/TiN, (b) TiN/TiO$_x$/Pt and (c) TiN/TiO$_x$/TiN devices (positive forming voltage).

ratio, 48 devices are measured to evaluate the forming voltage and the resistance of HRS, both of which are the key to low power operation of RRAM based CIMs. The results are summarized in Table. 2.3. The forming voltage decreases with th decrease of the oxygen ratio, which is consistent with the literature results. Devices with 4 sccm and 6 sccm share similar forming voltage, but the resistance of HRS of the 6 sccm devices is higher than that of the 4 sccm devices, which suggest that TiO$_x$ with lower oxygen ratio suffers from leakage current. Devices with 1.5 sccm show the lowest forming voltage, but the resistance of HRS is higher than that of the 4 sccm devices, indicating that the oxygen vacancy concentration of the dielectric layer is not the only factor that determines the resistance of HRS.

Oxygen ratio	Forming voltage (V)	HRS (Ω)
1.5 sccm	1.17	70 K
4 sccm	1.5	43.4 K
6 sccm	1.49	116.2 K

Table 2.3: Effects of oxygen ratio on the forming voltage and HRS resistance (The thickness of TiO$_x$ is 5 nm).

Devices with lower oxygen concentration ($Ar/O_2 = 30:1$ and $50:2$) are also fabricated but found to be leaky and cannot be formed. To combine the advantages of low forming voltage of devices with low oxygen ratio and the high resistance of HRS of devices with high oxygen ratio, a hetero-structure of TiO$_x$ with different oxygen ratio is proposed. The

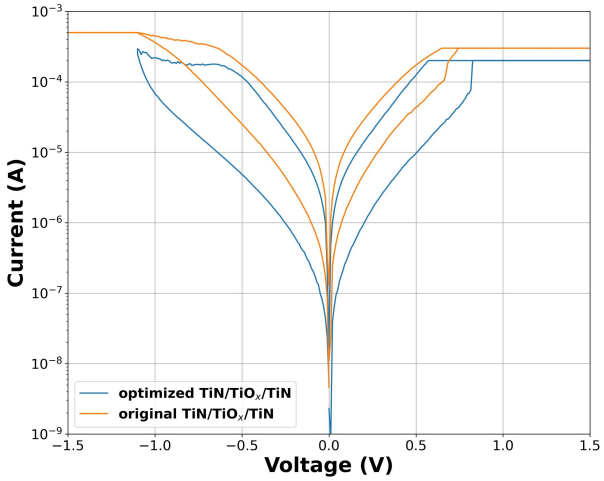


Figure 2.7: Switching I-V characteristic of TiN/TiO_x/TiN devices before and after interface engineering.

thickness of the TiO_x layer is fixed at 4 nm, the bottom TiO_x layer is deposited with oxygen ratio of being 6 sccm (Ar = 30 sccm) while the top of the TiO_x layer is deposited with oxygen ratio of being 2 sccm (Ar = 50 sccm). Same methodology is used to evaluate the performance and the results are summarized in Table. 2.4. It can be observed from the results that the forming voltage of the hetero-structure is lower than that of the pure 6 sccm devices, and the resistance of HRS is at a similar level of the 6 sccm devices. Compared to devices with 5 nm TiO_x layer described above, the benefit of thickness scaling is more dominant than the oxygen profile tuning. These finding promotes us to further investigate the scaling of the forming voltage and its influence on the resistance states, the details will be discussed in the next chapter.

Material Stack	Forming voltage (V)	HRS (Ω)
TiO_x (30 : 6) 4 nm	1.19	393.7 K
TiO_x (30 : 6) 3 nm + TiO_x (50 : 2) 1 nm	1.13	352.1 K
TiO_x (30 : 6) 2 nm + TiO_x (50 : 2) 2 nm	1.13	228.1 K
TiO_x (30 : 6) 3 nm	1.13	241.5 K

Table 2.4: Comparison of the forming voltage and HRS resistance of devices with different heterostructure.

2.4 Influence of Physical Dimensions

The study on the influence of oxygen ratio has revealed the importance of thickness scaling of the switching layer. To investigate the influence of physical dimensions on the resistive switching behavior of the devices, devices with different area and thickness are fabricated. The thickness of the TiO_x layer ranges from 2 nm to 4 nm, and the device area ranges from 600 nm × 600 nm to 4 μm × 4 μm. The oxygen ratio of the TiO_x layer is fixed at ($Q_{Ar} : Q_{O_2} = 30 : 12$, and the bottom electrode is the optimized TiN ($Q_{Ar} : Q_{N_2} = 30 : 1.5$, back sputtering power = 10 W). For each combination of thickness and area, 48 devices are measured to evaluate the forming voltage, the HRS resistance and the HRS breakdown voltage, which is the maximum RESET voltage that the device in HRS can withstand. The results are summarized in Figure. 2.8.

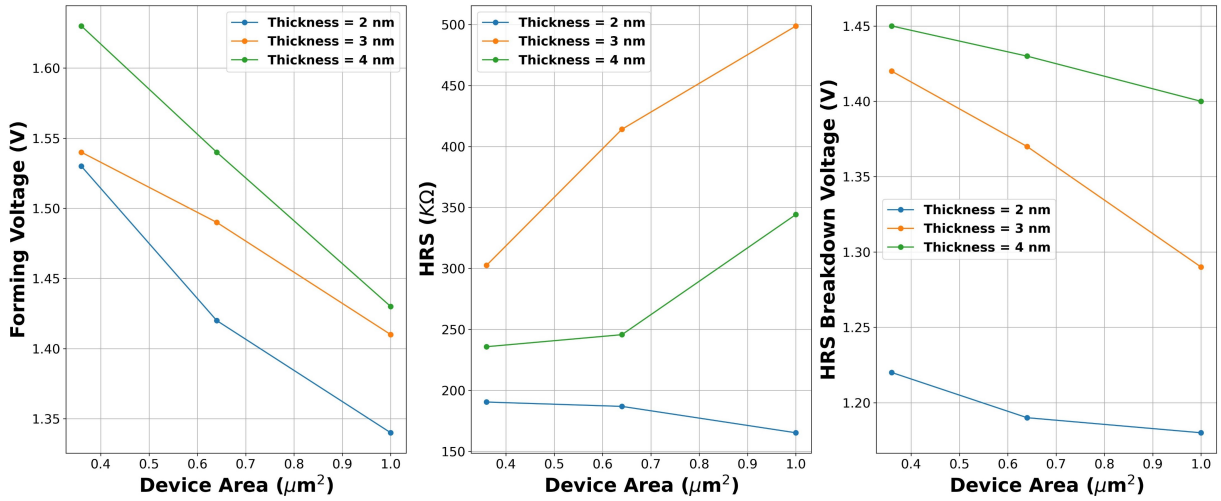


Figure 2.8: Influence of thickness and area on the forming voltage, HRS resistance and HRS breakdown voltage.

It can be observed from the results that the forming voltage decreases with the decrease of the thickness of the TiO_x layer, increases with the decrease of the device area, both of which are consistent with the literature results. The HRS resistance is dependent on the device area when the thickness of the TiO_x layer is above 2 nm (Figure. 2.8(b)), the resistance of HRS increases with the decrease of the device area. However, when the thickness of the TiO_x layer is below 2 nm, weak dependence on the device area is observed, which suggest that different switching mechanism is involved for ultra thin TiO_x layer

(more discussions on this topic in the next chapter). The HRS breakdown voltage follows the same trend as the forming voltage.

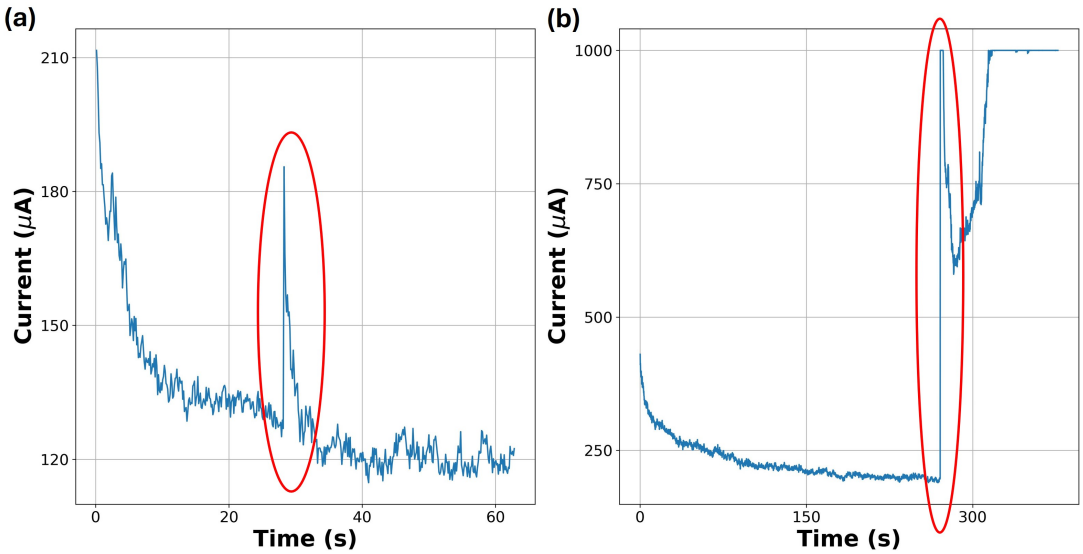


Figure 2.9: Transient of RESET process: (a) Competition between RESET oxidation and breakdown. (b) Breakdown of the device in HRS.

To get a better understanding of the HRS breakdown, transient analysis is conducted by applying constant voltage stress to the devices in HRS and monitor the current change. The amplitude of the voltage stress is chosen to be close to the HRS breakdown voltage, and SMU is chosen to sample the current for better current resolution and the maximum sampling rate is around 125 Hz. The results (Figure. 2.9) that the breakdown could happen during the RESET process, devices could recover from the breakdown and maintain the HRS resistance (Figure. 2.9(a)) or suffer from permanent degradation of the resistance states (Figure. 2.9(b)). It is also observed the time to breakdown is dependent on the amplitude of the voltage stress, the higher the amplitude the shorter the time to breakdown. In Figure. 2.8(c), the HRS breakdown voltage of 4 nm is larger than that of 3 nm, the HRS resistance of 4 nm, however, is smaller than that of 3 nm. The above results suggest that the RESET breakdown process is a electric-field driven process, like a dielectric breakdown process.

Analysis of HRS breakdown indicates that the RESET process of the TiO_x based devices is a competition between the oxidation of the conducting filament and the breakdown of the dielectric layer. Larger HRS breakdown voltage is preferred as it provides wider operation

window for the devices. To increase the HRS breakdown voltage, apart from the physical scaling of the device, the interface engineering between the electrode and the switching layer is also reported to be effective in increasing the HRS breakdown voltage. There are mainly two types of interface engineering:

- Reducing the defect concentration on the bottom electrode interface: Noble metals including Pt, Ir and Au[48], which are oxidation resistant, can be used to increase the HRS breakdown voltage. However, our previous study on Pt electrode shows that the forming voltage as well as the operation voltage of the devices is high, which is not desired in this study.
- Increase the dielectric breakdown strength of the switching layer: Using high bandgap dielectric layer such as HfO_2 and Al_2O_3[47] as insertion layer between the electrode and the switching layer is reported to be effective in increasing the HRS breakdown voltage. However, the high band gap dielectric layer could also increase the forming voltage of the devices, which requires further optimization.

Al_2O_3 insertion is chosen as the interface engineering method in this study, apart from the increase of the HRS breakdown voltage, various other benefits are observed with the Al_2O_3 insertion, which will be discussed in the next chapter.

2.5 Retention and Endurance

Retention of the resistance states is evaluated by applying a read voltage to the devices, the resistance of the devices is measured at a fixed interval of time. To minimize the influence of read-disturb, the read voltage is fixed at -0.2 V, the same polarity as that of the RESET. The devices are SET (RESET) to LRS (HRS) using the DC sweeping method, both LRS and HRS have been evaluated for up to 10^4 s at room temperature, which shows no degradation of the resistance states. Attempts have been made to evaluate the retention of the devices at elevated temperature, however, the devices are found to be unstable at temperature higher than 100 °C, permanent degradation of the resistance states is observed after annealing at 100 °C for 1 hour. The degradation mechanism is still under investigation, but our later study on the interface engineering of the devices suggest that the oxidation of the TiN top electrode may be the main reason for the degradation.

By varying the stop voltage of RESET during DC sweeping, the devices are found to be multi-bit capable (Fig. 2.11(a)), the resistance states of the devices can be tuned by

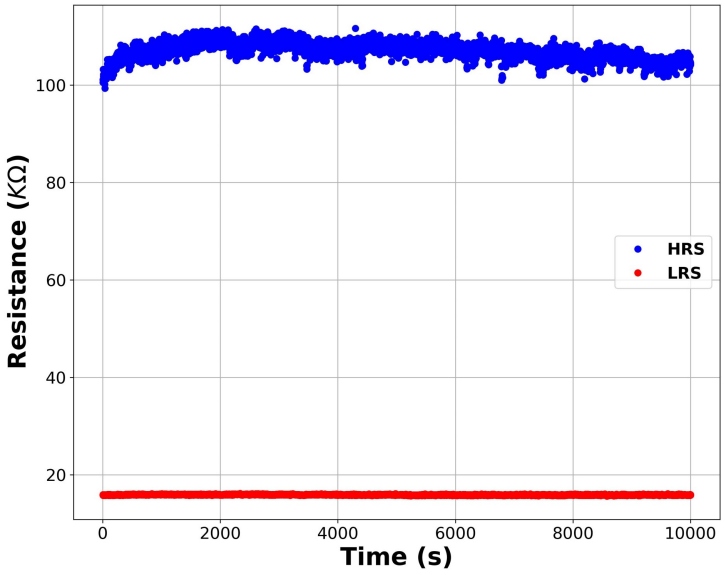


Figure 2.10: Retention of the devices at room temperature.

the amplitude of the RESET voltage. Retention test is also conducted on the multi-bit devices, the results show that the resistance states of the devices are stable (Fig. 2.11(b)).

For devices RESET to higher HRS ($>$ 200K Ω), the retention test have demonstrated that there exists random telegraph noise (RTN) behavior in the conduction of HRS (Figure. 2.12), very few devices in LRS show similar RTN behavior. The RTN suggest that there exists a two level system in the electron transport, the disappearance of the RTN in LRS suggests that such two level system exist between the conducting filament tip and the electrode. The conductance difference between the two states ($\sim 0.001G_0$) is found to be much smaller than the conductance quantum (G_0), indicating that the RTN could be a trap-assisted tunneling process. Different read voltage bias have been applied to evaluate the influence of voltage amplitude on RTN, the results show that the state transition frequency (tunneling rate) is dependent on the amplitude of the read voltage(Figure. 2.13), which further supports the assumption that the RTN is due to the trap-assisted tunneling process.

Endurance of the devices is evaluated by applying repeated DC sweeping to the devices, compliance current is applied during SET to prevent the breakdown of the devices. The devices are found to be able to maintain the resistance states after 10^3 cycles of SET/RESET. To speed up the endurance test, pulse based measurement have been ap-

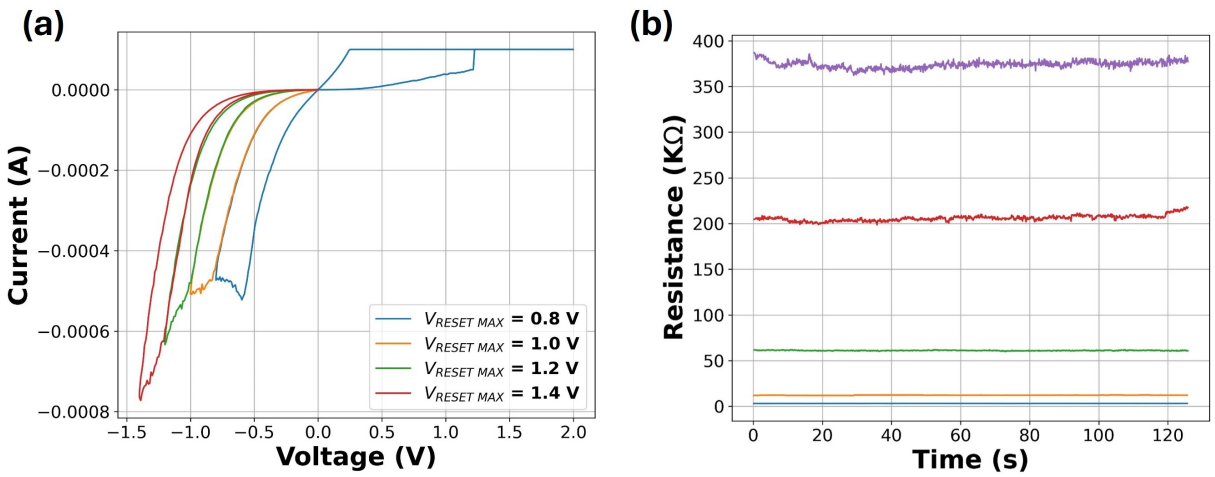


Figure 2.11: (a)Switching I-V of the device with different stop voltage of RESET. (b) Retention of the varied resistance states.

plied to the devices. The results show that the devices suffer stuck-at-HRS after only 10^2 cycles of SET/RESET if the pulse amplitude and width are fixed. Pulse based write-verify algorithm is developed to solve the stuck-at-HRS issue, which helps improve the endurance of the devices to 10^3 cycles. Limited by the device fabrication facility, there still exists a gap between the endurance of the devices and the state-of-the-art RRAM devices, which requires further optimization of the device fabrication process and the device structure.

2.6 Summary

In this chapter, the device fabrication and characterization of the TiO_x based resistive switching devices are presented. The influence of the electrode material, the oxygen ratio of the TiO_x layer, the physical dimensions of the devices and the retention and endurance of the devices are investigated. The results point out the importance of the interface engineering in device optimization, and the necessity of the scaling of the device dimensions for low power operation of the devices. The findings of this chapter provide a guideline for the following study on the interface engineering of the devices. Retention and endurance of the devices are also evaluated, the results show that the devices are able to maintain the resistance states after 10^4 s at room temperature and 10^3 cycles of SET/RESET, researches on transient analysis of RESET process and the RTN behavior of the devices are also

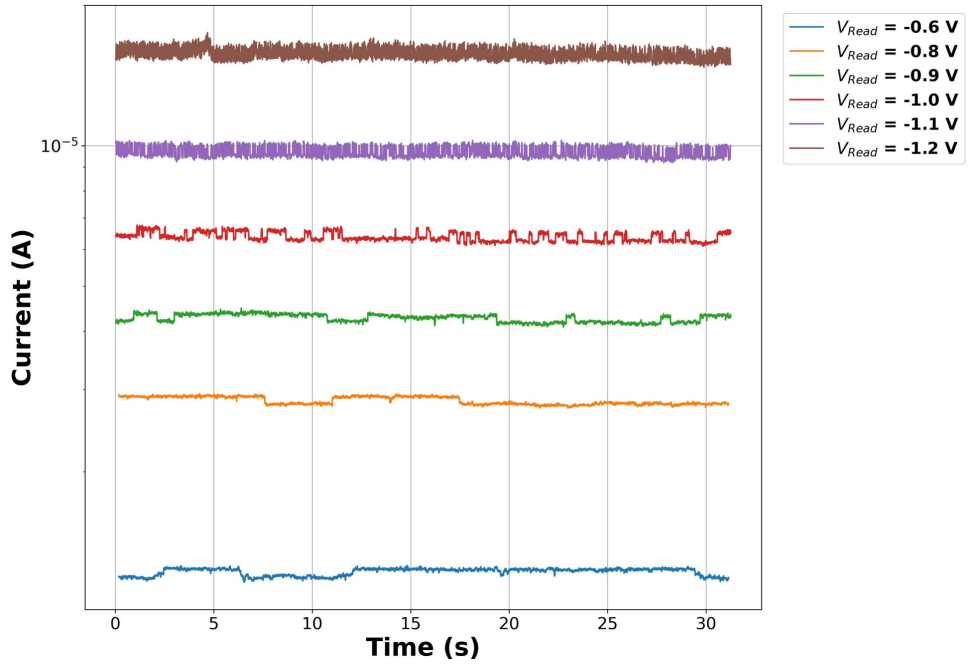


Figure 2.12: RTN behavior of the devices in HRS under different read voltage.

presented. These findings provide a comprehensive understanding of the resistive switching behavior of the TiO_x based devices, and the optimization of the device performance for the following study on the interface engineering of the devices.

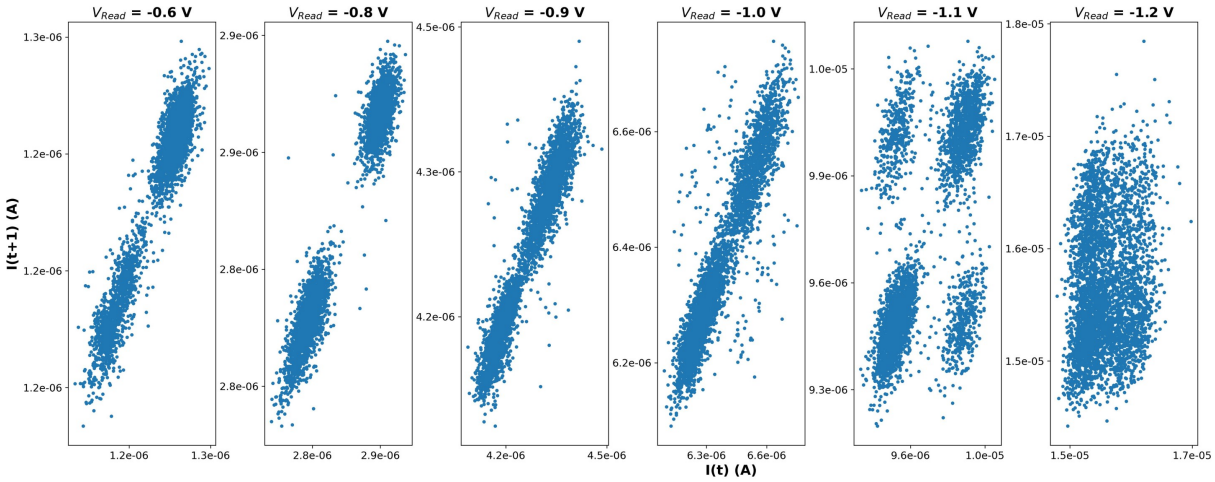


Figure 2.13: I_t vs I_{t+1} of RTN behavior under different read voltage.

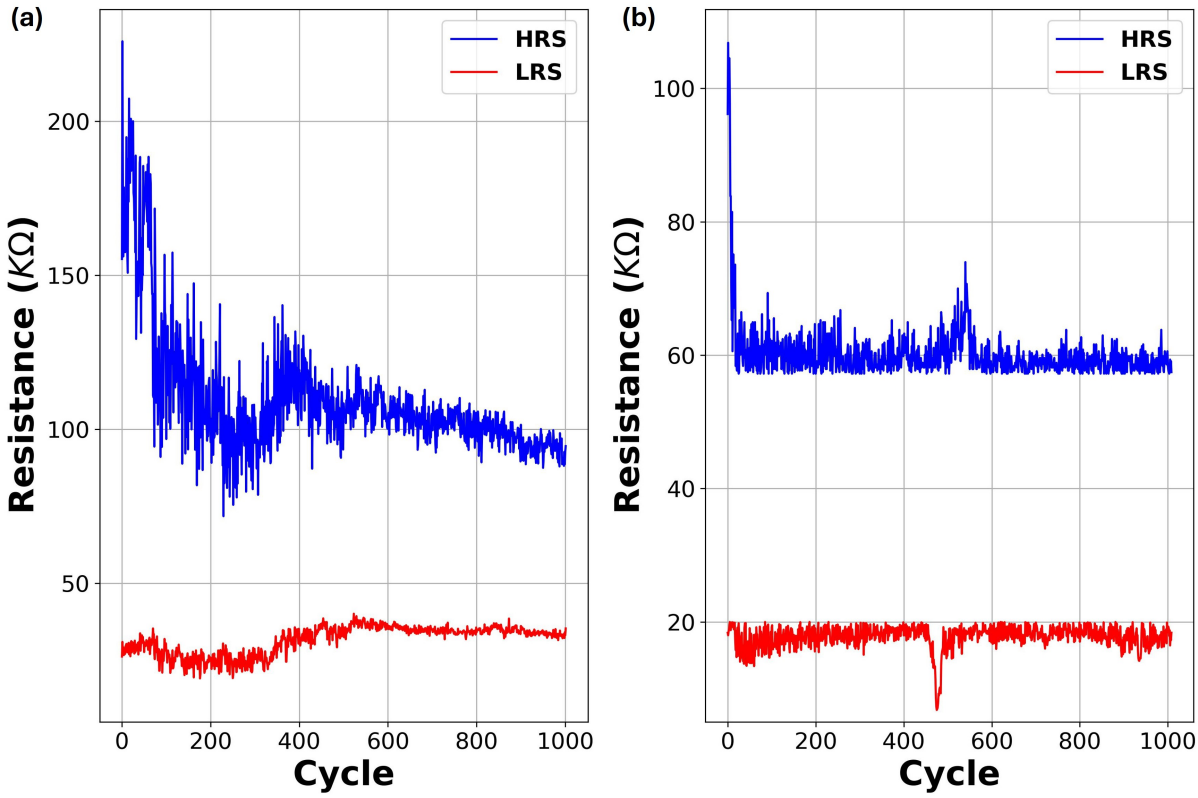


Figure 2.14: Endurance of the devices under (a) DC sweeping and (b) pulse based measurement.

Chapter 3

Device Optimization Through Interface Engineering

To increase the compute density and reduce the power consumption, integration of RRAM devices into advanced technology nodes is essential, which requires the scaling down of the device area as well as its operational voltage (Fig. 3.1 (a))[112, 113]. Studies on device area scaling show that smaller device area would lead to higher forming voltage[114], which is essential to form the conducting filament in the switching layer. Multiple efforts including insertion of active metal layer[24, 23] and localized doping[36] have been proposed to lower the forming voltage. However, previous efforts have been mostly focused on the top interface of the switching layer, the influence of the bottom electrode and the interface between the bottom electrode and the switching layer has not been fully explored. Another challenge for RRAM based CIM is the increase of bit-line current when scaling up the CIM array size, which would lead to the increase of the power consumption and non-ideal effects like IR-drop[115]. The bit line current is principally governed by the resistance value programmed to represent "1" (or "11...1" in multi-bit configurations). To reduce bit line current, the resistance for "1" can be programmed to exceed the standard LRS value, provided the sensing margin remains within specification - a flexibility enabled by RRAM's multi-bit programmability. Consequently, the selection of resistance values for "1" is bounded by HRS. The correlation between the forming voltage and the resistance states, especially HRS has not been fully explored, which is essential for the optimization of the RRAM devices for CIM applications.

In this chapter, the influence of bottom TiN electrode with different growth conditions has been explored, a strong dependence between forming voltage and the nitrogen concentration of TiN has been observed. The relationship between forming voltage and the

resistance of LRS and HRS is also explored, which shows that the amplitude of forming voltage determines the resistance of HRS: at a given range, the lower the forming voltage the higher the resistance of HRS. The scaling of forming is firstly achieved through decreasing the thickness of switching layer, then an opposite trend is observed in both LRS and HRS when the thickness is below 2 nm due to the increase of leakage current. N_2 Plasma treatment of the bottom and top electrode interfaces, reducing defect concentration on the bottom TiN surface and enhancing defect concentration on the top TiO_x surface, has been found to be effective in decreasing the forming voltage without sacrificing the resistance of HRS.

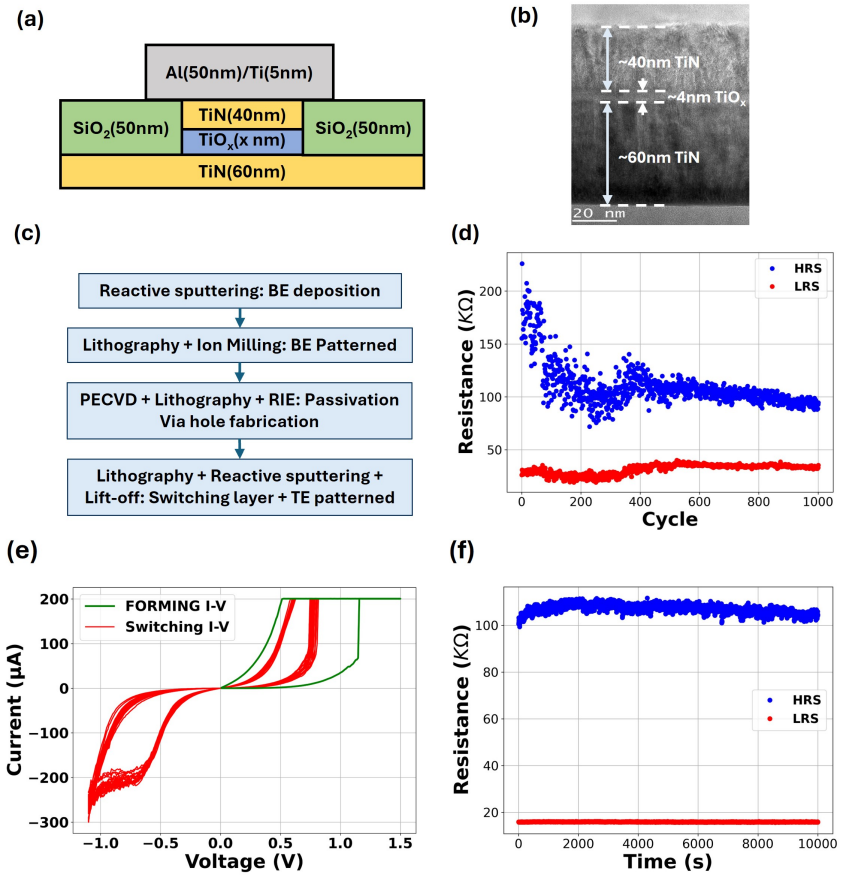


Figure 3.1: (a) Schematic of the fabricated TiN/TiO_x/TiN devices. (b) Cross-section TEM image of the device. (c) Fabrication process flow of the devices. (d) Endurance test of the devices. (e) Forming I-V (green) and switching I-V (red) of the formed devices. (f)Retention test of LRS and HRS.

3.1 Scaling of Forming Voltage

The schematic of the fabricated TiN/TiO_x/TiN devices is shown in Fig. 3.1(a). The bottom TiN electrode is grown on the SiO_2/Si substrate using DC reactive magnetron sputtering, back bias is applied during growth to improve the film quality. The fabrication process (Fig. 3.1(c)) begins with patterning a via structure (minimum opening area: 600 nm × 600 nm) on the bottom electrode. The TiO_x switching layer was then deposited via reactive sputtering of a pure Ti target in an oxygen/argon mixture, with an optimized gas flow ratio (Q_{Ar} : Q_{O_2}= 30:12 sccm) determined by oxygen flow variation at fixed Ar flow. This oxygen ratio was maintained consistently throughout the experiments. The top TiN electrode was subsequently sputtered without vacuum break. The cross-section TEM image of the device is show in Fig. 3.1(b). The fabricated devices are electrically formed by applying a positive DC sweeping voltage (compliance current applied) to the top electrode (Fig. 3.1(e)). The formed devices exhibited stable bipolar resistive switching, as confirmed by endurance (Fig. 3.1(d)) and retention (Fig. 3.1(f)) tests. Limited by the in-house fabrication and deposition facilities, the endurance performance of fabricated devices suffer from stuck-at-LRS degradation.

The forming voltage is varied by changing the thickness of the switching layer during fabrication. The compliance current as well as the voltage sweeping speed during forming are kept constant. Formed devices are RESET to HRS by applying a negative sweeping voltage to the top electrode, no compliance current is applied during RESET and the amplitude of maximum RESET sweeping voltage is adjusted based on the thickness of the switching layer. It is assumed that the resistance of HRS measured is the maximum resistance of HRS as further increase of RESET voltage would lead to HRS breakdown (RESET-induced breakdown), causing non-reversible degradation of HRS. For each thickness of switching layer, 50 devices are measured to evaluate the device-to-device variation. The resistance of LRS and HRS are measured at -0.2 V, the resistance of LRS and HRS versus forming voltage is summarized in Fig. 3.2(b). The results show that the resistance of LRS is independent of the forming voltage when the thickness of switching layer is larger than 2 nm. Thinner switching layer (<= 2 nm and forming voltage < 1.5 V) leads to the increase of resistance of LRS, which is desired for more energy efficient CIM applications, but degraded ON/OFF ratio may increase the design complexity due to smaller read margin.

The forming process can be described as a two-step convoluted process[40]: nucleation step (Fig. 3.3(a)): the oxygen vacancy is driven under the electric field, moving from top electrodes towards bottom electrodes and form the initial conducting filament region; filament growth step (Fig. 3.3(b)): further diffusion of oxygen vacancy would lead to the

expansion of the conducting filament region, leading to the decrease of the resistance of the devices until the compliance current is reached. For thinner switching layer, due to the presence of leakage current flowing through the dielectric layer, the forming would mostly stop at the nucleation step when the compliance current is reached. Overall the resistances are more dominated by the leakage current, thus both LRS and HRS are merging towards same values. The resistance of LRS can be decreased by increasing the compliance current during forming or during switching, but it would decrease the resistance of HRS as well.

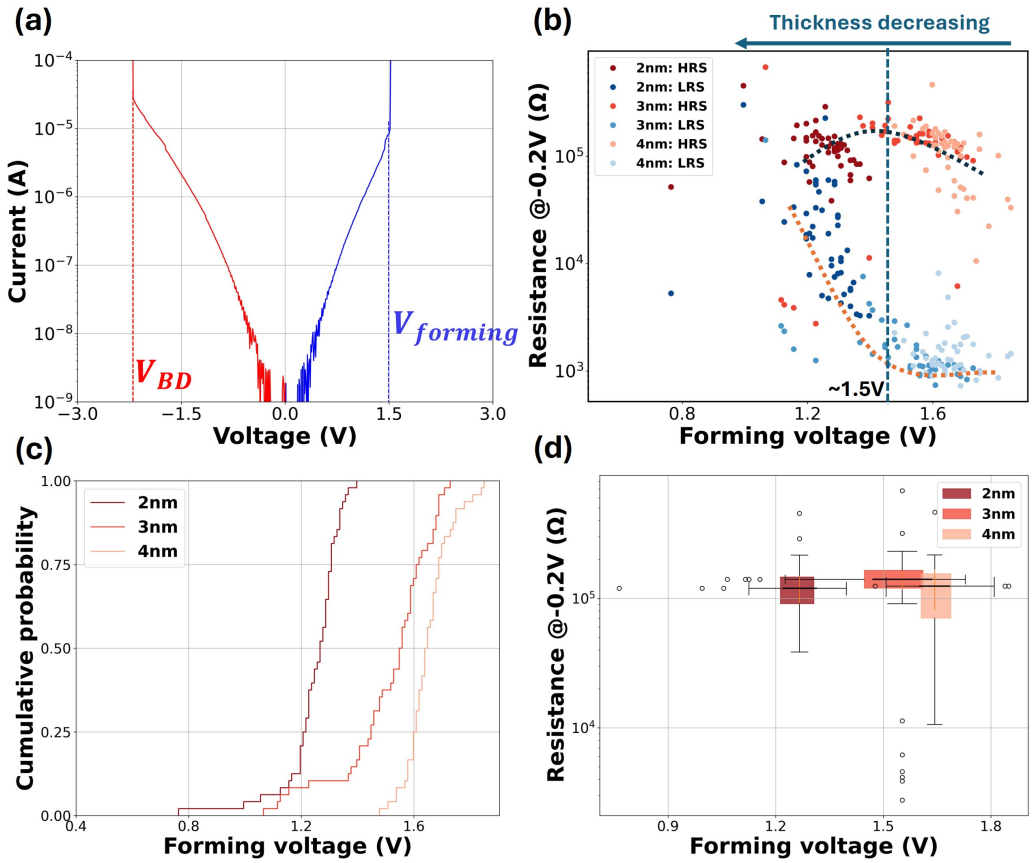


Figure 3.2: (a) Forming and breakdown I-V of fabrication devices. Electrical characterization of devices with varying oxide thicknesses: Dependence of LRS and HRS resistance on forming voltage across different thicknesses (2-4 nm). (c) Cumulative distribution function (CDF) of forming voltages. (d) Two-dimensional boxplot analysis correlating HRS resistance with forming voltage.

While for HRS, the resistance of maximum HRS would increase with the forming voltage

first and then decrease when the forming voltage is above 1.5 V. It is generally accepted that the conducting filament have a cone shape, with more defects concentrated at the top interface and less defects clustered at the bottom interface. Under ideal scenario, the drifting of oxygen vacancy would stop when the compliance current is reached. However, the capacitance discharging immediately after forming would lead to additional current flowing through conducting filament, which can be larger than the compliance current (current overshoot)[29]. The discharge current would heat up the filament region, leading to diffusion of oxygen defects under high temperature and expanding the diameter of the conducting filament, especially at the bottom interface. The increase of HRS below 1.5 V is mainly due to the suppression of leakage current with thicker dielectric layers. The decrease above 1.5 V, however, indicates that the formed conducting filaments suffer more from the current overshoot due to the increase of forming voltage. The current overshoot results in expansion of the diameter of the conducting filament, thus leading to the decrease of the resistance of HRS (Fig. 3.3(d)).

These observations illustrate the challenges of scaling down the forming voltage. It should also be noted that scaling down dielectric thickness is not the "silver bullet" as thinner dielectric would suffer from larger leakage current, poor stability and vanished ON/OFF ratio.

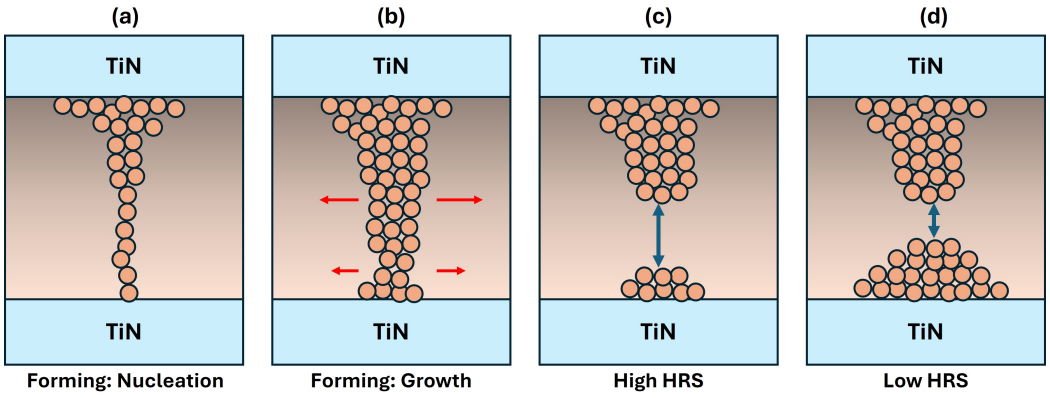


Figure 3.3: (a) Nucleation step of forming. (b) Growth step of forming. Defect profile of High HRS (c) and Low HRS (d) devices.

3.2 Influence of TiN Bottom Electrode

To explore additional methods for scaling down the forming voltage, TiN bottom electrodes with varying nitrogen concentration were deposited using the reactive sputtering. The total gas pressure and the flow rate of carrier gas Ar (30 sccm) were kept constant, while the flow rate of N_2 was varied and a constant RF back bias was applied during the deposition. X-ray diffraction (XRD) was performed on as-deposited TiN films to analyze the crystal structure (Fig. 3.4). The crystal orientation of TiN shows a dependence on the flow rate of N_2: both (200) and (111) phases co-exists when the flow rate of N_2 is low; as the N_2 flow rate increases, the (111) phase intensity becomes stronger and the full width at half maximum (FWHM) decreases, indicating better crystallization. Phi-scan has been done to confirm that all TiN films are polycrystalline. The crystallization dependence of TiN on the flow rate of N_2 may correlate with the deposition rate, higher N_2 flow rates result in lower deposition rates (TiN deposition rate of $Q_{N_2} = 5\ sccm$ is half of that of $Q_{N_2} = 1.5 sccm$). (111) face of cubic structure, which has the highest atomic density, is more resistive against the etching from RF substrate biasing, leading to the increase of the (111) phase intensity.

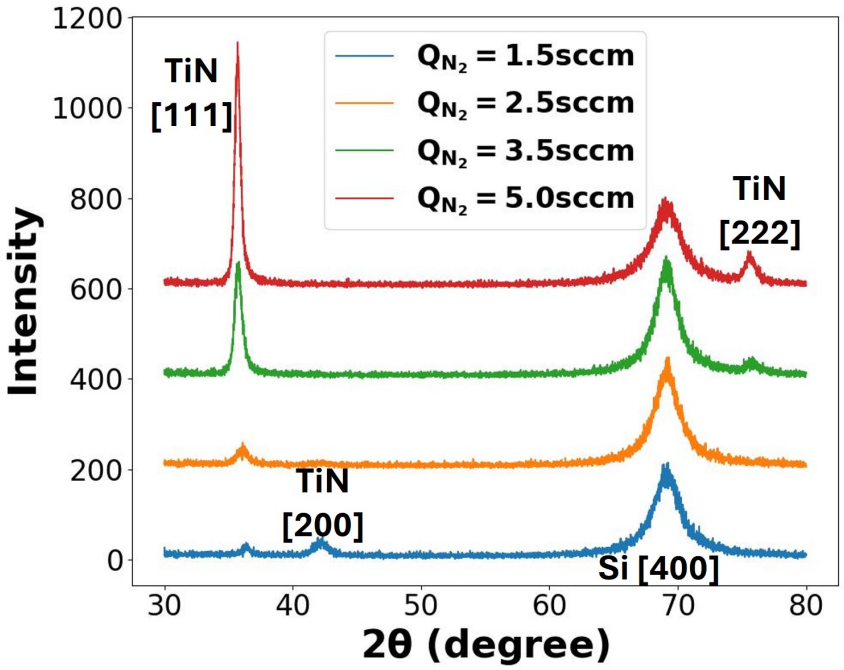


Figure 3.4: X-ray diffraction patterns of TiN films deposited with varying N_2 flow rates.

The varied TiN bottom electrodes were fabricated using the same process flow, and were loaded together for the identical deposition of TiO_x and the top electrode. The I-V characteristics of fresh devices with varied TiN bottom electrodes were summarized in Fig. 3.5(a): there is a sharp change of I-V behavior from 2.5 sccm to 3.5 sccm, which also aligns with the transition of crystal orientation preference. The polarity of the asymmetry of the I-V is also shifted with the bottom electrode composition (Fig. 3.5(c)). Same transition is also observed in the amplitude of the forming voltage (Fig. 3.5(b)), TiN bottom electrode with lower N_2 flow rate requires much lower forming voltage ($\sim 0.6V$).

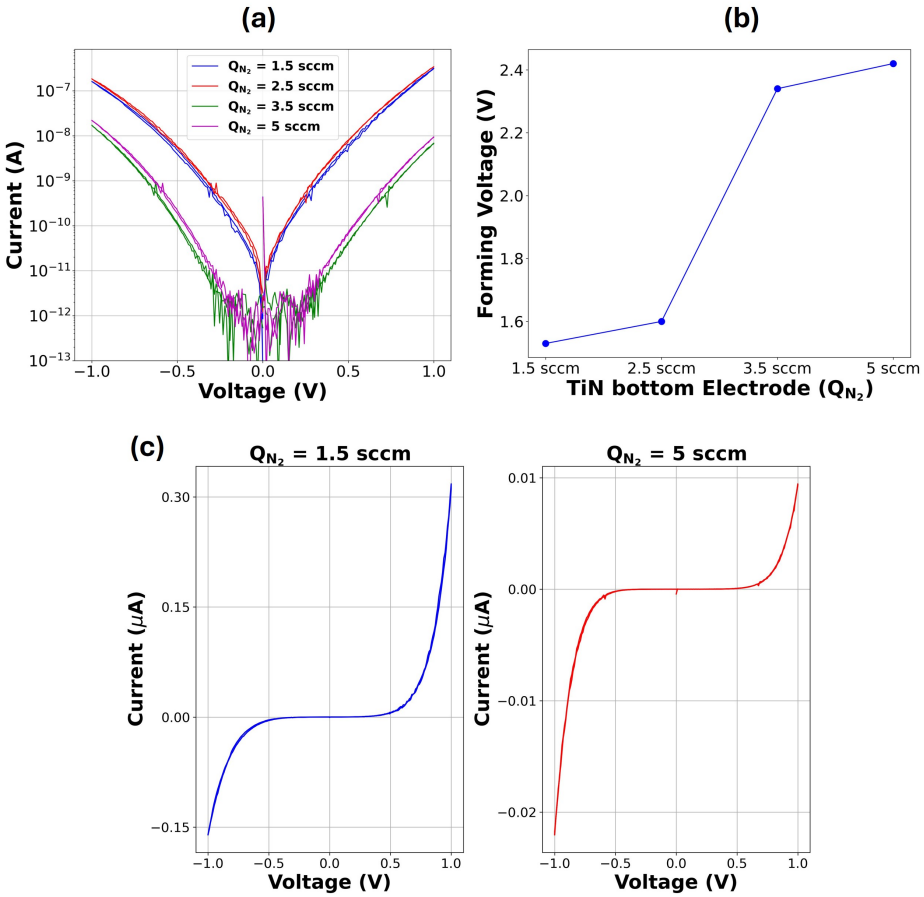


Figure 3.5: (a) I-V of fresh devices with varied TiN bottom electrode in log scale. (b) Forming voltage with varied TiN bottom electrode. (c) Comparison of I-V of fresh device between $Q_{N_2} = 1.5sccm$ and $Q_{N_2} = 5sccm$ in linear scale.

Previous studies have shown that the TiN (111) phase exhibits significantly higher oxidation resistance compared to other crystallographic orientations [116, 117], confirming that oxidation during fabrication or TiO_x deposition is not responsible for the reduced current density and elevated forming voltage observed in TiN electrodes with higher nitrogen content. Instead, the primary mechanism stems from nitrogen-dependent work function modulation: gate electrode studies demonstrate that higher nitrogen concentrations in TiN lead to an increased work function [118], thereby raising the Schottky barrier height at the TiN/TiO_x interface. This conclusion is further supported by the polarity shift in I-V asymmetry, which directly reflects a nitrogen-induced modification of interfacial conduction barriers.

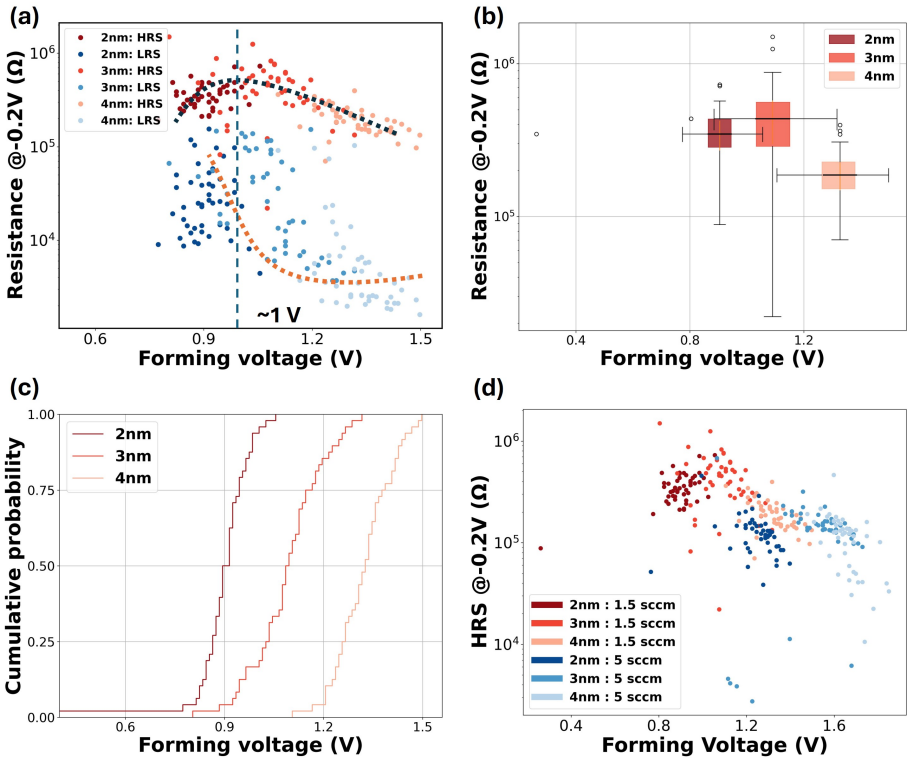


Figure 3.6: Electrical characteristics of devices with 1.5 sccm N_2 TiN bottom electrodes: (a) Resistance of LRS and HRS versus forming voltage . (b) Two-dimensional boxplot analysis correlating HRS resistance with forming voltage. (c) Cumulative distribution function (CDF) of forming voltages. (d) Comparison of resistance of HRS versus forming voltage for varied interface.

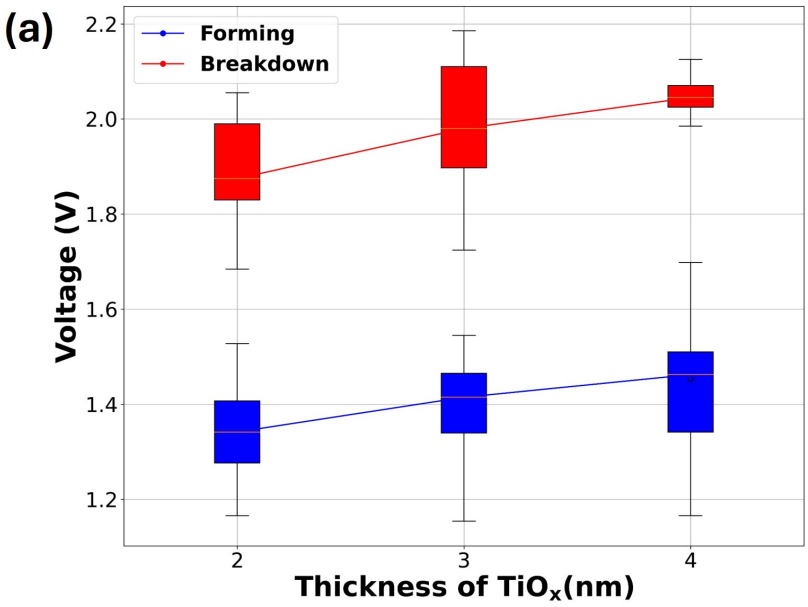
Additionally, TiN with lower nitrogen concentrations—which have a smaller fraction of the oxidation-resistant (111) phase—are more prone to oxidation. This likely results in a tapered oxygen profile near the interface, effectively reducing the effective barrier height. In contrast, nitrogen-rich TiN, with its dominant (111) orientation and higher oxidation stability, maintains a sharper interface and thus a higher barrier. This complementary effect further explains why higher nitrogen concentrations lead to both increased Schottky barriers and higher forming voltages.

The thickness of TiO_x on lower nitrogen concentration ($Q_{N_2} = 1.5\ sccm$) TiN bottom electrode is also varied to explore the dependence of HRS on the forming voltage (Fig. 3.6(a)), which shows similar trends as the one with higher nitrogen concentration. For the same thickness switching layer thickness, bottom TiN with lower nitrogen concentration requires lower forming voltage and has higher HRS resistance (Fig. 3.6(d)) while maintaining a similar level of LRS. As is illustrated in Fig. 3.8(a), for TiN with higher nitrogen concentration, a significant voltage drop is concentrated at the bottom interface due to the barrier height. For the same thickness of TiO_x, the electric field intensity required to trigger the forming is the same, the added voltage drop (V_ϕ) at the bottom interface would lead to the increase of the total voltage applied to the switching layer, thus enhancing the current overshoot after forming and leading to the decrease of the resistance of HRS. The results suggest that the work function of TiN should be decreased to lower down the forming voltage across the whole switching layer.

3.3 Influence of Al_2O_3 Insertion

The fabricated devices can only be formed by applying a positive bias to the top electrode. Forming with a negative bias leads to hard breakdown and the amplitude of negative breakdown voltage is larger than that of the positive forming voltage by > 0.5 V despite the symmetric device structure (Fig. 3.7(a)). Devices with different TiN bottom electrodes show similar asymmetry between forming voltage and breakdown voltage, which indicates that the asymmetry is not because of the difference of work function between the top and the bottom electrode, but due to the uneven distribution of defects in the switching layer. The deposition rate of TiO_x has been pre-calibrated and is found to be constant after a 10-minute conditioning period, suggesting that the asymmetry arises from the interfacial reactions between the switching layer and the electrodes. Possible interfacial reactions include the oxidation of bottom TiN during the TiO_x growth and the nitridation of TiO_x during the top TiN electrode deposition. To decouple the joint effects of these two possibilities, a thin aluminum oxide layer (1 nm) is deposited at top or bottom interface

using E-beam evaporation. The combined thickness of the dielectric layer ($TiO_x + Al_2O_3$) is kept at 3 nm.



(b)

Device Stack (BE/TiO_x/TE)	Forming Voltage (V)	Breakdown Voltage (V)
TiN/TiO_x/TiN	1.41	-1.98
TiN/Al_2O_3/TiO_x/TiN	1.47	-2.37
TiN/TiO_x/Al_2O_3/TiN	2.36	-1.96

Figure 3.7: (a) Forming voltage and breakdown voltage of device with varied thickness of switching layer (averaged over 50 devices). (b) Forming voltage and breakdown voltage of varied interface. (averaged over 50 devices)

The results (Fig. 3.7(b)) show that the insertion of Al_2O_3 layer at the bottom interface leads to the increase of the breakdown voltage but the asymmetry trend still exists, while the insertion of Al_2O_3 layer at the top interface leads to the increase of the forming voltage while the breakdown voltage remains the same. Less asymmetry is observed in the devices with Al_2O_3 layer at the top interface, which indicates that the nitridation of TiO_x during top TiN deposition is the main reason for the asymmetry and it can be used to lower down the forming voltage.

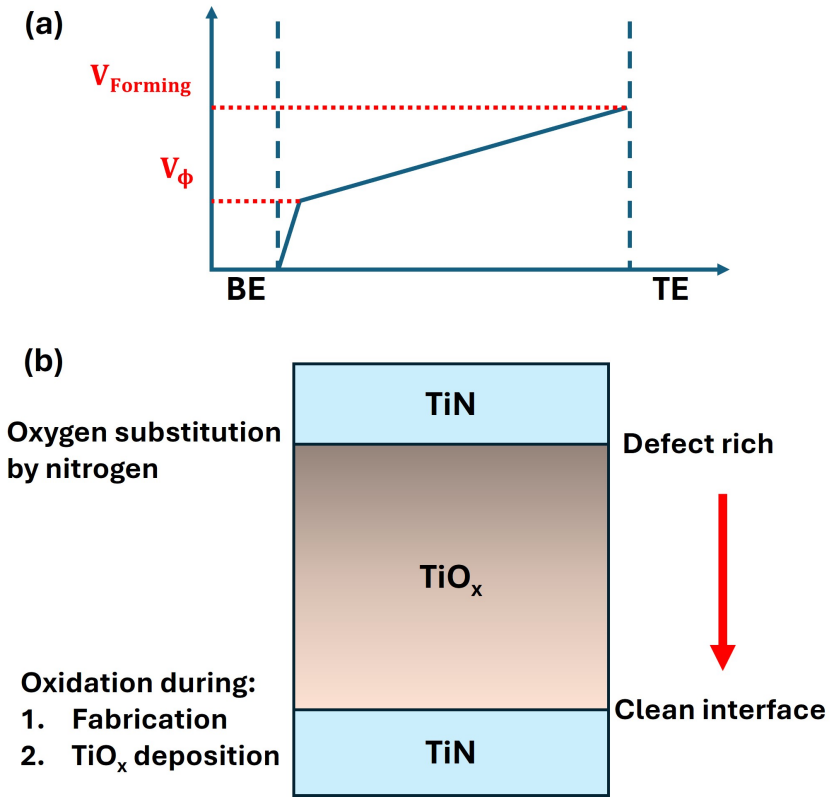


Figure 3.8: (a) Voltage drop diagram of devices under forming bias. (b) Defect profile of devices and possible hazards during material growth and fabrication.

Based on the results of Al_2O_3 insertion, nitrogen plasma treatment is applied to the top surface of TiO_x to enhance the generation of defects. After the deposition of TiO_x, a in-situ N_2 plasma treatment is conducted inside the sputtering chamber using RF back bias. To minimize the etching effect of the back bias, a small wattage (10 W) is applied under the pure nitrogen atmosphere. The result shows that a 3 nm TiO_x switching layer after the N_2 plasma treatment requires smaller forming voltage compared to even the 2 nm TiO_x devices, the average resistance of HRS of the top interface plasma treated 3 nm devices is higher than that of untreated 2 nm devices (Table. 3.1). Similar to other doping methods including the insertion of active metal layer or H_2 annealing, the N_2 plasma treatment on top interface of TiO_x (Fig. 3.8(b)) effectively substitutes O with N therefore introducing enough defects in the switching layer, lowering down the forming voltage without sacrificing the resistance of HRS.

Material Stack	Forming voltage (V)	HRS (Ω)
TiO_x (2 nm)	1.34 ($\sigma = 0.092$)	144 K ($\sigma = 57$ K)
TiO_x (3 nm)	1.41 ($\sigma = 0.087$)	358 K ($\sigma = 181$ K)
TiO_x (3 nm)/N_2 plasma	1.25 ($\sigma = 0.089$)	301 K ($\sigma = 132$ K)

Table 3.1: Effects of N_2 plasma treatment on top interface.

The insertion of Al_2O_3 layer at both interfaces illustrates the effectiveness of N_2 plasma generated defects at top interface and their role in conducting filament formation, however, the possible influence of oxidation of bottom TiN during TiO_x growth has not been ruled out. For (111) oriented TiN, the nitrogen-terminated interface has better oxidation resistivity compared to Ti-terminated interface[119]. Based on this, after the Ar plasma etching on bottom electrode, N_2 plasma treatment is applied before TiO_x deposition to passivate the surface with nitrogen termination. The results (Table. 3.2) show that despite the intrinsic difference between varied TiN bottom electrodes, the N_2 plasma treatment further lowers down the forming voltage and the resistance of maximum HRS is largely maintained, indicating that N_2 passivation of the bottom interface help mitigate the oxidation during TiO_x deposition. As expected, the already (111) oriented TiN ($Q_{N_2} = 5$ $sccm$) only sees little changes.

TiN BE(Q_{N_2})	Forming voltage (V)	HRS (Ω)
1.5 sccm	1.28 ($\sigma = 0.19$)	216 K ($\sigma = 112$ K)
1.5 sccm (N_2 treated)	1.09 ($\sigma = 0.17$)	208 K ($\sigma = 91$ K)
5 sccm	1.61 ($\sigma = 0.22$)	90 K ($\sigma = 33$ K)
5 sccm (N_2 treated)	1.58 ($\sigma = 0.17$)	92 K ($\sigma = 57$ K)

Table 3.2: Effect of N_2 plasma treatment on bottom electrode.

3.4 Summary

In this chapter, the dependence between the resistance of LRS and HRS and the amplitude of forming voltage is explored. The results show that the HRS resistance is inversely correlated with the forming voltage when the thickness of the switching layer is above 2 nm. Lower forming voltage is favored to achieve higher HRS resistance, which is essential for more energy-efficient CIM applications. The findings emphasize the importance of minimizing current overshoot during forming, which is the main reason for the decrease

of the resistance of HRS at higher forming voltage. Additionally, the peak amplitude of current overshoot is dependent on the forming voltage and the capacitance of the dielectric layer, indicating that future optimization should consider the dielectric constant of the switching layer. The scaling of forming voltage through thickness also shows that the leakage current of thinner dielectric layers is not negligible, leading to decreased HRS resistance, vanished ON/OFF ratio and poorer device stability.

Further scaling of the forming voltage of TiO_x based RRAM devices is achieved through the optimization of bottom TiN electrode and nitrogen plasma treatment on the electrode and the switching layer: The strong presence of (111) phase under higher nitrogen concentration in polycrystalline TiN leads to the increase of work function of TiN, which further increases the Schottky barrier height between the bottom electrode and the switching layer. Our results suggest that the work functions of TiN should be decreased to lower the forming voltage. Nitrogen plasma treatment of the top interface of TiO_x is effective in enhancing the generation of defects, which is essential for the formation of conducting filament. The nitrogen plasma treatment of the bottom interface can help passivate the TiN surface, which helps mitigate the oxidation during TiO_x growth, which further lowers down the forming voltage. We found that the best approach to reduce the forming voltage without sacrificing the HRS/LRS ratio is to use a lower N_2 concentration (e.g., 5%) in the bottom TiN growth to reduce the electrode work function of the bulk state, combined with N_2 plasma treatments before and after TiO_x deposition to passivate the bottom interface and enhance defect generation at the top interface. Our results proves the effectiveness of interface engineering in scaling down the forming voltage of TiO_x based RRAM devices, the results provide a guideline for the optimization of transition oxide based RRAM devices for advanced CMOS technology node integration and RRAM-based CIM applications.

Compared to other works (Table 3.3), the optimized devices in this work show lower forming voltage, lower operation voltage and higher HRS resistance, which is essential for more energy-efficient CIM applications. Futher work is needed to improve the endurance and retention performance of the devices, which is limited by the device fabrication process and possibly the material properties of TiO_x.

	This Work	UCSD[120]	IBM[114]	Weebit-Nano[121]	CAS[112]	Intel[60]
Switching layer	TiO_x	TiO_x	HfO_x	SiO_x	HfO_x	TaO_x
Device Area (μm^2)	0.36	0.49	0.0225	0.25	0.022	0.049
Forming Voltage (V)	< 1.5	~ 2	~ 3.3	< 4.6	~ 1.5	-
HRS (Ω)	1 M	1M	300 K	$<$ 100 K	100 K	40 K
LRS (Ω)	10 K	10 K	10 K	$<$ 10 K	10 K	5 K
Endurance	10^3	-	-	10^6	10^5	10^4
Retention	10^4 s @ RT	10^4 s @ RT	-	10 Year @ 85°C	10 Year @ 85°C	10 Year @ 85°C

Table 3.3: Comparison of optimized RRAM devices with other works.

Chapter 4

State-Aware Multi-bit Write Algorithm

As a potential candidate for non-volatile memory in sub-40 nm CMOS technologies, previous optimizations of RRAM technologies have mostly focused on single bit storage[122, 58], which limits its application scenario given the growing demand for memory density. The programming of multi-bit RRAM devices generally requires repeated write-verify steps until the device conductance reaches the target conductance range. Due to the limited endurance of RRAM devices, the number of write-verify steps needs to be minimized to prolong the device lifetime and increase the programming speed. Different strategies have been proposed to decrease the programming steps/time needed, and they can be divided into the following categories:

- *Incremental step pulse programming (ISPP)*: The SET/RESET voltage[74, 70] or the gate voltage in 1TxR configuration is increased after each write-verify step until the target conductance range is reached.
- *Range dependent tuning*: Performing coarse tuning[85, 83] first to bring the device conductance closer to the target conductance range and fine tuning (ISPP-like) later. The coarse tuning is normally achieved by mapping between the gate voltage and target conductance range in a 1TxR structure, using a fixed combination of gate voltage and SET voltage.

This chapter introduces our investigation into mitigating fast relaxation and overshoot issues in multi-bit RRAM programming. While minimizing write-verify cycles is essential

for efficient multi-bit storage, the rapid conductance decay (within 1 s) after programming remains a major challenge. Here, we characterize the fast relaxation behavior, highlighting its dependence on both the conductance state and the writing direction. Our findings suggest that terminating programming with a RESET operation can reduce fast relaxation; however, this approach alone tends to increase programming time due to frequent overshoots resulting from the continual ramp-up of bias voltage in ISPP-like schemes.

To overcome these limitations, we explore the state-dependent modulation of RESET by varying pulse amplitude and duration, thereby revealing that voltage amplitude is best suited for coarse tuning while the logarithmic dependence of pulse time offers improved fine tuning. Leveraging these insights, we propose a state-aware multi-bit write algorithm that combines the advantages of modified ISPP for SET operations with a tailored RESET scheme. The RESET process incorporates dual tuning phases: coarse adjustment via voltage amplitude adaptation based on the conductance difference, and fine tuning via pulse duration calculated from logarithmic time dependence, supported by status flag feedback.

Additionally, we introduce an electrical annealing method to further enhance the long-term stability of multi-bit programming, demonstrating its effectiveness in extending the refresh period for CIM applications.

4.1 Write Direction Dependent Fast Relaxation

To characterize the fast relaxation after programming, a read train pulse is applied after the write-verify step to capture the conductance change (as illustrated in Fig. 4.1(a)). The sampling time of each read pulse is 10 μs and the read period is 140 μs, 1000 read pulses are applied. The target conductance states (G_{Target}) are ranged from 10 μS to 50 μS (read at 0.2 V) to evaluate the influence of conductance states on fast relaxation. The measurement was conducted on over 30 devices, and the results are categorized based on the terminated write direction. The summarized results are shown in Fig. 4.1(b). It can be observed that fast relaxation is dependent on the terminated write direction, with the conductance drift of SET being more severe compared to that of RESET. Additionally, fast relaxation also depends on the conductance state, which indicates that fast relaxation is correlated with the defects (oxygen vacancy) contained in the switching layer [79, 81, 82]. The observations suggest that oxygen vacancies generated by SET favor recombination with oxygen ions[80], leading to conductance drift post-writing.

The measurement results suggest that RESET should be used to terminate the programming to mitigate the influence of fast relaxation. As discussed in [80], a performance

penalty is expected if directly incorporating this strategy into multi-bit programming algorithms like ISPP or RADAR [85], due to the overshoot of RESET. Therefore, a better understanding of conductance modulation of RESET is needed to minimize the likelihood of overshoot.

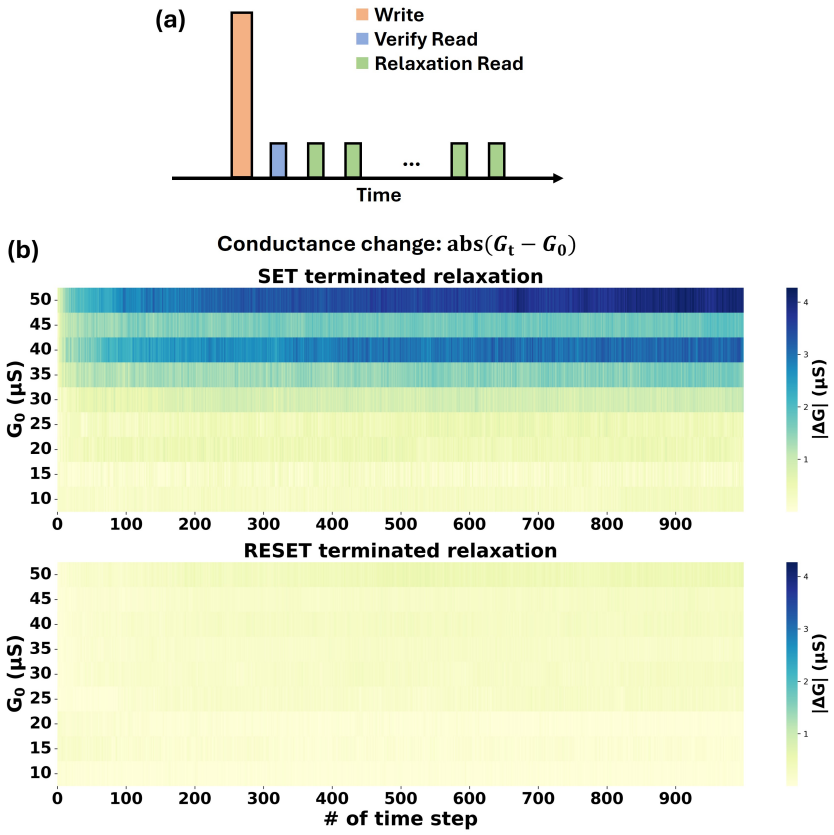


Figure 4.1: (a) Schematic of pulse train used for characterization of fast relaxation. (b) Average conductance change after programming with varied G_{Target}. Top figure is programming terminated with SET while the bottom one is terminated with RESET.

4.2 State Dependent Conductance Modulation of RESET

To explore the conductance modulation of RESET, devices are first programmed to a certain conductance state, then a RESET pulse is applied, followed by a read pulse to evaluate the conductance change. Various combinations of initial conductance states, pulse voltage amplitudes and pulse time of RESET are applied. The initial conductance states are ranged from 10 μS to 50 μS with step size being 10 μS, the voltage amplitudes are ranged from -0.8 V to -1.2 V with step size being 0.1 V and pulse time is ranged from 100 ns to 100 μs for each order of magnitude. Measurements were conducted on over 30 devices, a three-dimensional mapping of the average conductance after RESET versus pulse voltage amplitude and pulse time is plotted in Fig. 4.2, the average value of the conductance after RESET is represented by the colormap and three-dimensional surfaces for different initial conductance states are stacked for comparison.

The average conductance change versus pulse time is plotted in Fig. 4.3(a), showing a logarithmic time dependence of the conductance change: $\Delta G \propto log(t/t')$, it should also be noted that the amplitude of the conductance change is dependent on the voltage applied. As shown in Fig. 4.3(b), there exists a correlated relationship between the ratio of conductance change and the applied voltage amplitude. A similar linear relationship is also observed when comparing the amplitude of conductance change with the initial conductance state (Fig. 4.3(c)), the fitting slope of the linear dependence also shows a linear dependence on applied voltage ($k \propto (V - V_0)$). The relationship between ΔG, G_0 and V can be generalized to: $\frac{\Delta G}{G_0} \propto (V - V_0)$.

The results indicates that the conductance change before and after RESET not only depends on the RESET voltage amplitude and time but also on the initial conductance state. If the current conductance state (G_0) and target conductance state are known, the required voltage amplitude can be estimated based on the linear relationship mentioned above. Following this strategy, the pulse time can also be estimated based on the previous RESET operations applied to the devices. Similar dependence on voltage amplitude and pulse time of RESET have also be observed on HfO_2[123, 124] and TaO_x[26] based RRAM devices, indicating that this strategy can be generalized to other metal oxide based RRAM technologies.

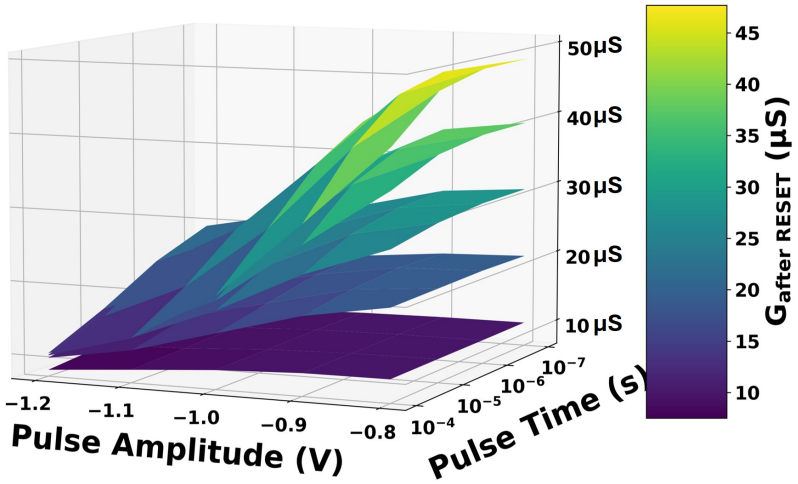


Figure 4.2: 3D mapping of average conductance after RESET versus pulse amplitude and pulse time, with varying initial conductance states.

4.3 Design of State Aware Multi-bit Write Algorithm

Based on the analysis in the previous section, we propose a state-aware multi-bit write algorithm, as illustrated in Fig. 4.4. The proposed algorithm is composed of three modules:

- Main body (Fig. 4.4(a)). The main body follows the design of ISPP but taking the mitigation of fast relaxation into consideration. The programming loop starts with a read operation to determine the writing direction. If the read conductance state (G_n) is within the target margin but was terminated with a SET operation, it is further SET to higher conductance state in the next programming cycle. If RESET is needed, the writing parameters for RESET are output from the RESET parameter calculation module. After the SET or RESET operation, the status flag update module would update the status flags and tune the writing parameters accordingly. The writing is marked as success only when the conductance state is within the target margin and the last writing direction (WD) is RESET.

- Status flag update module (Fig. 4.4(b)): This module is designed to fine tune the programming parameters (mainly for SET operations). It consists of two status flags: Oscillation and Monotone. Oscillation is defined as when the read current oscillates around the target current with SET/RESET pairs applied to the devices, indicating that the bias amplitude of either SET or RESET is too large to fine-tune

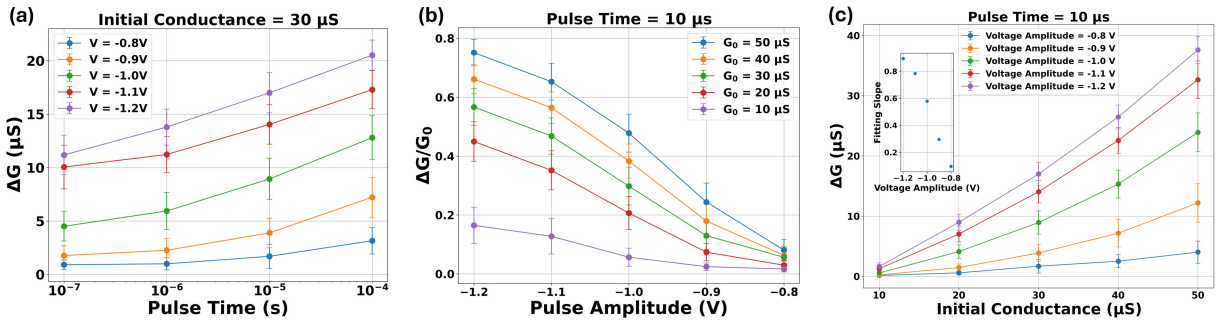


Figure 4.3: (a) Conductance change (ΔG) versus pulse time, initial conductance state before RESET is 30 μS. (c) The conductance change ratio versus the pulse amplitude. (d) Conductance change versus the initial conductance with the inset figure being the fitting slope versus the voltage amplitude (average over 30 devices).

the resistance state. Monotone is defined as when the read current consistently stays above or below the target value, which indicates that the bias amplitude of applied SET (or RESET) pulse is insufficient to trigger the resistance change. Bias amplitude of SET/RESET is tuned accordingly if either of the status flag is triggered.

- RESET voltage and time calculation module (Fig. 4.4(c)). This module calculates the optimal RESET parameters based on the linear relationships observed in the conductance modulation study. For each RESET writing, if the conductance difference, defines as $\Delta G = |G_{target} - G_n|$, is within the predefined fine-tune conductance range (G_{FT}) and the previous write direction was SET, the voltage remains unchanged. If ΔG exceeds G_{FT}, the voltage amplitude of RESET is calculated based on the linear relationship between $\frac{\Delta G}{G_0}$ and V. In both cases, pulse time is set as t_min. If ΔG is within G_{FT} and the previous write direction is also RESET, the new pulse time is calculated based on the observed logarithmic dependence of pulse time.

An example of the proposed algorithm is shown in Fig. 4.4(d), the monotone and oscillation flags are triggered in the presented examples, and the tuning of RESET voltage amplitude as well as the RESET pulse time based on state-dependent calculation is also illustrated.

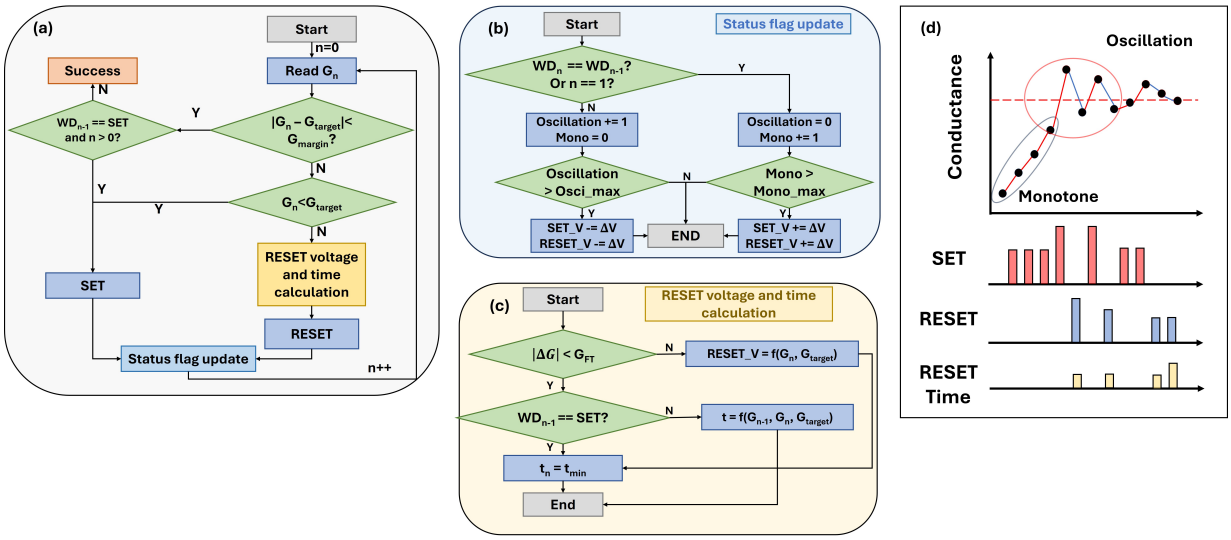


Figure 4.4: (a) Flowchart of the proposed multi-bit writing algorithm. (b) Flowchart of status flag update module. (c) Flowchart of RESET voltage and time calculation module. (d) Example of programming process. The bar height of SET/RESET (RESET time) represents the voltage amplitude (pulse time) applied. The triggering of monotone and oscillation status flags are illustrated in the conductance change plot.

4.4 Benchmark of Multi-bit Writing

To evaluate the performance of our proposed state-aware multi-bit write algorithm, we conducted benchmarks using TiO_x devices with initial conductance states randomly assigned between 10 μS and 50 μS. The ISPP algorithm used in this work shares the design of our proposed algorithm but without the RESET parameter calculation module. The pulse time of RESET in ISPP is fixed at 1 μs. The shared parameters between ISPP and this proposed algorithm are listed in Table. 4.1.

For the RESET parameter calculation module in this proposed algorithm: the fine tune conductance margin (G_{FT}) is set as 2.5 μS; the fitting equation used for RESET voltage calculation is:

$$\frac{|G_n - G_{target}|}{G_n} = k_v * (V_{RESET} - V_{min}) \tag{4.1}$$

where $k_v = -2\ V^{-1}$ and $V_{min} = -0.6\ V$ based on the measurements mentioned in previous

RESET start voltage	-1 V	Minimum conductance state	10 μS
RESET voltage step	0.05 V	Maximum conductance state	50 μS
SET start voltage	0.9V	Conductance step	2.5 μS
SET voltage step	0.05 V	Conductance margin	0.5 μS
SET pulse time	10 μs	Oscillation count max	3
Read voltage	0.2 V	Monotone count max	4
Maximum writing trials	300		

Table 4.1: Shared parameters between ISPP and proposed algorithm.

section; the fitting equation used for RESET time calculation is:

$$\Delta G = k_t * log(t) + b \tag{4.2}$$

the fitting slope k_t can be calculated from the last RESET writing:

$$k_t = \frac{G_n - G_{n-1}}{log(t_{n-1} + t_h) - log(t_h)} \tag{4.3}$$

Where G_n is current conductance state, G_{n-1} is the conductance state before last RESET operation and t_{n-1} is the pulse time of last RESET operation. t_h is set to 100 ns, which is assumed that no change in conductance would be expected within this time frame for each writing. The calculated time for RESET would be:

$$t_n = (e^{\frac{G_n - G_{target}}{k_t}} - 1) * t_h \tag{4.4}$$

If the writing direction before RESET is SET, the pulse time of RESET is set to $t_{min} =$ 100 ns. To eliminate the influence of read noise and writing variation, moving average of the slope k_t is used for the calculation of RESET time and the maximum pulse time applied is capped at 10 μs.

To benchmark the multi-bit writing performance of proposed algorithm versus ISPP, a bit pattern for 4-bit encoding is used. In the designed bit pattern, devices are encoded from (0000) to (1111) back and forth with varied step size being 1, 2, 4 and 8, this would help evaluate the performance of multi-bit writing algorithm with respect to state jump, especially for the intermediate states between (0000) and (1111). The conductance state to represent (0000) is set as 10 μS and that of (1111) is set as 50 μS, each conductance state step (one bit step) is 2.5 μS. The distribution of read conductance after programming is

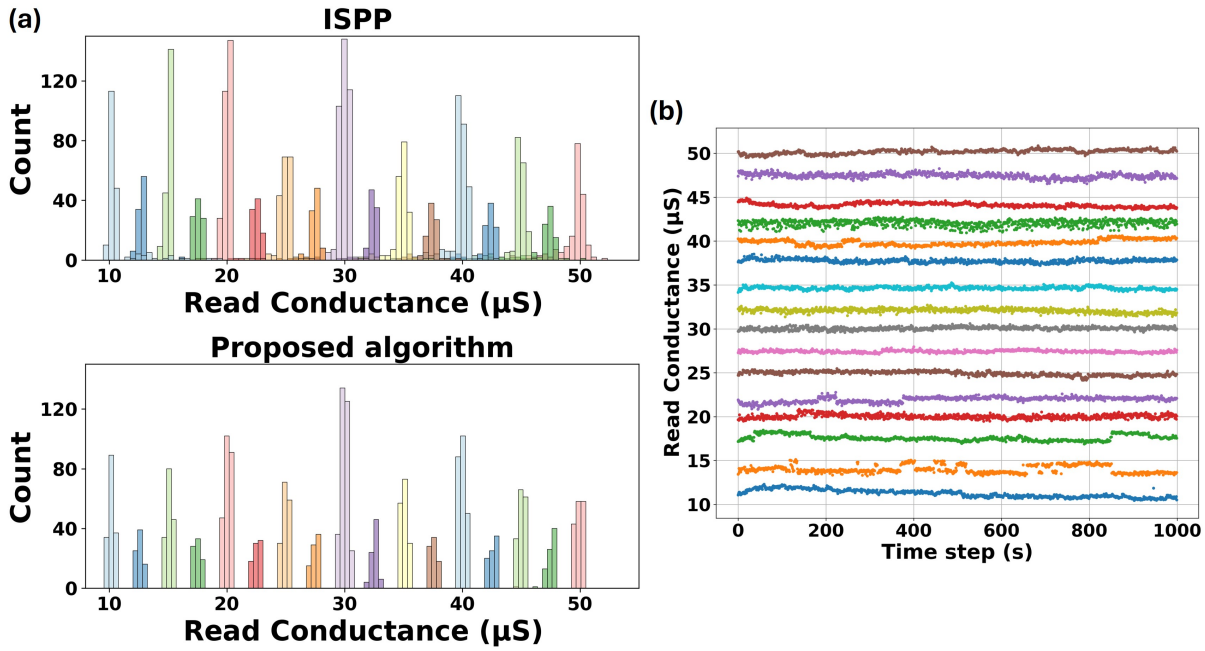


Figure 4.5: (a) After programming state distribution of ISPP versus proposed algorithm. (b) Retention test of 16 conductance states programmed using the proposed algorithm.

shown in Fig. 4.5(a): The proposed algorithm exhibits clear boundaries between states whereas ISPP shows state overlap at both HCS and LCS. The state overlap in ISPP is mainly due to reaching the maximum writing trials before the target conductance state is reached, which originates from repeated writing between SET and RESET to make sure the programming is RESET-terminated. Multi-bit retention measurement (Fig. 4.5(b)) has also been conducted to demonstrate the long time stability of programmed conductance states.

Apart from the conductance state distribution, two performance metrics are used to benchmark the performance: the total RESET time and the total number of writing steps. The total RESET time of ISPP is calculated as: # of RESET $\times$ 1 μs, while that of the proposed algorithm is the sum of RESET time applied. The total programming steps is the sum of total SET and RESET operations applied to the devices. Both performance metrics are measured and categorized according to bit steps and encoding directions. For each combination of bit step and encoding direction, the worst-case scenario (the maximum number of programming steps or the longest RESET time) of the averaged performance metric is used for comparison. Fig. 4.6 presents the summarized barplots (grey and orange

bars) of the two performance metrics against bit steps, with the corresponding encoding direction indicated below the x-axis of each chart. Compared to ISPP, the proposed algorithm reduces the number of programming steps by more than 2.4X and decreases the total RESET time needed by more than 2.2X. ISPP incurs a performance penalty when the step size is small (1, 2 or 4) or when the encoding direction is from HCS to LCS, highlighting the importance of precise control over writing parameters, particularly the RESET process, to achieve optimal performance.

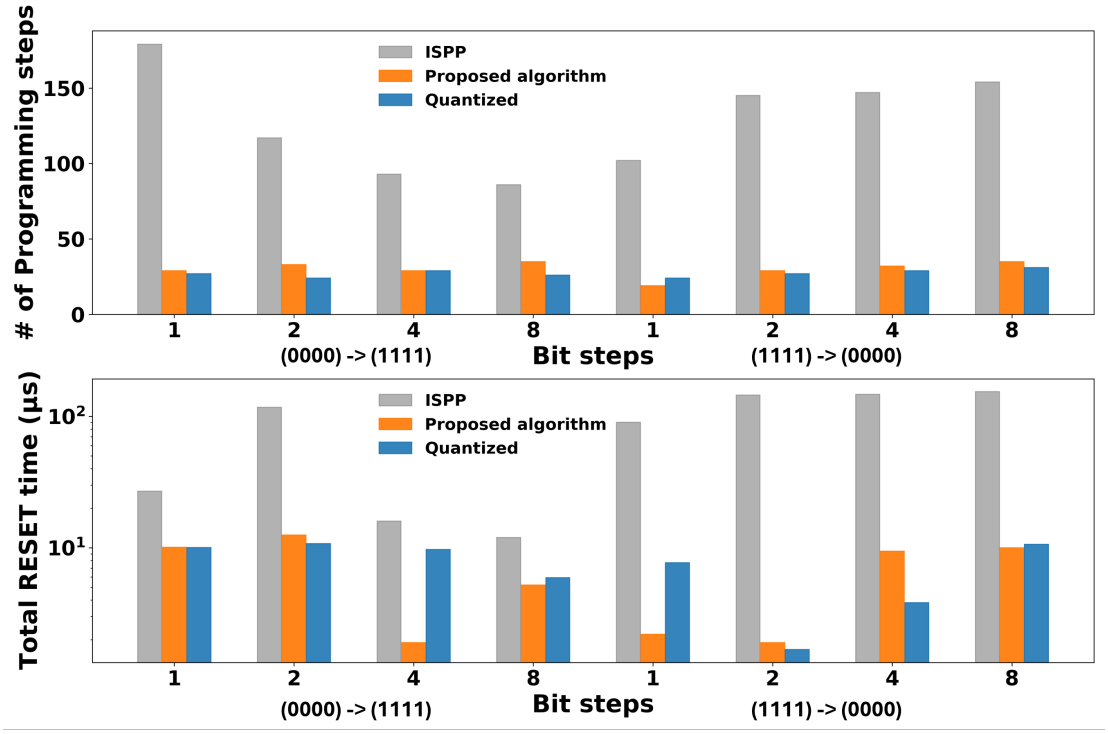


Figure 4.6: Benchmark of number of programming steps and total RESET time between ISPP, proposed algorithm and quantized proposed algorithm

The proposed algorithm outperforms the ISPP due to its state-aware calculation of RESET parameters, which is implemented using a parameter analyzer and a PC. The parameter analyzer samples the conductance state (current) with a 14-bit analog-to-digital converter (ADC), outputs the voltage waveform through a 12-bit digital-to-analog converter (DAC), and stores all data in a 32-bit floating-point format, all of which are resource-intensive in real hardware implementations. To assess the influence of ADC/DAC quantiza-

tion on performance and evaluate hardware overhead, we consider a multi-bit programming module comprising:

- ADC for current sampling: A 9-bit ADC with a quantization resolution of 50 nA and a measurement range from 50 nA to 20 μA.
- DAC for voltage output: A 6-bit DAC with a quantization resolution of 0.01 V and an output range from -0.6 V to -1.2 V.
- Pulse time control: Managed by a RISC-V CPU[121] controlling the number of clock cycles, with a clock period of 20 ns and an output range from 20 ns to 10 μs.

The quantized version of our proposed algorithm is implemented by applying ceiling function to all parameters (sampled current, pulse voltage amplitude and pulse time) during programming. The performance of the quantized algorithm is evaluated using the same performance metrics as described above, and the results (Fig. 4.6, blue bars) indicate that similar performance is achieved with ADC/DAC quantization loss taken into consideration. These results demonstrate that the proposed algorithm maintains its advantages even with limited hardware precision, requiring no extra hardware overhead. Therefore, it is practical for real-world implementations and can be efficiently integrated into existing systems.

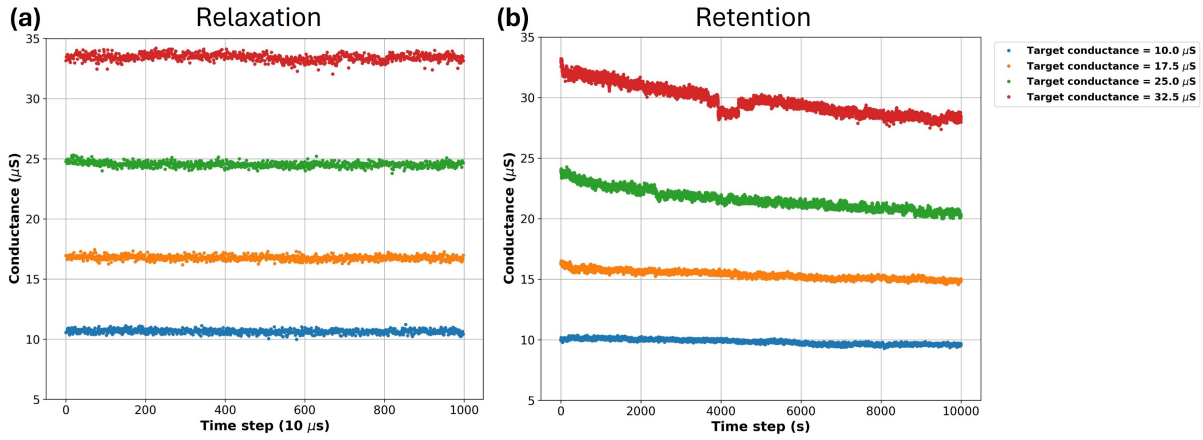


Figure 4.7: (a) Fast relaxation results and (b) retention results of conductance states after programming.

4.5 Electrical Annealing to Mitigate Retention Loss

The proposed algorithm mainly focuses on the optimization of multi-bit programming and fast relaxation after programming, however, the long term retention loss of the conductance states has not been addressed. The devices are programmed to 4 different conductance states (10 μS, 17.5 μS, 25 μS and 32.5 μS) using the state aware multi-bit programming algorithm. Read voltage is set to - 0.2 V as more severe RTN behavior is observed when using read voltage of 0.2 V (the read voltage polarity dependence of RTN remains to be investigated). The relaxation results and the retention results are summarized in Fig. 4.7. Minor decay of conductance states is observed during the fast relaxation test, while a significant retention loss is observed after 10000 s. The retention loss is found to be state dependent, with the larger conductance states showing more severe retention loss. The retention loss is also found to be unidirectional, all the states are shifted towards the lower conductance states (higher resistance states).

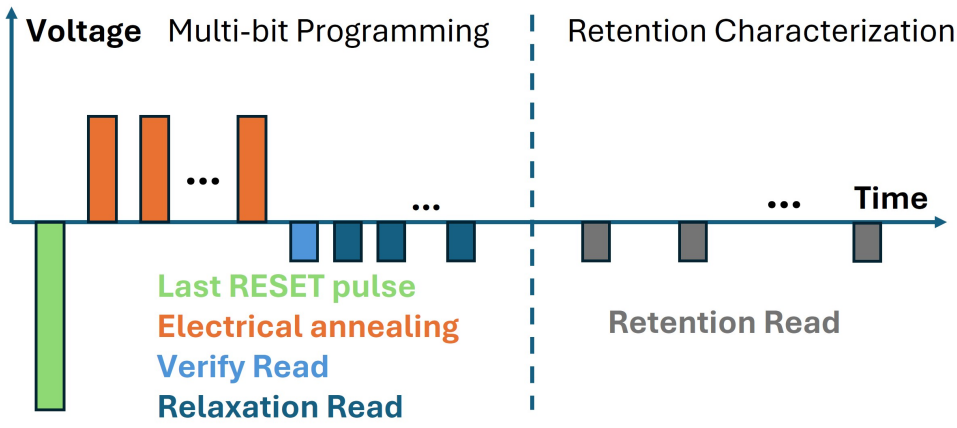


Figure 4.8: Customized voltage waveform for electrical annealing.

Similar to the SET induced fast relaxation decay[80, 81, 125], the monotonic drift and state dependence of the retention loss indicates that the retention loss is related to the oxygen vacancy around the conducting filament region, especially the dangling ones. The spontaneous oxidation of the dangling oxygen vacancies is thermally activated, which leads to the retention loss. Since the process is thermally driven, thermal annealing could significantly accelerate the retention loss, particularly in the initial stages, where the decay rate is fastest due to the anti-exponential time dependence of conductance changes. By expediting this early-stage relaxation, annealing leads to a more stable conductance state,

making subsequent readings more reliable. However, annealing is usually performed at elevated temperature, which not practical for real applications.

Therefore, an electrical annealing method based on voltage pulses is proposed to relieve the retention loss. To overcome this, we propose electrical annealing via sub-SET voltage pulses, which locally heats the filament region to mimic thermal effects without global temperature increases. As shown in Fig. 4.8, these pulses are applied after the final RESET pulse in the programming loop. This method accelerates the relaxation of dangling vacancies early on, stabilizing conductance faster while avoiding unintended conductance changes.

The sub-SET voltage pulse width is set to 10 μs, the voltage amplitude (0.4 V $\sim$ 0.75 V) and the number of sub-SET pulses (10 $\sim$ 100) are varied. The results show that the retention loss can be greatly mitigated when the voltage amplitude is set around 0.7 V and the number of sub-SET pulses is set to 10 (total biasing time = 100 μ s). The 4-bit retention results without and with electrical annealing are shown in Fig. 4.9. Electrical annealing significantly mitigates retention loss, with conductance states showing minimal degradation after 1000 seconds. Curve fitting with exponential decay function ($\frac{C(t)}{C(0)} = exp(-\frac{t}{t_0})$, t_0 is the time constant) is applied to the retention results, the fitting results show that the time constant of the retention loss is increased by around 2.5 X after electrical annealing.

The retention loss is mainly caused by the oxidation of the dangling oxygen vacancies around the conducting filament region. The dangling oxygen vacancies are thermally activated, the sub-SET voltage pulses can induce thermal effect to speed up the localized oxidation of the dangling oxygen vacancies. Moreover, the electric filed direction of the sub-SET voltage pulses is opposite to the spontaneous diffusion direction of the oxygen vacancies, which helps to stabilize the dangling oxygen vacancies. The combined effect of the thermal annealing and the electric field helps to speed up the retention loss and stabilize the dangling oxygen vacancies, which leads to a more reliable verify reading, and thus provides a more accurate programming of the conductance states.

4.6 Benchmark of CIM Performance

The electrical annealing helps to mitigate the retention loss of the conductance states, however, the retention loss, especially for higher conductance states, is still observed. If the multi-bit devices are used for CIM applications, a periodic refresh operation is needed to maintain the conductance states. Frequent refresh operations would degrade the endurance of the devices as well as increase the complexity of the CIM systems. To evaluate the

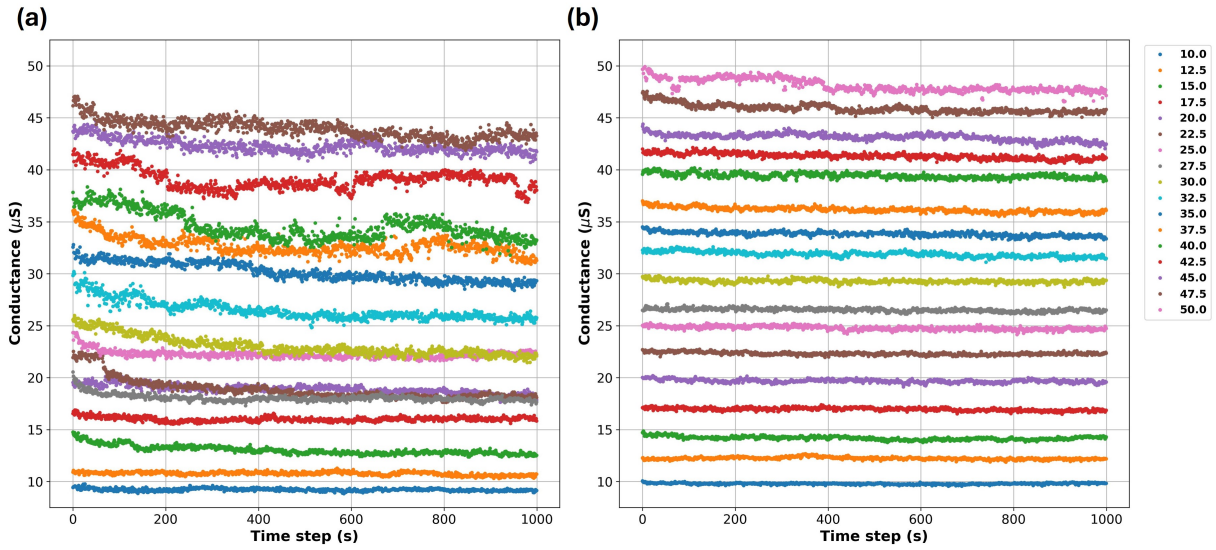


Figure 4.9: 4-bit Retention results (a) without and (b) with electrical annealing.

influence of retention loss in CIM applications and how the electrical annealing helps to elongate the refresh period, a CIM simulation based on NeuroSim is conducted.

It is assumed in the simulation that the CIM chips consist of compute tiles, control logic and global buffers. Inside the compute tile, the accumulation module and tile buffer are used to sum up partial sums of processing elements (PE) and output the summing results. The PE consists of several RRAM arrays, accumulation module, PE buffer and control logic. Inside the RRAM array: RRAM devices sharing the same wordline are grouped to represent the weights of the neural network (size of the group depends on the bit precision of the weights and RRAM devices), the input data is applied to the bitline and the output summing current is fed to ADC for current/voltage sampling. Limited by the area of ADC, each ADC is shared by several source lines, the access of which are controlled by the sourceline switch matrix and column mux. The output of the ADC is fed to the accumulation module, which carries the shift-add operation and output the partial sums to the PE buffer for further processing.

In the simulation setup (Fig. 4.10(a), the influence of ADC quantization, the size of the RRAM array, the shift-add operations inside the RRAM array and summing of the partial sums inside the PE are considered. The weights of the neural network are quantized to 8-bit and then represented by the conductance states of the RRAM devices (G_{max} being 1, G_{min} being 0). The weight mapping process (Fig. 4.10(b)) includes:

- Convert the 8-bit floating weight ($[-1, 1]$) to 8-bit integer ($[0, 255]$) through linear mapping.
- Map the 8-bit integer to ($floor(8, n)$) devices (n is the number of bits stored in one single RRAM device), with each RRAM devices represent n bit. For example, if the RRAM device stores 2-bit per cell, 4 devices are used to represent the 8-bit integer. The n-bit integer (w) to be mapped is converted to the conductance states of the RRAM devices, where the conversion equation is $G = \frac{G_{max}-G_{min}}{2^n-1} * w + G_{min}, G \in [G_{min}, G_{max}]$.
- If the retention loss is considered during the simulation, a conductance degradation model is used to scale the mapped conductance. Variation model based on gaussian distribution is also applied to the mapped conductance state based on the input variation parameters.
- Computation (matrix multiplication) based on mapped conductance states is conducted, and then recover the computation result (based on mapped conductance state) to the results represented using 8-bit integer ($[0, 255]$) first, and then to results based on 8-bit floating number ($[-1, 1]$).

During the simulation setup, a custom retention model is applied to the conductance states to emulate the retention loss, the retention loss is assumed to be conductance state dependent ($\frac{G_t}{G_0} = exp(-\frac{t}{t_0})$) and the time constant (t_0) of the retention loss is the same across all the devices. The dataset used for the simulation is CIFAR-10, which has 60000 32x32 RGB images, 10000 images are used as the test dateset. Pre-trained VGG8 model (Table. 4.2) is used for the simulation, the baseline accuracy of the pretrained VGG8 model is around 92 %. The retention time t is varied from 0.1 s to 10000 s to evaluate the influence of retention loss on classification accuracy, retention with and without electrical annealing is compared. The results is summarized in Fig. 4.11, the classification accuracy of the CIFAR-10 dataset is plotted against the retention time, and the electrical annealing method helps to elongate the refresh period by 3 X (the refresh standard is set when the accuracy drops to 80%). The results indicate that the electrical annealing is effective in mitigating the retention loss of the conductance states and helps to elongate the refresh period in CIM applications.

The ADC precision and the storage bits of the RRAM devices are also varied to evaluate the influence of hardware resources on the classification accuracy. The results are summarized in Fig. 4.12, the classification accuracy is plotted against the retention time with ADC precision bits being 5 and 8. The results show that the for multi-bit RRAM

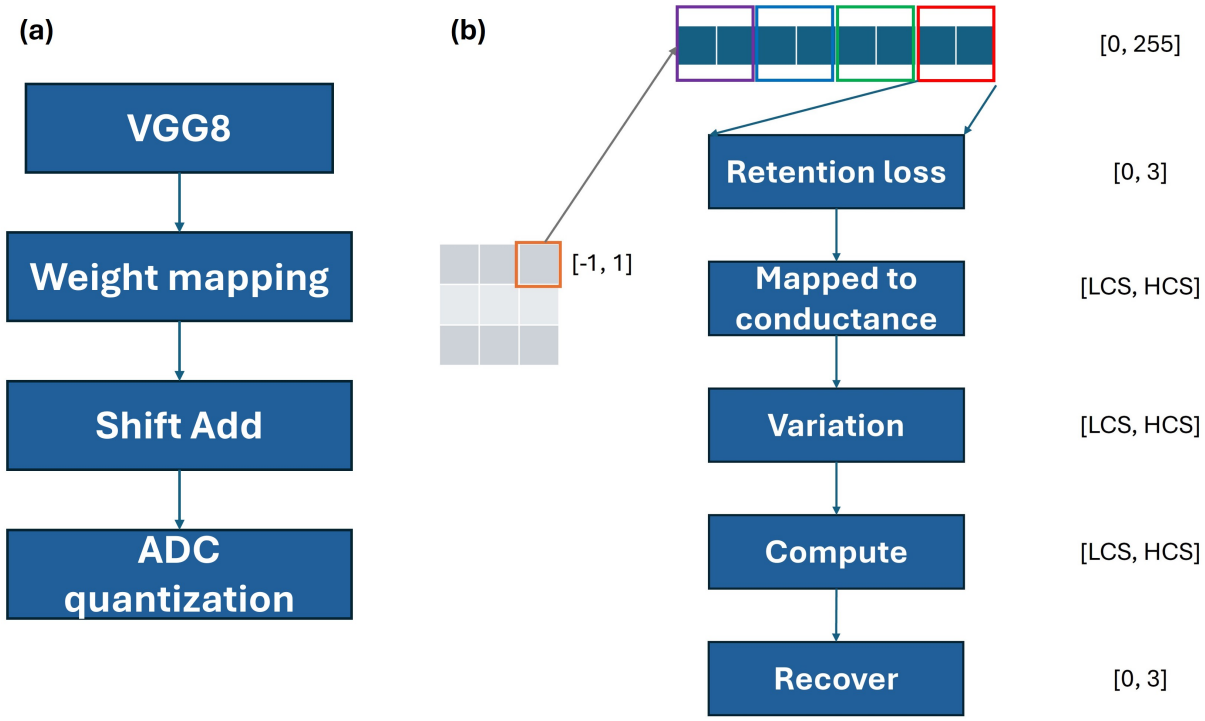


Figure 4.10: (a) Schematic of the CIM simulation setup. (b) Schematic of the weight mapping.

devices (bits >= 1), the classification accuracy is less sensitive to the ADC precision, and the accuracy is mainly determined by the retention time. However, for single-bit RRAM devices, the classification accuracy is more sensitive to the ADC precision. With the same retention loss, the accuracy is significantly improved when the ADC precision is increased from 5 to 8 bits. This is mainly due to the position of ADC quantization in the computation pipeline: the current flowing through the devices is sampled by the ADC first, then the quantized results will be shift-add in the accumulation module.

- For 4-bit devices, the summed result is $2^4 \times [I_1] + [I_0]$.
- for single-bit devices, the summed result is $\sum_{n=0}^{7} 2^i \times [I_i]$.

$[I_i]$ is the quantized current sampled by the ADC where i represent the bit position, I_0 is the least significant bit. For each sampled current I_i, the quantization error $(I_i - [I_i])$ of the ADC would be amplified by 2^i in the shift-add accumulation module. Single bit devices

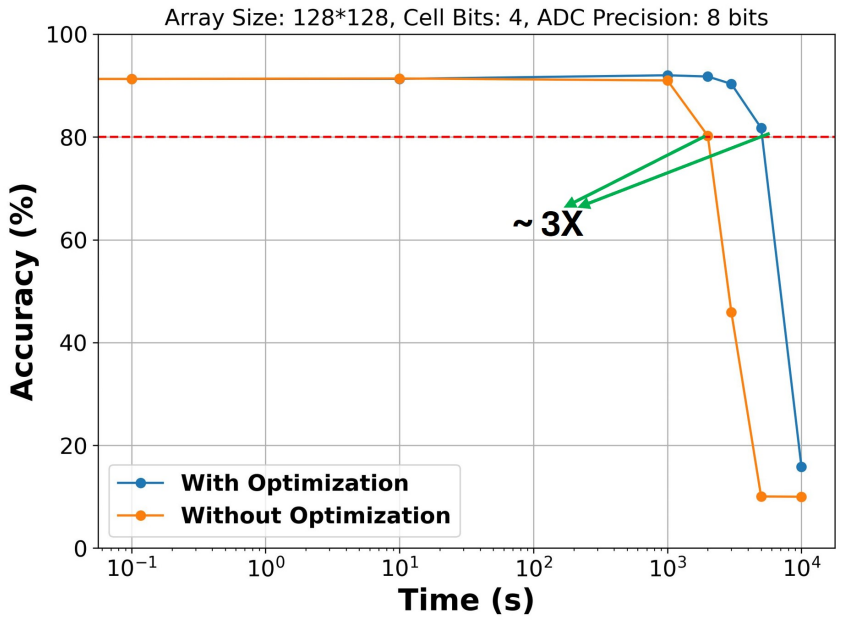


Figure 4.11: CIM simulation results of classification accuracy versus retention time.

would have more quantization error compared to multi-bit devices, leading to the increased sensitivity of the classification accuracy to the ADC precision. The results suggest that the adoption of multi-bit RRAM devices would lower the requirement of the ADC precision, which would increase the compute parallelism and reduce the energy consumption of the CIM systems.

4.7 Summary

In this chapter, fast relaxation after multi-bit programming is found to be depend on the write-direction, with RESET preferred as the termination writing direction to minimize the influence of fast relaxation. State dependent conductance modulation of RESET is explored, the observed conductance state dependence of voltage amplitude and logarithmic dependence of pulse time have been proposed to accelerate the multi-bit programming. Based on these findings, we proposed a state-aware multi-bit write algorithm that accelerates multi-bit programming by incorporating state-dependent calculations of RESET parameters. Unlike ISPP, our algorithm adjusts the RESET parameters based on the cur-

Layer (type)	Kernel Size / Number	Output Shape	Param #
Conv2d	3x3, 128	[1, 128, 32, 32]	3,584
Conv2d	3x3, 128	[1, 128, 32, 32]	147,584
MaxPool2d	2x2	[1, 128, 16, 16]	–
Conv2d	3x3, 256	[1, 256, 16, 16]	295,168
Conv2d	3x3, 256	[1, 256, 16, 16]	590,080
MaxPool2d	2x2	[1, 256, 8, 8]	–
Conv2d	3x3, 512	[1, 512, 8, 8]	1,180,160
Conv2d	3x3, 512	[1, 512, 8, 8]	2,359,808
MaxPool2d	2x2	[1, 512, 4, 4]	–
Linear	–	[1, 1024]	8,389,632
Linear	–	[1, 10]	10,250
Total params: 12,976,266			

Table 4.2: VGG8 Architecture Summary

rent conductance state, target conductance state, and conductance difference. This precise parameter selection effectively eliminates oscillations between SET and RESET operations.

Experimental validation using 4-bit encoding and designed bit patterns demonstrates that the proposed algorithm outperforms the ISPP approach, achieving over 2.4 times reduction in programming steps and over 2.2 times reduction in total RESET time. Furthermore, we assessed the influence of ADC/DAC quantization loss on the proposed algorithm, showing that it performs effectively without requiring additional hardware overhead. The proposed state-aware multi-bit write algorithm provides a practical and efficient solution for enhancing the performance of RRAM devices in high-density memory applications.

Additionally, an electrical annealing technique using sub-SET voltage pulses is proposed to mitigate the retention loss of the conductance states, demonstrating a 3X extension of the refresh period in CIM applications. The proposed algorithm and electrical annealing method provide a comprehensive solution for enhancing the performance and stability of multi-bit RRAM devices, making them a promising candidate for CIM applications.

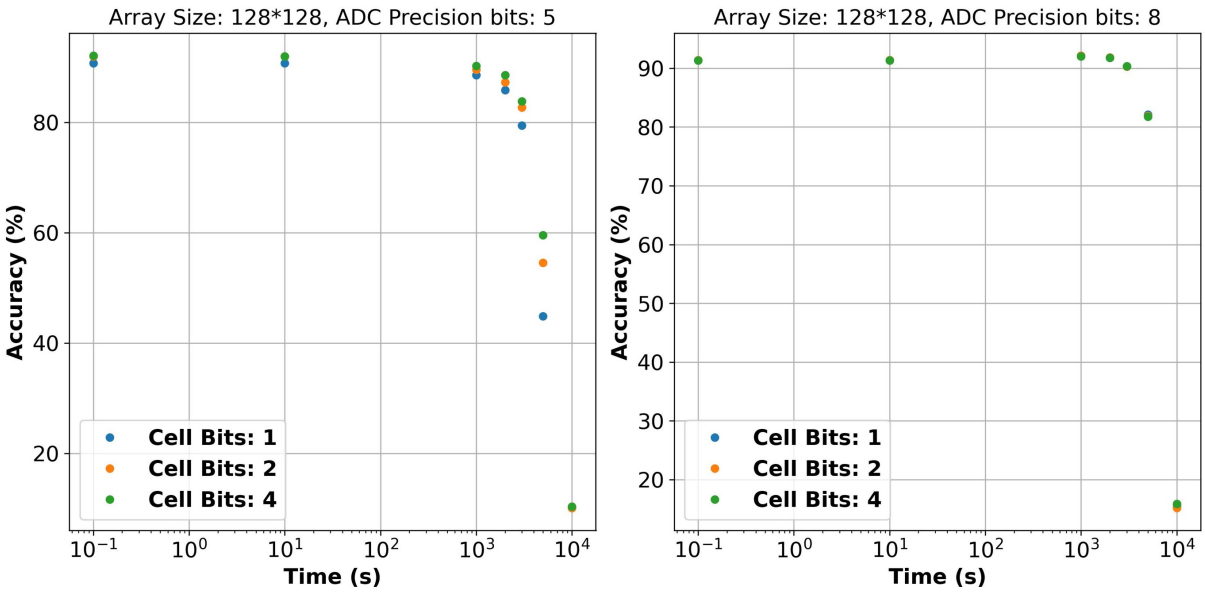


Figure 4.12: CIM simulation results of classification accuracy with ADC precision bits being (a)5 and (b)8.

Chapter 5

BEOL Integration of TiO_x based RRAM Devices on CMOS Chips

5.1 Circuit Design of the Chip

At the design stage of this CMOS chip, the device performance of TiO_x based devices has not been fully optimized, and the device characteristics are not well understood. Therefore, the design of this chip is meant to be a test vehicle for the integration of RRAM devices on CMOS chips, and the focus is on the feasibility of the integration process. Therefore, no sensing circuit is included in the design. TThe chip includes two memory arrays: a 1K-bit array and a 2K-bit array. The 1K-bit (2K-bit) array consists of 128 word lines, 8 (16) bit lines, and one source line.The word line is controlled by a 7-bit decoder, and the bit line is connected to the external signal source. The source line is connected to the ground pin of the chip.

A two-stage decoder was implemented to minimize input fan-in, the first stage consists of a 3-to-8 decoder (A0, A1, A2), implemented using 3-input NAND gates, and a 4-to-16 decoder (A3, A4, A5, A6), implemented using a 4-input NAND gate. The second stage are implemented using 3-input NAND gates, which assembles the combinations of the first stage outputs and the enable signal to control the output of word lines. The word line is connected to the gate of the transistor. The transistor has a two-finger structure with total width being 10 μm to match the programming current of the RRAM devices (the amplitude of programming current of device at the design stage is around 1 mA). The RRAM devices is designed to connect with the drain terminal of the transistor, and the source terminal of the transistor is connected to the source line, which is connected to the common ground

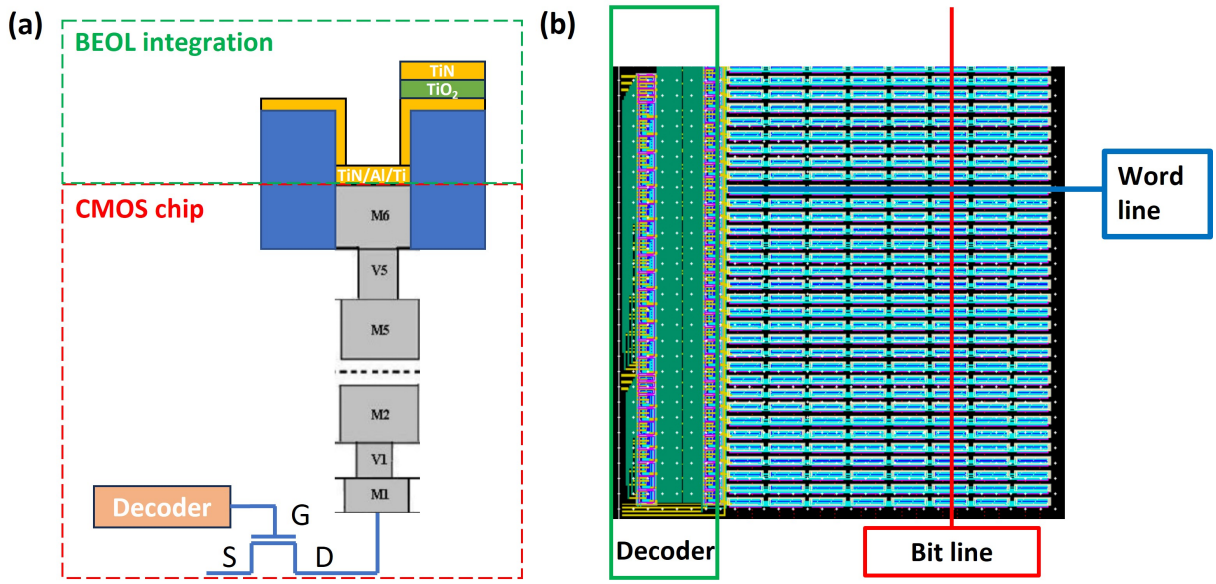


Figure 5.1: (a) Schematic of Integrated RRAM device on CMOS chip (b) Layout of the 1T1R array.

pin of the chip to minimize the number of pins. During the operation of the chip, the decoder is enabled by the micro-controller, and the word line is enabled by the decoder to control the transistor, which in turn controls the current flowing through the RRAM devices. The bit line with RRAM integrated is probed by the parameter analyzer, the common ground line is also connected to the parameter analyzer through wire connection. The voltage applied to the 1T1R structure is controlled by the parameter analyzer, and the current flowing through the RRAM devices is measured by the parameter analyzer. To accommodate the need of RESET, the polarity of the voltage applied to the RRAM devices can be reversed by the parameter analyzer (applying positive voltage to the source terminal of the transistor and ground the bit line). The supply voltage of the decoder is also connected to the parameter analyzer to control the saturation current of the transistor, which in turn controls the compliance current of the RRAM devices.

The chip is designed using the TSMC 180 nm technology, with 4 1K-bit array and 2 2K-bit array included in the chip. The chip is designed to have a size of 3 mm $\times$ 4 mm, the minimum size that can be handled in the EBL system. To leave space for fabrication handling, all the pins and arrays are placed on the three sides of the chip. Each array has independent decoder signal (control bits, enable signal and supply voltage), and the source

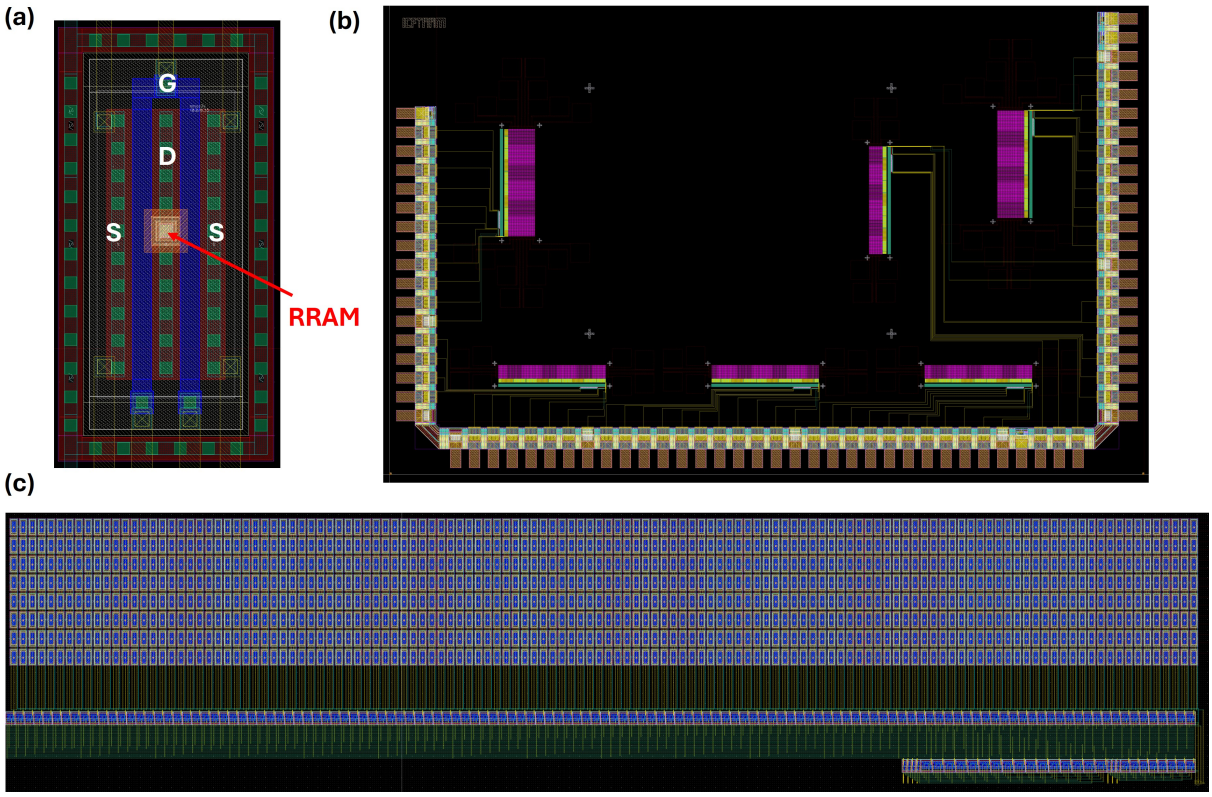


Figure 5.2: (a) Layout of the two-finger transistor and the position of the RRAM device; (b) Layout of the whole chip; (c) Layout of the 1T1R array.

line is shared by all the arrays. Except for the pin enablement signal, all the pins used are analog pins. The pitch size of the pin is chosen to be 60 μm, which is the maximum size provided by the process library, for easy wire bonding.

5.2 BEOL Integration Fabrication Process Flow

In this design, the RRAM devices are designed to be deposited on top of the M6 layer of the CMOS chip (Due to miscommunication with the foundry support engineer, a 1 μm $\times$ μm hole was etched through the passivation layer to form connection with the M6 pad, the high aspect ratio of the hole as well as the non-uniform surface of the passivation layer have greatly decreased the yield of the devices, the future work should remove all the passivation

layer in the array region and have M6 pad directly exposed for integration).

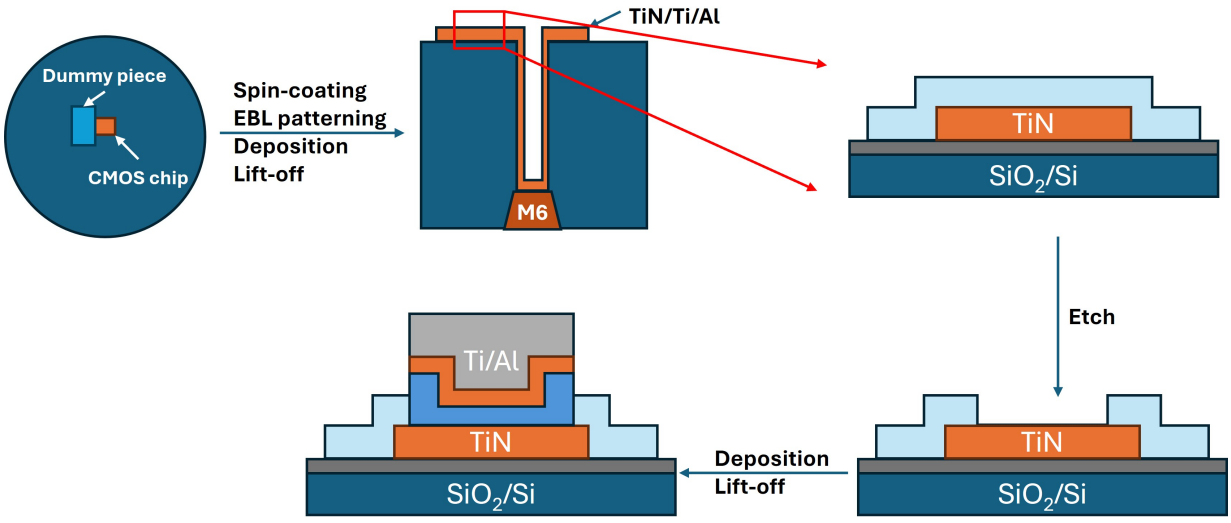


Figure 5.3: BEOL integration process flow of TiO_x based RRAM devices on CMOS chips.

The fabrication process flow of the BEOL integration is shown in Fig. 5.3. The CMOS chip is ultrasonic cleaned using Acetone/IPA and loaded on the carrier wafer for spin coating of photoresist. Dummy wafer pieces are assembled around the CMOS chip to mitigate the edge bead of photoresist during spin coating. Both the chip and the dummy wafer pieces are fixed on the carrier wafer using E-beam photoresist (baked after assembly) to avoid contamination of the EBL system. Coated chip is removed from the carrier wafer (back side cleaned using the acetone) and loaded to the EBL system for patterning. The bottom electrode is defined using EBL, and the bottom electrode is deposited using a combination of Al, Ti and TiN. After lift-off of the bottom electrode, a layer of SiO_2 is deposited using PECVD on top of the bottom electrode and then patterned using EBL to define the junction area. Reactive-ion etching (RIE) is used to etch the patterned SiO_2 to form the via structure. Another EBL patterning is done to define the shape of the top electrode, functional layers TiO_x and top electrode TiN are deposited on top of the pre-defined via using reactive sputtering. Lift-off as well as ultrasonic cleaning are done afterwards to prepare the chip for packaging.

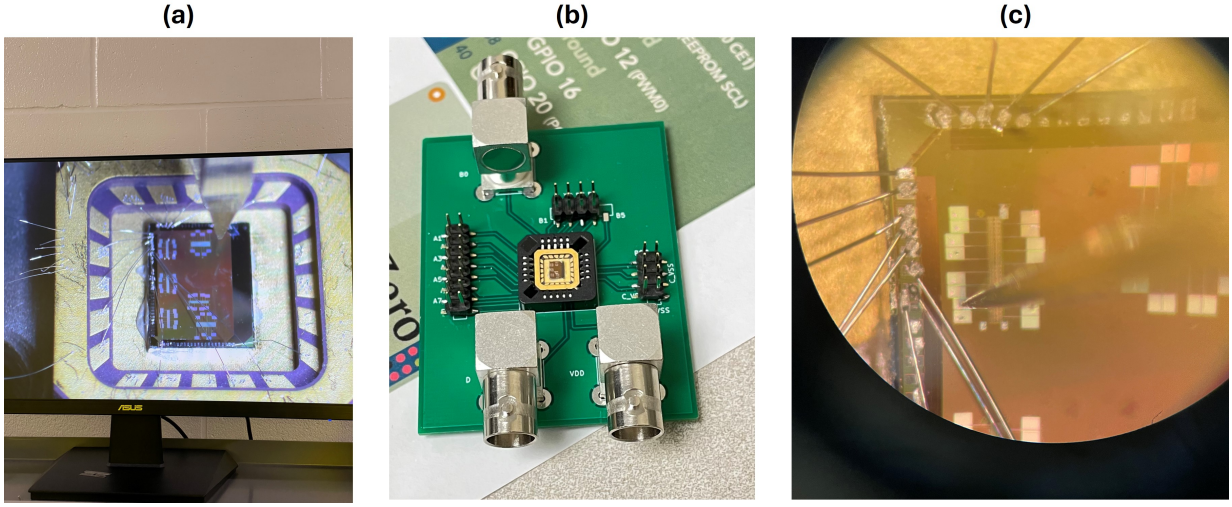


Figure 5.4: (a) Wire bonding of the chip; (b) PCB board for the chip; (c) Microscope image of the probed chip.

5.3 Packaging and Characterization

Fabricated chip are loaded to the 20-pin chip carrier (fixed with the silver paste) and wire-bonded for pin connection. A custom PCB board is designed to mount the chip carrier and wire the digital pins to the micro-controller (Raspberry Pi). Power supply as well as the analog pins are connected through the BNC ports on PCB board to the parameter analyzer (Keithley 4200A). The whole assembled PCB board is loaded on the probe station, and the top electrode (bitline) of the integrated devices is probed using the probe station.

During measurement, the micro-controller sends out signals first to enable the pins of the chip and the decoder, the parameter analyzer is then used to apply supply voltage to the decoder. The source terminal of the enabled transistor is grounded, a sweeping voltage is applied to the top electrode (bit line) of the integrated device to electrically form the device first. As can be observed from Figure. 5.5(a), after the current jump of forming, the voltage drop across the device is shifted to the transistor due to the sudden resistance decrease, making the transistor work in its saturation region and help limit the current flowing through the device. By adjusting the supply voltage and bias direction, the integrated devices can be RESET (Figure. 5.5(b)) or SET (Figure. 5.5(c)) with the transistor functioning as the compliance current source.

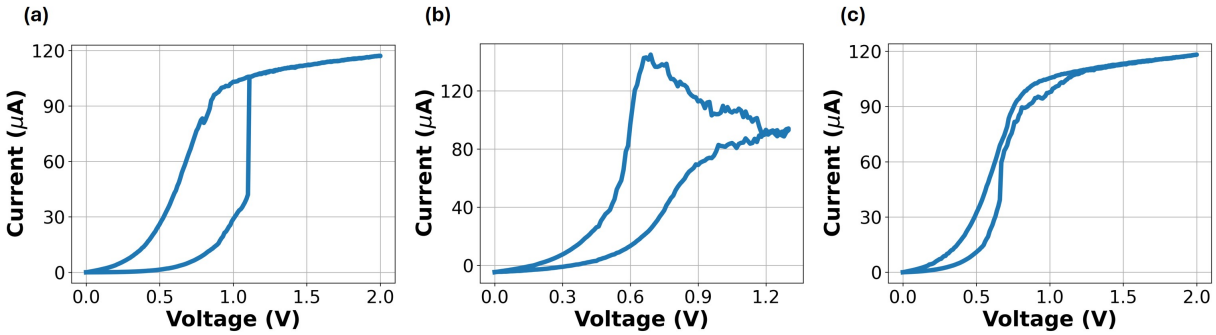


Figure 5.5: I-V characteristics of the integrated RRAM devices. (a) Forming process of the device; (b) RESET process of the device; (c) SET process of the device.

5.4 Summary

This chapter presents the back-end-of-line (BEOL) integration of TiO_x-based RRAM devices on CMOS chips. The chip was designed using TSMC 180 nm technology, with simplified circuitry tailored for integration feasibility rather than functional operation. Detailed fabrication steps—including EBL patterning, metal deposition, via formation, and RRAM stack deposition—were successfully developed and implemented.

Electrical characterization of the 1T1R structures demonstrates reliable switching behavior, validating both the device integration process and the role of transistor in current compliance. The presented BEOL methodology can be generalized to integrate other emerging memory or sensor devices with CMOS platforms, enabling broader adoption of post-CMOS hybrid systems for compute-in-memory applications.

Chapter 6

Conclusions and Future work

The work presented in this thesis has focused on the development of TiO_x based resistive switching devices, with a special focus on its applications in Compute-In-Memory. The project starts from the continuous optimization of TiO_x based devices, which includes device structure, fabrication process flow, material growth and interface engineering. With over 200 batches of devices fabricated, the project has achieved a high yield of devices with good performance, starting from a 30 μ m $\times$ 30 μ m Pt/TiO_x/Pt device with yield less than 10 %, forming voltage larger than 5 V and operation current larger than 1 mA, ending with a 400 nm $\times$ 400 nm TiN/TiO_x/TiN with yield larger than 95 %, forming voltage less than 1.5 V and operation current smaller than 100 μA. Based on well-optimized devices, a state aware multi-bit programming algorithm have been proposed and demonstrated, which can achieve 4-bit per cell programming with programming speed more than 2.2 times faster than the state-of-the-art algorithm. The developed multi-bit writing algorithm has enabled us to evaluate the multi-state stability of TiO_x based devices, based on which a electrical annealing method is proposed to mitigate the long term retention loss of multi-bit programming, which is essential to the application of CIM. The project has also explored the integration of TiO_x based devices with CMOS technology, and demonstrated the feasibility of BEOL integration of TiO_x devices on tape-out CMOS chips using university fabrication facilities. The BEOL integration method paves the way for the future development of CIM, and can be generalized to other technologies that are not compatible with CMOS process.

Despite the achievements, there are still many challenges to implement RRAM based CIM in real-world applications. In the following sections, future works will be discussed in three aspects: device optimization, hardware aware CIM design and system level integration.

6.1 Device Optimization

The primary focus of the device optimization of this work has been on increasing the resistance of the resistance states to lower down the energy consumption during CIM computing, another important metric that need to be focused is the programming current, which determines the write energy consumption as well as the size of the transistor in 1T1R structure (larger area is needed to accommodate the larger programming current). Scaling of programming current based on previous efforts (including thickness scaling and interface engineering) has come to a bottleneck, where marginal benefit is achieved despite complex design of device parameters. New insights are needed to further reduce the programming current.

Based on previous work on scaling of forming voltage, the formation of conducting filaments is a two step process, where the first step is the nucleation of oxygen vacancy and second step is the horizontal expansion of conducting filaments. To minimize the operation current, it is preferred that the forming stops at the first step, with which movement of a only few atoms is needed to control the resistance state. However, the two step process is regarded as a convolution process, making it hard to decouple them simply using external parameters like the amplitude of forming voltage. One possible solution is to insert a thin layer of insulating material below the top electrode, the pin-hole of the thin layer will act as a nucleation site for the conducting filament, which can be used to control the diameter of the conducting filament. Our previous trial of inserting Al_2O_3 at the top electrode has already shown promising results, the programming current of both SET and RESET operation is reduced by 1 order of magnitude (SET: $\sim$ 200μA -> $\sim$ 20μ A, RESET: $\sim$ 300μA -> $\sim$ 10μA). One additional benefit is that the resistance of HRS and LRS is also increased, which further proves our assumption that the diameter of the conducting filament is reduced. However, the insertion of Al_2O_3 also brings some challenges, including the increase of the forming voltage and instability of the device performance. Optimizations including the thickness of Al_2O_3 and the interface between Al_2O_3 and TiO_x need to be further studied.

Another possible solution to reduce the programming current is through thermal confinement method. The thermal confinement method is based on the fact that the rupture of the conducting filament is thermal assisted process, where the temperature of the filament is increased by the Joule heating. By using electrode with lower thermal conductivity, the heat generated by the Joule heating is confined in the device region, which can help increase the localized temperature of the filament and reduce the programming current needed. Inspired by the thermal confinement work of phase change memory[126, 127], TaN can be used as the potential candidate for the bottom electrode, which has lower thermal con-

ductivity compared to TiN. The thermal confinement method can be combined with the insertion of Al_2O_3 to further reduce the programming current. One interesting fact is that the latest device structure of TSMC's RRAM is using TaN as both the top and bottom electrodes, although the motivation behind is not clear, it may be related to the thermal confinement effect.

Apart from the material stack optimization, the device optimization can also be achieved through the optimization of the device structure and fabrication process flow. Via structure is chosen as current device structure due to its scalability and fast prototyping speed. However, there are still some drawbacks of such design:

- The presence of photoresist for top stack (TiO_x and top TiN) patterning limits the possibility of using high temperature process for the switching layer and the top electrode, such as using atomic layer deposition (ALD) for the switching layer and the top electrode.
- The contamination of the bottom electrode during the fabrication remains a puzzle despite the efforts of cleaning the substrate using Ar and N_2 plasma.
- SiO_2 is used as the passivation layer between the bottom electrode and the top electrode. The presence of oxygen in the SiO_2 layer may cause the oxidation of the TiO_x layer during the operation of the device, which may result in the degradation of the device performance.

To solve such issues, pillar structure can be adopted for material stacks that are well explored, such as the TiN/TiO_x/TiN stack. Fabrication recipe, especially the whole stack etching, needs to be optimized to ensure the high yield of the devices. The pillar structure can help enable the high temperature process for the switching layer and the top electrode, and the contamination of the bottom electrode can be avoided if all the materials can be deposited in one chamber (such as in the sputtering chamber). Passivation layer like SiN can be used to replace SiO_2 to avoid the sidewall oxidation of the TiO_x layer.

One last direction of device optimization is to use TiO_x based devices as a platform for new physics, especially the RTN behavior observed in HRS. The RTN behavior indicates that the conduction of HRS consists of a two-level system, deeper understanding of such behavior can help provide new insights for the design of RRAM devices. Another interesting direction is to enable the operation of TiO_x based devices at cryogenic temperature. The MIM structure of TiO_x based devices is similar to the tunnel junction of SQUID, which is a key component of quantum computing. Since the TiN electrode is a superconductor, if the device can operate at cryogenic temperature, we can tune the Josephson

effects between the two electrodes by switching the resistance state of the device. Our initial trial of operating the device at cryogenic temperature has shown that the no switching behavior is observed possibly because the oxygen vacancy migration is suppressed at low temperature ($\sim 300mK$). No Josephson effect is observed in the device, which may be due to the poor quality of the TiN electrode. Further optimization of the device structure is needed, additional materials like NbTiN can be used as the electrode to enhance the superconducting properties of the device.

6.2 Hardware Aware CIM Design

Our proposed electrical annealing method help to mitigate the retention loss of multi-bit programming, however, the fitting results show that the long term retention loss is still not fully mitigated, the degradation of the classification accuracy still exist as the electrical annealing method cannot change the intrinsic properties of the device. Despite the efforts of the device optimization, the retention loss of multi-bit programming is still a challenge for the application of CIM.

For convolution neural network, conventional method would map the kernel weights (filter) to the CIM array, during which process the weight values are quantized to the resistance states of the device. The quantization process normally uses a linear mapping method, where the resistance states are equally divided to represent the weight values. However, the distribution of the weights of a well-train neural network is not uniform (most following Gaussian distribution), the majority of the weights are concentrated around zero. The linear mapping method would result in a large number of weights being mapped to only one or two resistance states, the retention loss of these resistance states would shift the whole distribution of the mapped weights, which would result in a large classification accuracy loss. Compared to linear mapping, a non-linear weight mapping are needed to mitigate such issue. One possible solution is to use LRS, which suffers more from the retention loss, to represent weights close to 1 or -1, and use HRS to represent weights close to 0. Based on this assumption, instead of having one 1T1R cell to represent one weight, two 1T1R cells or a 2T2R cell are needed to represent the negative and positive value of the weight.

Another challenge for RRAM based CIM is the re-configurability for different applications. The researches on RRAM based CIM starts in the era of CNN and ResNet, which are the most popular neural network architecture for image recognition, with the core structure of neural network being the convolutional layer. However, the trend of neural network architecture is rapidly changing, with the introduction of new architecture

like Transformer, which is the core structure of LLMs, how can the RRAM based CIM be re-utilized for new architecture is still a open question. Compared to digital ASICs, redesigning RRAM based CIM requires not only the change in architecture but also the change in the peripheral circuits, which is time consuming and costly. To solve such issue, more efforts are needed in the compiler side, where the compiler can automatically map the neural network architecture to the RRAM based CIM, and optimize the mapping based on the hardware properties of RRAM devices.

6.3 System Level Integration

One of the motivations behind our efforts of BEOL integration of TiO_x RRAM devices on CMOS chips is to enable the silicon demonstration of RRAM based CIM, in the form of a CIM Macro that contains the RRAM array, peripheral circuits and the communication interface:

- RRAM array: The base unit of the RRAM array is expected to be the 1T1R cell, other variants like 2T2R cell and 1T4R cell can be explored based on the upper-level CIM design. The device area of the RRAM devices is expected to be around 200 nm $\times$ 200 nm, with the total area of one 1T1R cell being around 1 μm $\times$ 1 μm (if the CMOS chip is fabricated using the 180 nm technology node). Limited by the yield of the RRAM devices, the size of the smallest sub-array is expected to be 32 $\times$ 32. Larger array size can be achieved by tiling the sub-array, in the form of H-tree or mesh network.

- Peripheral circuits: The peripheral circuits of RRAM based CIM include the sense amplifier, the ADC/DACs, the write driver, the read driver and the address decoder. Researches on RRAM based CIM[121, 128] have shown that the peripheral circuits takes up a large portion of the total area of the CIM Macro, with the sense amplifier and the ADC/DACs being the most area consuming components, even larger than the RRAM array. Options like self-terminated writing can be explored to minimize the area of the peripheral circuits.

Apart from the design of the CIM Macro, the development for the test setup for the CIM Macro is still missing. Our previous trial of using the parameter analyzer to test the 1T1R cell has been proved to be inefficient, lack of programming flexibility and pulse measurement channels (only 2 channels are available) limits the characterization speed of the devices. A FPGA based test setup is needed to enable the high speed testing of the

CIM Macro as well as the characterization of single devices. The FPGA based test setup is expected to have the following features:

- Integrated sensing circuits: Dedicated high speed and high precision ADCs are needed to enable the high speed sensing and writing of the RRAM devices. The ADCs should be able to support the operation of the RRAM devices with speed larger than 100 MHz (10 ns sample time), and the resolution of the ADC/DACs should be larger than 10 bits to support the multi-bit programming of the RRAM devices.
- Integrated writing circuits: The writing circuits should be able to generate the arbitrary pulse shape needed for the programming of the RRAM devices (such as electrical annealing method). The writing circuits should be able to support the operation of the RRAM devices with speed larger than 1 GHz (minimum time segment being 1 ns), with the output voltage of the writing circuits being larger than 3 V and the output current being larger than 1 mA.
- Integrated row/column decoder: Before the integration of the RRAM devices with the CMOS chip, array level of testing is needed to characterize the yield of the RRAM devices, the row/column decoder is needed to enable the high speed testing of the RRAM array.

The channel number of the sensing and writing circuit should be larger than 3 to enable the biasing of the source, drain and gate terminal of the 1T1R structure.

References

[1] J.C. Liu, S. Mukhopadhyay, Amit Kundu, S.H. Chen, H.C. Wang, D.S. Huang, J.H. Lee, M.I. Wang, Ryan Lu, S.S. Lin, Y.M. Chen, H.L. Shang, P.W. Wang, H.C. Lin, Geoffrey Yeap, and Jun He. A Reliability Enhanced 5nm CMOS Technology Featuring 5^{th} Generation FinFET with Fully-Developed EUV and High Mobility Channel for Mobile SoC and High Performance Computing Application. In *2020 IEEE International Electron Devices Meeting (IEDM)*, pages 9.2.1–9.2.4, San Francisco, CA, USA, December 2020. IEEE.

[2] Chih-Hao Chang, V.S. Chang, K.H. Pan, K.T. Lai, J. H. Lu, J.A. Ng, C.Y. Chen, B.F. Wu, C.J. Lin, C.S. Liang, C.P. Tsao, Y.S. Mor, C.T. Li, T.C. Lin, C.H. Hsieh, P.N. Chen, H.H. Hsu, J.H. Chen, H.F. Chen, J.Y. Yeh, M.C. Chiang, C.Y. Lin, J.J. Liaw, C.H. Wang, S.B. Lee, C.C. Chen, H.J. Lin, R. Chen, K.W. Chen, C.O. Chui, Y.C. Yeo, K.B. Huang, T.L. Lee, M.H. Tsai, K.S. Chen, Y.C. Lu, S.M. Jang, and S.-Y. Wu. Critical Process Features Enabling Aggressive Contacted Gate Pitch Scaling for 3nm CMOS Technology and Beyond. In *2022 International Electron Devices Meeting (IEDM)*, pages 27.1.1–27.1.4, San Francisco, CA, USA, December 2022. IEEE.

[3] H.-S. Philip Wong, Heng-Yuan Lee, Shimeng Yu, Yu-Sheng Chen, Yi Wu, Pang-Shiu Chen, Byoungil Lee, Frederick T. Chen, and Ming-Jinn Tsai. Metal–Oxide RRAM. *Proceedings of the IEEE*, 100(6):1951–1970, June 2012.

[4] H. Y. Lee, P. S. Chen, T. Y. Wu, Y. S. Chen, C. C. Wang, P. J. Tzeng, C. H. Lin, F. Chen, C. H. Lien, and M.-J. Tsai. Low power and high speed bipolar switching with a thin reactive Ti buffer layer in robust HfO2 based RRAM. In *2008 IEEE International Electron Devices Meeting*, pages 1–4, San Francisco, CA, USA, December 2008. IEEE.

[5] B. Govoreanu, A. Redolfi, L. Zhang, C. Adelmann, M. Popovici, S. Clima, H. Hody, V. Paraschiv, I.P. Radu, A. Franquet, J.-C. Liu, J. Swerts, O. Richard, H. Bender, L. Altimime, and M. Jurczak. Vacancy-modulated conductive oxide resistive RAM (VMCO-RRAM): An area-scalable switching current, self-compliant, highly nonlinear and wide on/off-window resistive switching cell. In *2013 IEEE International Electron Devices Meeting*, pages 10.2.1–10.2.4, Washington, DC, USA, December 2013. IEEE.

[6] N. Raghavan, R. Degraeve, A. Fantini, L. Goux, S. Strangio, B. Govoreanu, D. J. Wouters, G. Groeseneken, and M. Jurczak. Microscopic origin of random telegraph noise fluctuations in aggressively scaled RRAM and its impact on read disturb variability. In *2013 IEEE International Reliability Physics Symposium (IRPS)*, pages 5E.3.1–5E.3.7, Anaheim, CA, April 2013. IEEE.

[7] Shimeng Yu, Ximeng Guan, and H.-S. Philip Wong. On the stochastic nature of resistive switching in metal oxide RRAM: Physical modeling, monte carlo simulation, and experimental characterization. In *2011 International Electron Devices Meeting*, pages 17.3.1–17.3.4, Washington, DC, USA, December 2011. IEEE.

[8] Yangyin Chen. ReRAM: History, Status, and Future. *IEEE Transactions on Electron Devices*, 67(4):1420–1433, April 2020.

[9] M. Prezioso, F. Merrikh-Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev, and D. B. Strukov. Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature*, 521(7550):61–64, May 2015.

[10] Xiaoyu Sun, Shihui Yin, Xiaochen Peng, Rui Liu, Jae-sun Seo, and Shimeng Yu. XNOR-RRAM: A scalable and parallel resistive synaptic architecture for binary neural networks. In *2018 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 1423–1428, Dresden, Germany, March 2018. IEEE.

[11] Weier Wan, Rajkumar Kubendran, Clemens Schaefer, Sukru Burc Eryilmaz, Wenqiang Zhang, Dabin Wu, Stephen Deiss, Priyanka Raina, He Qian, Bin Gao, Siddharth Joshi, Huaqiang Wu, H.-S. Philip Wong, and Gert Cauwenberghs. A compute-in-memory chip based on resistive random-access memory. *Nature*, 608(7923):504–512, August 2022.

[12] Shimeng Yu and Pai-Yu Chen. Emerging Memory Technologies: Recent Trends and Prospects. *IEEE Solid-State Circuits Magazine*, 8(2):43–56, 2016.

[13] Tetsuo Endoh, Hiroaki Honjo, Koichi Nishioka, and Shoji Ikeda. Recent Progresses in STT-MRAM and SOT-MRAM for Next Generation MRAM. In *2020 IEEE Symposium on VLSI Technology*, pages 1–2, Honolulu, HI, USA, June 2020. IEEE.

[14] H.-S. Philip Wong, Simone Raoux, SangBum Kim, Jiale Liang, John P. Reifenberg, Bipin Rajendran, Mehdi Asheghi, and Kenneth E. Goodson. Phase Change Memory. *Proceedings of the IEEE*, 98(12):2201–2227, December 2010.

[15] Qiqiao Wu, Yue Cao, Qing Luo, Haijun Jiang, Zhongze Han, Yongkang Han, Chunmeng Dou, Hangbing Lv, Qi Liu, Jianguo Yang, and Ming Liu. A 9-Mb HZO-Based Embedded FeRAM With 10$ˆ{12}$-Cycle Endurance and 5/7-ns Read/Write Using ECC-Assisted Data Refresh and Offset-Canceled Sense Amplifier. *IEEE Journal of Solid-State Circuits*, pages 1–11, 2023.

[16] Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, Rick Boyle, Pierre-luc Cantin, Clifford Chao, Chris Clark, Jeremy Coriell, Mike Daley, Matt Dau, Jeffrey Dean, Ben Gelb, Tara Vazir Ghaemmaghami, Rajendra Gottipati, William Gulland, Robert Hagmann, C. Richard Ho, Doug Hogberg, John Hu, Robert Hundt, Dan Hurt, Julian Ibarz, Aaron Jaffey, Alek Jaworski, Alexander Kaplan, Harshit Khaitan, Daniel Killebrew, Andy Koch, Naveen Kumar, Steve Lacy, James Laudon, James Law, Diemthu Le, Chris Leary, Zhuyuan Liu, Kyle Lucke, Alan Lundin, Gordon MacKean, Adriana Maggiore, Maire Mahony, Kieran Miller, Rahul Nagarajan, Ravi Narayanaswami, Ray Ni, Kathy Nix, Thomas Norrie, Mark Omernick, Narayana Penukonda, Andy Phelps, Jonathan Ross, Matt Ross, Amir Salek, Emad Samadiani, Chris Severn, Gregory Sizikov, Matthew Snelham, Jed Souter, Dan Steinberg, Andy Swing, Mercedes Tan, Gregory Thorson, Bo Tian, Horia Toma, Erick Tuttle, Vijay Vasudevan, Richard Walter, Walter Wang, Eric Wilcox, and Doe Hyun Yoon. In-Datacenter Performance Analysis of a Tensor Processing Unit. In *Proceedings of the 44th Annual International Symposium on Computer Architecture*, pages 1–12, Toronto ON Canada, June 2017. ACM.

[17] Yu-Hsin Chen, Tushar Krishna, Joel S. Emer, and Vivienne Sze. Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks. *IEEE Journal of Solid-State Circuits*, 52(1):127–138, January 2017.

[18] Xueyao Huang, Huaqiang Wu, Deepak C. Sekar, Steve N. Nguyen, Kun Wang, and He Qian. Optimization of TiN/TaOx/HfO2/TiN RRAM Arrays for Improved Switching and Data Retention. In *2015 IEEE International Memory Workshop (IMW)*, pages 1–4, Monterey, CA, USA, May 2015. IEEE.

[19] Dmitri B. Strukov, Gregory S. Snider, Duncan R. Stewart, and R. Stanley Williams. The missing memristor found. *Nature*, 453(7191):80–83, May 2008.

[20] Kyungho Hong, Kyung Kyu Min, Min-Hwi Kim, Suhyun Bang, Tae-Hyeon Kim, Dong Keun Lee, Yeon Joon Choi, Chae Soo Kim, Jae Yoon Lee, Sungjun Kim, Seongjae Cho, and Byung-Gook Park. Investigation of the Thermal Recovery From Reset Breakdown of a SiN $_x$ -Based RRAM. *IEEE Transactions on Electron Devices*, 67(4):1600–1605, April 2020.

[21] Kaichen Zhu, Sebastian Pazos, Fernando Aguirre, Yaqing Shen, Yue Yuan, Wenwen Zheng, Osamah Alharbi, Marco A. Villena, Bin Fang, Xinyi Li, Alessandro Milozzi, Matteo Farronato, Miguel Muñoz-Rojo, Tao Wang, Ren Li, Hossein Fariborzi, Juan B. Roldan, Guenther Benstetter, Xixiang Zhang, Husam N. Alshareef, Tibor Grasser, Huaqiang Wu, Daniele Ielmini, and Mario Lanza. Hybrid 2D–CMOS microchips for memristive applications. *Nature*, 618(7963):57–62, June 2023.

[22] L Goux, M Jurczak, and D J Wouters. On the Gradual Unipolar and Bipolar Resistive Switching of TiN\HfO2\Pt Memory Systems. *Electrochemical and Solid-State Letters*, 2010.

[23] B. Govoreanu, G.S. Kar, Y-Y. Chen, V. Paraschiv, S. Kubicek, A. Fantini, I.P. Radu, L. Goux, S. Clima, R. Degraeve, N. Jossart, O. Richard, T. Vandeweyer, K. Seo, P. Hendrickx, G. Pourtois, H. Bender, L. Altimime, D.J. Wouters, J.A. Kittl, and M. Jurczak. 10×10nm^2 Hf/HfOx crossbar resistive RAM with excellent performance, reliability and low-energy operation. In *2011 International Electron Devices Meeting*, pages 31.6.1–31.6.4, Washington, DC, USA, December 2011. IEEE.

[24] DooSung Lee, YongHun Sung, InGun Lee, JongGi Kim, Hyunchul Sohn, and Dae-Hong Ko. Enhanced bipolar resistive switching of HfO2 with a Ti interlayer. *Applied Physics A*, 102(4):997–1001, March 2011.

[25] B. Govoreanu, D. Crotti, S. Subhechha, L. Zhang, Y.Y. Chen, S. Clima, V. Paraschiv, H. Hody, C. Adelmann, M. Popovici, O. Richard, and M. Jurczak. A-VMCO: A novel forming-free, self-rectifying, analog memory cell with low-current operation, nonfilamentary switching and excellent variability. In *2015 Symposium on VLSI Technology (VLSI Technology)*, pages T132–T133, Kyoto, Japan, June 2015. IEEE.

[26] Yu-Fen Wang, Yen-Chuan Lin, I-Ting Wang, Tzu-Ping Lin, and Tuo-Hung Hou. Characterization and Modeling of Nonfilamentary Ta/TaOx/TiO2/Ti Analog Synaptic Device. *Scientific Reports*, 5(1):10150, May 2015.

[27] L. Goux, A. Fantini, B. Govoreanu, G. Kar, S. Clima, Y.-Y. Chen, R. Degraeve, D. J. Wouters, G. Pourtois, and M. Jurczak. Asymmetry and Switching Phenomenology in TiN\(Al2O3)\HfO2\Hf Systems. *ECS Solid State Letters*, 1(4):P63–P65, August 2012.

[28] Taeyoon Kim, Gwangho Baek, Seungmo Yang, Jung Yup Yang, Kap Soo Yoon, Soo Gil Kim, Jae Yeon Lee, Hyun Sik Im, and Jin Pyo Hong. Exploring oxygen-affinity-controlled TaN electrodes for thermally advanced TaOx bipolar resistive switching. *Scientific Reports*, 8(1):8532, June 2018.

[29] D. C. Gilmer, G. Bersuker, H.-Y. Park, C. Park, B. Butcher, W. Wang, P. D. Kirsch, and R. Jammy. Effects of RRAM Stack Configuration on Forming Voltage and Current Overshoot. In *2011 3rd IEEE International Memory Workshop (IMW)*, pages 1–4, Monterey, CA, USA, May 2011. IEEE.

[30] L Goux, A Fantini, R Degraeve, N Raghavan, R Nigon, S Strangio, G Kar, D J Wouters, Y Y Chen, M Komura, F De Stefano, V V Afanas'ev, and M Jurczak. Understanding of the intrinsic characteristics and memory trade-offs of sub-µA filamentary RRAM operation. *IEEE Symposium on VLSI Technology*, 2013.

[31] T. Diokh, E. Le-Roux, S. Jeannot, M. Gros-Jean, P. Candelier, J. F. Nodin, V. Jousseaume, L. Perniola, H. Grampeix, T. Cabout, E. Jalaguier, M. Guillermet, and B. De Salvo. Investigation of the impact of the oxide thickness and RESET conditions on disturb in HfO2-RRAM integrated in a 65nm CMOS technology. In *2013 IEEE International Reliability Physics Symposium (IRPS)*, pages 5E.4.1–5E.4.4, Anaheim, CA, April 2013. IEEE.

[32] Yu-Sheng Chen, Heng-Yuan Lee, Pang-Shiu Chen, Tai-Yuan Wu, Ching-Chiun Wang, Pei-Jer Tzeng, Frederick Chen, Ming-Jinn Tsai, and Chenhsin Lien. An Ultra-thin Forming-Free $\hbox{HfO}_{x}$ Resistance Memory With Excellent Electrical Performance. *IEEE Electron Device Letters*, 31(12):1473–1475, December 2010.

[33] Pei-Yi Gu, Yu-Sheng Chen, Heng-Yuan Lee, Pang-Shiu Chen, Wen-Hsing Liu, Wei-Su Chen, Yen-Ya Hsu, Frederick Chen, and Ming-Jinn Tsai. Scalability with silicon nitride encapsulation layer for Ti/HfOx pillar RRAM. In *Proceedings of 2010 International Symposium on VLSI Technology, System and Application*, pages 146–147, Hsin Chu, Taiwan, 2010. IEEE.

[34] S. Park, J. Kim, W. Koo, Y. Kwon, D. Suh, S. Lee, S. Lee, K. Kim, E. Han, J. Lee, J. Kim, K. Kim, Y. Lim, Y. Ahn, E. Park, S. Ryu, S. Kim, N. Park, H. Em, S. Kang,

J. Cheon, E. Kim, J. Yi, K. Kim, T. Kim, S. Lee, M. Na, and S. Cha. A Holistic Methodology Toward Large-scale AI Implementation using Realistic ReRAM based ACiM from Cell to Architecture. In *2023 International Electron Devices Meeting (IEDM)*, pages 1–4, San Francisco, CA, USA, December 2023. IEEE.

[35] H.Y. Lee, Y.S. Chen, P.S. Chen, T.Y. Wu, F. Chen, C.C. Wang, P.J. Tzeng, M.-J. Tsai, and C. Lien. Low-Power and Nanosecond Switching in Robust Hafnium Oxide Resistive Memory With a Thin Ti Cap. *IEEE Electron Device Letters*, 31(1):44–46, January 2010.

[36] Yang Yin Chen, L. Goux, J. Swerts, M. Toeller, C. Adelmann, J. Kittl, M. Jurczak, G. Groeseneken, and D. J. Wouters. Hydrogen-Induced Resistive Switching in TiN/ALD $ \hbox{HfO}_{2}$/PEALD TiN RRAM Device. *IEEE Electron Device Letters*, 33(4):483–485, April 2012.

[37] C.Y. Chen, L. Goux, A. Fantini, R. Degraeve, A. Redolfi, G. Groeseneken, and M. Jurczak. Stack optimization of oxide-based RRAM for fast write speed (<1 μs) at low operating current (<10 μA). *Solid-State Electronics*, 125:198–203, November 2016.

[38] C. Y. Chen, L. Goux, A. Fantini, A. Redolfi, G. Groeseneken, and M. Jurczak. Oxygen chemical potential profile optimization for fast low current (<10μA) resistive switching in oxide-based RRAM. In *2016 International Symposium on VLSI Technology, Systems and Application (VLSI-TSA)*, pages 1–2, Hsinchu, April 2016. IEEE.

[39] Xiaoxin Xu, Lu Tai, Tiancheng Gong, Jiahao Yin, Peng Huang, Jie Yu, Da Nian Dong, Qing Luo, Jing Liu, Zhaoan Yu, Xi Zhu, Xiu Long Wu, Qi Liu, Hangbing Lv, and Ming Liu. 40× Retention Improvement by Eliminating Resistance Relaxation with High Temperature Forming in 28 nm RRAM Chip. In *2018 IEEE International Electron Devices Meeting (IEDM)*, pages 20.1.1–20.1.4, San Francisco, CA, December 2018. IEEE.

[40] N. Raghavan, A. Fantini, R. Degraeve, P.J. Roussel, L. Goux, B. Govoreanu, D.J. Wouters, G. Groeseneken, and M. Jurczak. Statistical insight into controlled forming and forming free stacks for HfOx RRAM. *Microelectronic Engineering*, 109:177–181, September 2013.

[41] Xueqi Li, Liyang Pan, Junyi Wang, Bin Gao, Jianshi Tang, He Qian, and Huaqiang Wu. Methods to Realize Low-BER and High-Reliability RRAM Chip With Fast Page-Forming Capability. *IEEE Transactions on Electron Devices*, pages 1–7, 2023.

[42] Gennadi Bersuker, Dmitry Veksler, David M. Nminibapiel, Pragya R. Shrestha, Jason P. Campbell, Jason T. Ryan, Helmut Baumgart, Maribeth S. Mason, and Kin P. Cheung. Toward reliable RRAM performance: macro- and micro-analysis of operation processes. *Journal of Computational Electronics*, 16(4):1085–1094, December 2017.

[43] Alessandro Grossi, Cristian Zambelli, Piero Olivo, Enrique Miranda, Valeriy Stikanov, Christian Walczyk, and Christian Wenger. Electrical characterization and modeling of pulse-based forming techniques in RRAM arrays. *Solid-State Electronics*, 115:17–25, January 2016.

[44] Xiaoyong Xue, Wenxiang Jian, Jianguo Yang, Fanjie Xiao, Gang Chen, Shuliu Xu, Yufeng Xie, Yinyin Lin, Ryan Huang, Qingtian Zou, and Jingang Wu. A 0.13 µm 8 Mb Logic-Based Cu$_{\rm x}$Si$_{\rm y}$O ReRAM With Self-Adaptive Operation for Yield Enhancement and Power Reduction. *IEEE Journal of Solid-State Circuits*, 48(5):1315–1322, May 2013.

[45] L. Goux, P. Czarnecki, Y. Y. Chen, L. Pantisano, X. P. Wang, R. Degraeve, B. Govoreanu, M. Jurczak, D. J. Wouters, and L. Altimime. Evidences of oxygen-mediated resistive-switching mechanism in TiN\HfO2\Pt cells. *Applied Physics Letters*, 97(24):243509, December 2010.

[46] D. C. Gilmer, G. Bersuker, S. Koveshnikov, M. Jo, A. Kalantarian, B. Butcher, R. Geer, Y. Nishi, P. D. Kirsch, and Raj Jammy. Asymmetry, Vacancy Engineering and Mechanism for Bipolar RRAM. In *2012 4th IEEE International Memory Workshop*, pages 1–4, Milan, Italy, May 2012. IEEE.

[47] L. Goux, A. Fantini, G. Kar, Y.-Y. Chen, N. Jossart, R. Degraeve, S. Clima, B. Govoreanu, G. Lorenzo, G. Pourtois, D.J. Wouters, J.A. Kittl, L. Altimime, and M. Jurczak. Ultralow sub-500nA operating current high-performance TiN\Al2O3\HfO2\Hf\TiN bipolar RRAM achieved through understanding-based stack-engineering. In *2012 Symposium on VLSI Technology (VLSIT)*, pages 159–160, Honolulu, HI, USA, June 2012. IEEE.

[48] Jaesung Park, Jiyong Woo, Amit Prakash, Sangheon Lee, Seokjae Lim, and Hyunsang Hwang. Improved reset breakdown strength in a HfO_x -based resistive memory by introducing RuO_x oxygen diffusion barrier. *AIP Advances*, 6(5):055114, May 2016.

[49] H. Y. Lee, Y. S. Chen, P. S. Chen, P. Y. Gu, Y. Y. Hsu, S. M. Wang, W. H. Liu, C. H. Tsai, S. S. Sheu, P. C. Chiang, W. P. Lin, C. H. Lin, W. S. Chen, F. T. Chen, C. H. Lien, and M.-J. Tsai. Evidence and solution of over-RESET problem for HfOX based resistive memory with sub-ns switching speed and high endurance. In *2010 International Electron Devices Meeting*, pages 19.7.1–19.7.4, San Francisco, CA, USA, December 2010. IEEE.

[50] Umberto Celano, Ludovic Goux, Robin Degraeve, Andrea Fantini, Olivier Richard, Hugo Bender, Malgorzata Jurczak, and Wilfried Vandervorst. Imaging the Three-Dimensional Conductive Channel in Filamentary-Based Oxide Resistive Switching Memory. *Nano Letters*, 15(12):7970–7975, December 2015.

[51] Stefano Larentis, Federico Nardi, Simone Balatti, David C. Gilmer, and Daniele Ielmini. Resistive Switching by Voltage-Driven Ion Migration in Bipolar RRAM—Part II: Modeling. *IEEE Transactions on Electron Devices*, 59(9):2468–2475, September 2012.

[52] Dmitri B. Strukov and R. Stanley Williams. Exponential ionic drift: fast switching and low volatility of thin-film memristors. *Applied Physics A*, 94(3):515–519, March 2009.

[53] A. Rodriguez-Fernandez, C. Cagli, J. Sune, and E. Miranda. Switching Voltage and Time Statistics of Filamentary Conductive Paths in HfO $_2$ -Based ReRAM Devices. *IEEE Electron Device Letters*, 39(5):656–659, May 2018.

[54] Yoshifumi Nishi, Stephan Menzel, Karsten Fleck, Ulrich Bottger, and Rainer Waser. Origin of the SET Kinetics of the Resistive Switching in Tantalum Oxide Thin Films. *IEEE Electron Device Letters*, 35(2):259–261, February 2014.

[55] Daniele Ielmini, Federico Nardi, and Simone Balatti. Evidence for Voltage-Driven Set/Reset Processes in Bipolar Switching RRAM. *IEEE Transactions on Electron Devices*, 59(8):2049–2056, August 2012.

[56] S. Koveshnikov, K. Matthews, K. Min, D. C. Gilmer, M.G. Sung, S. Deora, H. F. Li, S. Gausepohl, P. D. Kirsch, and R. Jammy. Real-time study of switching kinetics in integrated 1T/ HfOx 1R RRAM: Intrinsic tunability of set/reset voltage and trade-off with switching time. In *2012 International Electron Devices Meeting*, pages 20.4.1–20.4.3, San Francisco, CA, USA, December 2012. IEEE.

[57] Peng Huang, Yijiao Wang, Haitong Li, Bin Gao, Bing Chen, Feifei Zhang, Lang Zeng, Gang Du, Jinfeng Kang, and Xiaoyan Liu. Analysis of the Voltage–Time Dilemma of

Metal Oxide-Based RRAM and Solution Exploration of High Speed and Low Voltage AC Switching. *IEEE Transactions on Nanotechnology*, 13(6):1127–1132, November 2014.

[58] Yi-Cheng Huang, Shang-Hsuan Liu, Hsu-Shun Chen, Hsin-Chang Feng, Chih-Feng Li, Chou-Ying Yang, Wei-Keng Chang, Chang-Feng Yang, Chun-Yu Wu, Yen-Cheng Lin, Tsung-Tse Yang, Chih-Yang Chang, Wen-Ting Chu, Harry Chuang, Yih Wang, Yu-Der Chih, and Tsung-Yung Jonathan Chang. 15.7 A 32Mb RRAM in a 12nm FinFet Technology with a $0.0249\mu m^2$ Bit-Cell, a 3.2GB/S Read Throughput, a 10KCycle Write Endurance and a 10-Year Retention at 105°C. In *2024 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 288–290, San Francisco, CA, USA, February 2024. IEEE.

[59] Chung-Cheng Chou, Zheng-Jun Lin, Chien-An Lai, Chin-I Su, Pei-Ling Tseng, Wei-Chi Chen, Wu-Chin Tsai, Wen-Ting Chu, Tong-Chern Ong, Harry Chuang, Yu-Der Chih, and Tsung-Yung Jonathan Chang. A 22nm 96KX144 RRAM Macro with a Self-Tracking Reference and a Low Ripple Charge Pump to Achieve a Configurable Read Window and a Wide Operating Voltage Range. In *2020 IEEE Symposium on VLSI Circuits*, pages 1–2, Honolulu, HI, USA, June 2020. IEEE.

[60] O. Golonzka, U. Arslan, P. Bai, M. Bohr, O. Baykan, Y. Chang, A. Chaudhari, A. Chen, J. Clarke, C. Connor, N. Das, C. English, T. Ghani, F. Hamzaoglu, P. Hentges, P. Jain, C. Jezewski, I. Karpov, H. Kothari, R. Kotlyar, B. Lin, M. Metz, J. Odonnell, D. Ouellette, J. Park, A. Pirkle, P. Quintero, D. Seghete, M. Sekhar, A. Sen Gupta, M. Seth, N. Strutt, C. Wiegand, H. J. Yoo, and K. Fischer. Non-Volatile RRAM Embedded into 22FFL FinFET Technology. In *2019 Symposium on VLSI Technology*, pages T230–T231, Kyoto, Japan, June 2019. IEEE.

[61] B. Chen, Y. Lu, B. Gao, Y.H. Fu, F.F. Zhang, P. Huang, Y.S. Chen, L.F. Liu, X.Y. Liu, J.F. Kang, Y.Y. Wang, Z. Fang, H.Y. Yu, X. Li, X.P. Wang, N. Singh, G. Q. Lo, and D. L. Kwong. Physical mechanisms of endurance degradation in TMO-RRAM. In *2011 International Electron Devices Meeting*, pages 12.3.1–12.3.4, Washington, DC, USA, December 2011. IEEE.

[62] Yang Yin Chen, Robin Degraeve, Sergiu Clima, Bogdan Govoreanu, Ludovic Goux, Andrea Fantini, Gouri Sankar Kar, Geoffrey Pourtois, Guido Groeseneken, Dirk J. Wouters, and Malgorzata Jurczak. Understanding of the endurance failure in scaled HfO2-based 1T1R RRAM through vacancy mobility degradation. In *2012 International Electron Devices Meeting*, pages 20.3.1–20.3.4, San Francisco, CA, USA, December 2012. IEEE.

[63] Yang Yin Chen, Andrea Fantini, Gouri Sankar Kar, Dirk J Wouters, Guido Groeseneken, Jorge A Kittl, Malgorzata Jurczak, and Laith Altimime. Balancing SET/RESET Pulse for > 1010 Endurance in HfO2/Hf 1T1R Bipolar RRAM. *IEEE TRANSACTIONS ON ELECTRON DEVICES*, 59(12):7, 2012.

[64] Yang Yin Chen, Masanori Komura, Robin Degraeve, Bogdan Govoreanu, Ludovic Goux, Andrea Fantini, Naga Raghavan, Sergiu Clima, Leqi Zhang, Attilio Belmonte, Augusto Redolfi, Gouri Sankar Kar, Guido Groeseneken, Dirk J. Wouters, and Malgorzata Jurczak. Improvement of data retention in HfO2/Hf 1T1R RRAM cell under low operating current. In *2013 IEEE International Electron Devices Meeting*, pages 10.1.1–10.1.4, Washington, DC, USA, December 2013. IEEE.

[65] Qishen Wang, Yuhang Yang, Zongwei Wang, Shengyu Bao, Jingwei Sun, Linbo Shan, Lin Bao, Yi Gao, Haisu Zhang, Yaotian Ling, Wuzhi Zhang, Yansheng Wang, Yimao Cai, and Ru Huang. A Logic-Process Compatible RRAM with 15.43 Mb/mm ^{2} Density and 10years@150°C retention using STI-less Dynamic-Gate and Self-Passivation Sidewall. In *2023 International Electron Devices Meeting (IEDM)*, pages 1–4, San Francisco, CA, USA, December 2023. IEEE.

[66] Wonbo Shim, Yandong Luo, Jae-sun Seo, and Shimeng Yu. Impact of Read Disturb on Multilevel RRAM based Inference Engine: Experiments and Model Prediction. In *2020 IEEE International Reliability Physics Symposium (IRPS)*, pages 1–5, Dallas, TX, USA, April 2020. IEEE.

[67] Nagarajan Raghavan, Robin Degraeve, Andrea Fantini, Ludovic Goux, Dirk J. Wouters, Guido Groeseneken, and Malgorzata Jurczak. Modeling the Impact of Reset Depth on Vacancy-Induced Filament Perturbations in ${\rm HfO}_{2}$ RRAM. *IEEE Electron Device Letters*, 34(5):614–616, May 2013.

[68] Wun-Cheng Luo, Kuan-Liang Lin, Jiun-Jia Huang, Chung-Lun Lee, and Tuo-Hung Hou. Rapid Prediction of RRAM RESET-State Disturb by Ramped Voltage Stress. *IEEE Electron Device Letters*, 33(4):597–599, April 2012.

[69] Wun-Cheng Luo, Jen-Chieh Liu, Hsien-Tsung Feng, Yen-Chuan Lin, Jiun-Jia Huang, Kuan-Liang Lin, and Tuo-Hung Hou. RRAM SET speed-disturb dilemma and rapid statistical prediction methodology. In *2012 International Electron Devices Meeting*, pages 9.5.1–9.5.4, San Francisco, CA, USA, December 2012. IEEE.

[70] Ligang Gao, Pai-Yu Chen, and Shimeng Yu. Programming Protocol Optimization for Analog Weight Tuning in Resistive Memories. *IEEE Electron Device Letters*, 36(11):1157–1159, November 2015.

[71] Ming-Hsiu Lee, K.C. Hsu, F.M. Lee, Y.Y. Lin, E.K. Lai, J.Y. Wu, D.Y. Lee, and K.Y. Hsieh. Resistance control of transition metal oxide resistive memory (TMO ReRAM). In *2016 13th IEEE International Conference on Solid-State and Integrated Circuit Technology (ICSICT)*, pages 182–185, Hangzhou, China, October 2016. IEEE.

[72] Takeo Ohno, Tsuyoshi Hasegawa, Tohru Tsuruoka, Kazuya Terabe, James K. Gimzewski, and Masakazu Aono. Short-term plasticity and long-term potentiation mimicked in single inorganic synapses. *Nature Materials*, 10(8):591–595, August 2011.

[73] Valerio Milo, Elisabetta Chicca, and Daniele Ielmini. Brain-inspired recurrent neural network with plastic RRAM synapses. In *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5, Florence, 2018. IEEE.

[74] F. M. Puglisi, C. Wenger, and P. Pavan. A Novel Program-Verify Algorithm for Multi-Bit Operation in HfO_2 RRAM. *IEEE Electron Device Letters*, 36(10):1030–1032, October 2015.

[75] Chieh-Pu Lo, Wen-Zhang Lin, Wei-Yu Lin, Huan-Ting Lin, Tzu-Hsien Yang, Yen-Ning Chiang, Ya-Chin King, Chrong-Jung Lin, Yu-Der Chih, Tsung-Yung Jonathon Chang, and Meng-Fan Chang. A ReRAM Macro Using Dynamic Trip-Point-Mismatch Sampling Current-Mode Sense Amplifier and Low-DC Voltage-Mode Write-Termination Scheme Against Resistance and Write-Delay Variation. *IEEE Journal of Solid-State Circuits*, 54(2):584–595, February 2019.

[76] Alexandre Levisse, Marc Bocquet, Marco Rios, Mouhamad Alayan, Mathieu Moreau, Etienne Nowak, Gabriel Molas, Elisa Vianello, David Atienza, and Jean-Michel Portal. Write Termination Circuits for RRAM: A Holistic Approach From Technology to Application Considerations. *IEEE Access*, 8:109297–109308, 2020.

[77] Zongwu Wang, Zhezhi He, Rui Yang, Shiquan Fan, Jie Lin, Fangxin Liu, Yueyang Jia, Chenxi Yuan, Qidong Tang, and Li Jiang. Self-Terminating Write of Multi-Level Cell ReRAM for Efficient Neuromorphic Computing. In *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 1251–1256, Antwerp, Belgium, March 2022. IEEE.

[78] A. Fantini, G. Gorine, R. Degraeve, L. Goux, C. Y. Chen, A. Redolfi, S. Clima, A. Cabrini, G. Torelli, and M. Jurczak. Intrinsic program instability in HfO2 RRAM and consequences on program algorithms. In *2015 IEEE International Electron Devices Meeting (IEDM)*, pages 7.5.1–7.5.4, Washington, DC, USA, December 2015. IEEE.

[79] Meiran Zhao, Huaqiang Wu, Bin Gao, Qingtian Zhang, Wei Wu, Shan Wang, Yue Xi, Dong Wu, Ning Deng, Shimeng Yu, Hong-Yu Chen, and He Qian. Investigation of statistical retention of filamentary analog RRAM for neuromophic computing. In *2017 IEEE International Electron Devices Meeting (IEDM)*, pages 39.4.1–39.4.4, San Francisco, CA, USA, December 2017. IEEE.

[80] Yulin Feng, Peng Huang, Yudi Zhao, Yihao Shan, Yizhou Zhang, Zheng Zhou, Lifeng Liu, Xiaoyan Liu, and Jinfeng Kang. Improvement of State Stability in Multi-Level Resistive Random-Access Memory (RRAM) Array for Neuromorphic Computing. *IEEE Electron Device Letters*, 42(8):1168–1171, August 2021.

[81] Zhixing Jiang, Yue Xi, Jianshi Tang, Yuyao Lu, Ruihua Yu, Ruofei Hu, Qi Hu, Bin Gao, He Qian, and Huaqiang Wu. COPS: An Efficient and Reliability-Enhanced Programming Scheme for Analog RRAM and On-Chip Implementation of Denoising Diffusion Probabilistic Model. In *2023 International Electron Devices Meeting (IEDM)*, pages 1–4, San Francisco, CA, USA, December 2023. IEEE.

[82] Yue Xi, Jianshi Tang, Bin Gao, Feng Xu, Xinyi Li, Yuyao Lu, He Qian, and Huaqiang Wu. The Impact of Thermal Enhance Layers on the Relaxation Effect in Analog RRAM. *IEEE Transactions on Electron Devices*, 69(8):4254–4258, August 2022.

[83] Jingwei Sun, Zongwei Wang, Jiajun Gao, Linbo Shan, Qishen Wang, Yuhang Yang, Yimao Cai, and Ru Huang. ASAP: An Efficient and Reliable Programming Algorithm for Multi-level RRAM Cell. In *2024 IEEE International Reliability Physics Symposium (IRPS)*, pages 1–4, Grapevine, TX, USA, April 2024. IEEE.

[84] Wangxin He, Jian Meng, Sujan Kumar Gonugondla, Shimeng Yu, Naresh R. Shanbhag, and Jae-sun Seo. PRIVE: Efficient RRAM Programming with Chip Verification for RRAM-based In-Memory Computing Acceleration. In *2023 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 1–6, Antwerp, Belgium, April 2023. IEEE.

[85] Binh Q. Le, Akash Levy, Tony F. Wu, Robert M. Radway, E. Ray Hsieh, Xin Zheng, Mark Nelson, Priyanka Raina, H.-S. Philip Wong, Simon Wong, and Subhasish Mitra. RADAR: A Fast and Energy-Efficient Programming Technique for Multiple Bits-Per-Cell RRAM Arrays. *IEEE Transactions on Electron Devices*, 68(9):4397–4403, September 2021.

[86] Qing Dong, Mahmut E. Sinangil, Burak Erbagci, Dar Sun, Win-San Khwa, Hung-Jen Liao, Yih Wang, and Jonathan Chang. 15.3 A 351TOPS/W and 372.4GOPS

Compute-in-Memory SRAM Macro in 7nm FinFET CMOS for Machine-Learning Applications. In *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, pages 242–244, San Francisco, CA, USA, February 2020. IEEE.

[87] Young-Cheon Kwon, Suk Han Lee, Jaehoon Lee, Sang-Hyuk Kwon, Je Min Ryu, Jong-Pil Son, O Seongil, Hak-Soo Yu, Haesuk Lee, Soo Young Kim, Youngmin Cho, Jin Guk Kim, Jongyoon Choi, Hyun-Sung Shin, Jin Kim, BengSeng Phuah, Hyoung-Min Kim, Myeong Jun Song, Ahn Choi, Daeho Kim, SooYoung Kim, Eun-Bong Kim, David Wang, Shinhaeng Kang, Yuhwan Ro, Seungwoo Seo, JoonHo Song, Jaeyoun Youn, Kyomin Sohn, and Nam Sung Kim. 25.4 A 20nm 6GB Function-In-Memory DRAM, Based on HBM2 with a 1.2TFLOPS Programmable Computing Unit Using Bank-Level Parallelism, for Machine Learning Applications. In *2021 IEEE International Solid- State Circuits Conference (ISSCC)*, pages 350–352, San Francisco, CA, USA, February 2021. IEEE.

[88] X. Sun, W. S. Khwa, Y. S. Chen, C. H. Lee, H. Y. Lee, S. M. Yu, R. Naous, J. Y. Wu, T. C. Chen, X. Bao, M. F. Chang, C. H. Diaz, H.-S. P. Wong, and K. Akarvardar. PCM-Based Analog Compute-In-Memory: Impact of Device Non-Idealities on Inference Accuracy. *IEEE Transactions on Electron Devices*, 68(11):5585–5591, November 2021.

[89] Shimeng Yu, Zhiwei Li, Pai-Yu Chen, Huaqiang Wu, Bin Gao, Deli Wang, Wei Wu, and He Qian. Binary neural network with 16 Mb RRAM macro chip for classification and online training. In *2016 IEEE International Electron Devices Meeting (IEDM)*, pages 16.2.1–16.2.4, San Francisco, CA, USA, December 2016. IEEE.

[90] Cheng-Xin Xue, Ting-Wei Chang, Tung-Cheng Chang, Hui-Yao Kao, Yen-Cheng Chiu, Chun-Ying Lee, Ya-Chin King, Chrong-Jung Lin, Ren-Shuo Liu, Chih-Cheng Hsieh, Kea-Tiong Tang, Wei-Hao Chen, Meng-Fan Chang, Je-Syu Liu, Jia-Fang Li, Wei-Yu Lin, Wei-En Lin, Jing-Hong Wang, Wei-Chen Wei, and Tsung-Yuan Huang. Embedded 1-Mb ReRAM-Based Computing-in- Memory Macro With Multibit Input and Weight for CNN-Based AI Edge Processors. *IEEE Journal of Solid-State Circuits*, 55(1):203–215, January 2020.

[91] Shimeng Yu, Jiale Liang, Yi Wu, and H-S Philip Wong. Read/write schemes analysis for novel complementary resistive switches in passive crossbar memory arrays. *Nanotechnology*, 21(46):465202, November 2010.

[92] An Chen. Analysis of Partial Bias Schemes for the Writing of Crossbar Memory Arrays. *IEEE Transactions on Electron Devices*, 62(9):2845–2849, September 2015.

[93] Jiyong Woo and Shimeng Yu. Two-Step Read Scheme in One-Selector and One-RRAM Crossbar-Based Neural Network for Improved Inference Robustness. *IEEE Transactions on Electron Devices*, 65(12):5549–5553, December 2018.

[94] Haitong Li, Hong-Yu Chen, Zhe Chen, Bing Chen, Rui Liu, Gang Qiu, Peng Huang, Feifei Zhang, Zizhen Jiang, Bin Gao, Lifeng Liu, Xiaoyan Liu, Shimeng Yu, H.-S. Philip Wong, and Jinfeng Kang. Write disturb analyses on half-selected cells of cross-point RRAM arrays. In *2014 IEEE International Reliability Physics Symposium*, pages MY.3.1–MY.3.4, Waikoloa, HI, USA, June 2014. IEEE.

[95] Geoffrey W. Burr, Rohit S. Shenoy, Kumar Virwani, Pritish Narayanan, Alvaro Padilla, Bülent Kurdi, and Hyunsang Hwang. Access devices for 3D crosspoint memory. *Journal of Vacuum Science & Technology B, Nanotechnology and Microelectronics: Materials, Processing, Measurement, and Phenomena*, 32(4):040802, July 2014.

[96] Kai-Shin Li, Ming-Taou Lee, Min-Cheng Chen, Cho-Lun Hsu, J. M. Lu, C. H. Lin, C. C. Chen, B. W. Wu, Y. F. Hou, C. Yi. Lin, Y. J. Chen, T. Y. Lai, M. Y. Li, I. Yang, C. S. Wu, Fu-Liang Yang, and W. K. Yeh. Study of sub-5 nm RRAM, tunneling selector and selector less device. In *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 385–388, Lisbon, Portugal, May 2015. IEEE.

[97] D. Alfaro Robayo, D. Deleruyelle, E. Vianello, N. Castellani, L. Ciampolini, B. Giraud, C. Cagli, G. Ghibaudo, E. Nowak, G. Molas, G. Sassine, J. Minguet Lopez, L. Grenouillet, A. Verdy, G. Navarro, M. Bernard, E. Esmanhotto, and C. Carabasse. Reliability and Variability of 1S1R OxRAM-OTS for High Density Crossbar Integration. In *2019 IEEE International Electron Devices Meeting (IEDM)*, pages 35.3.1–35.3.4, San Francisco, CA, USA, December 2019. IEEE.

[98] Yexin Deng, Hong-Yu Chen, Bin Gao, Shimeng Yu, Shih-Chieh Wu, Liang Zhao, Bing Chen, Zizhen Jiang, Xiaoyan Liu, Tuo-Hung Hou, Yoshio Nishi, Jinfeng Kang, and H.-S. Philip Wong. Design and optimization methodology for 3D RRAM arrays. In *2013 IEEE International Electron Devices Meeting*, pages 25.7.1–25.7.4, Washington, DC, USA, December 2013. IEEE.

[99] Yang Yin Chen, Ludovic Goux, Sergiu Clima, Bogdan Govoreanu, Robin Degraeve, Gouri Sankar Kar, Andrea Fantini, Guido Groeseneken, Dirk J. Wouters, and Malgorzata Jurczak. Endurance/Retention Trade-off on $\hbox{HfO}_{2}/\hbox{Metal}$ Cap 1T1R Bipolar RRAM. *IEEE Transactions on Electron Devices*, 60(3):1114–1121, March 2013.

[100] Linfang Wang, Wang Ye, Jinru Lai, Jing Liu, Jianguo Yang, Xin Si, Changxing Huo, Chunmeng Dou, Xiaoxin Xu, Qi Liu, Dashan Shang, Feng Zhang, Hangbing Lv, Meng-Fan Chang, Hiroshi Iwai, and Ming Liu. A 14nm 100Kb 2T1R Transpose RRAM with >150X resistance ratio enhancement and 27.95% reduction on energy-latency product using low-power near threshold read operation and fast data-line current stabling scheme. *Symposium on VLSI Technology Digest of Technical Papers*, page 2, 2021.

[101] Jianguo Yang, Xiaoyong Xue, Xiaoxin Xu, Hangbing Lv, Feng Zhang, Xiaoyang Zeng, Meng-Fan Chang, and Ming Liu. A 28nm 1.5Mb Embedded 1T2R RRAM with 14.8 Mb/mm ^{2} using Sneaking Current Suppression and Compensation Techniques. In *2020 IEEE Symposium on VLSI Circuits*, pages 1–2, Honolulu, HI, USA, June 2020. IEEE.

[102] E. R. Hsieh, X. Zheng, B. Q. Le, Y. C. Shih, R. M. Radway, M. Nelson, S. Mitra, and S. Wong. Four-Bits-Per-Memory One-Transistor-and-Eight-Resistive-Random-Access-Memory (1T8R) Array. *IEEE Electron Device Letters*, 42(3):335–338, March 2021.

[103] E.R. Hsieh, X. Zheng, M. Nelson, B.Q. Le, H.-S.P. Wong, S. Mitra, S. Wong, M. Giordano, B. Hodson, A. Levy, S.K. Osekowsky, R.M. Radway, Y.C. Shih, W. Wan, and T. F. Wu. High-Density Multiple Bits-per-Cell 1T4R RRAM Array with Gradual SET/RESET and its Effectiveness for Deep Learning. In *2019 IEEE International Electron Devices Meeting (IEDM)*, pages 35.6.1–35.6.4, San Francisco, CA, USA, December 2019. IEEE.

[104] Sugil Lee, Giju Jung, Mohammed E. Fouda, Jongeun Lee, Ahmed Eltawil, and Fadi Kurdahi. Learning to Predict IR Drop with Effective Training for ReRAM-based Neural Network Hardware. In *2020 57th ACM/IEEE Design Automation Conference (DAC)*, pages 1–6, San Francisco, CA, USA, July 2020. IEEE.

[105] Jong-Hyeok Yoon, Muya Chang, Win-San Khwa, Yu-Der Chih, Meng-Fan Chang, and Arijit Raychowdhury. A 40-nm, 64-Kb, 56.67 TOPS/W Voltage-Sensing Computing-In-Memory/Digital RRAM Macro Supporting Iterative Write With Verification and Online Read-Disturb Detection. *IEEE Journal of Solid-State Circuits*, 57(1):68–79, January 2022.

[106] Hongwu Jiang, Wantong Li, Shanshi Huang, and Shimeng Yu. A 40nm Analog-Input ADC-Free Compute-in-Memory RRAM Macro with Pulse-Width Modulation between Sub-arrays. In *2022 IEEE Symposium on VLSI Technology and Circuits*

(VLSI Technology and Circuits), pages 266–267, Honolulu, HI, USA, June 2022. IEEE.

[107] Jinshan Yue, Yongpan Liu, Xiaoyu Feng, Yifan He, Jingyu Wang, Zhe Yuan, Mingtao Zhan, Jiaxin Liu, Jian-Wei Su, Yen-Lin Chung, Ping-Chun Wu, Li-Yang Hong, Meng-Fan Chang, Nan Sun, Chunmeng Dou, Xueqing Li, Ming Liu, and Huazhong Yang. An Energy-Efficient Computing-in-Memory NN Processor With Set-Associate Blockwise Sparsity and Ping-Pong Weight Update. *IEEE Journal of Solid-State Circuits*, pages 1–16, 2023.

[108] Tim Kempen, Rainer Waser, and Vikas Rana. 50x Endurance Improvement in TaOx RRAM by Extrinsic Doping. In *2021 IEEE International Memory Workshop (IMW)*, pages 1–4, Dresden, Germany, May 2021. IEEE.

[109] Deok-Hwang Kwon, Kyung Min Kim, Jae Hyuck Jang, Jong Myeong Jeon, Min Hwan Lee, Gun Hwan Kim, Xiang-Shu Li, Gyeong-Su Park, Bora Lee, Seungwu Han, Miyoung Kim, and Cheol Seong Hwang. Atomic structure of conducting nanofilaments in TiO2 resistive switching memory. *Nature Nanotechnology*, 5(2):148–153, February 2010.

[110] J. Joshua Yang, Matthew D. Pickett, Xuema Li, Douglas A. A. Ohlberg, Duncan R. Stewart, and R. Stanley Williams. Memristive switching mechanism for metal/oxide/metal nanodevices. *Nature Nanotechnology*, 3(7):429–433, July 2008.

[111] Ashwani Kumar, Yucheng Zhou, Sai Praneeth Potladurthy, Jeong-Hoon Kim, Weihong Xu, Flavio Ponzina, Seonghyun Kim, Ertugrul Cubukcu, Tajana Rosing, Gert Cauwenberghs, and Duygu Kuzum. Filament-Free Bulk RRAM with High Endurance and Long Retention for Neuromorphic Few-Shot Learning On-Chip. In *2024 IEEE International Electron Devices Meeting (IEDM)*, pages 1–4, San Francisco, CA, USA, December 2024. IEEE.

[112] Jianguo Yang, Xiaoyong Xue, Xiaoxin Xu, Qiao Wang, Haijun Jiang, Jie Yu, Danian Dong, Feng Zhang, Hangbing Lv, and Ming Liu. 24.2 A 14nm-FinFET 1Mb Embedded 1T1R RRAM with a 0.022µm ^{2} Cell Size Using Self-Adaptive Delayed Termination and Multi-Cell Reference. In *2021 IEEE International Solid- State Circuits Conference (ISSCC)*, pages 336–338, San Francisco, CA, USA, February 2021. IEEE.

[113] Xiaoxin Xu, Jie Yu, Tiancheng Gong, Jianguo Yang, Jiahao Yin, Da Nian Dong, Qing Luo, Jing Liu, Zhaoan Yu, Qi Liu, Hangbing Lv, and Ming Liu. First Demonstration

of OxRRAM Integration on 14nm FinFet Platform and Scaling Potential Analysis towards Sub-10nm Node. In *2020 IEEE International Electron Devices Meeting (IEDM)*, pages 24.3.1–24.3.4, San Francisco, CA, USA, December 2020. IEEE.

[114] Ernest Wu, Takashi Ando, and Paul Jamison. Reversing a decades-long scaling law of dielectric breakdown for ReRAM forming voltage reduction - Modeling competition among defect generation and annihilation processes (invited). In *2023 International Electron Devices Meeting (IEDM)*, pages 1–4, San Francisco, CA, USA, December 2023. IEEE.

[115] Samuel D. Spetalnick, Muya Chang, Shota Konno, Brian Crafton, Ashwin S. Lele, Win-San Khwa, Yu-Der Chih, Meng-Fan Chang, and Arijit Raychowdhury. A 2.38 MCells/mm ^{2} 9.81 -350 TOPS/W RRAM Compute-in-Memory Macro in 40nm CMOS with Hybrid Offset/I $_{off}$ Cancellation and I $_{cell}$ R $_{blsl}$ Drop Mitigation. In *2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, pages 1–2, Kyoto, Japan, June 2023. IEEE.

[116] Ruyi Zhang, Qian-Ying Ma, Haigang Liu, Tian-Yu Sun, Jiachang Bi, Yang Song, Shaoqin Peng, Lingyan Liang, Junhua Gao, Hongtao Cao, Liang-Feng Huang, and Yanwei Cao. Crystal Orientation-Dependent Oxidation of Epitaxial TiN Films with Tunable Plasmonics. *ACS Photonics*, 8(3):847–856, March 2021.

[117] Beom Yong Kim, Baek Su Kim, Seung Dam Hyun, Ho Hyun Kim, Yong Bin Lee, Hyun Woo Park, Min Hyuk Park, and Cheol Seong Hwang. Study of ferroelectric characteristics of Hf0.5Zr0.5O2 thin films grown on sputtered or atomic-layer-deposited TiN bottom electrodes. *Applied Physics Letters*, 117(2):022902, July 2020.

[118] Jörgen Westlinder, Gustaf Sjöblom, and Jörgen Olsson. Variable work function in MOS capacitors utilizing nitrogen-controlled TiNx gate electrodes. *Microelectronic Engineering*, 75(4):389–396, November 2004.

[119] Arrigo Calzolari and Alessandra Catellani. Controlling the TiN Electrode Work Function at the Atomistic Level: A First Principles Investigation. *IEEE Access*, 8:156308–156313, 2020.

[120] Gina C. Adam, Brian D. Hoskins, Mirko Prezioso, Farnood Merrikh-Bayat, Bhaswar Chakrabarti, and Dmitri B. Strukov. 3-D Memristor Crossbars for Analog and Neuromorphic Computing Applications. *IEEE Transactions on Electron Devices*, 64(1):312–318, January 2017.

[121] Bastien Giraud, Sebastien Ricavy, Cyrille Laffond, Ilan Sever, Valentin Gherman, Florent Lepin, Mariam Diallo, Khadija Zenati, Sylvain Dumas, Olivier Guille, Maxim Vershkov, Alessandro Bricalli, Giuseppe Piccolboni, Jean-Philippe Noel, Anass Samir, Gaël Pillonnet, Yvain Thonnart, and Gabriel Molas. Smart Write Algorithm to Enhance Performances and Reliability of an RRAM Macro. *IEEE Journal of Solid-State Circuits*, pages 1–13, 2024.

[122] Hangbing Lv, Xiaoxin Xu, Peng Yuan, Danian Dong, Tiancheng Gong, Jing Liu, Zhaoan Yu, Peng Huang, Kun Zhang, Changxing Huo, Chanbing Chen, Yuanlu Xie, Qing Luo, Shibing Long, Qi Liu, Jinfeng Kang, Daisy Yang, Simon Yin, Shengfen Chiu, and Ming Liu. BEOL Based RRAM with one extra-mask for low cost, highly reliable embedded application in 28 nm node and beyond. In *2017 IEEE International Electron Devices Meeting (IEDM)*, pages 2.4.1–2.4.4, San Francisco, CA, USA, December 2017. IEEE.

[123] Erika Covi, Stefano Brivio, Jacopo Frascaroli, Marco Fanciulli, and Sabina Spiga. (Invited) Analog HfO_2 -RRAM Switches for Neural Networks. *ECS Transactions*, 75(32):85–94, January 2017.

[124] Jiyong Woo, Kibong Moon, Jeonghwan Song, Sangheon Lee, Myounghun Kwak, Jaesung Park, and Hyunsang Hwang. Improved Synaptic Behavior Under Identical Pulses Using AlO_x /HfO_2 Bilayer RRAM Array for Neuromorphic Systems. *IEEE Electron Device Letters*, 37(8):994–997, August 2016.

[125] Wangxin He, Wonbo Shim, Shihui Yin, Xiaoyu Sun, Deliang Fan, Shimeng Yu, and Jae-sun Seo. Characterization and Mitigation of Relaxation Effects on Multi-level RRAM based In-Memory Computing. In *2021 IEEE International Reliability Physics Symposium (IRPS)*, pages 1–7, Monterey, CA, USA, March 2021. IEEE.

[126] J.Y. Wu, M. Breitwisch, S. Kim, T.H. Hsu, R. Cheek, P.Y. Du, J. Li, E. K. Lai, Y. Zhu, T.Y. Wang, H.Y. Cheng, A. Schrott, E.A. Joseph, R. Dasaka, S. Raoux, M.H. Lee, H.L. Lung, and C. Lam. A low power phase change memory using thermally confined TaN/TiN bottom electrode. In *2011 International Electron Devices Meeting*, pages 3.2.1–3.2.4, Washington, DC, USA, December 2011. IEEE.

[127] Aditya Sood, Sukru Burc Eryilmaz, Rakesh Jeyasingh, Jungwan Cho, Mehdi Asheghi, H.-S. Philip Wong, and Kenneth E. Goodson. Thermal characterization of nanostructured superlattices of TiN/TaN: Applications as electrodes in Phase

Change Memory. In *Fourteenth Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, pages 765–770, Orlando, FL, USA, May 2014. IEEE.

[128] Junjie Wang, Teng Zhang, Shuang Liu, Yihe Liu, Yuancong Wu, Shaogang Hu, Tupei Chen, Yang Liu, Yuchao Yang, and Ru Huang. Design and Implementation of a Hybrid, ADC/DAC-Free, Input-Sparsity-Aware, Precision Reconfigurable RRAM Processing-in-Memory Chip. *IEEE Journal of Solid-State Circuits*, pages 1–10, 2023.