RRAM in High-speed TCAM Design and Its Applications with New Switching Materials

by

Kangqiang Pan

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2022

© Kangqiang Pan 2022

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

With the continuous scaling of transistor devices reaching their physical limits, emerging non-volatile memory (eNVM) devices such as resistive random-access memory (RRAM) is considered one of the alternatives to maintain growth in semiconductor technology as predicted by Moore's law. Meanwhile, it is becoming difficult for the traditional von Neuman architecture to meet the continuous growing demand of computation power in integrated circuit (IC) applications. New data processing scheme such as neuromorphic network are attracting great interests in both research and industry. RRAM, as a type of eNVM and resistive switching device, possesses the advantages of compact size, high switching speed, low programming voltage, large ON/OFF resistance ratio and compatibility with current complementary metal-oxide-semiconductor (CMOS) fabrication process. It has been studied extensively in implementing large scale random-access memory (RAM) array and artificial neural networks (ANNs). In this thesis, a novel RRAM-based circuit is presented to achieve high density and highly energy efficient memory system, with tunable delay element (TDE) for reference signal generation. A parallel-RRAM structure is proposed to address the serious issue of RRAM intra-cell and inter-cell switching variations in terms of programming voltage and resultant resistance after the programming process. RRAM-based neuromorphic network is also explored through device and circuit level innovations.

For the RRAM-based memory system, a current race (CR)-based ternary content addressable memory (TCAM) circuit design is proposed using RRAM technology. The suggested design adopts a match-line (ML) booster feature in sensing amplifier to improve search speed and tolerance to RRAM switching variations. Two cascading schemes, direct cascading (DC) and SR-latch cascading (SRC), are proposed to further improve performance and energy efficiency for large TCAM array. The DC structure features high noise margin while SRC structure improves search speed. Additionally, a same clock phase cascading (SCPC) scheme is proposed to reduce latency in cascading structure, by placing evaluation phase of all stages in the same clock phase. With the suggested ML booster, the 64-bit 1-stage design has speed and energy consumption matching the best performance reported by other eNVM-based TCAM design. The proposed 128-bit 2-stage design also has comparable speed and energy to SRAM-based TCAM design with significantly more compact size (90% reduction) and non-volatility.

Meanwhile, a TDE design with delay range from ∼100ps to ∼1ns is proposed, which can be used in TCAM design for reference signal generation. Impacts of RRAM resistance on delay range and power consumption of the circuit are analyzed. An improved parallel RRAM TDE circuit is also proposed to reduce impact of switching variation of RRAM device and provide finer tunable delay resolution.

The last part of this study is focusing on RRAM-based neuromorphic networks, two RRAM devices are presented and reviewed: Al_2O_3-based and CuZnSe (CZSe)-based. The capacitive-coupled Al_2O_3-based RRAM is used in design simulation of a leaky-integrate and fire (LIF) neuron circuit. It can provide post-fabrication tunability of leakage rate, improving flexibility of circuit design. The CZSe-based device demonstrates concurrent resistive switching and light activated conduction effect. Its synaptic behavior is investigated and used in simulating an ANN for pattern recognition. The simulated results indicate high output accuracy from the ANN.

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisors, Professor Lan Wei and Professor Norman Zhou, for their invaluable advice and support throughout my Master's study at the University of Waterloo, especially during the difficult time of this global pandemic. With their feedback during our regular meetings, I continuously found room to improve myself and refine my research work.

I would also like to thank the rest of my thesis committee: Professor Ajoy Opal and Professor Yiming Wu for their insightful feedback and encouragement. I also owe thanks to Professor Yiming Wu for providing valuable suggestion on paper writing.

I would also like to express my gratitude to Dr. Amr Tosson for providing me with useful feedback on my research projects as well as the amount of time he spent on reviewing and providing comments to my research papers. I would also like to thank Tao Guo for giving me advice and supporting me on my lab experiments.

Last but not least, I would like to thank my family and friends who have supported me throughout my life.

Dedication

This is dedicated to my parents and my brother for their love and support.

Table of Contents

List of Figures x

List of Tables xvi

List of Abbreviations xviii

1 Introduction 1

1.1 Scope of Research 4

1.2 Organization 5

2 Related Work 6

2.1 Memristor and RRAM 6

2.2 Applications of RRAM 8

2.2.1 Large Scale RAM Array Integration 9

2.2.2 In-memory Computing and Neuromorphic Computing 10

2.3 Switching Variation of RRAM 13

3 TCAM Design Incorporating RRAM Devices 18

3.1 Introduction 18

3.2 2T2R TCAM Cell Structure 22

3.3 Current Racing Sensing Scheme 23

3.4 Impact of Stored Data on MLSA Performance 28

3.5 RRAM Variation and Impact on V_{ML} 29
3.6 ML Booster 30
3.7 Performance Comparison with Other eNVM-based TCAM Designs 33
3.8 TDE Design Using RRAM 35
3.9 Parallel RRAM Configuration in TDE 37
3.10 Summary 43

4 TCAM Multi-stage Cascading Design **45**

4.1 Introduction 45
4.2 Layout and Area Consideration 46
4.3 2-stage with Direct Cascading 47
4.4 2-stage with SR-latch Cascading 50
4.5 Performance Comparison of 1-stage and 2-stage TCAM Design 53
4.6 Same Clock Phase Cascading 55
4.7 Further Staging and Competing with SRAM-based TCAM 60
4.8 Summary 65

5 Resistive Switching with Innovative Materials and Potential Neuromorphic Applications **67**

5.1 Introduction 67
5.2 Capacitive Coupled Effect in Al_2O_3-based RRAM Device 68
5.3 LIF Neuron Circuit Design with Al2O3-based RRAM Device 71
5.4 Effect of Illumination in CZSe-based RRAM Device 74
5.5 Synaptic Behavior in CZSe-based RRAM Device 77
5.6 Summary 79

6 Conclusions and Future Work **81**

6.1 Conclusions 81
6.2 Future Work 82

References 84

APPENDICES 94

A List of Publications 95

List of Figures

1.1 Growth microprocessor in terms of transistor counts, performance, frequency, power and number of logic cores. Figure is taken from [6]. 2

1.2 CMOS scaling trend: size of SRAM, Contacted Gate Pitch (CGP) and Metal 1 pitch. Figure is taken from [7]. 2

1.3 von Neuman architecture 4

2.1 I-V curve of (a) unipolar switching RRAM: V_{SET} and V_{RESET} have different polarities. (b) bipolar switching RRAM: V_{SET} and V_{RESET} have different polarities. The red curve represents SET process and the blue curve represents RESET process. Figure is adapted from [26]. 7

2.2 Conduction mechanism of (a) OxRAM (Figure is taken from [27]): CF formed by oxygen vacancies. (b) CBRAM (Figure is taken from [30]): CF formed by accumulation of reduced metal ions. 8

2.3 1T1R and 1S1R array structure for implementing large scale RAM. (Figure is taken from [31]. SL here stands for Source-Line) 10

2.4 Structure of a IMPLY gate consisting of two RRAMs P and Q with reference resistor R_G with resistance between LRS and HRS of the RRAM device. Condition voltage $|V_{COND}|< |V_{SET}|$. (Figure is taken from [34].) 11

2.5 RRAM used as synapse to emulate neuron system. (Figure is taken from [20].) 12

2.6 LIF neuron working mechanism: input pulse signals are integrated as membrane potential over time. A output spike is generated if the membrane potential exceeds the spiking threshold. (Figure is taken from [20].) 13

2.7 Physical origin of resistance dispersion in LRS: (A) CF radius defined by fluctuations in number of particles and (B) fluctuation in CF constriction geometry. (Figure is taken from [41].) 14

2.8 RRAM switching variation from experiment and simulation, where I_0, v_0, and γ_0 are model fitting parameters. (Figure is taken from [31].) 15

2.9 Conventional and innovative method of fabricating vertical RRAM array. (Figure is taken from [46].) 16

3.1 Example TCAM system circuit block diagram: TCAM array compares input from SL with store data patterns. Search results are generated by MLSA and forward to PE, which determines the highest priority match. (Figure is taken from [55].) 19

3.2 (a) NAND-type ML structure: all cells need to match to discharge ML to ground and output a match. (b) NOR-type ML structure: all cells need to match to minimized pull down current and output a match. (Figure is taken from [57].) 20

3.3 (a) CSI TDE technique by adjusting pull-up and pull-down current of inverter (b) SCI TDE technique by adjusting capacitance attached to intermediate node between inverters. (Figure is taken from [66].) 21

3.4 A n-bit 2T2R TCAM cell array structure. (e.g., the ith cell is loaded with state '0' based on mapping in Table 3.1, R_{ia} in HRS and R_{ib} in LRS.) . . . 23

3.5 TCAM sensing amplifier types: (a) Conventional Pre-charge, (b) Current-race, (c) charge-redistribution, (d) charge-injection. (Figure is taken from [55]). 24

3.6 (a) Basic CR-MLSA circuit implementation (b) Clock gating feature that turns off TCAM circuits to save power. 25

3.7 Simulated functional waveform of CR-MLSA. When clk=0V, MLSA is in pre-discharge phase. When $clk = V_{DD}$, MLSA is in evaluation phase: V_{ML} increases at different speed based on match result until en=0V. t_{search} is search delay between rising edge of clk and V_{OUT}, t_{charge} is ML charging window and NM is noise margin of the MLSA. 27

3.8 Simulated V_{ML} waveform of proposed CR-MLSA in S1 and S2. Due to difference in current charging C_{ML}, V_{ML} rises gradually in S1 and rapidly at the beginning following by a plateau region in S2. 29

3.9 Gaussian Distribution of HRS and LRS resistance through Monte-carlo simulation of the ASU RRAM model with variation in model parameter. . . . 30

3.10 CR-MLSA with ML booster: the ML booster consists of $N2$ and $M8$ use a feedback mechanism to provide I_{boost} to ML. 31

3.11 Simulated functional waveform of CR-MLSA with ML booster. In a match case, $V_{ML,match}$ surpasses $V_{th,N2}$ and activate the ML booster as indicated by the turning point in $V_{ML,match}$ waveform. t_{search}, t_{charge} and NM are same as defined in Section 3.3. 32

3.12 Structure of PE using Latch-and-Reset Approach. (Figure is taken from [75].) 34

3.13 TDE circuit with single RRAM. In normal operation mode, programming transistors $MN4$, $MN5$, $MP4$ and $MP5$ are off. In programming mode, signal *prog* disables $MN2$ and $MP2$ while the RRAM can be SET (using $MP4$ and $MN4$) or RESET (using $MP5$ and $MN5$). 35

3.14 Simulation of pulse programming applied on ASU RRAM to achieve quasi-analog switching: (a) Waveform of voltage applied across RRAM device and (b) RRAM resistance versus time corresponding to each programming pulse. 36

3.15 Simulation results of (a) propagation delay t_p vs RRAM resistance and (b) power consumption vs Propagation delay t_p for single RRAM TDE with 2 RRAM devices models listed in Table 3.5. 37

3.16 Simulated effect of programming two RRAM cells in parallel: (a) I_{PROG} for RRAMs RESET in parallel vs individually: a portion of I_{PROG} for programming the nominal device is reduced and compensated to the device with slower RESET process. (b) Change of RRAM resistance of in parallel RESET vs individual RESET. 38

3.17 TDE circuit with 2 RRAMs. Two transmission gates are added to connect the two RRAMs in parallel during normal operation and RESET programming mode. They are disabled during the SET programming mode. 39

3.18 Parallel RRAM TDE circuit vs single RRAM TDE circuit through simulation data: (a) propagation delay t_p vs RRAM resistance: Range of t_p is slightly reduced in the parallel RRAM TDE. (b) power consumption vs propagation delay: power consumption of parallel RRAM TDE increases due to extra overhead circuit. 40

3.19 (a) t_{pr} and t_{pf} of single RRAM TDE with 2 RRAM devices models listed in Table 3.5. (b) t_{pr} and t_{pf} of parallel RRAM TDE vs single RRAM TDE with RRAM device #1. 41

3.20 Result of Monte-carlo simulation considering RRAM switching variations using ASU RRAM model. (a-d) number of programming pulse N_{pulse} required to program RRAM from LRS to HRS considering RRAM switching variation: (a) single RRAM TDE with normal V_{PROG}, (b) parallel RRAM TDE with normal V_{PROG}, (c) parallel RRAM TDE with V_{PROG} increased by 0.1V and (d) parallel RRAM TDE with V_{PROG} pulse width increased by 10%. (e-h) Programming energy E_{PROG} for programming RRAMs from LRS to HRS corresponding to each of (a-d) respectively. 43

4.1 (a) Example layout of a array of 16 2T2R TCAM cells in 4×4 fashion. With a 128-bit TCAM array, the area ratio between TCAM array and CR-MLSA is ∼25:1. (b) Example layout of a CR-MLSA with ML booster. 46

4.2 Overall TCAM circuit area and ratio of 2T2R TCAM array area with increasing number of cascaded stages. 47

4.3 2-stage CR-MLSA TCAM design using DC with a D-latch ($L1_{DC}$). V_{OUT1} is latched by $L1_{DC}$, which generate the 2nd stage activation signal act. . . . 48

4.4 Simulated unctional waveform of 2-stage cascaded CR-MLSA: (1) Full match: $act = V_{DD}$ enabling the 2nd because of a 1st stage match. Singal $act = V_{DD}$ maintains during evaluation phase of the 2nd stage. $V_{OUT2} = V_{DD}$ because of a 2nd match. (2) 1st stage mismatch: act = 0V and the 2nd stage is not activated. (3) 2nd stage mismatch: $act = V_{DD}$ but V_{OUT2} = 0V. 49

4.5 2-stage CR-MLSA with SRC structure. $L1_{SRC}$ starts with state C (reset) in each new cycle. If 1st stage search is a match, $L1_{SRC}$ enters state B (set) then state D (hold), enabling the 2nd stage. Otherwise, $L1_{SRC}$ enter state D directly from state C, with the 2nd stage disabled. 50

4.6 Flowchart of NAND-based SR latch operation states. 51

4.7 Simulated functional waveform of 2-stage cascaded MLSA with SRC method: (1) Full match: $L1_{SRC}$ enters state B (set) then state D (hold) because V_{OUT1} = 0V, $en'_2 = V_{DD}$ enables the 2nd. $V_{OUT2} = V_{DD}$ because of a 2nd match. (2) 1st stage mismatch: $L1_{SRC}$ stays at state C (reset) and enters state D (hold) directly. en'_2 = 0V and the 2nd stage is not activated. (3) 2nd stage mismatch: $en'_2 = V_{DD}$ but V_{OUT2} = 0V. 52

4.8 Simulated functional waveform of SRC structure of a 2-stage TCAM. V_{OUT1} is active low during evaluation phase. Compared to DC, SRC has lower NM because of simplified circuit structure. 54

4.9 2-stage CR-MLSA with DC structure and SCPC: clock signals clk, $\overline{clk}$ are modified to adapt for SCPC such that both stages are in synchronization in clock phase. Both stages can share the same RSG but use different TDEs. D-latch is removed since there is no need to store 1st output. 56

4.10 2-stage CR-MLSA with SRC structure and SCPC: both stages can share the same RSG but use different TDEs, similarly to DC structure. SR-latch $L1_{SRC}$ is still required for storing V_{OUT1}. 57

4.11 Simulated functional waveform of 2-stage cascaded CR-MLSA with SCPC: (a) DC structure: V_{OUT1} can propagate to the 2nd stage when $clk = V_{DD}$ and continue with the rest of the search during the same clock phase, (b) SRC structure: SR latch is still required to store V_{OUT1}, otherwise 2nd stage does not finish evaluation because V_{OUT1} resets before en_2 = 0V. (1) Full match, (2) First stage mismatch and (3) Second stage mismatch. 58

4.12 Topology of a 4-stage cascading TCAM. 60

4.13 Average E_{search} and t_{search} of 2-stage and 4-stage vs 1-stage 128-bit TCAM design from simulation result. With increasing W_{charge}, Multi-stage design yields lower t_{search} than 1-stage design with E_{search} reduction of $\geq 20\%$ for SRC structure, $\sim 18\%$ for 2-stage design and only $\sim 6\%$ for 4-stage design with DC structure. 62

4.14 Average E_{search} and t_{search} of 2-stage and 4-stage vs 1-stage 128-bit TCAM design. With different activation rate scenarios (increasing r_{subseq}) and W_{charge} = 300nm, With high enough r_{subseq}, average E_{search} of multi-stage TCAM design can exceed E_{search} of the 1-stage design. 65

5.1 Measurement results of (a) I-V curve of Ag/Al_2O_3/Al device with V = -2 to 2V; (b) I-V curve of Ag/Al_2O_3/Al device with V = -4 to 4V; (c) V-t and corresponding I-t curves of Ag/Al_2O_3/Al device with V = -4 to 4V; (d) I-V curve of Al/Al_2O_3/Al device. 69

5.2 Genetic Algorithm (GA) calculation process in a flow chart 70

5.3 LIF neuron circuit implemented using capacitive coupled memory device . 72

5.4 Simulated LIF neuron circuit performance: (a) Input pulse signal with increasing frequency; (b) V_{MEM} with $M1$ in LRS; (c) V_{OUT} with $M1$ in LRS; (d) V_{MEM} with $M1$ in HRS; (e) V_{OUT} with $M1$ in HRS; (h) Input pulse signal with increasing amplitude; (i) V_{MEM} with $M1$ in LRS; (j) V_{OUT} with $M1$ in LRS; (k) V_{MEM} with $M1$ in HRS; (l) V_{OUT} with $M1$ in HRS. Due to higher leakage rate of $M1$ in LRS than HRS, the neuron circuit with $M1$ in LRS requires higher input pulse signal amplitude and/or frequency to generate output spike signals. 74

5.5 ITO/Ag/CZSe/Mo device measured performance (a) I-V curve of (b) Effect of light on low-voltage conductance 76

5.6 Effect of light on ITO/Ag/CZSe/Mo device switching behavior observed in resistive switching measurement: (a) from dark to shining light (b) from shining light to dark 77

5.7 Ag/CZSe/Mo (Annealed) I-V curve of 100 sweeping cycles in with y-axis (current) in (a) linear scale and (b) log scale. 78

5.8 (a) Synaptic behavior (LTP and LTD) of Ag/CZSe/Mo device shown through measurement by applying sequence of SET and RESET voltage pulse to program resistance. (b) Simulation result of a 3-layer feedforward ANN constructed in Python using the Ag/CZSe/Mo device as synapses. 79

List of Tables

2.1 Operations of an RRAM-based IMPLY gate (Logic 0 is encoded by HRS of the RRAM and logic 1 is encoded by LRS) 12

3.1 TCAM cells resistance state and SL/$\overline{\text{SL}}$ voltage corresponding to different data bits 23
3.2 Summary of TCAM design component parameters for simulation with Spectre in Cadence 26
3.3 Simulated effect of using a ML booster in a 64-bit 2T2R TCAM CR-MLSA in S1 and S2 33
3.4 Performance comparison with other eNVM-based TCAMs (ML word size = 64-bit) 34
3.5 RRAM device characteristics and TDE circuit setup 37

4.1 Operation states of the cascading SR latch $L1_{SRC}$ 51
4.2 Performance comparison of 1-stage and 2-stage TCAM design through simulation with 2 cascading approaches 53
4.3 Operation states of the cascading SR latch $L1_{SRC}$ with SCPC 59
4.4 Performance comparison of 1-stage and 2-stage TCAM design through simulation with 2 cascading approaches and SCPC 60
4.5 Performance comparison with TCAM design using 65nm technology. (ML word size >64-bit) 64
4.6 Three activation cases for evaluating average E_{search} in simulation 64

5.1 Decoupled memristor and capacitor parameter of the fabricated Al_2O_3-based RRAM device 71

5.2 Decomposition of solution used for electro-deposition of CZSe 75

List of Abbreviations

1S1R 1-selector-1-RRAM 9, 10, 20

1T1R 1-transistor-1-RRAM 9, 10, 20, 22

2T2R 2-transistor-2-RRAM 22, 29, 43–46, 53, 63, 65, 81

ANN Artificial Neural Network 5, 68, 77–80, 82

BCAM Binary Content-addressable Memory 18

BEOL Back End of Line 46

BL Bit-line 9, 10, 18

CAM Content-addressable Memory 18

CBRAM Conductive Bridge RRAM 7, 68

CF Conductive Filament 7, 8, 14–16, 20, 22, 71, 75, 76

CMOS Complementary Metal Oxide Semiconductor 3, 4, 9, 10, 21, 26, 46, 50, 83

CNN Convolutional Neural Network 12

CPU Central Processing Unit 3

CR Current Race 4, 5, 22–24, 45, 81

CR-MLSA Current Race based Match-line Sensing Amplifier 22, 24, 26, 27, 30, 33, 43–47, 50, 54, 55, 63, 65

CSI Current-starved Inverter 21

CZSe CuZnSe 5, 68, 74–76, 78, 79, 82, 83

CZTSe Kesterite ($Cu_2ZnSnSe_4$) 68, 75

DC Direct Cascading 5, 45, 47, 50, 52–55, 57, 60–62, 65, 81

DLL Digital Delay-locked Loop 20

DNN Deep Neural Network 12

DRAM Dynamic Random-access Memory 3, 8, 13

DTCO Design-technology Co-optimization 1

eNVM emerging Non-volatile Memory 3, 6, 9–11, 20–22, 25, 33, 65, 81

FET Field Effect Transistor 1

GA Genetic Algorithm xiv, 67, 70, 71, 79

HRS High Resistance State 7, 9–11, 14, 22, 25, 28, 38, 41, 68, 69, 73, 75

IC Integrated Circuit 1, 4, 5, 8

IF Integration-and-fire 13

IMC In-memory Computing 11

ITO Indium Tin Oxide 75

LIF Leaky Integration-and-fire 5, 12, 13, 67, 68, 71–73, 79, 82

LRS Low Resistance State 7, 9–12, 14, 23, 25, 28, 41, 68, 69, 73, 75

LTD long-term Depression 78

LTP Long-term Potentiation 78

ML Match-line 4, 5, 18, 19, 22–24, 26–28, 30–33, 43–47, 50, 51, 53, 54, 63, 65, 81

MLC Multi-level Cell 10, 14

MLSA Match-line Sensing Amplifier 4, 5, 18–20, 23–26, 28, 30, 44, 46, 55, 81

MOSFET Metal-oxide-semiconductor Field Effect Transistor 1, 3

NC Neuromorphic Computing 12, 14, 77

NMOS N-type Metal-oxide-semiconductor 63

NVM Non-volatile Memory 11

OxRAM Oxide-RRAM 7, 15, 68, 76

PCRAM Phase-change Random-access Memory 6, 9, 13, 25

PDK Process Design Kit 25, 35, 46, 73

PE Priority Encoder 19, 33

PLL Phase-locked Loop 20

PMOS P-type Metal-oxide-semiconductor 63

PnE Pre-charge and Evaluate 20, 23, 24, 35, 55, 81

RAM Random-access Memory 9, 10, 14, 16, 18, 20, 22

RRAM Random-access Memory 3–17, 20–23, 25–30, 35–44, 46, 63, 67, 68, 74, 76, 79, 81–83

RSG Reference Signal Generator 24, 26, 46, 48, 51, 52, 55, 82

RTA Rapid Thermal Annealing 77, 78, 80, 82

SA Sensing Amplifier 9, 14, 16, 63

SCI Shunt-capacitor Inverter 21

SCPC Same Clock Phase Cascading 5, 45, 55, 57, 59, 60, 63, 65, 81

SL Search-line 18, 82

SNN Spiking Neural Network 12, 13, 67, 72, 82

SRAM Static Random-access Memory 8, 13, 20, 45, 63, 65, 81

SRC SR-latch Cascading 5, 45, 50, 53–55, 57, 60–63, 65, 81

SrcL Source-line 9

STBA Schimitt Trigger based Amplifier 72, 73

STTMRAM Spin-transfer-torque Magnetic Random-access Memory 6, 9, 13

TCAM Ternary Content-addressable Memory 4, 5, 18–23, 27–29, 31–35, 44–47, 49, 50, 53, 57, 59–61, 63, 65, 66, 81, 82

TDE Tunable Delay Element 4, 5, 20–22, 24, 35, 36, 38–42, 44, 46, 48, 52, 55, 59, 63, 81–83

TS Threshold Selector 9, 10, 72

UV ultraviolet 76

VCO Voltage-controlled Oscillator 20

WL Word-line 9, 10, 18

Chapter 1

Introduction

During the last few decades, the continuous scaling of Metal-oxide-semiconductor Field Effect Transistor (MOSFET) has led to exponential growth of the numbers of transistors on dense Integrated Circuit (IC) with a drastic decrease in cost per transistor, as predicted by Moore's law. This was enabled firstly by dimensional scaling of the transistor geometries and then followed by effective scaling through the introduction of technologies such as strain engineering [1], high-k dielectric materials [2], multi-gate structures [3]. However, even with the technological innovations in materials, process, devices, as well as Design-technology Co-optimization (DTCO), technology scaling is inevitably coming to an end due to many challenges. For example, aggressive scaling leads to shortening of Field Effect Transistor (FET) conducting channel and severe short-channel effects, which result in increase of sub-threshold slope and increase of leakage current [4]. Consequentially, power and power density in an IC chip are rapidly increasing, resulting in degraded power-performance trade-off as well as severe thermal and reliability issues. Process variations also increase to a level that designers need to deal with them in physical design as scaling continues [5].

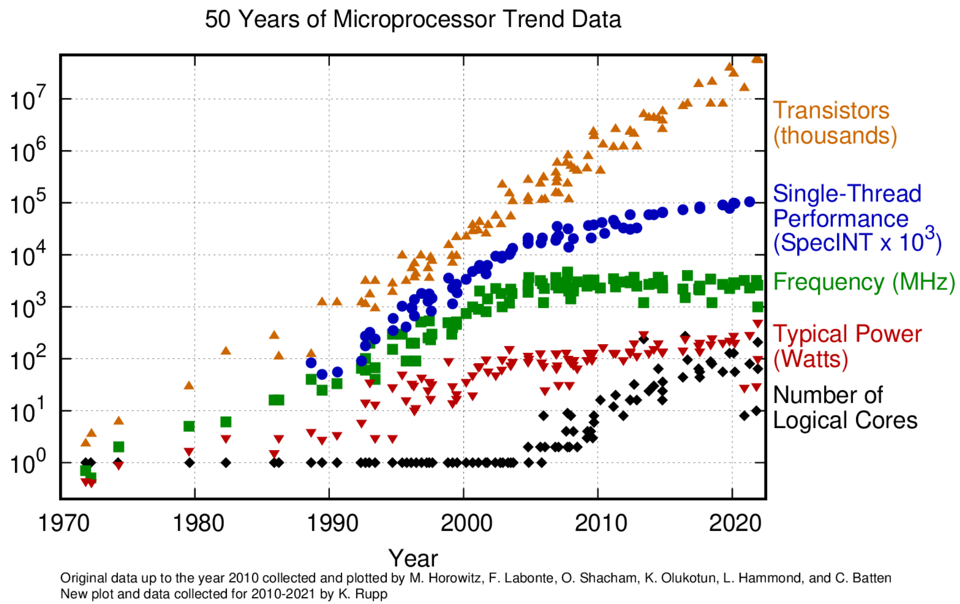

Figure 1.1: Growth microprocessor in terms of transistor counts, performance, frequency, power and number of logic cores. Figure is taken from [6].

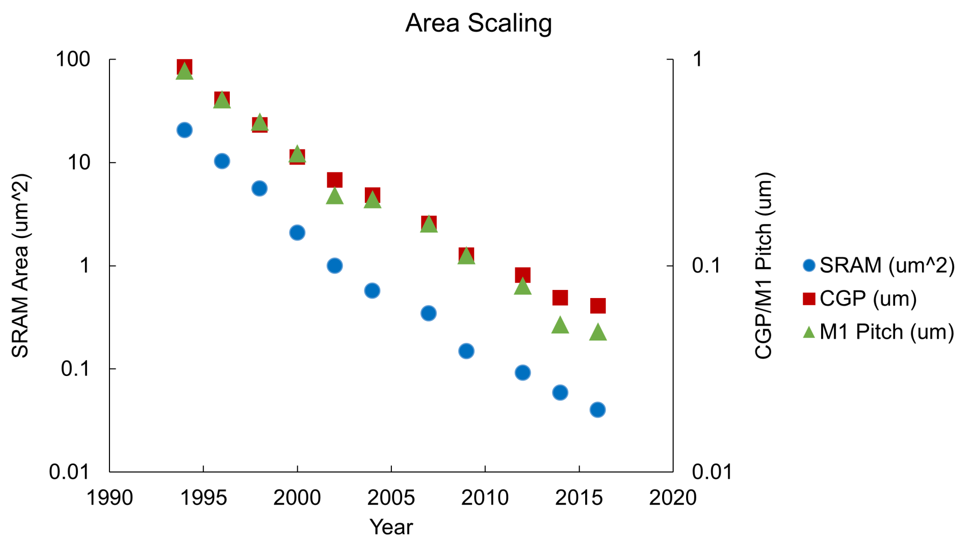

Figure 1.2: CMOS scaling trend: size of SRAM, Contacted Gate Pitch (CGP) and Metal 1 pitch. Figure is taken from [7].

Different technologies [1, 2, 3, 4, 5, 8, 9, 10] have been proposed to resolve issues introduced by MOSFET scaling in order to sustain the growth predicted by Moore's law. Meanwhile, researchers are also actively exploring solutions at every level, such as new devices, novel materials and integration processes, innovated circuit design and system architectures, as well as co-design and co-optimization across these levels [11]. Among all the efforts, Random-access Memory (RRAM) stands out as a promising target of research. As a type of emerging Non-volatile Memory (eNVM) device, RRAM's main advantages can be summarized as follow, which demonstrates that it has great potential to be used in large memory array integration:

- Low programming voltage [12],
- Low write time [13],
- Large high/low resistance ratio [14],
- Small cell area with potential of high density integration [12],
- Compatibility with Complementary Metal Oxide Semiconductor (CMOS) transistor technology [11].

In addition, RRAM has the potential in implementing innovative computing architecture. In the conventional von Neuman Architecture, the Central Processing Unit (CPU) needs to be constantly loading data from and storing data back to a Dynamic Random-access Memory (DRAM). Even though the multiple level of cache memory can be implemented on the CPU side to improve latency performance, the limited bandwidth data transmission between CPU and DRAM is still constraining the overall performance of a system built with this architecture. Furthermore, the performance scaling of CPU (2× per 2 years) and memory (2× per 10 years) are out of balance [15]. This performance gap is another roadblock for further improvements of von Neuman computing system as a whole.

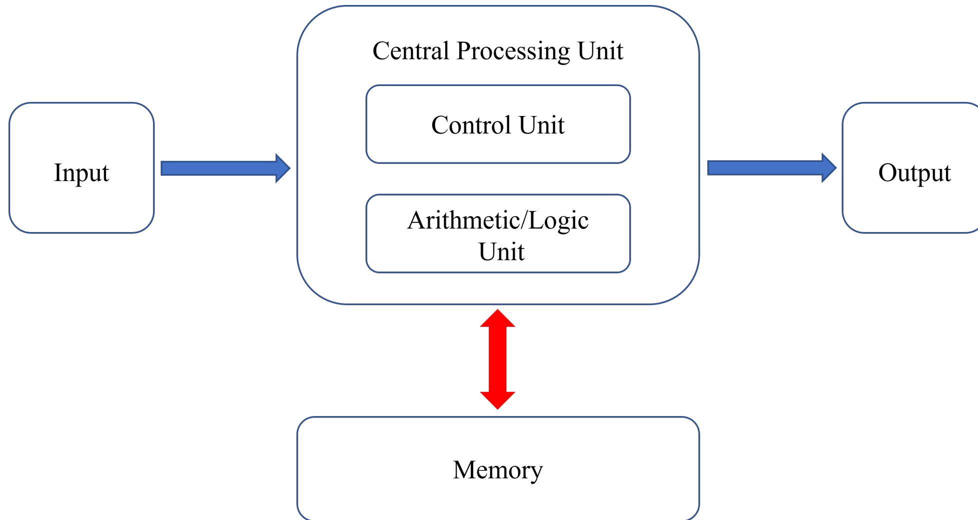

Figure 1.3: von Neuman architecture

While the bottleneck of von Neuman Architecture is becoming a major drawback preventing further advance of computation system, neuromorphic computing based processors have been claimed as an alternative. Examples are IBM TrueNorth [16], Intel Loihi [17] and Google TPU [18]. These processors have their architecture optimized for matrix multiplications in neuromorphic computers. They are known to have energy-delay product performance well exceeding traditional computing architecture. As a non-volatile memory device with resistive switching behavior, RRAM are observed by studies to have synaptic behavior [19, 20, 21]. Furthermore, it is compatible with the current complementary CMOS fabrication process. Therefore, it becomes a sound candidate for building neuromophic computing based IC solution.

1.1 Scope of Research

In this study, I have explored RRAM-based device and circuit for memory (in particular, for Ternary Content-addressable Memory (TCAM) systems) and neuromorphic applications. For the memory application, a TCAM with Current Race (CR)-based Match-line Sensing Amplifier (MLSA) is proposed with high performance, energy efficiency and area benefits. To improve tolerance to RRAM switching variation, a compact Match-line (ML) booster is introduced. Meanwhile, an RRAM-based Tunable Delay Element (TDE) circuit is presented, which can be used in TCAM design to tune the timing of reference control

signal. An improved parallel RRAM TDE is also proposed, to reduce impact of RRAM switching variation to TDE performance and improve achievable delay resolution.

Design methodology of cascading multiple stages of TCAM is also investigated for implementing RRAM-based TCAM with large word size. Two different cascading hardware structures are proposed and investigated: MLSA Direct Cascading (DC) and SR-latch Cascading (SRC). A Same Clock Phase Cascading (SCPC) technique is also introduced to reduce output latency of TCAM circuit with cascading structure. Design of the TCAM and TDE circuit is done using the Cadence Virtuoso design platform. Performance of the circuit designs is evaluated through simulation with Spectre circuit simulator in Cadence.

For neuromorphic applications, RRAM devices with novel behaviors besides the conventional resistive switching are explored as part of this thesis. A Al_2O_3-based RRAM with intrinsic capacitive effect is studied and analyzed for usage in neuron model circuit design. Device characteristics are extracted from measurement of physically fabricated device. The results are then used to design Leaky Integration-and-fire (LIF) neuron circuit in Cadence Virtuoso and evaluated through simulation with Spectre. A CuZnSe (CZSe)-based device is also evaluated in synaptic behavior for usage in building Artificial Neural Network (ANN). Device characteristics are also obtained from measurement of fabricated device. The measurement data is then used to simulate synaptic behavior in ANN model developed in Python. Effect of illumination on this CZSe-based device is also reviewed.

1.2 Organization

This rest of this study is organized as followed. Chapter 2 consists of review of current research trend in application of RRAM in traditional IC and innovative applications. In Chapter 3, a CR-based TCAM design using RRAM with a ML booster is proposed with an RRAM-based TDE circuit proposed to tune timing of reference control signal in TCAM system. In Chapter 4, the TCAM design proposed in chapter 3 is further explored by cascading multiple TCAM stages and aiming for high-speed and energy-efficient applications. In chapter 5, a Al_2O_3-based and a CZSe-based device are analyzed in their potential for usage in neuromorphic computing. Effect of light on the CZSe-based device is also reviewed.

Chapter 2

Related Work

2.1 Memristor and RRAM

The idea of memristor was first proposed by Chua in 1971 as a theoretical two-terminal electronic device in addition to the existing three classical elements: resistor, capacitor and inductor [22]. Its voltage-current relationship can be characterized by the following Equation (2.1). Memristance $M(q)$ is characterized by Equation (2.2) where flux $\varphi(t)$ and charge $q(t)$ represent accumulation of voltage and current respectively. Therefore, $M(q)$ has same unit Ohm (Ω) as resistance and has the memory of voltage applied to and current flowing through the memristor device.

$$v(t) = M(q(t))i(t) \tag{2.1}$$

$$M(q) = \frac{\mathrm{d}\varphi(q)}{\mathrm{d}q} \tag{2.2}$$

The first memristor device was reported by the HP labs in 2008, which was fabricated in a Pt/TiO_2/Pt multi-layer structure [23]. In this device, migration of oxygen vacancies under external voltage bias acts as a deterministic factor of non-volatile resistive switching. This device is known to be an RRAM device. Ever since then, RRAM and other eNVM devices, such as Spin-transfer-torque Magnetic Random-access Memory (STTM-RAM) [24] and Phase-change Random-access Memory (PCRAM) [25] have attracted enormous amount of interest in developing different kinds of eNVM applications.

Resistance state of an RRAM can be changed by applying external voltage bias across its two electrodes. When a SET voltage V_{SET} is applied during the SET process, a Conductive Filament (CF) is formed in the dielectric material between two electrode and the device enters a Low Resistance State (LRS). When a RESET voltage V_{RESET} is applied during the RESET process, the CF is ruptured and the device returns to High Resistance State (HRS). Depending on the polarities of V_{SET} and V_{RESET}, RRAM can be divided into two groups: unipolar (V_{SET} and V_{RESET} share the same polarity) and bipolar switching (V_{SET} and V_{RESET} have different polarities), as shown in Figure 2.1. For the rest of this study, bipolar switching RRAMs will be the focus of discussion since it is more widely studied and used.

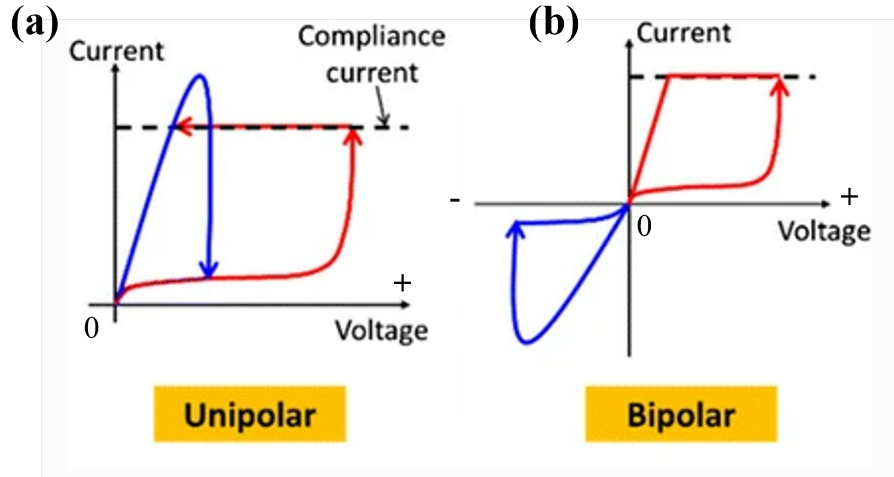

Figure 2.1: I-V curve of (a) unipolar switching RRAM: V_{SET} and V_{RESET} have different polarities. (b) bipolar switching RRAM: V_{SET} and V_{RESET} have different polarities. The red curve represents SET process and the blue curve represents RESET process. Figure is adapted from [26].

RRAM can also be divided into two subcategories of Oxide-RRAM (OxRAM) and Conductive Bridge RRAM (CBRAM) depending on the composition of the formed CF [12]. For a OxRAM, the CF is formed by oxygen vacancies left in the oxide dielectric material after oxygen ions are pulled towards the active electrode when V_{SET} is applied. Under the effect of V_{RESET}, oxygen ions migrates back to the dielectric, recombining with the oxygen vacancies and rupturing the CF [27]. As for a CBRAM, applying V_{SET} will force metal ions such as Ag^{+} and Cu^{2+} to move towards electrode with lower potential. These ions are then reduced to metal atoms at the interface between the dielectric material and

electrode. Accumulation of such metal atoms leads to formation of CFs. With a V_{RESET} applied, the metal atoms clusters are reduced in size and number, causing the CF to rupture. Both the SET and RESET process are repeatable for RRAM. The state-of-the-art RRAM can achieve endurance of $> 10^{12}$ cycles [28] and programmed state retention of $> 10^6$ seconds [29]. Active efforts are still being made for continuous improvement of RRAM performance.

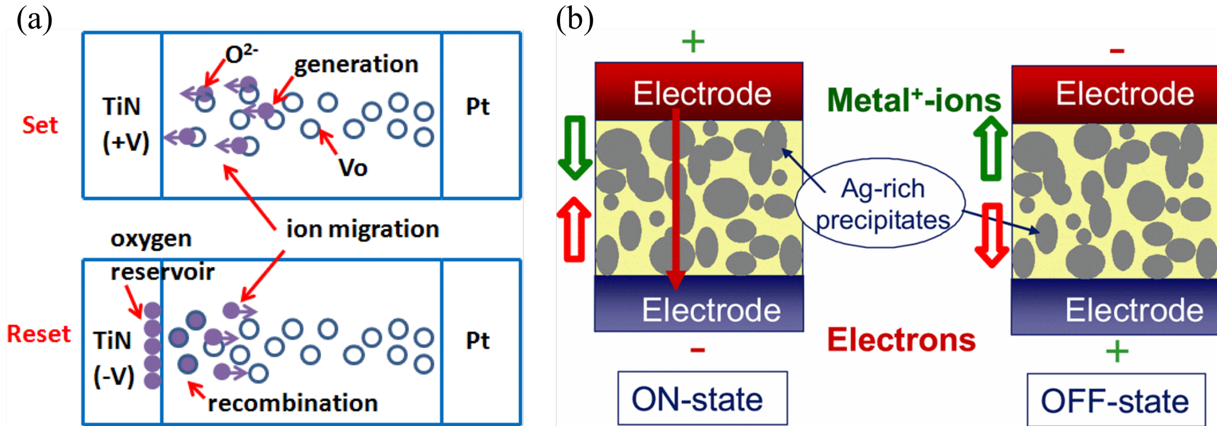

Figure 2.2: Conduction mechanism of (a) OxRAM (Figure is taken from [27]): CF formed by oxygen vacancies. (b) CBRAM (Figure is taken from [30]): CF formed by accumulation of reduced metal ions.

2.2 Applications of RRAM

RRAM has multiple advantages that make it a candidate for building IC applications:

- Low programming voltage (<3V) [12],
- Low write time (<0.1ns) [13] compared to Static Random-access Memory (SRAM) (∼1ns) [12],
- Large high/low resistance ratio ($> 10^4$) [14],
- Scalability with minimum cell area down to $4F^2$, where F is feature size of the lithography (For comparison, DRAM cell area is $6F^2$.) [12],

- Compatibility with CMOS transistor fabrication process [11].

Moreover, as a 2-terminal eNVM devices, RRAM, STTMRAM and PCRAM share similar switching behavior. It makes them interchangeable in a lot of applications already developed on one eNVM technology. This can broaden the scope of potential applications for RRAM. In this section, several perspectives of these applications are reviewed.

2.2.1 Large Scale RAM Array Integration

As a memory device, one natural and widely explored application of RRAM is to build large scale Random-access Memory (RAM) array. Two common cell configurations are 1-transistor-1-RRAM (1T1R) and 1-selector-1-RRAM (1S1R) as shown in Figure 2.3. For the 1T1R configuration, top electrode of the RRAM is connected to the Bit-line (BL) and bottom electrode is connected to drain of a transistor. Source of this transistor is connected to a Source-line (SrcL). Whenever a read/write operation is initiated, the transistor is turned on with voltage applied on the Word-line (WL) connected to the gate of the transistor. In a write operation, depending on the data bit (either 0 or 1) to be written, $|V_{SET}|$ or $|V_{RESET}|$ is applied to BL or SrcL respectively while the other line is grounded. In a read operation, the transistor is also turned on. A read voltage V_{read} significantly smaller than the programming voltage is applied to BL while SrcL is grounded to avoid modifying the existing resistance state. Meanwhile a Sensing Amplifier (SA) can differentiate the stored data bit by sensing the difference between a HRS and a LRS of read current flowing through the targeted RRAM. Because of the low sub-threshold current of a transistor, this configuration can resolve the sneak path problem. Meanwhile, the transistor can also be used for providing a compliance current when programming the RRAM since the maximum current that can flow through the transistor is controlled using gate-to-source voltage (V_{gs}) of a transistor.

In a 1S1R crossbar array, the transistor is replaced with a 2-terminal Threshold Selector (TS). This TS has very low leakage current when voltage bias applied across the device is under a threshold value. In this configuration, top electrode of the RRAM is connected to WL and its bottom electrode is connected to the TS, which is connected to BL on the other end. Since this configuration has no separate control terminal, the unselected WL and BL must be kept at a non-zero potential to avoid the sneak path problem. In a write operation, only the selected cell has programming voltage $V_{write}(|V_{SET}|$ or $|V_{RESET}|)$ applied on one terminal (WL or BL), while the other terminal is grounded. In a $V_{write}/2$ scheme, unselected BL's and WL's are connected to $V_{write}/2$. In a $V_{write}/3$ scheme, unselected

BL's are connected $2V_{write}/3$ and unselected WL's are connected to $V_{write}/3$. As a result, voltage applied across unselected cells is either 0V (in the $V_{write}/2$ scheme) or $V_{write}/3$ (in $V_{write}/3$ scheme) to avoid turning on the TS devices and eliminate sneak paths. In a read operation, all BL's and unselected WL's are connected to V_{read}. Only the selected WL is grounded. Therefore, an entire row of cells are read in parallel. In many cases, the TS is implemented through adding an extra layer of materials in RRAM device, and thus a 1S1R cell is essentially built as one device. This configuration in principle can achieve a smaller cell size than the 1T1R configuration and a higher integration density in RAM array architecture [12].

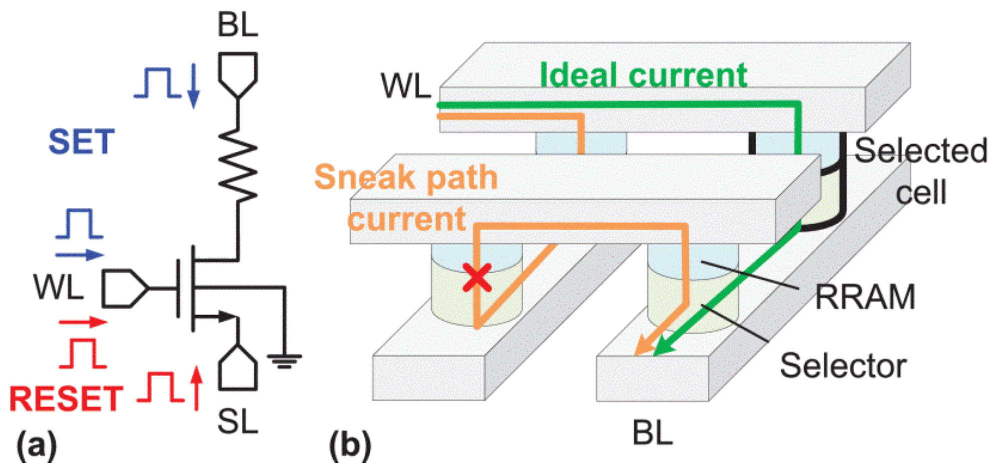

Figure 2.3: 1T1R and 1S1R array structure for implementing large scale RAM. (Figure is taken from [31]. SL here stands for Source-Line)

Studies have also been done to use RRAM as a Multi-level Cell (MLC) to increase amount of data that the memory array can stored. This is done so by utilizing the intermediate resistance states in addition to LRS and HRS. For a 1T1R configuration, it has been shown that different WL signal pulse widths [32] and V_{RESET} amplitudes [33] are two methods to program a single RRAM to 4 different resistance states and store 2-bit of information per RRAM cell.

2.2.2 In-memory Computing and Neuromorphic Computing

The non-linear I-V characteristic of eNVM devices enable the possibility of implementing arithmetic logic computation in a different way from conventional CMOS logic gates.

Furthermore, eNVM devices can store information from logic computation due to their non-volatility, providing an alternative of implementing sequential logic circuit in addition to conventional methods of using data registers. Moreover, the RRAM-based array provides a natural mapping for matrix computation with wide range of applications in image processing and neuromorphic computing. In recent years, RRAM-based In-memory Computing (IMC) has become an active research field, with the hope of breaking the data bottleneck in the conventional von Neuman architecture.

In the arithmetic logic front, an IMPLY logic gate can be constructed using RRAM [34], as shown in Figure 2.4. Logic 0 is encoded by HRS of the RRAM and logic 1 is encoded by LRS. Reference resistor R_G is selected to have resistance between LRS and HRS of the RRAM device. Condition voltage $|V_{COND}| < |V_{SET}|$ is used for logic gate operation. Operation of this IMPLY logic is summarized as listed in Table 2.1. With inputs P and Q initially stored in RRAMs, the IMPLY gate stores the output back to RRAM Q. The IMPLY logic operation is known to be logic complete. Thus, multiple IMPLY logic gates can be cascaded to build more complicated structure such as adders. Similar methods of utilizing the non-linearity and Non-volatile Memory (NVM) characteristics of eNVMs for logic computations include memristor-aided logic (MAGIC) [35], memristor ratioed logic (MRL) [36] and even ternary logic computation [37].

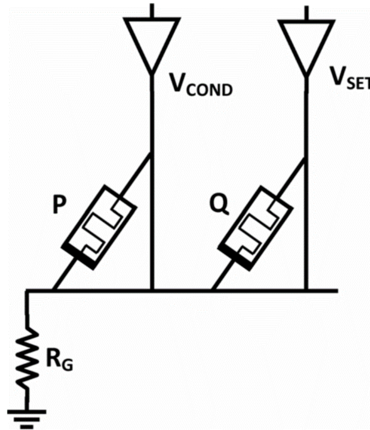

Figure 2.4: Structure of a IMPLY gate consisting of two RRAMs P and Q with reference resistor R_G with resistance between LRS and HRS of the RRAM device. Condition voltage $|V_{COND}| < |V_{SET}|$. (Figure is taken from [34].)

Table 2.1: Operations of an RRAM-based IMPLY gate (Logic 0 is encoded by HRS of the RRAM and logic 1 is encoded by LRS)

P	Q	$Q_{new} = P \rightarrow Q$	Detail
0	0	0	Voltage on Q is $\sim V_{SET}$, Q is SET
0	1	1	Q is already in LRS, no change
1	0	0	P in LRS, voltage across Q is $V_{SET} - V_{COND}$
1	1	1	Q is already in LRS, no change

RRAM has also been extensively used in study of Neuromorphic Computing (NC) to build Deep Neural Network (DNN) such as Convolutional Neural Network (CNN) [38] and Spiking Neural Network (SNN) [19]. In these applications, RRAMs can be used to model synapses connecting between neurons, as shown in Figure 2.5. Its adjustable resistance is used to represent synaptic weights and can be tuned by using different learning algorithms during training processes with sets of training data. With a trained network, new data can be fed in and evaluated in order to generate result about whether or not the input data match any of the trained pattern. These neural network applications have been widely used in pattern recognition.

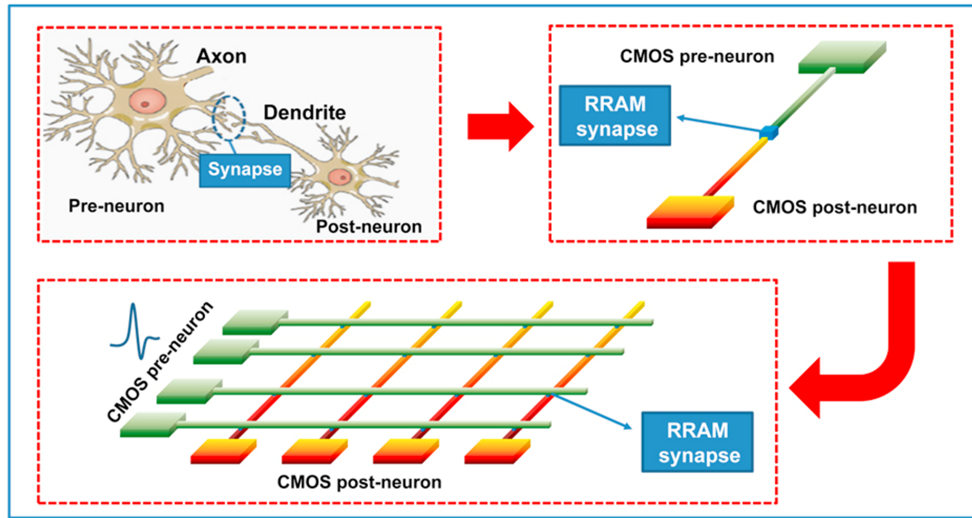

Figure 2.5: RRAM used as synapse to emulate neuron system. (Figure is taken from [20].)

Although not as popular as its applications as synapse model, RRAM-based neuron circuit designs have also been proposed. RRAM is mainly used to implement a LIF neuron

or Integration-and-fire (IF) used in a SNN. The two neuron models are considered to be simplified version of the Hodgkin–Huxley neuron model. A simplified operation model of an LIF is shown in Figure 2.6. In these neuron models, the input synaptic pulse, which can be used for encoding information, are integrated by the neuron circuit as the stored membrane potential increases. If the membrane potential exceeds a built-in threshold value, an output pulse will be generated and propagates to other connected neurons. In addition to this LIF mechanism, a LIF neuron also models the leakage effect which gradually reduces the stored membrane potential, representing previously stored information, over time. In [39], the quasi-analog conductivity increase of RRAM during SET process is used to model the integration behavior of an LIF neuron. In combination with a read (equivalent to neuron firing) and a RESET operation during the same clock period, an LIF neuron can be modelled by a single RRAM. Other neurons designs, such as in [40], typically use RRAM in conjunction with active circuit components (e.g., transistors) to utilize RRAM's switching behavior to mimic neuron firing actions.

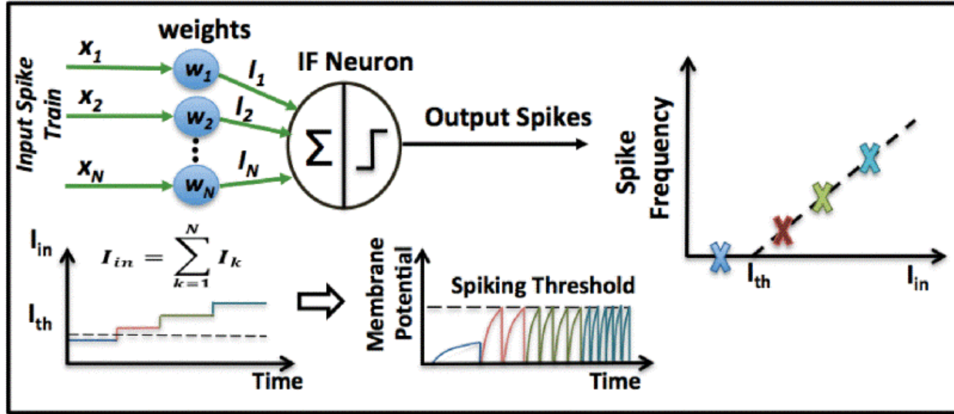

Figure 2.6: LIF neuron working mechanism: input pulse signals are integrated as membrane potential over time. A output spike is generated if the membrane potential exceeds the spiking threshold. (Figure is taken from [20].)

2.3 Switching Variation of RRAM

As an emerging memory technology, RRAM is constantly improving in terms of material, device and fabrication process. While RRAM has various benefits over other volatile (e.g., SRAM, DRAM) and non-volatile (e.g., PCRAM, STTMRAM) memories, one big problem

that RRAM is facing is switching variation, due to the stochastic nature of CF formation based on oxygen vacancies or metal ion migration. This can be further categorized into inter-cell (i.e., device-to-device) and intra-cell (i.e., cycle-to-cycle) variation. Resistance of RRAM can be impacted by variation in programming current and duration of the programming pulses [41]. The origin of such variation in LRS comes from fluctuation in CF radius and constriction geometry, as illustrated in Figure 2.7. As for HRS, the rupture of a CF is a stochastic event, which can either lead to a narrow filament or a ruptured filament [41]. In addition to the intrinsic variation, extrinsic switching variation can also be caused by process induced impurities [42].

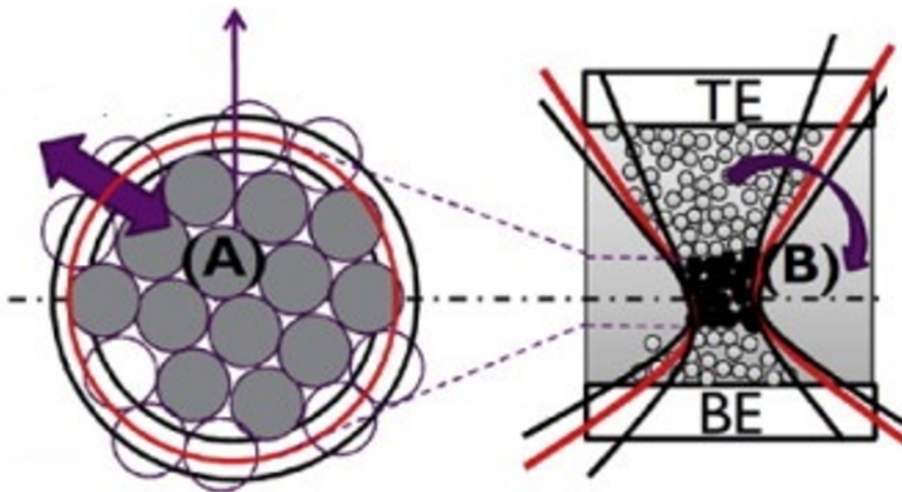

Figure 2.7: Physical origin of resistance dispersion in LRS: (A) CF radius defined by fluctuations in number of particles and (B) fluctuation in CF constriction geometry. (Figure is taken from [41].)

Impact of such variations can be shown in Figure 2.8. Even with the same RESET pulse applied on RRAMs fabricated with the same process, the resultant RRAM HRS resistance can significantly vary among the devices. Furthermore, variation in RRAM resistance can also be observed in LRS. The fluctuation can bring challenges for circuit design to detect RRAM resistance states. For example, in RAM arrays and NC applications, with RRAM resistance fluctuating in both LRS and HRS, difficulty of designing a SA increases since its operation has to account for such resistance variations. In applications involving MLC, the difficulty of programming and sensing circuits further increases when each RRAM stores more than just two resistance states.

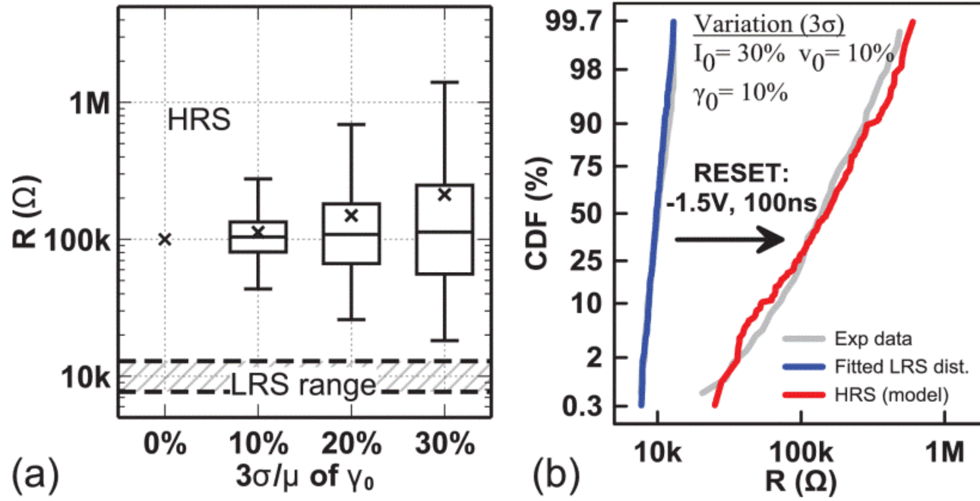

Figure 2.8: RRAM switching variation from experiment and simulation, where I_0, v_0, and γ_0 are model fitting parameters. (Figure is taken from [31].)

Switching variation of RRAMs can be tackled in both fabrication process and application perspectives. Several categories of fabrication methods have shown to be effective in limiting RRAM switching variation. For OxRAM, one method is to introduce dopants in the dielectric materials in order to reduce and control oxygen vacancy formation energy to form stable CF [43, 44]. A metallic or thin oxide buffer layer insertion in RRAM structure can also have the effect in controlling the shape of formed CF, increasing switching stability [45]. Innovative structure is also a solution to mitigate switching variation. For example, an innovative vertical RRAM fabrication process is proposed by [46] as shown in Figure 2.9. A well constrained sidewall oxidation method can effectively prevent unwanted diffusion of oxide material in conventional fabrication method, which is believed to be a main reason of causing switching variation in RRAMs.

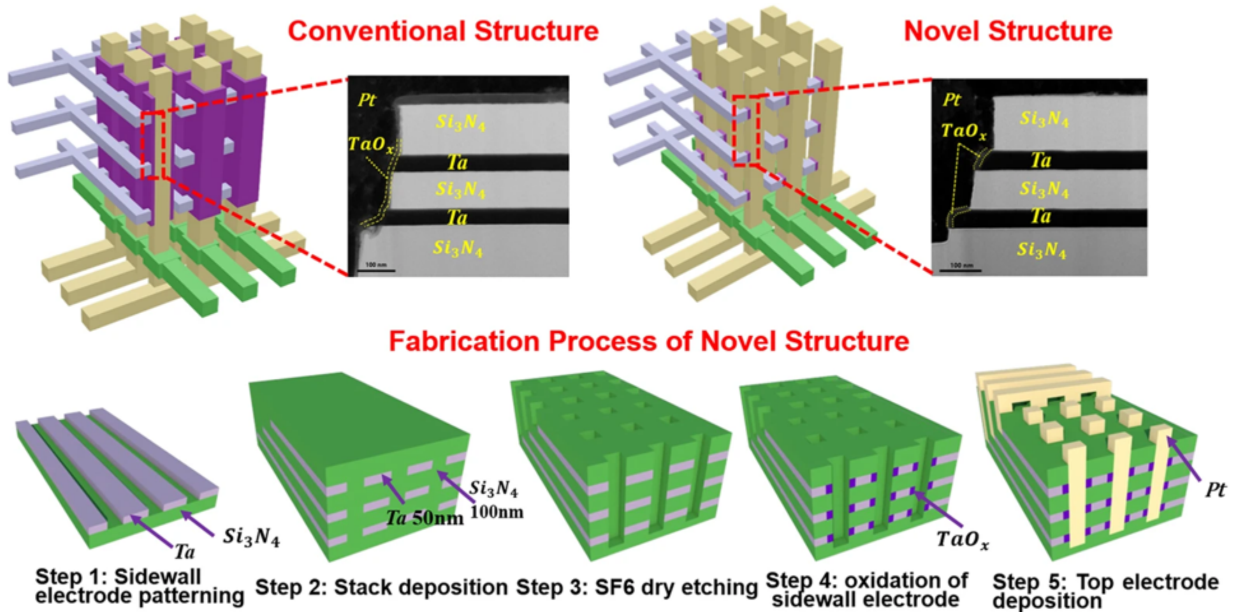

Figure 2.9: Conventional and innovative method of fabricating vertical RRAM array. (Figure is taken from [46].)

As from the application perspective, a verify-after-write mechanism can be incorporated to reduce impact of switching variation [47, 48]. Such mechanism requires a SA to detect resistance of RRAM and a reference scheme of determining if an error (resistance outside of expected range) exists after the programming process. In some applications such as RAM array, such SA circuits are already included in design in order to read and extract the stored information. But space is still needed to be set aside for allocating the reference system. This adds extra overhead to the existing circuit design and costs chip area. Due to the extra overhead system to ensure accuracy, additional power consumption and computation delay are also inevitable.

The problem of RRAM switching variation also impacts RRAM data retention and endurance. If a formed CF is too thin, this can lead to problematic data retention of data stored in an RRAM [49]. If a programming error occurs, which can be identified using verify-after-write mechanism, a re-program is required. This re-programming process negates initial programming result and degrades endurance of the device since the effective number of switching process decreases. Variation can be also observed in a CF forming process for those devices that requires an initial CF forming. Even though a failed forming process could be resolved by using a retry-forming mechanism, the endurance of these

affected RRAM devices is reportedly degraded [42].

Chapter 3

TCAM Design Incorporating RRAM Devices

3.1 Introduction

Content-addressable Memory (CAM) is a memory technology that serves data look-up purpose for different applications. It compares the input data content with stored content and return the address of the best match. Primary applications of CAMs are mainly in the field of data routing in network routers [50, 51, 52]. It can also be found in applications such as image processing [53] and neural network acceleration [54]. Depending on the information stored in each CAM cell, CAM can be further classified into Binary Content-addressable Memory (BCAM) cell and TCAM cell. As their names suggest, a BCAM cell stores only binary states of '0' or '1' while a TCAM cell can also stores an extra 'x' (don't care) state for partial match of data packets. Due to the additional state to be stored, a TCAM cell requires 2-bit encoding and hence twice the amount of hardware of a BCAM cell. Yet a TCAM cell allows a partially matching result to be generated, which is useful in packet forwarding and classification in data routing [55].

An example of TCAM system circuit macro is shown in Figure 3.1. The WL and BL peripherals located on the left and right edges are used to perform read and write operations to the TCAM cell arrays, similarly to operations in RAM. The Search-line (SL) peripherals at the top and bottom are used for data searching purpose. During a new data search cycle, the target data pattern, 144-bit in this example, is presented at the SL's and compared to all 128 stored words, each of which has a ML connected to a MLSA. If a match occurs, the MLSA will generate a match signal. All MLSAs have their output connected

to a Priority Encoder (PE). The role of PE is to determine the top priority match result when there are multiple matches forwarded from MLSAs. The final search result will then be forwarded to expected destinations.

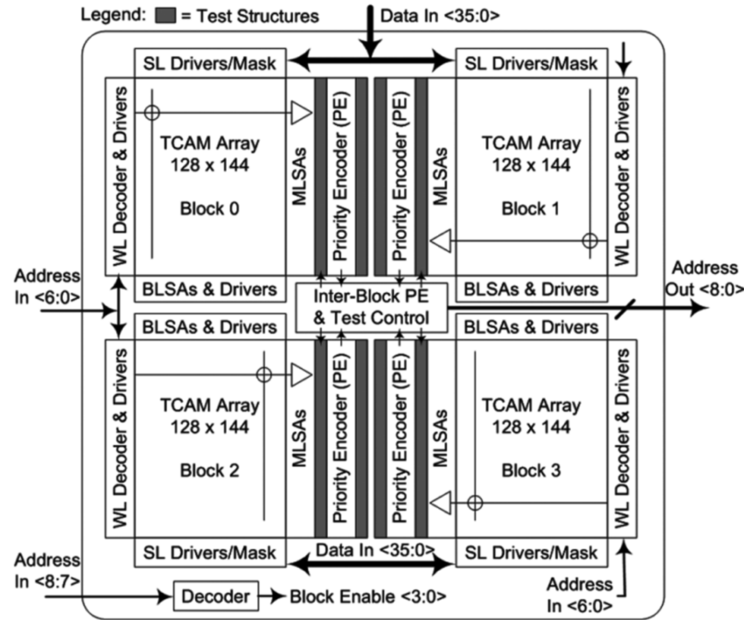

Figure 3.1: Example TCAM system circuit block diagram: TCAM array compares input from SL with store data patterns. Search results are generated by MLSA and forward to PE, which determines the highest priority match. (Figure is taken from [55].)

ML structure can be classified into NAND-type and NOR-type as shown in Figure 3.2. In a NAND-type ML structure, all TCAM cells need to have stored bits matching the input pattern in order to drive ML low and output a match signal. Otherwise, the pull-down path attached to the ML is interrupted. As for a NOR-type ML structure, any bit mismatch would cause total pull-down current to increase. When a match case is detected, the pull-down current is minimum. This feature is utilized by the MLSA to generate output signal based on search results. Between the two types of ML structure, NAND-type is known to have lower power consumption with the cost of lower speed while but NOR-type is known to have higher speed with a cost of higher power consumption [56].

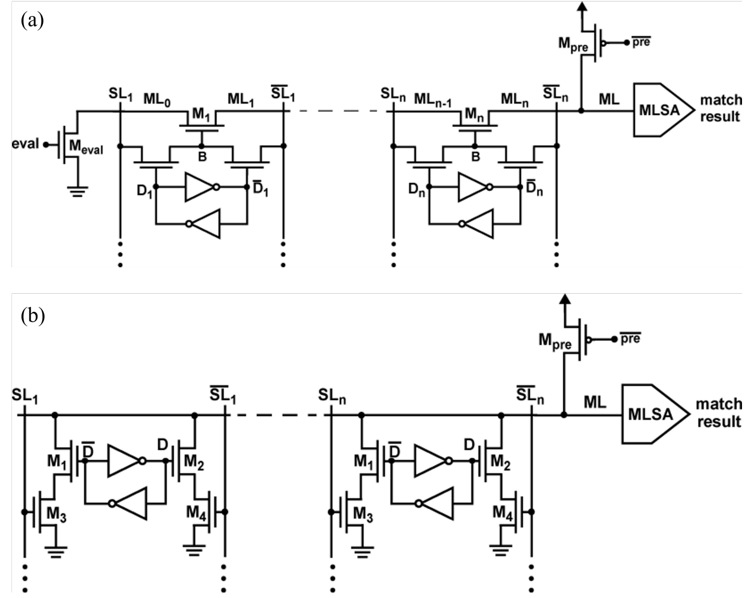

Figure 3.2: (a) NAND-type ML structure: all cells need to match to discharge ML to ground and output a match. (b) NOR-type ML structure: all cells need to match to minimized pull down current and output a match. (Figure is taken from [57].)

Ever since the discovery of eNVM devices, they are considered strong candidates for implementing large scale RAM arrays in 1T1R or 1S1R configuration. The small size of a single memory cell is considerably great advantage for scalability, compared to traditional SRAM cells which consist of 6 transistors. TCAM circuit system, which has traditionally been constructed using SRAM cells, has also been designed using eNVM devices [58, 59, 60, 61, 62]. Most of them adopt MLSA of the Pre-charge and Evaluate (PnE) scheme, for which there are energy efficient alternatives. Furthermore, given the stochastic nature in formation and rupture of CF in RRAM, device switching variation is a concern needed to be addressed in RRAM systems [63, 64]. However, there is still a lack of counter approaches at TCAM circuit design level.

TDE circuits are commonly used to correct timing violation in sequential logic circuit in order to guarantee correctness of digital circuit operation with the targeted clock frequency. It can be found in applications of Digital Delay-locked Loop (DLL), Phase-locked Loop (PLL) and Voltage-controlled Oscillator (VCO)) [65]. As a matter of fact, TDE can be also used in TCAM system design for delaying reference cell signal delivered to each MLSA. This delay signal is used as a guidance of how much time is allocated for a match result to be generated during a search, in order to save search time and energy consumption. A

well designed TDE circuit is important for TCAM design to ensure high performance and energy efficiency.

With the CMOS transistor technology, there are two common methods of implementing TDE circuits. They are Current-starved Inverter (CSI) technique and Shunt-capacitor Inverter (SCI) technique, as shown in Figure 3.3. The CSI technique is realized by using control signal vectors to adjust pull-up and pull-down current of inverters and change signal propagation delay. On the other hand, the SCI technique uses control signal vectors to adjust capacitance connected to the signal propagation path between inverters to tune amount of signal propagation delay. Major concerns of these design techniques are that power consumption and chip area can be significant since more components are involved in order to achieve finer tunable delay resolutions.

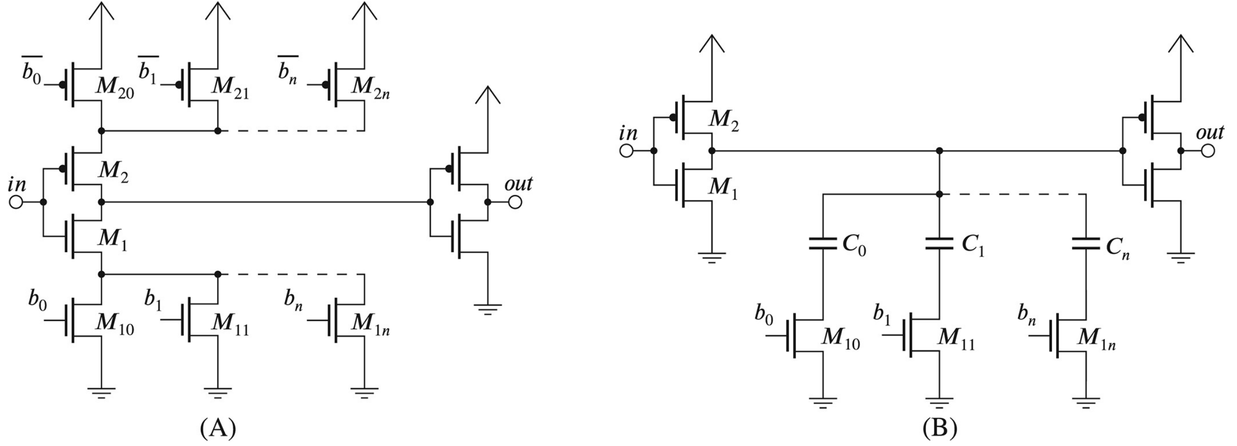

Figure 3.3: (a) CSI TDE technique by adjusting pull-up and pull-down current of inverter (b) SCI TDE technique by adjusting capacitance attached to intermediate node between inverters. (Figure is taken from [66].)

There have been several TDE circuits proposed recently using eNVM devices [67, 68, 69]. The common strategy of these designs is to use eNVM as quasi-analog switching devices to provide a range of propagation delay. Such quasi-analog switching can be accomplished by setting compliance current on a transistor or use pulse signal for programming. Because RRAM has multiple advantages including small size and ease of programming for large resistive range [12, 13, 14], it has great potential to be used for building compact TDE circuit. However, one major concern of using RRAM for quasi-analog switching is the switching variation, which has not been address by these previous proposed TDE design using eNVM devices.

The switching variation in RRAM originates from the randomness present in the CF formation and rupture process. It can be reflected in programming voltages and resultant resistance. There have been studies on device fabrication level to address this issue [43, 44, 45, 46]. On circuit application level, a voltage/current reference scheme is usually deployed to implement a verify-after-write mechanism [47, 48]. However, such application solutions are power-hungry and consume large silicon area for designing a reference circuit system. A energy- and area-efficient approach to address switching variation is still a roadblock for using RRAM or other eNVM devices in compact TDE circuit design.

In this chapter, a 2-transistor-2-RRAM (2T2R) CR based TCAM design is proposed with a compact and energy-efficient ML booster introduced to improve searching speed and counteract RRAM switching variation. This is the foundation work for scaling up word length of RRAM TCAM circuit design in the next chapter. Meanwhile, a TDE circuit design that can be incorporated in the TCAM system is proposed. A parallel configuration of RRAM in TDE circuit is also proposed to reduce impact of switching variation to the performance of the TDE circuit. Design of the TCAM and TDE circuit is accomplished using the Cadence Virtuoso design platform. Performance of the circuit design is evaluated through Spectre circuit simulator in Cadence. In summary, the main contributions of this thesis in this chapter are: (a) in-depth analysis of the 2T2R CR-based TCAM impacted by multiple factors, (b) proposing a compact ML booster to counteract negative impact from RRAM switching variations and improve Current Race based Match-line Sensing Amplifier (CR-MLSA) performance with negligible cost in extra energy consumption, (c) proposing a RRAM-based TDE and (d) using parallel RRAM configuration to improve achievable delay resolution and regulate impact of RRAM switching variations.

3.2 2T2R TCAM Cell Structure

2T2R TCAM cell structure proposed in [58] is used to build the TCAM array. It has a similar structure to the common 1T1R configuration used in RAM memory arrays. Programming circuit used for RAM can be used to write data into the RRAM pairs (R_{ia} and R_{ib}, i=0, 1..., n-1) in the 2T2R TCAM array according to mapping in Table 3.1. Such programming connection can be disconnected during normal operation mode. Since all RRAMs of the same ML are connected in parallel, multiple cells can be programmed at the same time. During a search of a n-bit word, each of the $\mathrm{SL}_i/\overline{\mathrm{SL}_i}$ (i=0, 1..., n-1) pairs is set based on mapping listed in Table 3.1 to corresponding bit in the input word. If the ith bit is a match, the ML is discharged through the path with RRAM in HRS in the ith TCAM cell, while the other path is turn off. If the ith bit is a mismatch, the ML is instead

discharged through the path with RRAM in LRS with much higher discharging current. If a 'x' is stored in the ith bit, the search of this bit always returns with a match result regardless of the actual input data bit. These TCAM cells are connected to the ML in a NOR-type ML fashion, which is described in the previous section.

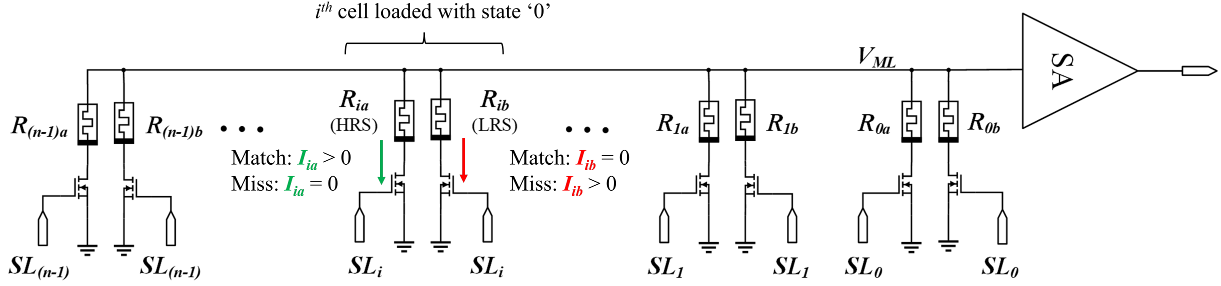

Figure 3.4: A n-bit 2T2R TCAM cell array structure. (e.g., the ith cell is loaded with state '0' based on mapping in Table 3.1, R_{ia} in HRS and R_{ib} in LRS.)

Table 3.1: TCAM cells resistance state and SL/$\overline{\text{SL}}$ voltage corresponding to different data bits

Stored bit	R_{ia} state	R_{ib} state
0	HRS	LRS
1	LRS	HRS
X	HRS	HRS
Search bit	SL (RRAM 1)	$\overline{\text{SL}}$ (RRAM 2)
0	V_{DD}	GND
1	GND	V_{DD}

3.3 Current Racing Sensing Scheme

There are multiple methods of implementing MLSA for TCAM design, as shown in Figure 3.5. The PnE scheme requires SL/$\overline{\text{SL}}$ input to each TCAM cell to be turned off during the pre-charge phase. Considering the number of bits in each word line and number of words in each array, constantly switching SL/$\overline{\text{SL}}$ in each clock cycle increases overall energy consumption. On the other hand, CR is a simple scheme that reduces energy consumption by only changing SL/$\overline{\text{SL}}$ based on input data pattern. In [55], applications up to 144-bit

are examined, showing that the CR scheme reduces energy consumption by 50% compared to the PnE scheme. Therefore, the current race scheme is selected for further discussion of implementation of MLSA in this study.

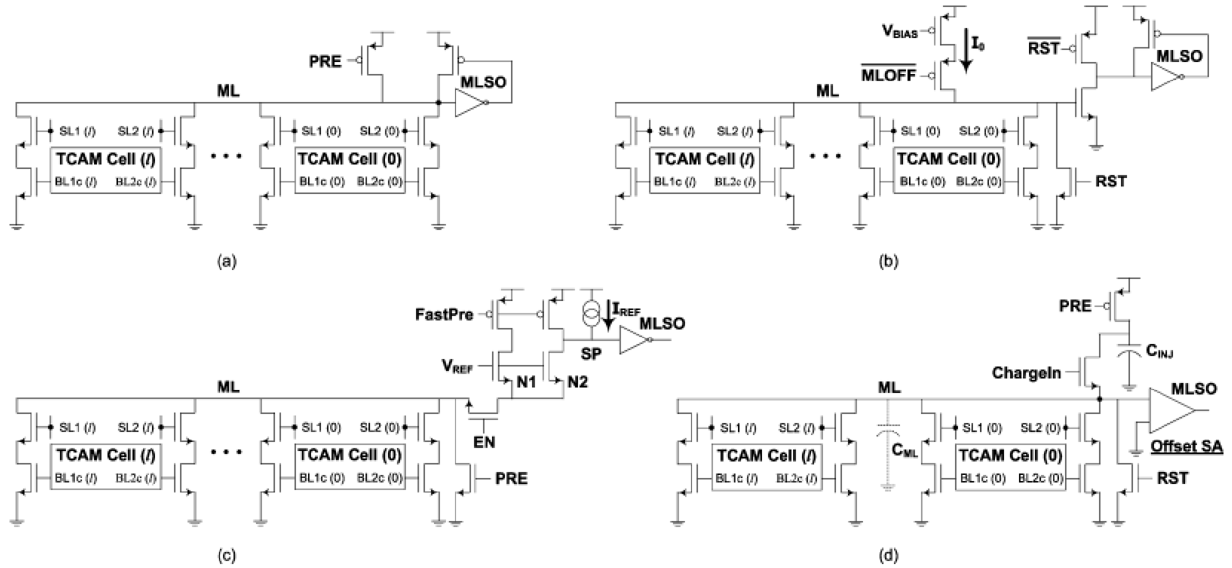

Figure 3.5: TCAM sensing amplifier types: (a) Conventional Pre-charge, (b) Current-race, (c) charge-redistribution, (d) charge-injection. (Figure is taken from [55]).

Circuit of the CR-MLSA [70] is shown in Figure 3.6(a). For correct operation of a CR-MLSA, a Reference Signal Generator (RSG) implemented as a dummy MLSA with no mismatch is used to generate reference signal *en*. This signal *en* is used during the evaluation phase to interrupt charging current I_{charge} to each ML and latch match results [70]. TDEs are used to adjust timing of signal *en* sent to each MLSA, accounting for any circuit variation from fabrication. A clock gating technique, as shown in Figure 3.6(b), can also be incorporated in the circuit such that the system energy consumption can be minimized during idle mode.

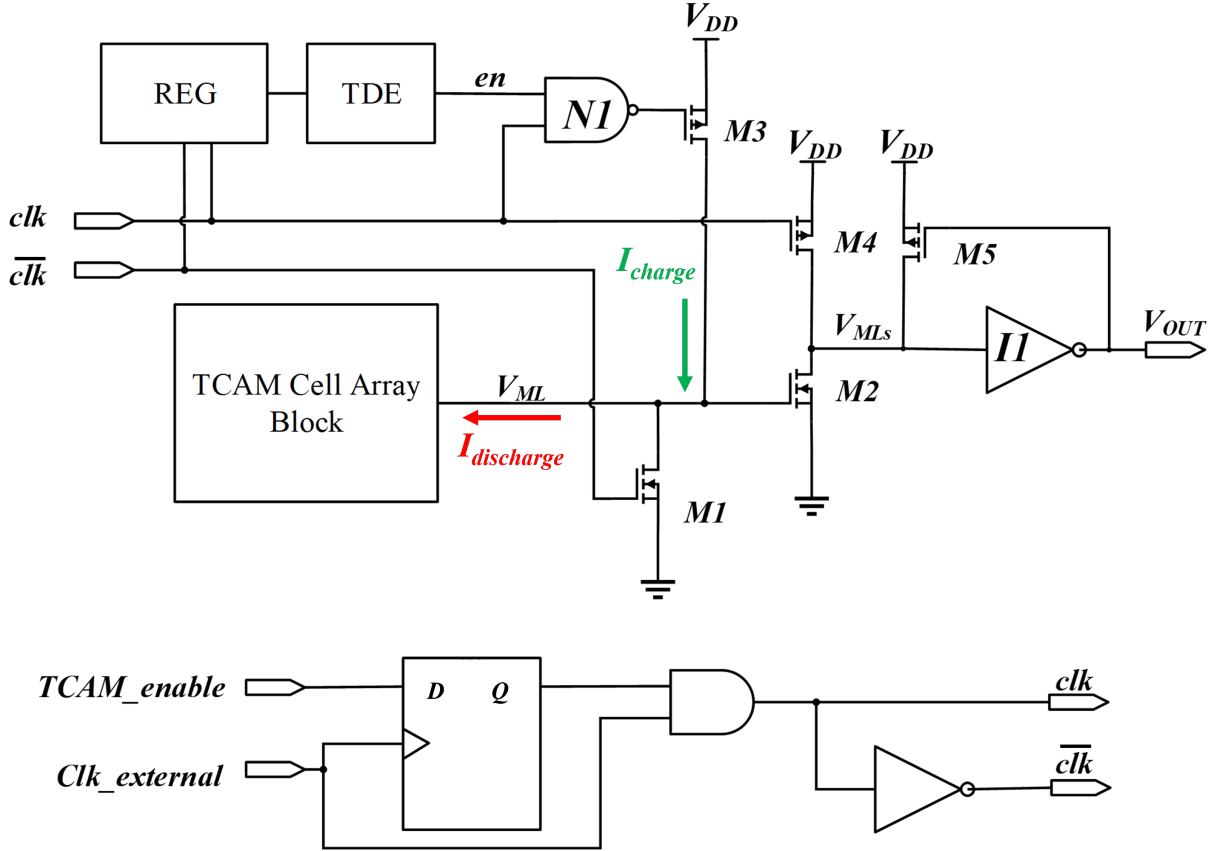

Figure 3.6: (a) Basic CR-MLSA circuit implementation (b) Clock gating feature that turns off TCAM circuits to save power.

In order to proceed with further analysis and evaluate performance of the circuit design proposed, TSMC 65nm Process Design Kit (PDK) and ASU RRAM model [71] fitted to data of IMEC HfO_x-based RRAM devices [72, 73] are used for simulation using Spectre in Cadence. Key design parameters are summarized in Table 3.2. The design scheme is generally applicable to eNVM devices including RRAM and PCRAM. However, it is imperative that the device programming voltage, i.e. V_{SET} and V_{RESET}, is higher than applied voltage during search operation ($V_{search} = V_{DD}$) to avoid altering saved device state unintentionally. Also, using an eNVM device with high HRS/LRS ratio is beneficial to increase the MLSA noise margin.

Table 3.2: Summary of TCAM design component parameters for simulation with Spectre in Cadence

Transistor			RRAM [71]		V_{DD}
Tech	Width	Length	V_{SET}/V_{RESET}	R_{HRS}/R_{LRS}	
TSMC 65nm	200nm	60nm	2/-1.2V	3.35M/10K	0.7V

The functional waveform of the proposed CR-MLSA is shown in Figure 3.7. Two operation phases of CR-MLSA are listed as follow:

- Pre-discharge ($clk = 0$V and $\overline{clk} = V_{DD}$): $M1$ is turned on to discharge ML voltage (V_{ML}) to 0V, which turns off $M2$. $M3$ is also turned off, so V_{ML} maintains at 0V. $M4$ is turned on to charge V_{MLs} to V_{DD}. This resets CR-MLSA output (V_{OUT}) to 0V.

- Evaluation ($clk = V_{DD}$ and $\overline{clk} = 0$V): $M1$ and $M4$ are turned off. I_{charge} from V_{DD} to ML is enabled by $M3$. Meanwhile, the ML is discharged by $I_{discharge}$ based on match results. If a match is detected, V_{ML} surpasses MLSA threshold voltage ($V_{th,match}$) to drive V_{MLs} to 0V and V_{OUT} to V_{DD} until the next search cycle. If a mismatch is detected, $V_{MLs} < V_{th,match}$ during the evaluation phase and V_{OUT}=0V.

During every evaluation phase, the RSG always mimics a MLSA with match result. Reference signal en is a delayed and inverted signal from RSG's output. During every evaluation phase, $en = V_{DD}$ until RSG finishes a search. Once en drops to 0V, $M3$ is disabled to cut off I_{charge} to each ML. Thus, the ML charging window (t_{charge}) is defined as the time period when clk and en are both at V_{DD}. Among all mismatch cases, a 1-bit miss yields the highest $V_{ML,miss}$ ($V_{ML,1-miss}$). Noise margin (NM) of a CR-MLSA is defined as difference between $V_{th,match}$ and maximum $V_{ML,1-miss}$ during t_{charge} [74], as shown in Figure 3.7. This indicates how much tolerance a design can provide to device variations in both CMOS transistors and RRAMs.

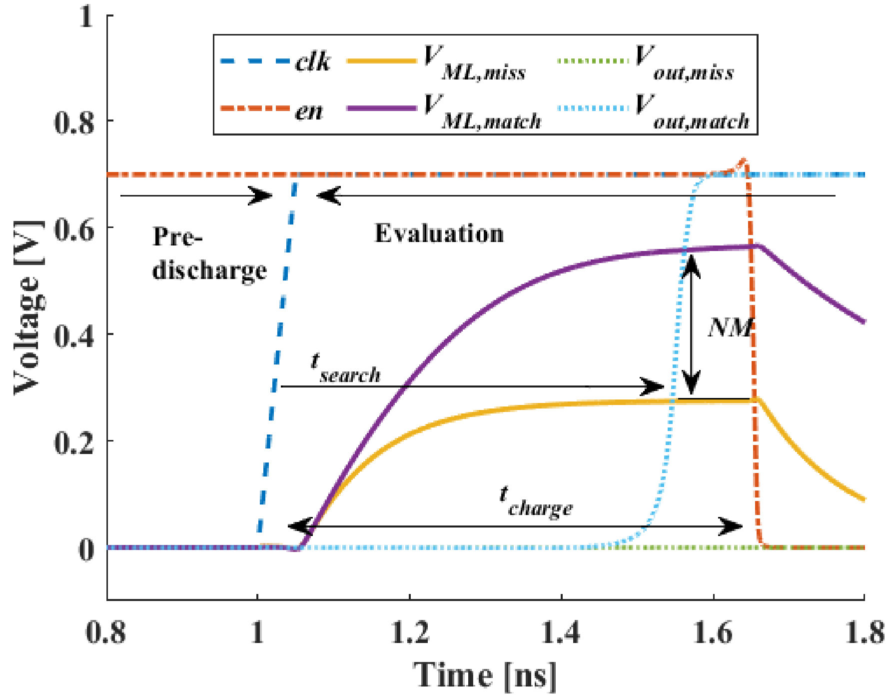

Figure 3.7: Simulated functional waveform of CR-MLSA. When clk=0V, MLSA is in pre-discharge phase. When $clk = V_{DD}$, MLSA is in evaluation phase: V_{ML} increases at different speed based on match result until en=0V. t_{search} is search delay between rising edge of clk and V_{OUT}, t_{charge} is ML charging window and NM is noise margin of the MLSA.

V_{ML} is a key to CR-MLSA performance. High enough $V_{ML,match}$ ensures CR-MLSA generates $V_{OUT} = V_{DD}$ within t_{charge}. Further increasing $V_{ML,match}$ reduces search delay (t_{search}) between rising edge of clk and V_{OUT} and improves NM. Meanwhile, low $V_{ML,miss}$ is desirable to limit I_{M2}. Three main factors from the TCAM cells that affect V_{ML} are listed below:

- Increasing ML word size of bits (n-bit),
- Data pattern stored in TCAM cells,
- Increasing RRAM resistance variations.

Increasing word size increases $I_{discharge}$ proportionally, which decreases V_{ML}.The other two factors are explained in more details in the following sections.

3.4 Impact of Stored Data on MLSA Performance

In a match or 1-bit miss search cases during evaluation phase, the range of V_{ML} rising trajectory is bounded by two extreme scenarios determined by data stored in the other (n-1) bits of TCAM cells:

(S1) all other (n-1) cells store definite value of '1' or '0' (i.e., each RRAM pair has one in LRS and one in HRS)

(S2) all other (n-1) cells store 'x' (all these RRAMs are in HRS)

As illustrated by simulation results in Figure 3.8, V_{ML} has a gradual and logarithmic increase in S1 but a rapid increase at the beginning followed by a plateau region in S2. While the load capacitance of ML (C_{ML}) from interconnect and device parasitic capacitance is relatively biased and state independent, the current charging C_{ML} varies due to the difference in RRAM resistance on the leakage paths through the (n-1) cells. C_{ML} is charged much faster in S1 with half of RRAMs in the (n-1) cells in LRS than in S2 where all RRAMs in the (n-1) cells are in HRS. In S2, if $V_{th,match}$ of a MLSA locates within the voltage range of this plateau region, t_{search} becomes unpredictable when considering variations of other circuit components. Therefore, $V_{ML,match}$ should ideally be pushed to well above required $V_{th,match}$ to avoid uncertainty that may arise in t_{search}.

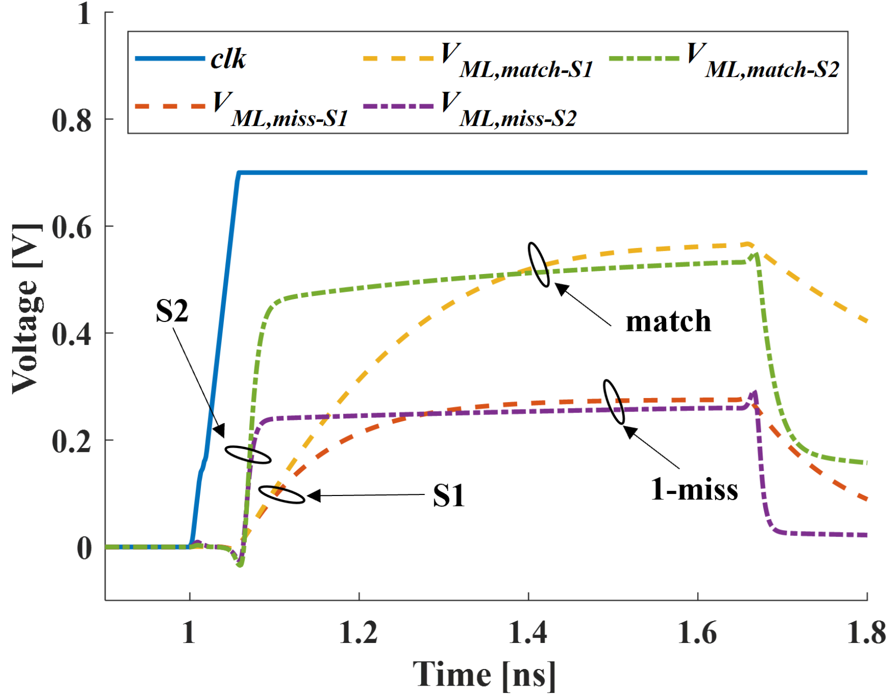

Figure 3.8: Simulated V_{ML} waveform of proposed CR-MLSA in S1 and S2. Due to difference in current charging C_{ML}, V_{ML} rises gradually in S1 and rapidly at the beginning following by a plateau region in S2.

3.5 RRAM Variation and Impact on V_{ML}

RRAM switching variation is a serious issue that requires attention for RRAM-based circuits. Variations in R_{HRS} and R_{LRS} of RRAM used in this TCAM design can be modelled by using the ASU RRAM model in Monte-carlo simulation and applying $3\sigma/\mu$ variations to model fitting parameters I_0 (30%), γ_0 (10%) and v_0 (10%) [31]. The results shown in Figure 3.9 indicate $3\delta/\mu$ variation of ~10% of nominal value is observed for both R_{HRS} and R_{LRS}. In a match case, the RRAM with lowest R_{HRS} dominates total pull-down resistance (R_{pd}) of the 2T2R array. 10% variation is equivalent to a 10% reduction in R_{pd} and increasing $I_{discharge}$. Such variation can be impacted by magnitude and duration of applied V_{RESET}. With increasing R_{HRS} variation, $V_{ML,match}$ is expected to decrease, which can result in increasing t_{search} and possibly failed search operation.

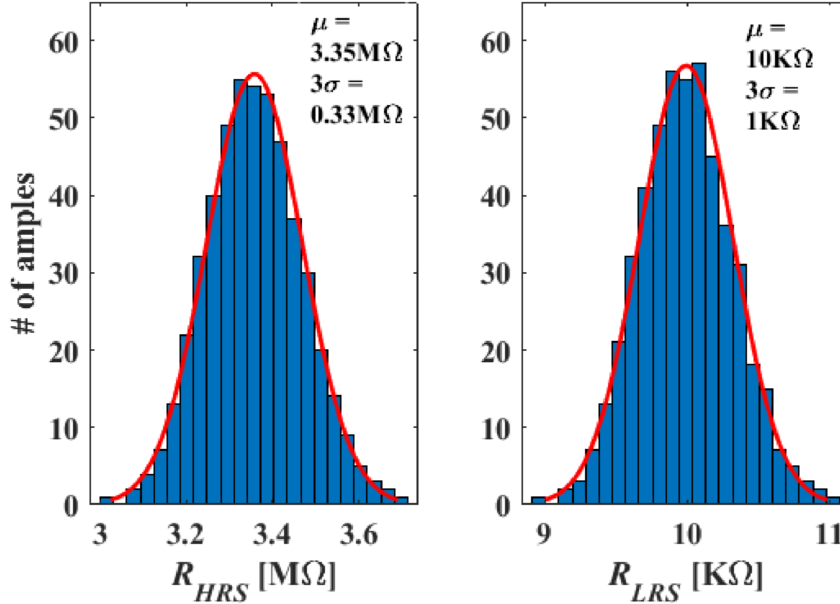

Figure 3.9: Gaussian Distribution of HRS and LRS resistance through Monte-carlo simulation of the ASU RRAM model with variation in model parameter.

3.6 ML Booster

To guarantee CR-MLSA functionality under any data storage pattern and considering RRAM variations, one method is to increase I_{charge} by increasing V_{DD} and/or transistor width of M3 (W_{charge}). But it also increases both $V_{ML,match}$ and $V_{ML,miss}$, which increase overall power consumption. Another method is to reduce $I_{discharge}$ by encoding SL/$\overline{\text{SL}}$ inputs [58]. However, it requires extra area to implement overhead circuit for the encoding feature. In this study, a compact ML booster as shown in Figure 3.10 is proposed to dynamically increase total charging current to ML and raise $V_{ML,match}$. It consists of a 2-NAND gate $N2$ with an exta charging transistor $M8$ to provide boosting charge current I_{boost}. It incorporates a feedback mechanism with one input of $N2$ connected to V_{ML}. $N2$ is designed to have threshold voltage $V_{th,N2}$ lower than $V_{th,match}$ of the original MLSA. Once it is turned on, $M8$ is enabled to increase I_{boost}. As for the mismatch cases, with $V_{ML,miss}$ kept below $V_{th,N2}$, $M8$ stays off and $I_{boost} \approx 0$A. Thus, no extra dynamic power is consumed. Increasing ML word size and RRAM resistance variation are expected to lower V_{ML}, which work in favour of keeping $M8$ off in mismatch cases. However, V_{DD} needs to

be restricted to avoid $V_{ML,miss}$ surpassing $V_{th,N2}$.

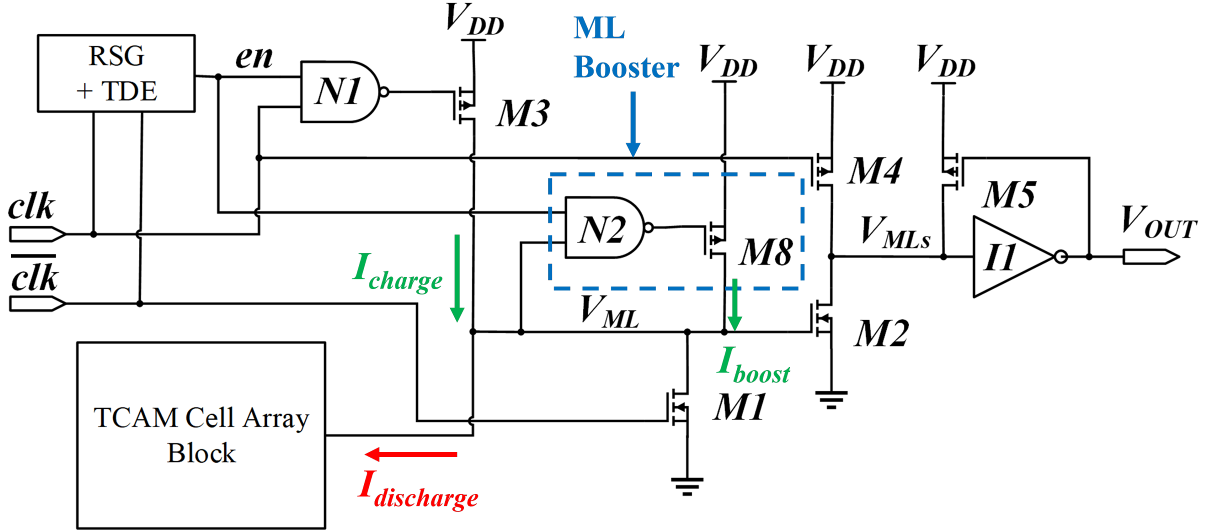

Figure 3.10: CR-MLSA with ML booster: the ML booster consists of $N2$ and $M8$ use a feedback mechanism to provide I_{boost} to ML.

The effect of using a ML booster with a 64-bit TCAM SA structure can be illustrated through simulated functional waveform shown in Figure 3.11. In a match case, when $V_{ML,match}$ surpasses $V_{th,N2}$, the ML booster is activated as a turning point can be observed in the waveform. This speeds up rising of $V_{ML,match}$ and boosts up final $V_{ML,match}$. Hence, the ML booster can reduce t_{search} and increase NM. In the 1-miss mismatch cases, there is no turning point observed in $V_{ML,1-miss}$ waveform, indicating that the ML booster is inactive. If there are more than 1 bit mismatch, $V_{ML,miss} < V_{ML,1-miss}$. Therefore, ML booster is kept inactive in all mismatch cases.

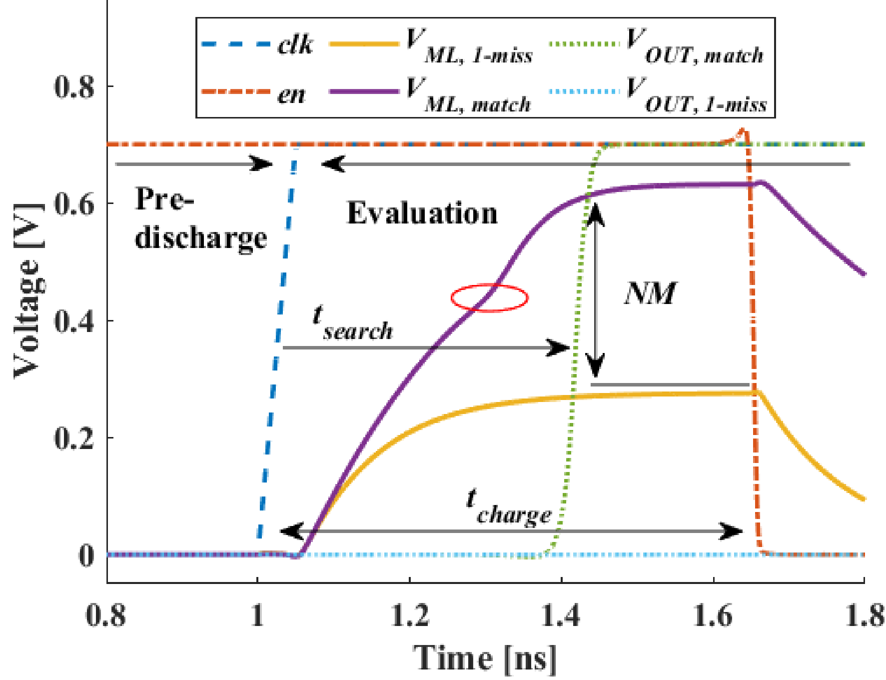

Figure 3.11: Simulated functional waveform of CR-MLSA with ML booster. In a match case, $V_{ML,match}$ surpasses $V_{th,N2}$ and activate the ML booster as indicated by the turning point in $V_{ML,match}$ waveform. t_{search}, t_{charge} and NM are same as defined in Section 3.3.

Effect of using a ML booster in a 64-bit TCAM is illustrated through simulation with the boundary cases S1 and S2 as summarized in Table 3.6. In S1, the ML booster reduces t_{search} and improves NM by more than 20% with almost no extra search energy consumption (E_{search}). In S2, the original design fails as no valid V_{OUT} is generated within t_{charge}, due to $V_{ML,match}$ being stuck at the previously discussed plateau region below $V_{th,match}$. With the ML booster, this issue is fixed and the search is successful. Based on the simulation results, it is clear that the ML booster efficiently improve search speed and allow circuit to be functional with small transistor sizes.

Table 3.3: Simulated effect of using a ML booster in a 64-bit 2T2R TCAM CR-MLSA in S1 and S2

S1 (match)	t_{search} (ns)	NM (V)	E_{search} (fJ/bit/search)
w/o booster	0.53	0.28	0.215
w/ booster	0.40	0.34	0.216
S2 (match)	t_{search} (ns)	NM (V)	E_{search} (fJ/bit/search)
w/o booster	Failed	Failed	0.255
w/ booster	0.20	0.31	0.204

3.7 Performance Comparison with Other eNVM-based TCAM Designs

The proposed CR-MLSA design with ML Booster is evaluated against other reported eNVM-based 64-bit TCAM design with results summarized in Table 3.4. The proposed design is evaluated with a pessimistic 50% variation in R_{HRS} (50% reduction of R_{HRS}). Loading effect of the next stage circuit is also an important consideration since it can impact both search delay and energy consumption. Since there are multiple ways that a search PE circuits can be implemented [55, 75], CR-MLSA performance will be impacted by actual circuit implementation of PE because of the loading effect. For evaluation in this study, the output terminal of the CR-MLSA is connected to an inverter to simulate the loading condition of a Latch-and-Reset PE circuit, as shown in Figure 3.12. Last but not least, the average E_{search} of the proposed design is calculated based on a 5% match rate.

Table 3.4: Performance comparison with other eNVM-based TCAMs (ML word size = 64-bit)

	TCAM cell	Technology	SA	V_{DD} (V)	t_{search} (ns)	Average E_{search} (fJ/search/bit)
[58]	2T2R	IBM 90nm + PCRAM	PnE	1.2	1.9	N/A
[60]	3T1R	90nm + RRAM	PnE	1	0.96	0.51
[59]	4T2R	180nm + RRAM	PnE	0.7	1	N/A
[61]	5T2R	45nm + RRAM	PnE	N/A	0.4	0.23
This work	2T2R	TSMC 65nm + RRAM	CR	0.7	0.58	0.21

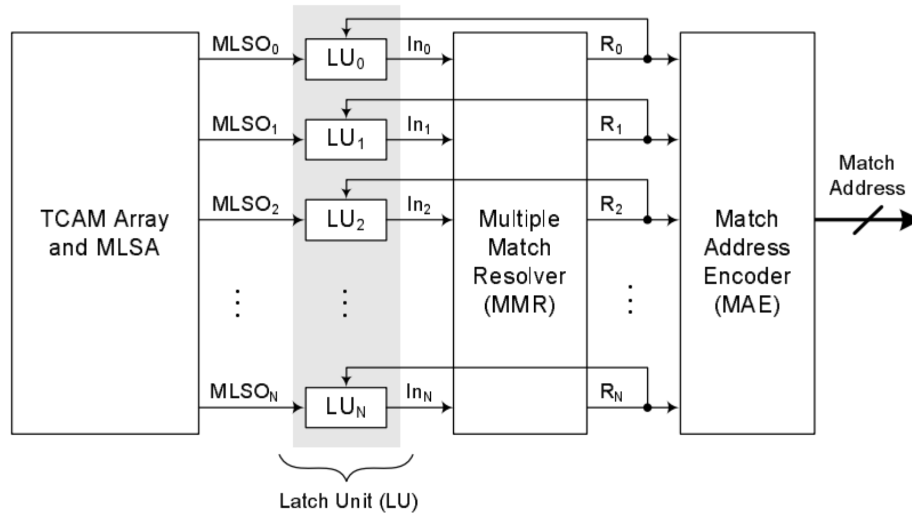

Figure 3.12: Structure of PE using Latch-and-Reset Approach. (Figure is taken from [75].)

As indicated in Table 3.4, the proposed design achieves much better t_{search} and lower E_{search} than designs using 90nm/180nm technology. Compared to simulation result of design in the more advanced 45nm technology, the proposed design has roughly the same

E_{search}, slightly higher t_{search} but a more compact TCAM cell. All other listed designs incorporate a PnE SA and ≥ 2 transistors per TCAM cell. Thus, the proposed design is proven to reduce E_{search} and cell area.

3.8 TDE Design Using RRAM

A simple TDE circuit using a single RRAM cell can be implemented as shown in Figure 3.13. Same as the TCAM circuit design, TSMC 65nm PDK and ASU RRAM model are used for circuit design in Cadence Virtuoso and simulation in Spectre. The circuit consists of a data buffer with a programmable RRAM device inserted between the two stages of inverters. The inverters can be potentially share with other circuit along the signal propagation path. During normal operation mode, programming transistor $MN4, 5$ and $MP4, 5$ are turned off. The amount of propagation delay (t_p) inserted in the circuit is mainly determined by resistance of RRAM. During programming mode, signal *prog* is used to disable transistor $MN2$ and $MP2$. Signals *set* (*set_n*, *set_p*) and *reset* (*reset_n*, *reset_p*) are used correspondingly to program resistance state of the RRAM.

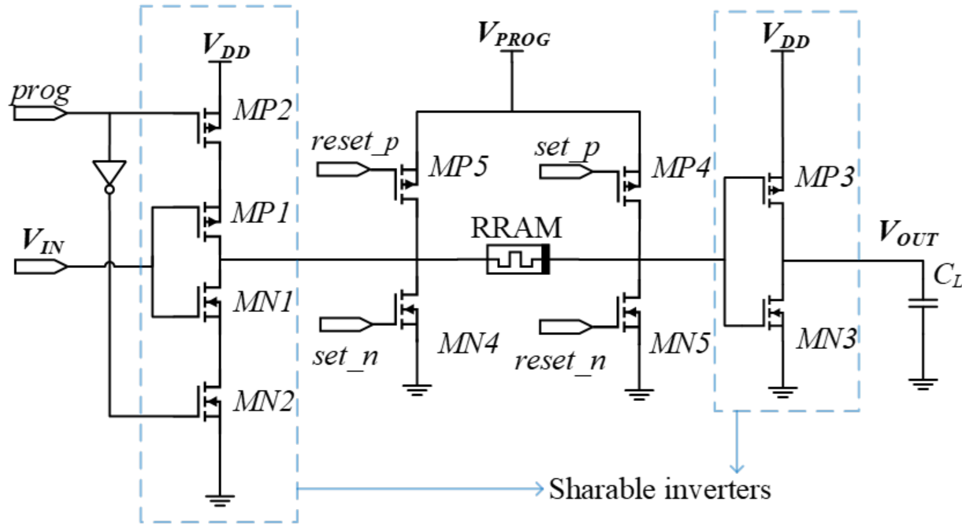

Figure 3.13: TDE circuit with single RRAM. In normal operation mode, programming transistors $MN4$, $MN5$, $MP4$ and $MP5$ are off. In programming mode, signal *prog* disables $MN2$ and $MP2$ while the RRAM can be SET (using $MP4$ and $MN4$) or RESET (using $MP5$ and $MN5$).

Pulse programming can be used to achieve quasi-analog switching as shown by simulation result in Figure 3.14. V_{PROG} is set to 2.1V and programming signal with pulse width of 1 μs is applied at $reset_n$ with $reset_p$ = 0V, change of resistance of the RRAM cell corresponding to each programming pulse is shown in Figure 3.14(b). The switching process shows a rather abrupt change in RRAM resistance at the early stage of the switching process. Afterwards the resistance changes become gradual and close to linearly increasing with the number of pulses applied increases.

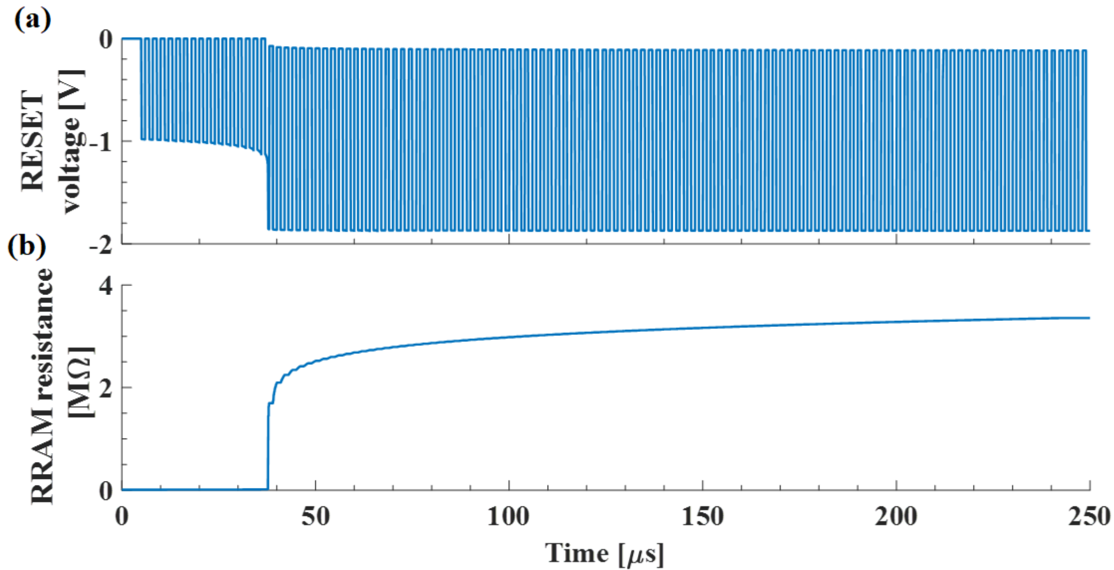

Figure 3.14: Simulation of pulse programming applied on ASU RRAM to achieve quasi-analog switching: (a) Waveform of voltage applied across RRAM device and (b) RRAM resistance versus time corresponding to each programming pulse.

The range of t_p and power consumption of the circuit is shown in Figure 3.15 When the circuit is operating in normal operation mode with V_{DD} = 1.2V and load capacitor set to 10fF. Here two RRAM devices fitted with the ASU RRAM model are used for performance simulation of the circuit in Spectre with details shown in Figure 3.5. Range of t_p is obviously dependent upon range of resistance of each device. Amount of t_p is approximately linearly proportional to resistance value of the RRAM device. Since device #1 has a higher R_{HRS} than device #2, TDE implemented with device #1 can also provide a wider range of t_p. Power consumption also increases as t_p increases. This is because, as resistance of RRAM device increases, more current is required from V_{DD} to charged up parasitic capacitor present at the gate node of $MP3$ and $MN3$.

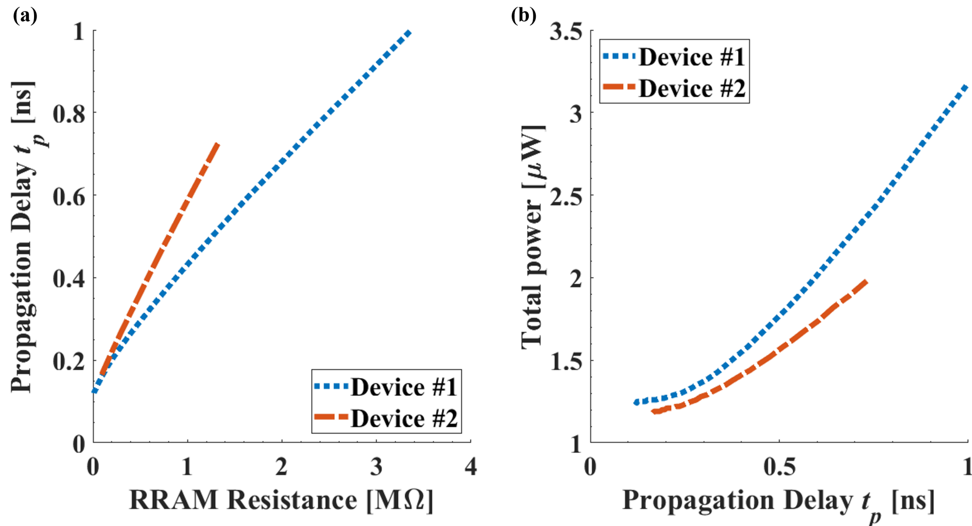

Figure 3.15: Simulation results of (a) propagation delay t_p vs RRAM resistance and (b) power consumption vs Propagation delay t_p for single RRAM TDE with 2 RRAM devices models listed in Table 3.5.

Table 3.5: RRAM device characteristics and TDE circuit setup

Device	V_{SET}/V_{RESET}	LRS/HRS Resistance	V_{PROG}
#1 [72, 73]	2V/-1.2V	10KΩ/3.35MΩ	2.2V
#2 [76]	1.6V/-1V	92KΩ/1.33MΩ	1.5V

3.9 Parallel RRAM Configuration in TDE

Instead of using a massive reference circuit system to address switching variation of RRAM as in [47, 48]. A parallel programming mechanism has been proposed in [77]. A quick verification of this idea can be done through simulation using capability of modeling switching variation from the ASU RRAM model [31]. By adjusting γ_0 value of an RRAM device instance to lower (0.9×) than nominal, the device takes a longer period time to be program with the same V_{PROG}. When programming this device in parallel with a nominal device, a difference of the programming process can be observed compared to programming both devices individually. The comparison result of RESET programming processes is shown in Figure 3.16. When both RRAM devices are RESET individually, the amount of programming time significantly varies from ∼0.2 to ∼1.2 μs, as indicated by both programming

current (I_{PROG}) and resistance of RRAM. But when they are RESET in parallel, I_{PROG} through the nominal device is reduced to compensate for the slower device in order for it to speed up. Two devices can synchronize in the middle of the RESET process and finish the rest of the programming process together.

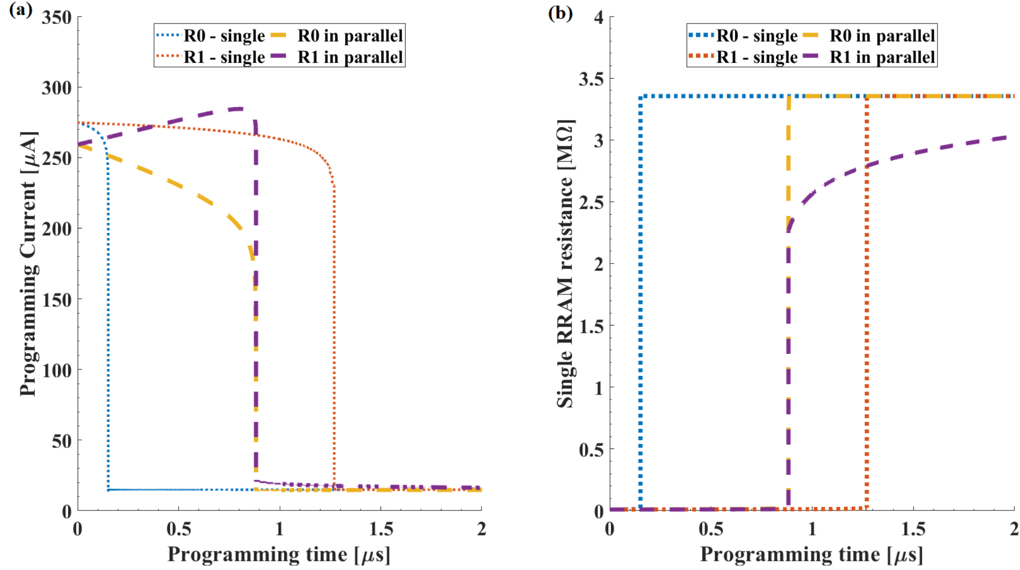

Figure 3.16: Simulated effect of programming two RRAM cells in parallel: (a) I_{PROG} for RRAMs RESET in parallel vs individually: a portion of I_{PROG} for programming the nominal device is reduced and compensated to the device with slower RESET process. (b) Change of RRAM resistance of in parallel RESET vs individual RESET.

A TDE circuit incorporating two RRAM devices in parallel configuration is shown in Figure 3.17. Two transmission gates are added to put RRAM $R0$ and $R1$ in parallel during normal operation mode and RESET programming mode. During SET programming mode, the parallel connection is removed by disabling the transmission gates. At the same time, the two devices are SET using $MN4a$ and $MN4b$ respectively. The reason is that, if two RRAM cells are SET in parallel, the faster cell will draw more programming current than the other cell. This will leave one of the RRAMs stuck in HRS, defeating the purpose of putting two RRAMs in parallel.

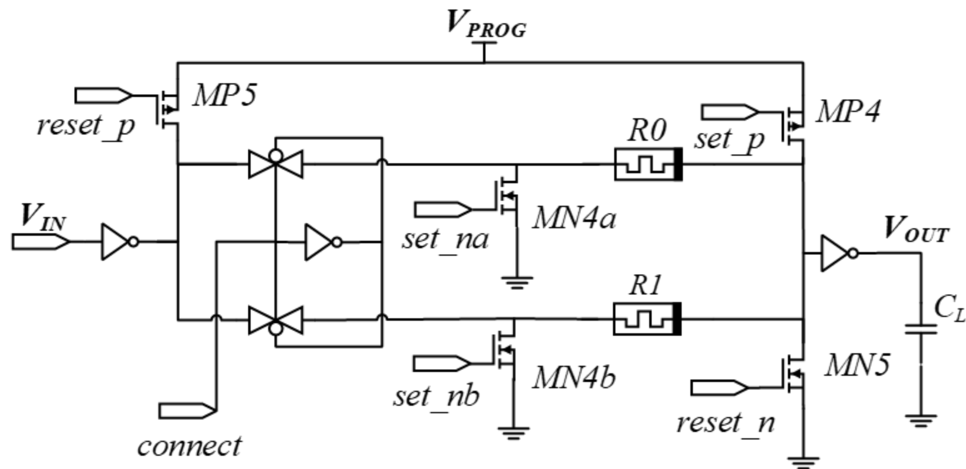

Figure 3.17: TDE circuit with 2 RRAMs. Two transmission gates are added to connect the two RRAMs in parallel during normal operation and RESET programming mode. They are disabled during the SET programming mode.

Performance of the parallel RRAM TDE circuit from simulation is shown and compared to the single RRAM TDE circuit in Figure 3.18. Notice that the data are from circuits both implemented using device #1 from Table 3.5. When putting two RRAMs in parallel, equivalent of resistance place in between the two inverters is reduced from a single RRAM. This reduces the range of t_p slightly instead of half t_p, likely due to the fact that parasitic capacitance introduced by extra components is becoming dominant in influencing t_p. On the other hand, the power consumption of the parallel RRAM TDE increases, also due to the introduction of more transistors to the parallel RRAM TDE circuit.

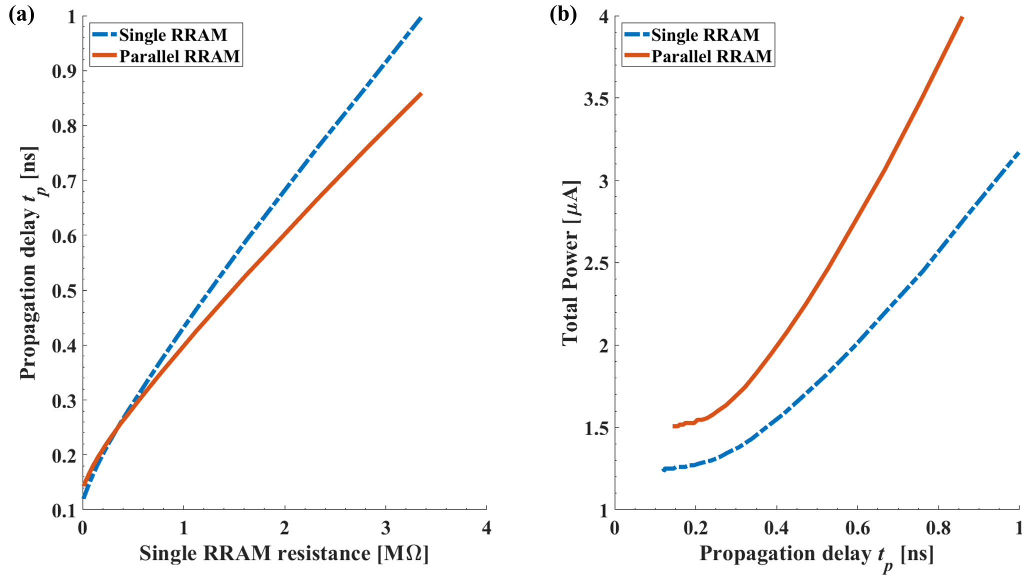

Figure 3.18: Parallel RRAM TDE circuit vs single RRAM TDE circuit through simulation data: (a) propagation delay t_p vs RRAM resistance: Range of t_p is slightly reduced in the parallel RRAM TDE. (b) power consumption vs propagation delay: power consumption of parallel RRAM TDE increases due to extra overhead circuit.

Another thing to notice in this TDE design is the problem of imbalance between rising (t_{pr}) and falling edge delay (t_{pf}) in both single and parallel RRAM TDE circuits. This is due to the fact that the circuit involve two power supply with different supply voltage V_{DD} and V_{PROG}. Even though the programming transistors are turned off during normal operation mode, leakage current is still expected from V_{PROG} to the circuit operation with V_{DD}. As shown in Figure 3.19(a), as resistance of RRAM increases, the imbalance Δt_p between t_{pr} and t_{pf} increases. Because of the wider range of t_p provided by device #1, when using TDE close to upper limit of device #1, $\Delta t_p \approx 30\% \times t_p$ can be observed. With the parallel RRAM TDE, equivalent resistance is reduced due to the parallel RRAM configuration compared to the single RRAM circuit. Amount of Δt_p is also proportionally reduced, as shown in Figure 3.19(b).

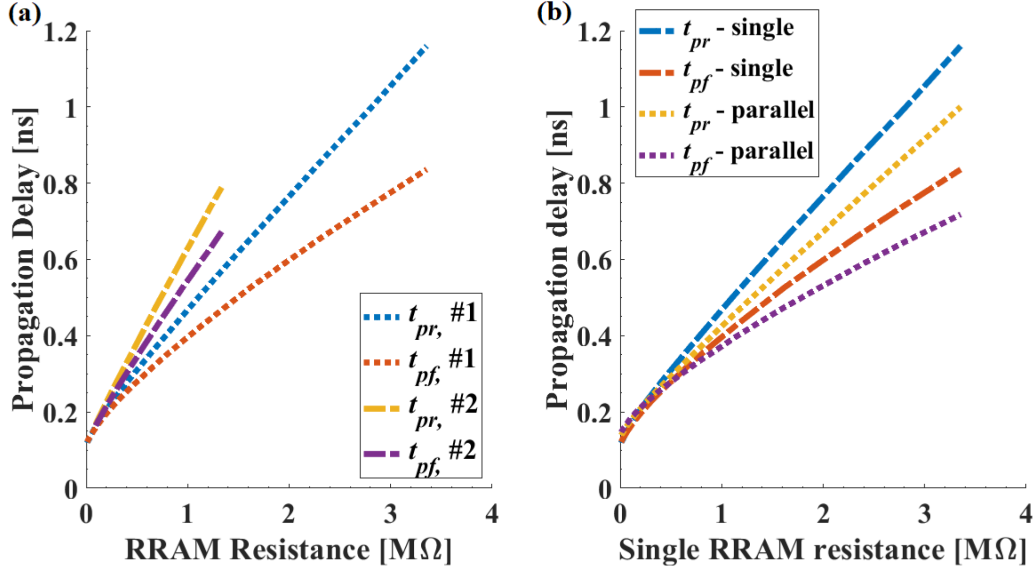

Figure 3.19: (a) t_{pr} and t_{pf} of single RRAM TDE with 2 RRAM devices models listed in Table 3.5. (b) t_{pr} and t_{pf} of parallel RRAM TDE vs single RRAM TDE with RRAM device #1.

To provide a broader view of how switching variations influence RESET programming process of RRAMs in parallel versus individual RRAM, a Monte-carlo analysis using Spectre and Cadence is run to RESET 1000 RRAMs in single configuration and 500 pairs of RRAMs with the parallel configuration. Switching variation is introduced by varying switching parameter of ASU RRAM model with device #1 as previously described. In the single RRAM TDE, programming transistor $MN4, 5$ and $MP4, 5$ are implemented with $W/L = 200$nm/60nm. In the parallel RRAM TDE transistors $MN5$ and $MP5$ used for parallel RESET are implemented with $W/L = 400$nm/60nm. The programming process is carried out by using programming pulse with pulse width of 1μs and $V_{PROG} = 2.1$V.

As shown in Figure 3.20 (a-d), the number of RESET pulses (N_{pulse}) to successfully program each RRAM from LRS to HRS is counted. The parallel RRAM TDE (Figure 3.20(b)) has approximately the same successful fully programming rate than the single RRAM TDE (Figure 3.20(a)). In the single RRAM configuration, >50% of the devices finish the LRS to HRS transition with N_{pulse} <10. Majority of the rest spreading out between 20 to 100. In the parallel configuration, N_{pulse} is spread out more evenly between 20 to 120. The parallel configuration slows down the RESET process, but overall programming time becomes more unpredictable. The effect of increasing amplitude (Figure 3.20(c)) and

pulse width (Figure 3.20(d)) of V_{PROG} are also investigated. By increasing V_{PROG} to 2.2V, N_{pulse} has a more concentrated distribution in the range of 10 to 50. On the other hand, increasing V_{PROG} pulse width by 10% only shifts the original distribution to the left (lower N_{pulse}) by a small amount.

The energy consumption (E_{PROG}) of the RESET process is also recorded respectively as shown in Figure 3.20 (e-h). Distribution of E_{PROG} for all four programming configuration follows the corresponding profile of N_{pulse}. When programming the RRAM in a standalone configuration, E_{PROG} is typically less than 10nJ. As for parallel RRAM TDE, average E_{PROG} increases to above 50nJ, with more complicated circuit and longer programming duration. With increasing V_{PROG} amplitude, the average E_{PROG} is reduced. Therefore, the advantage of shortening programming time outweighs increasing programming power. On the other hand, increasing V_{PROG} pulse width does not have obvious effect on E_{PROG}.

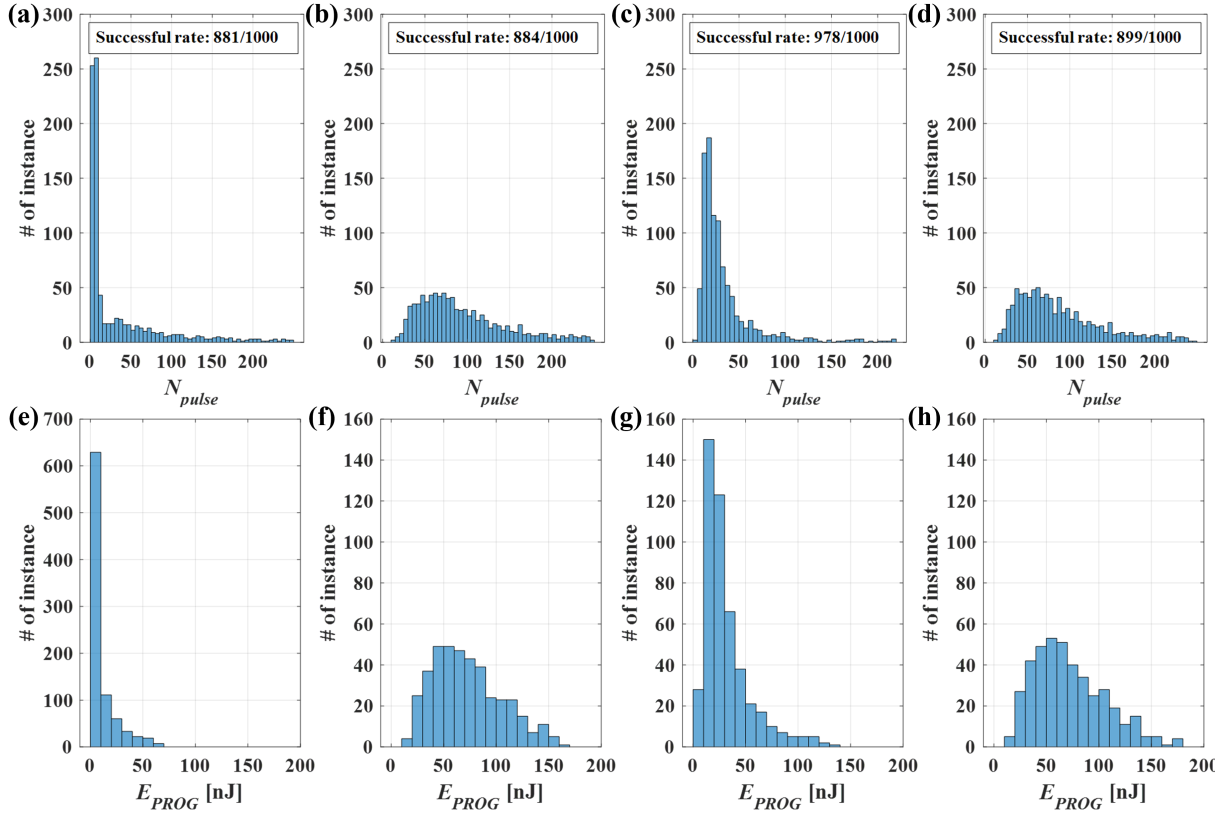

Figure 3.20: Result of Monte-carlo simulation considering RRAM switching variations using ASU RRAM model. (a-d) number of programming pulse N_{pulse} required to program RRAM from LRS to HRS considering RRAM switching variation: (a) single RRAM TDE with normal V_{PROG}, (b) parallel RRAM TDE with normal V_{PROG}, (c) parallel RRAM TDE with V_{PROG} increased by 0.1V and (d) parallel RRAM TDE with V_{PROG} pulse width increased by 10%. (e-h) Programming energy E_{PROG} for programming RRAMs from LRS to HRS corresponding to each of (a-d) respectively.

3.10 Summary

In this chapter, a 2T2R CR-MLSA structure is presented. Basic searching performance of this circuit structure is analyzed in terms of how it is impacted by increasing bit width attached to ML, variations of RRAM resistance, and the bit information stored in each cell, either the stored information is definite ('1/0') or 'x'. Approaches of how to counteract

negative impacts from such factors are also discussed with a new alternative proposed by using a compact ML booster design. By comparing simulated performance of the proposed 2T2R CR-MLSA design including ML booster with previous published design, the proposed design features in low search delay, low energy consumption and compact TCAM cell. A TDE circuit is also proposed in this section, which can be used in TCAM design for tuning reference signal sent each MLSA to ensure correct and efficient search operations. A parallel RRAM configuration is also proposed to reduce impact of RRAM switching variation to TDE circuit performance, which is verified through simulation.

Chapter 4

TCAM Multi-stage Cascading Design

4.1 Introduction

Because very low off-state current (I_{OFF}) of MOS transistor, SRAM-based TCAM can support large word size per ML (usually n≥128) [74, 78]. Segmentation of SRAM-based TCAM is proposed mainly to reduce E_{search}. With the 2T2R TCAM cells, segmentation can also limit $I_{discharge}$ by reducing total number of pull-down paths in each segment. This in return increase $V_{ML,match}$ and lowers t_{search} and improve NM in each ML segment. As compared to directly using output signal V_{OUT} of one TCAM segmentation stage to enable the next stage, the ML booster provides another option of cascading stages using a SR latch between two stages. This, as shown in this chapter, reduce both search delay and energy consumption compared to the former cascading structure. Meanwhile, the CR-MLSA allows same clock phase cascading to be implemented. This further improves performance of the proposed multi-stage design by reducing TCAM system output latency while maintaining advantages of search energy reduction of cascading structure. Development of the circuit is continued in this chapter using Cadence Virtuoso. Performance of the circuit in this chapter is also evaluated through simulation using Spectre in Cadence. The main contributions of this thesis in this chapter are: (a) proposing the DC and SRC structures for multi-stage TCAM design and compare the two in terms of design trade-off (among search speed, energy consumption and noise margin) and (b) proposing the SCPC scheme to reduce latency of the CR-based TCAM system and further improving the performance of the system.

4.2 Layout and Area Consideration

Layout of major components of the CR-MLSA circuit in Figure 3.10 using Cadence Virtuoso is shown in Figure 4.1, in a standard cell style. TSMC 65nm PDK is used for model of transistors in the layout. The layout excludes the RSG and TDE circuit since they can be shared by multiple MLSAs. As labelled in the figure, the layout of a block of 16 2T2R cells occupies around 4.3 × 5.1 μm^2 of chip area. For a design of 128-bit TCAM array, 8 blocks of 16 cells can be arranged in a 4×2 fashion and occupy a total of around 17.2 × 10.2 μm^2 of chip area. Since RRAM devices can be implemented in Back End of Line (BEOL) of CMOS fabrication process, they do not occupy chip area at the transistor layer [11]. All these 2T2R TCAM cells share a common reference point connected to ground during normal operation. Gate channel of each transistor is connected to corresponding SL/$\overline{\text{SL}}$ when integrated into the complete system. Drain of each of transistors is connected to the corresponding RRAM devices implemented in BEOL process. All these RRAM devices that share the same ML are then have the other ends connected to the ML route, which eventually is connected back to the CR-MLSA. The CR-MLSA occupies 2.2 × 3.2 μm^2 of chip area. Therefore, the chip area ratio of a 128-bit TCAM array to a MLSA is ∼25:1.

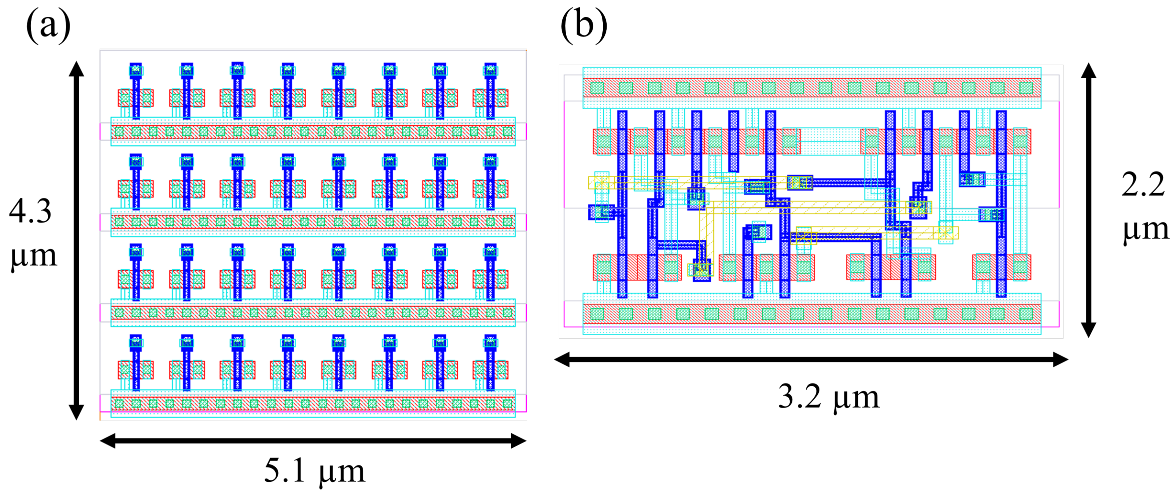

Figure 4.1: (a) Example layout of a array of 16 2T2R TCAM cells in 4 × 4 fashion. With a 128-bit TCAM array, the area ratio between TCAM array and CR-MLSA is ∼25:1. (b) Example layout of a CR-MLSA with ML booster.

Based on data of circuit layout area, a plot can be generated as shown in Figure 4.2

to estimate impact to chip area as the total number of cascaded stages increases. If the overall number of stages is limited to no more than 4, the overall area increases by 10% compared to overall area of a single stage TCAM design, including the 128-bit TCAM cell array. Meanwhile, the area occupied by the 4-stage CR-MLSA circuit can be limited to below 15% of the overall TCAM circuit area. In the rest of this chapter, the technique of cascading CR-MLSA stages is investigated in terms of E_{search}, t_{search} and NM with 2-stage and 4-stage designs. However, even if a 8-stages cascading structure is used, the total area of all the CR-MLSA circuit is still less than 25% of the overall TCAM circuit area.

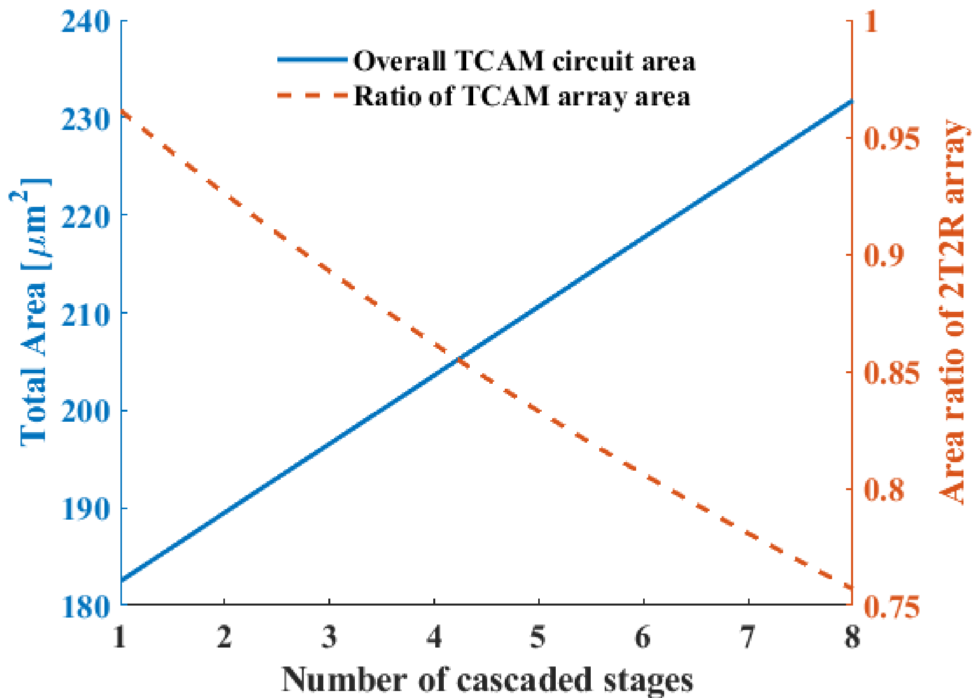

Figure 4.2: Overall TCAM circuit area and ratio of 2T2R TCAM array area with increasing number of cascaded stages.

4.3 2-stage with Direct Cascading

A straightforward way of segmenting a long ML is proposed in this study using CR-MLSA DC is shown in Figure 4.3 with a 2-stage TCAM design. TCAM with longer words can be

potentially segmented into > 2 stages, while principles of benefits still apply. The 1st stage has its output signal V_{OUT1} stored in a D-latch $L1_{DC}$ during its evaluation phase, which is also pre-discharge phase of the 2nd stage. Output Q of this D-latch is connected to a 3-NAND gate $N21$ used to enable $M23$ in the 2nd stage. The two other inputs of $N21$ are $\overline{clk}$ and reference signal en_2. If a match is found in the 1st stage, the 2nd stage is enabled and operate similarly to the 1st stage, with a half clock cycle delay. Otherwise, $N21$ does not enable $M23$, which forces the 2nd stage to stay at the pre-discharged state until the next search cycle. In the latter scenario, all the dynamic power consumption of the current stage is cut off with only the static power consumption remaining. Because of $L1_{DC}$, the 2nd stage always has a half clock cycle delay from the 1st stage. Therefore, a separate set of RSG and TDE circuits are required to generate reference signal en_2, which has a half clock cycle delay from en_1.

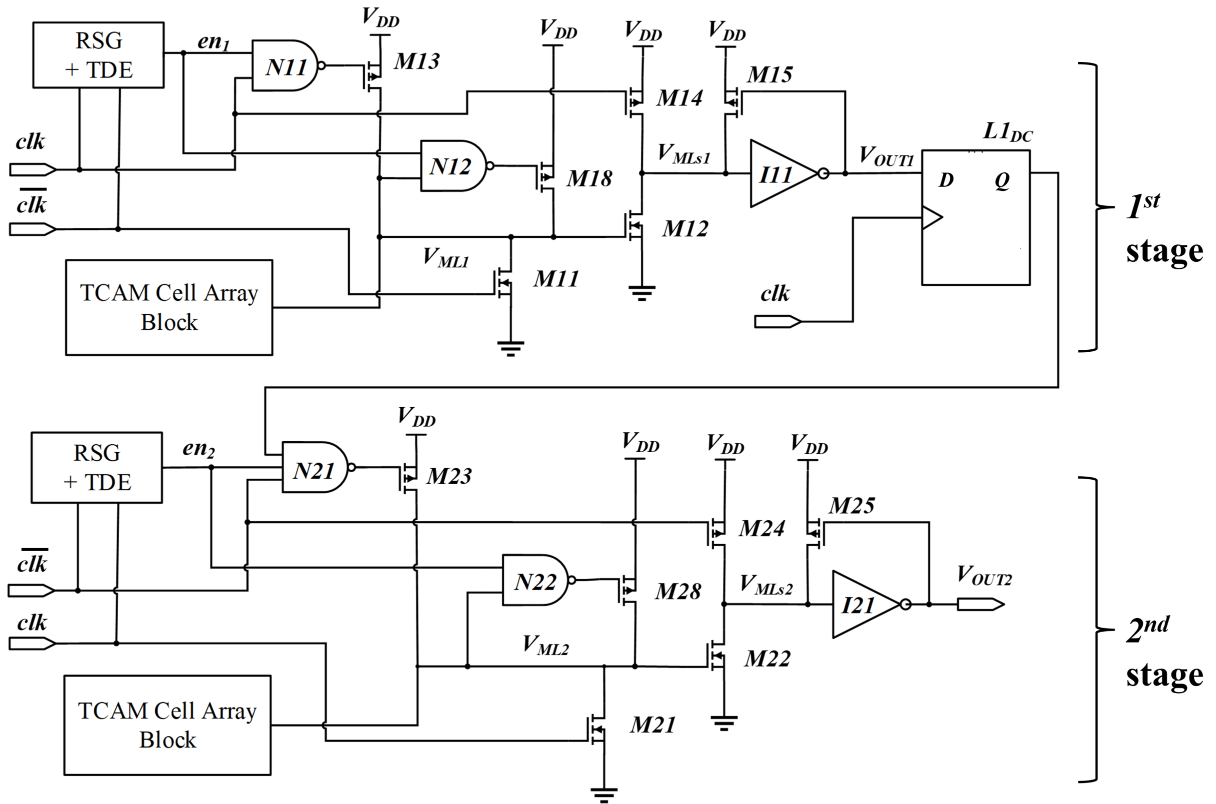

Figure 4.3: 2-stage CR-MLSA TCAM design using DC with a D-latch ($L1_{DC}$). V_{OUT1} is latched by $L1_{DC}$, which generate the 2nd stage activation signal *act*.

There are 3 type of search results from this cascaded TCAM configuration: (1) full match, (2) 1st stage mismatch and (3) 2nd stage mismatch. They can be explained with the waveform shown in Figure 4.4. When there is a 1st stage match detected, V_{OUT1} is raised to V_{DD} during the evaluation phase of the 1st stage when $clk = V_{DD}$. This is stored in $L1_{DC}$, which that output activation signal $act = V_{DD}$ for the 2nd stage. The 2nd stage then carries on the rest of the searching task during its evaluation phase when $clk = 0$V, while act maintains at V_{DD}. If V_{OUT2} is raised to V_{DD} at the end of the 2nd stage searching process, a full match is detected. Otherwise, V_{OUT2} stays at 0V, signaling a 2nd mismatch. On the other hand, if a 1st stage mismatch is detected in the first place, $act = 0$V. Therefore, $M23$ is disabled and the 2nd stage does not consume dynamic power.

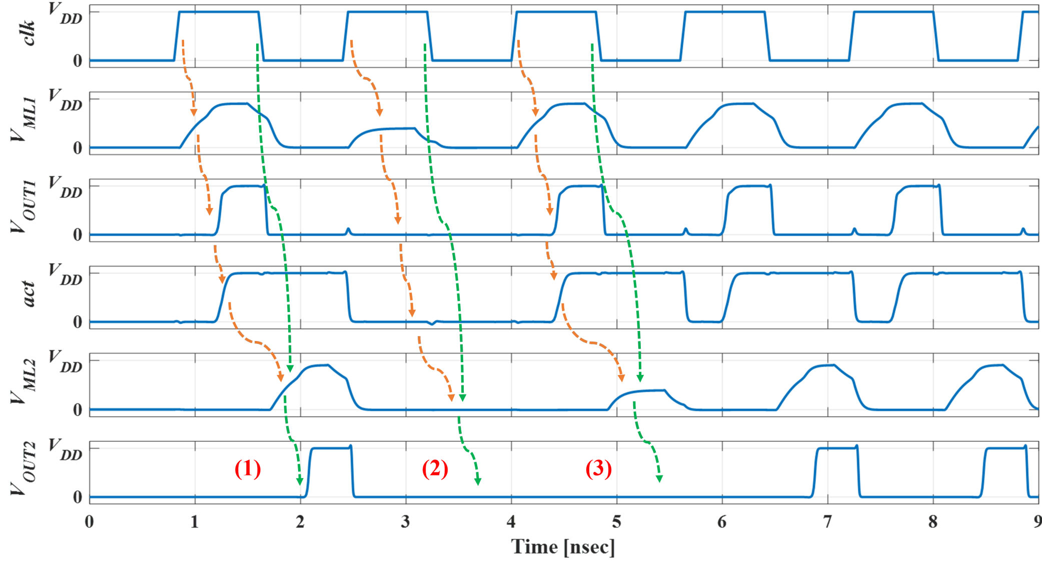

Figure 4.4: Simulated unctional waveform of 2-stage cascaded CR-MLSA: (1) Full match: $act = V_{DD}$ enabling the 2nd because of a 1st stage match. Singal $act = V_{DD}$ maintains during evaluation phase of the 2nd stage. $V_{OUT2} = V_{DD}$ because of a 2nd match. (2) 1st stage mismatch: $act = 0$V and the 2nd stage is not activated. (3) 2nd stage mismatch: $act = V_{DD}$ but $V_{OUT2} = 0$V.

4.4 2-stage with SR-latch Cascading

Compared to CR-MLSA DC method described in the previous section, an alternative SRC method is proposed that directly utilize the ML booster output can yield lower t_{search} and E_{search} but with a compromise of NM. Since $V_{ML,1-miss}$ is not expected to enable $N2$ of a 1-stage CR-MLSA shown in Figure 3.10. Output of $N2$ can be directly used as V_{OUT} of a cascading stage. The proposed cascading structure is shown in Figure 4.5 with a 2-stage TCAM design. Again, TCAM with longer words can be potentially segmented into > 2 stages. A NAND-based SR latch $L1_{SRC}$ is used to store the 2nd stage enabling signal en_2'. This NAND-based SR latch can be implemented using eight CMOS transistor.

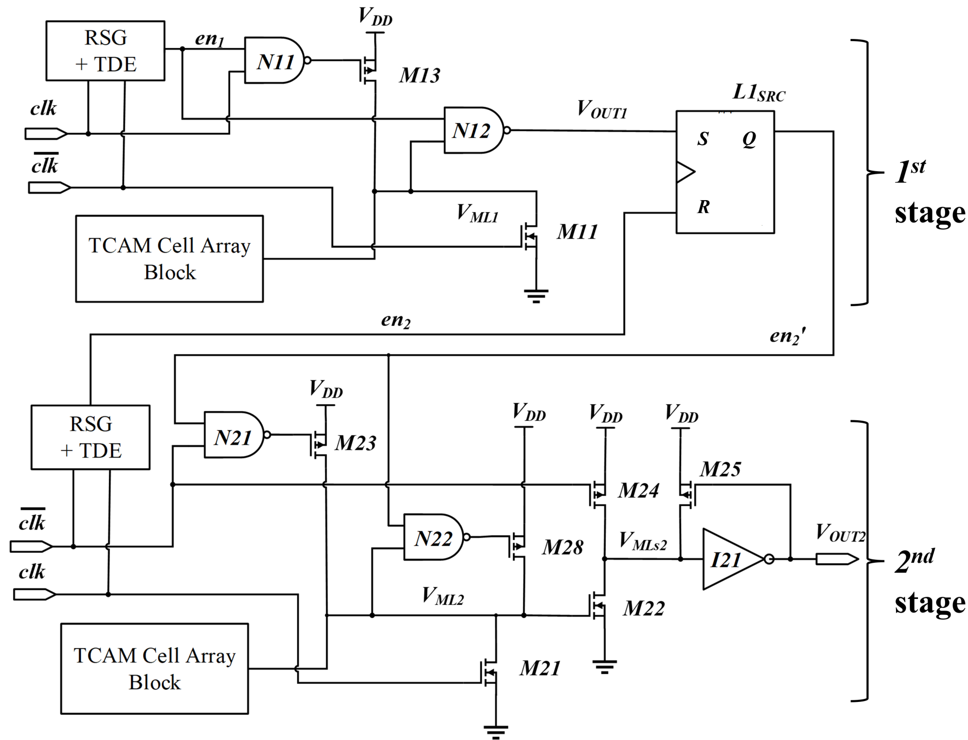

Figure 4.5: 2-stage CR-MLSA with SRC structure. $L1_{SRC}$ starts with state C (reset) in each new cycle. If 1st stage search is a match, $L1_{SRC}$ enters state B (set) then state D (hold), enabling the 2nd stage. Otherwise, $L1_{SRC}$ enter state D directly from state C, with the 2nd stage disabled.

Input S of $L1_{SRC}$ is connected to V_{OUT1} of the 1st stage, which controls $M8$ previously.

V_{OUT1} also becomes active low during evaluation phase. Input R of $L1_{SRC}$ is connected to reference signal en_2 from the 2nd stage. Output Q of $L1_{SRC}$ is connected to 2-NAND gates $N21$ and $N22$ in the 2nd stage, replacing en_2. Operation of $L1_{SRC}$ is summarized in Table 4.1. The flow chart of $L1_{SRC}$ state transition is shown in Figure 4.6.

In order to avoid meta-stable state of $L1_{SRC}$, state A, where both input S and R equal to 0V should be avoided. When input S=0V, the 1st stage is in evaluation phase and a match is encountered such that V_{OUT1} = 0V. Meanwhile, the 2nd stage is in pre-discharge phase. V_{OUT1} resumes to V_{DD} when this phase is finished. When input R=0V, the 2nd stage is in evaluation phase. The RSG finishes a mimicked match search and send out en_2 = 0V to cut off charge current from V_{DD} to ML. In the meantime, the 1st stage is in pre-discharge phase of the next search cycle. By the end of this phase, en_2 resumes to V_{DD}. Therefore, S=0V and R=0V are expected to appear in different clock phases by design and not appear at the same time to cause meta-stability.

Table 4.1: Operation states of the cascading SR latch $L1_{SRC}$

State	S (V_{OUT1})	R (en_2)	Scenario	Q_{new} (en'_2)
A	0	0	Avoided by design	Not allowed
B	0	1	1st stage: in evaluation (match) 2nd stage in pre-discharge	1 (set)
C	1	0	1st stage: in pre-discharge 2nd stage: evaluation complete	0 (reset)
D	1	1	1st stage: evaluation complete 2nd stage: pre-discharge or evaluation	$Q_{old}(hold)$

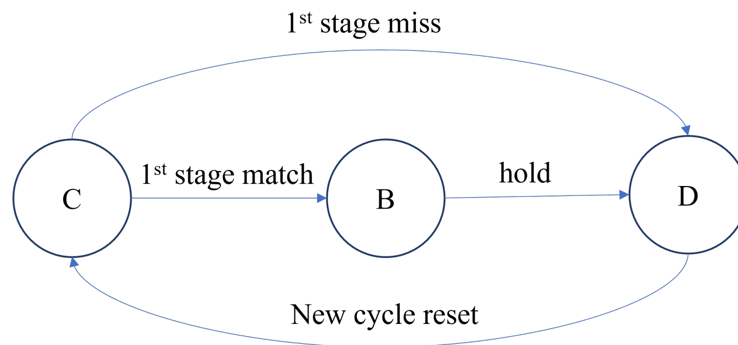

Figure 4.6: Flowchart of NAND-based SR latch operation states.

When a new search cycle begins, $L1_{SRC}$ is reset (state C) with a low en_2 from the 2nd stage. Then the 1st stage enters evaluation phase and en_2 rises to V_{DD}. If a match is detected in the 1st stage, V_{OUT1}=0V. $L1_{SRC}$ is then set (state B) with $en'_2 = V_{DD}$, enabling the 2nd stage. Then $L1_{SRC}$ enters state D where Q holds current value until the next search cycle begins. If a 2nd stage search results in a match $V_{OUT1} = V_{DD}$, as shown in Figure 4.7 (a). Otherwise $V_{OUT1} = 0$V (Figure 4.7 (c)). If a mismatch is detected in the 1st stage, $L1_{SRC}$ enters state D directly from state C with $en'_2 = 0$V and the 2nd stage disabled, which is shown in Figure 4.7 (b). Similar to the DC method, there is a half clock cycle delay between 1st and 2nd stage. Therefore, separate RSG and TDE are required to generate en_2 for the 2nd stage.

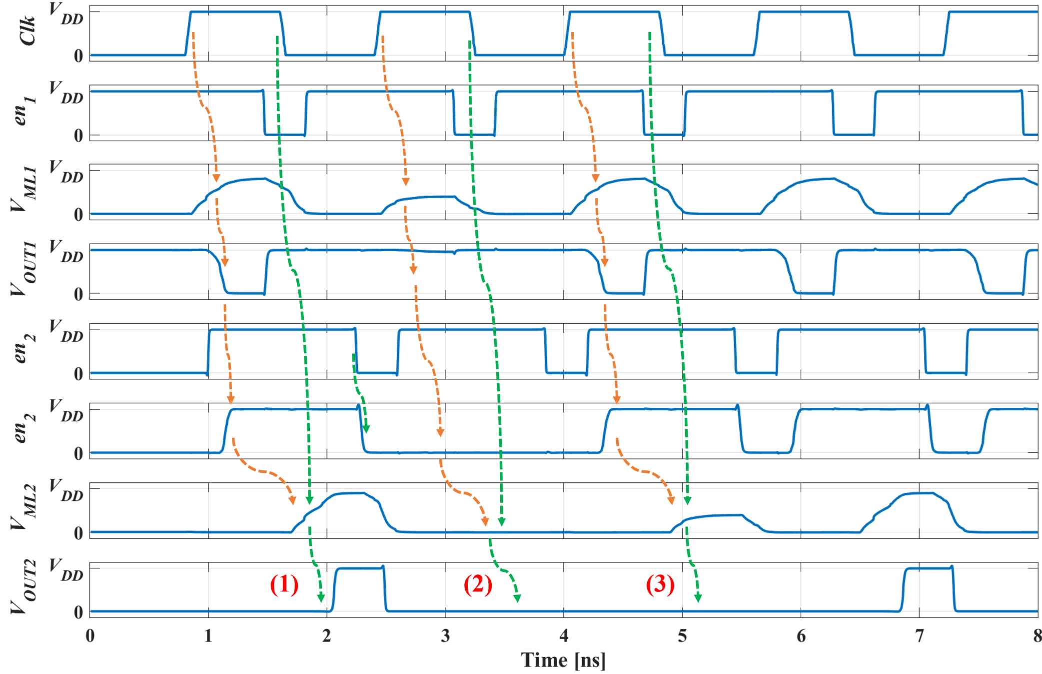

Figure 4.7: Simulated functional waveform of 2-stage cascaded MLSA with SRC method: (1) Full match: $L1_{SRC}$ enters state B (set) then state D (hold) because $V_{OUT1} = 0$V, $en'_2 = V_{DD}$ enables the 2nd. $V_{OUT2} = V_{DD}$ because of a 2nd match. (2) 1st stage mismatch: $L1_{SRC}$ stays at state C (reset) and enters state D (hold) directly. $en'_2 = 0$V and the 2nd stage is not activated. (3) 2nd stage mismatch: $en'_2 = V_{DD}$ but $V_{OUT2} = 0$V.

4.5 Performance Comparison of 1-stage and 2-stage TCAM Design

A performance comparison between 1-stage and 2-stage design (with both DC and SRC structure), implemented with a 128-bit 2T2R TCAM array, is summarized in Table 4.2 with data obtained through simulation in Spectre. All designs are evaluated with a pessimistic 50% variation in R_{HRS} (50% reduction of R_{HRS}). In the 1-stage design, W_{charge} is increased to 360nm and width of $M8$ (W_{boost}) is set to 600nm to ensure high enough $V_{ML,match}$ to generate valid V_{OUT} for a match case. Transistors in 2-stage design are all implemented with W/L=200nm/60nm for $NM > 150$mV in both stages.

For the 2-stage designs, t_{search} in each stage is reduced by ≥ 0.11ns in DC and ≥ 0.13ns in SRC due to increased $V_{ML,match}$ from segmentation. This allows system clock frequency (f_{clk}) to increase. SRC can achieve a further reduction in t_{search} in each stage compared to DC due to simplification of cascading structure. If the 2nd stage is inactive due to 1st stage mismatch, E_{search} of the 2-stage designs (both DC and SRC) is only 43% of the 1-stage design. Even with both stages activated, E_{search} of the 2-stage SRC design is only 84% of the 1-stage design while the 2-stage DC design is around the same as the 1-stage design. In case of ML word size >128-bit, E_{search} in 1-stage design is expected to increase more rapidly with bigger word size than multi-stage design because larger W_{charge} and W_{boost} are required for sufficient I_{charge} and I_{boost}.

Table 4.2: Performance comparison of 1-stage and 2-stage TCAM design through simulation with 2 cascading approaches

128-bit TCAM	t_{search} (ns)	E_{search} (fJ/bit/search)				NM (V)	
		Stage 1 miss	Stage 1 match	Stage 2 miss	Stage 2 match	Stage 1	Stage 2
1 stage	0.735	0.232	0.287	N/A	N/A	0.274	N/A
2 stage DC	0.608 + 0.623	0.104	N/A	0.245	0.281	0.337	0.324
2 stage SRC	0.404 + 0.601	0.103	N/A	0.203	0.240	0.151	0.331

Comparing the NM result between the 1-stage and 2-stage DC design, the 2-stage DC design improves NM by $\sim 20\%$ because of the increase in $V_{ML,match}$ from segmentation. This makes the multi-stage design more tolerant to device variations. The SRC structure comes with a disadvantage of decreasing NM in the cascading stage. Due to the change of cascading structure in SRC compared to DC, V_{OUT1} becomes active low during evaluation phase. Reusing definition of $V_{th,match}$ and NM of the 1 stage TCAM design, they can

be measured as shown in Figure 4.8. As indicated by data in Table 4.2, the NM of the cascading (1st) stage in 2-stage SRC design is reduced by $\sim 45\%$ from the 1-stage design and $\sim 55\%$ from the 2-stage DC design. Thus, the cost of improving t_{search} and E_{search} in SRC is reduction in NM. However, the high NM in 1-stage and 2-stage DC design comes from the fact that $V_{th,match}$ is high. This is because the CR-MLSA structure requires a very high $V_{ML,match}$ to generate $V_{OUT} = V_{DD}$. For example, with $V_{DD} = 0.7$V, $V_{th} > 0.5$V for 1-stage and 2-stage DC design. But for 2-stage SRC design, $V_{th,match} \approx 0.4$V with a simplification of structure. With the same word length of ML, cascading with >2 stages is expected to improve $V_{ML,match}$ by reducing $I_{discharge}$ in each segment. Impact of device variations is expected to have reduced influence. Therefore, it makes SRC more suitable to aim for high-speed and low-energy system design.

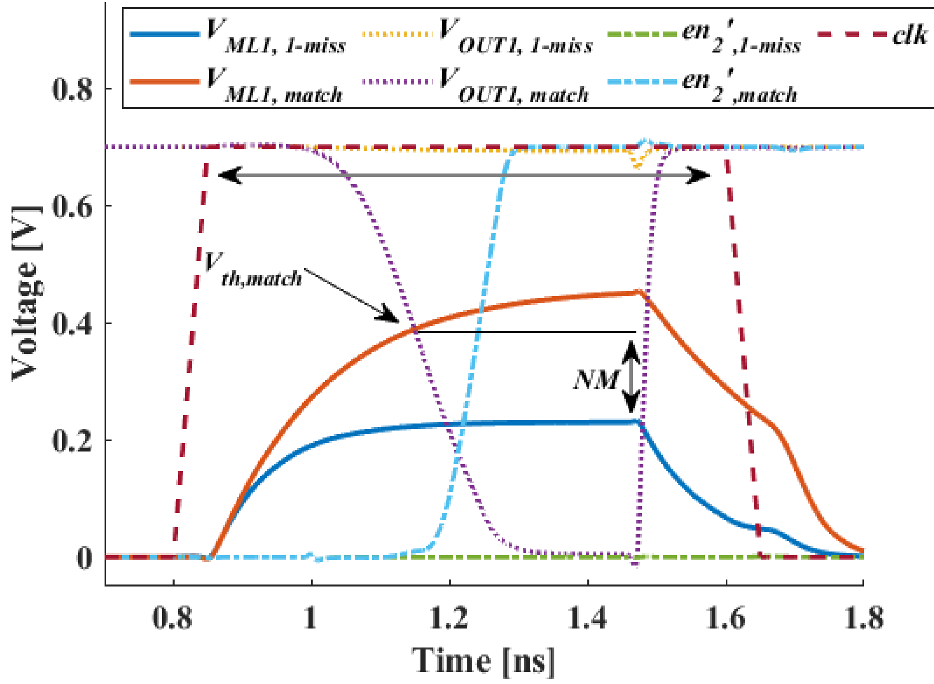

Figure 4.8: Simulated functional waveform of SRC structure of a 2-stage TCAM. V_{OUT1} is active low during evaluation phase. Compared to DC, SRC has lower NM because of simplified circuit structure.

4.6 Same Clock Phase Cascading

In PnE MLSA, a minimum delay of half a clock cycle exists between stages to reduce E_{search}. Because the pre-charge phase consumes dynamic power from V_{DD}, V_{OUT1} is needed to decide enabling pre-charge of the 2nd stage. However, pre-discharge phase of the CR-MLSA does not consume dynamic power from V_{DD}. For a 2-stage CR-MLSA, V_{OUT1} is not needed until evaluation phase of the 2nd stage. Therefore, both stages can be evaluated in the same clock phase.

In this SCPC approach, modifications need to be made to the two 2-stage cascading structure shown previously. Because both stages are now in synchronization in terms of operation phases, they can share the same clk and $\overline{clk}$ connections for the MLSA. The same RSG can be used for both the 1st and 2nd stages. However, during an evaluation phase of a 2-stage design, slightly larger t_{charge} should be allocated for the 2nd stage compared to the 1st stage to account for signal propagation between stages. Hence, separate TDEs are required to generate reference signal en_1 and en_2. The 2-stage CR-MLSA circuit using DC with SCPC is shown in Figure 4.9. In addition to the clock signal and reference signal connection, the D-latch $L1_{DC}$ used for storing intermediate search result can be removed since evaluation phase of both stages happen during the same clock phase. As for 2-stage CR-MLSA circuit using SRC with SCPC shown in Figure 4.10, the SR latch is still needed for storing V_{OUT1}.

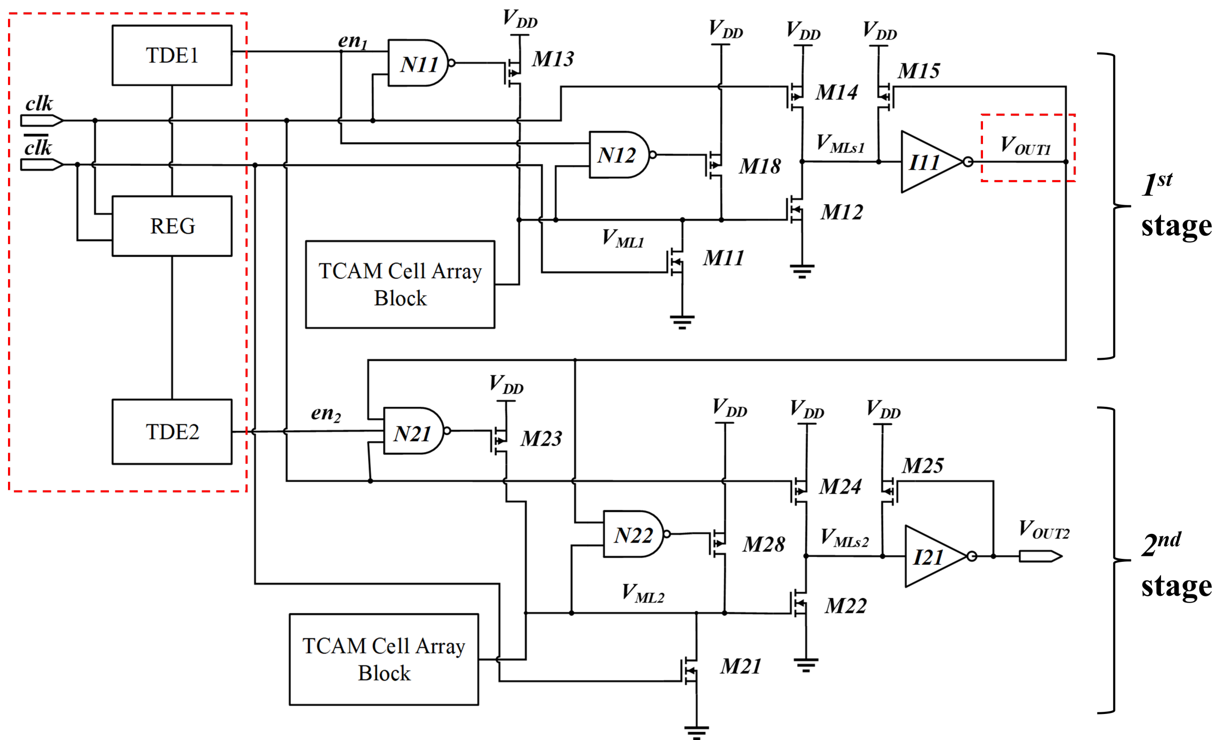

Figure 4.9: 2-stage CR-MLSA with DC structure and SCPC: clock signals clk, $\overline{clk}$ are modified to adapt for SCPC such that both stages are in synchronization in clock phase. Both stages can share the same RSG but use different TDEs. D-latch is removed since there is no need to store 1st output.

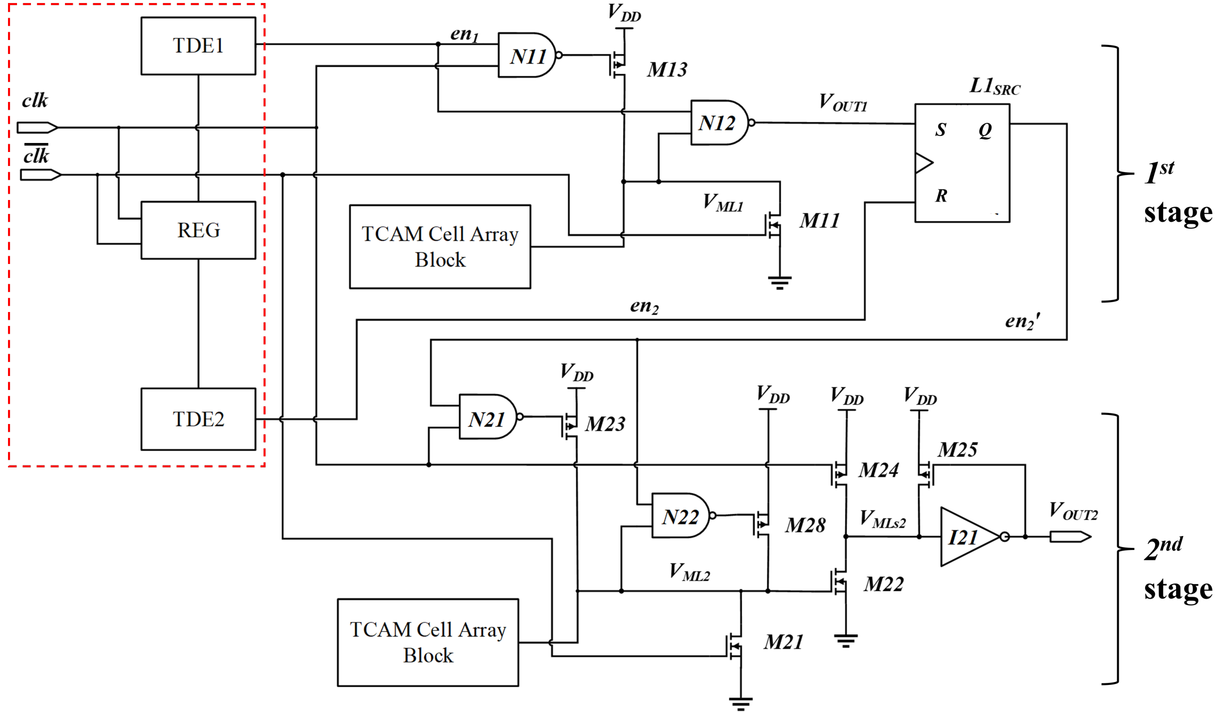

Figure 4.10: 2-stage CR-MLSA with SRC structure and SCPC: both stages can share the same RSG but use different TDEs, similarly to DC structure. SR-latch $L1_{SRC}$ is still required for storing V_{OUT1}.

Functional waveform of the two designs with SCPC is shown in Figure 4.11. With the SCPC, a complete search cycle in TCAM can be finished within one clock cycle. Both stages can be evaluated during the clock phase when $clk = V_{DD}$. Thus, overall search latency is reduced. For DC structure, V_{OUT1} can propagate to the 2nd during evaluation phase ($clk = V_{DD}$) such that the 2nd stage can continue with the rest of search process during the same clock phase. V_{OUT1} is only reset at the end of the clock cycle. Therefore, a D-latch is no longer required to store the intermediate search result V_{OUT1}. As for the SRC structure, V_{OUT1} is pulled down to 0V when $en_1 = 0$V, signifying 1st stage search completion. If V_{OUT1} is directly applied as control signal of the 2nd stage, the search process of the 2nd stage may not complete as V_{OUT1} may be reset before $en_2 = 0$V. Therefore, SR latch $L1_{SRC}$ is still required in SCPC to ensure that evaluation of the 2nd stage continues after $en_1 = 0$V and before $en_2 = 0$V.

In the normal cascading structure, $V_{OUT1} = 0$V and $en_2 = 0$V happens in different clock phases, by design, $L1_{SRC}$ does not become meta-stable. However, with SCPC, transition

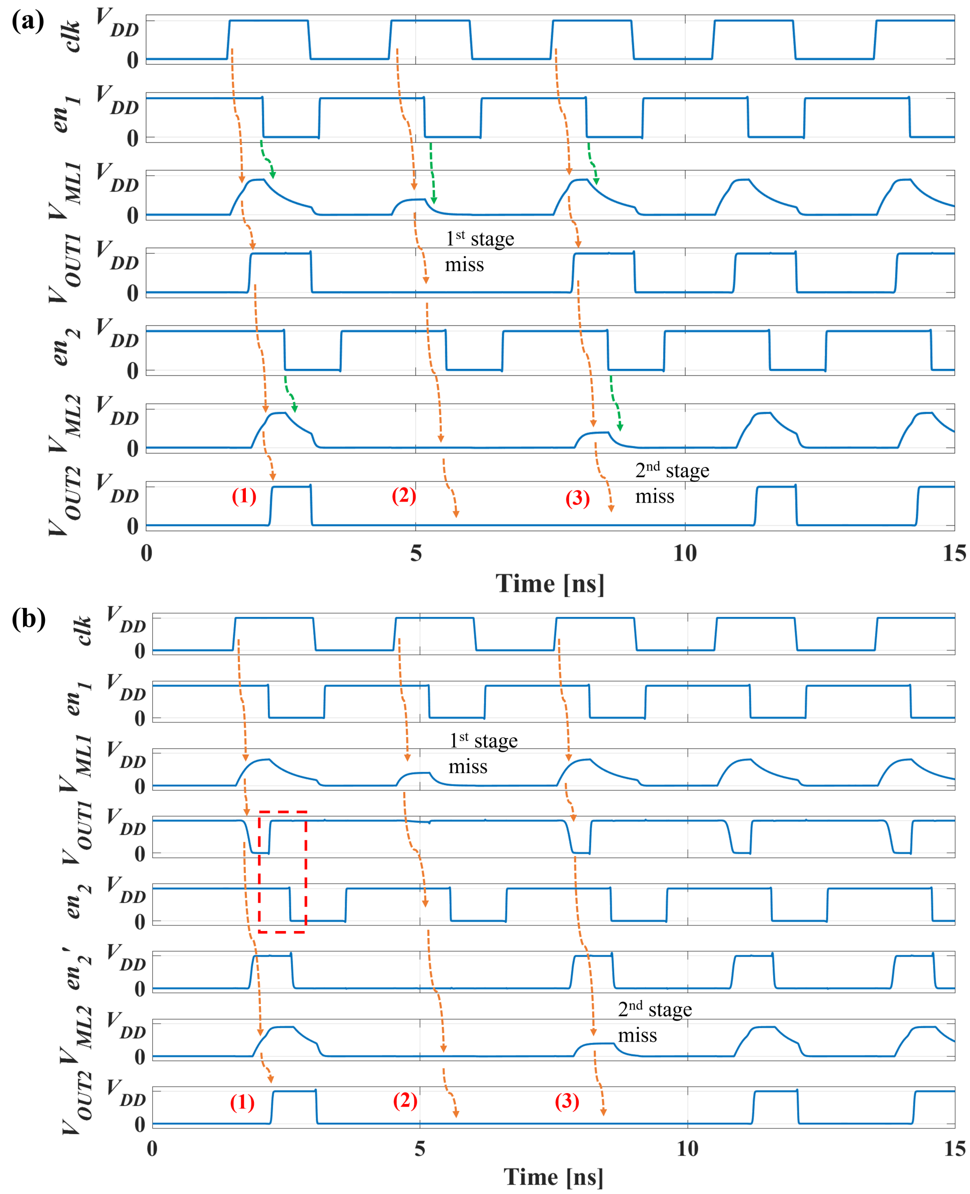

Figure 4.11: Simulated functional waveform of 2-stage cascaded CR-MLSA with SCPC: (a) DC structure: V_{OUT1} can propagate to the 2nd stage when $clk = V_{DD}$ and continue with the rest of the search during the same clock phase, (b) SRC structure: SR latch is still required to store V_{OUT1}, otherwise 2nd stage does not finish evaluation because V_{OUT1} resets before $en_2 = 0\text{V}$. (1) Full match, (2) First stage mismatch and (3) Second stage mismatch.

of both signals can happen within the same clock phase for evaluation, as shown in Figure 4.11. The updated states of $L1_{SRC}$ is shown as summarized in Table 4.3 with the same flow of state in Figure 4.6 still applies. However, with inattentive timing of signal en_2, $V_{OUT1} = 0$V and $en_2 = 0$V can happen at the same time causing meta-stability of $L1_{SRC}$. Therefore, TDE for generating en_2 should be carefully tuned to prevent meta-stability.

Table 4.3: Operation states of the cascading SR latch $L1_{SRC}$ with SCPC

State	S (V_{OUT1})	R (en_2)	Scenario	Q_{new} (en_2')
A	0	0	Avoided by tuning en_2	Not allowed
B	0	1	1st stage in evaluation (match) 2nd stage in evaluation (idle)	1 (set)
C	1	0	(1) Both stages in evaluation (complete) or (2) Both stages in pre-discharge	0 (reset)
D	1	1	1st stage evaluation complete 2nd stage in evaluation	$Q_{old}(hold)$

The simulated performance of 2-stage design TCAM design using SCPC, compared to the 1-stage 128-bit TCAM design, is shown in Table 4.4. The SCPC does not affect E_{search} or NM much compared to the original approach with evaluation of the 2 stages happen in different clock phases (Table 4.2). However, putting evaluation phase of 2 stages into one clock phase inevitably yield t_{search} higher than the 1-stage design without other modification to the circuit. This is because the 2nd stage search cannot begin until the 1^{stage} generates a result, in order to maintain E_{search} reduction. Increasing t_{search} implies reduction of f_{clk}, which is undesirable. Solutions to reduce t_{search} and increase f_{clk} are increasing I_{charge} by increasing W_{charge} or reducing $I_{discharge}$ by using SL/$\overline{\text{SL}}$ encoding feature [58]. In this study, the former approach is investigated as shown in the next section, due to its simplicity.

Table 4.4: Performance comparison of 1-stage and 2-stage TCAM design through simulation with 2 cascading approaches and SCPC

128-bit TCAM	t_{search} (ns)	E_{search} (fJ/bit/search)				NM (V)	
		Stage 1 miss	Stage 1 match	Stage 2 miss	Stage 2 match	Stage 1	Stage 2
1 stage	0.735	0.232	0.287	N/A	N/A	0.274	N/A
2 stage DC	1.12	0.103	N/A	0.248	0.284	0.328	0.328
2 stage SRC	0.96	0.103	N/A	0.213	0.244	0.157	0.334

4.7 Further Staging and Competing with SRAM-based TCAM

Cascading technique of DC and SRC can be used to further construct a 4-stage TCAM with topology as shown in Figure 4.12, The same topology applies for design with or without SCPC. For the rest of this section, discussion continues with designs using SCPC.

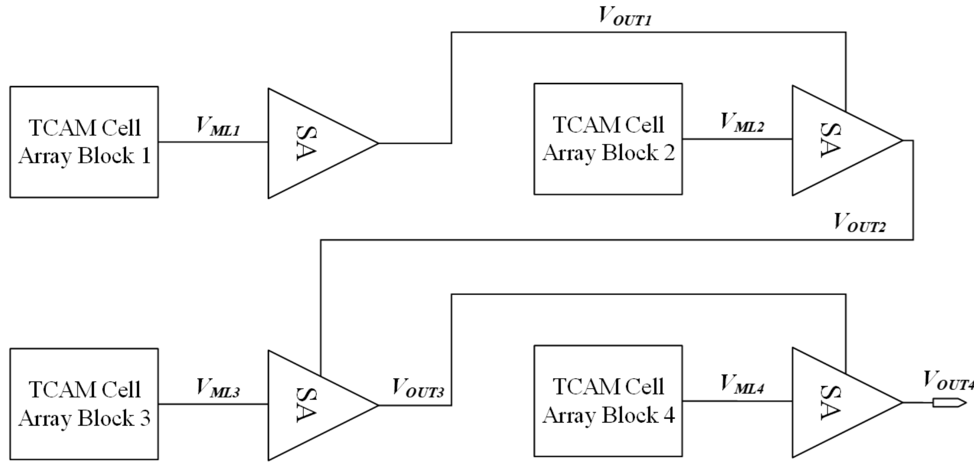

Figure 4.12: Topology of a 4-stage cascading TCAM.

To illustrate the performance of SCPC, a 128-bit TCAM is implemented in 2-stage and 4-stage with both DC and SRC structure using Cadence wiht the previously mentioned design and simulation environments setup. All transistors are implemented with

W/L=200nm/60nm. The simulated performance of each implementation is shown in Figure 4.13, where average E_{search} is calculated based on Equations 4.1 to 4.3. The equations corresponds to activation rates of each stage, which linearly decreases as stage number increases. Match rate of 5% is considered. For the SRC structure, average E_{search} reduction of both 2-stage and 4-stage TCAM is $\sim$ 30% of the 1-stage design. Yet, t_{search} of the 2-stage design is much larger than the 1-stage design. t_{search} of the 4-stage design is also slightly higher than the 1-stage design. With slight increase of W_{charge} = 200nm to 300nm in each stage, a significant decrease in t_{search} is observed with increasing average E_{search} in both 2-stage and 4-stage design. Both are able to generate t_{search} lower than the 1-stage design, while the average E_{search} reduction is still $\geq$ 20%.

As for the DC structure, due to the extra overhead circuit, average E_{search} and t_{search} of both 2-stage and 4-stage design are higher than their SRC counterparts. With W_{charge} = 200nm, the average E_{search} reduction is $\sim$ 23% for 2-stage design and $\sim$ 20% for 4-stage design. t_{search} of both designs are also much higher than the 1-stage design. Similar effect from increasing W_{charge} in SRC is observed in the DC structure: t_{search} reduction with increase average E_{search}. With W_{charge} = 300nm, t_{search} of both 2-stage and 4-stage designs are reduced to slightly below the 1-stage design. Reduction of average E_{search} is around $\sim$ 18% for 2-stage design and only $\sim$ 6% for 4-stage design.

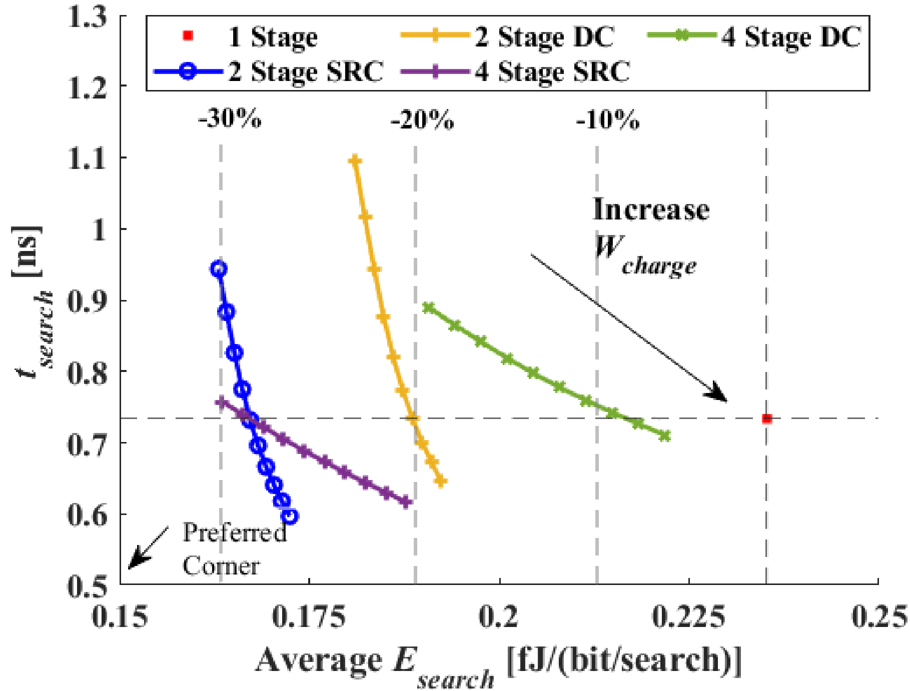

Figure 4.13: Average E_{search} and t_{search} of 2-stage and 4-stage vs 1-stage 128-bit TCAM design from simulation result. With increasing W_{charge}, Multi-stage design yields lower t_{search} than 1-stage design with E_{search} reduction of $\geq 20\%$ for SRC structure, $\sim 18\%$ for 2-stage design and only $\sim 6\%$ for 4-stage design with DC structure.

$$E_{1stage} = E_{miss} \times 0.95 + E_{match} \times 0.05 \tag{4.1}$$

$$E_{2stage} = E_{stage1,miss} \times 0.5 + E_{stage2,miss} \times 0.45 + E_{match} \times 0.05 \tag{4.2}$$

$$\begin{aligned} E_{4stage} = {} & E_{stage1,miss} \times 0.25 + E_{stage2,miss} \times 0.25 \\ & + E_{stage3,miss} \times 0.25 + E_{stage4,miss} \times 0.2 + E_{match} \times 0.05 \end{aligned} \tag{4.3}$$

Overall, multi-stage cascading design using SRC structure is a better option than DC structure considering only t_{search} and E_{search}. However, the SRC structure has lower NM in the cascading stages compared to the DC structure, as explained in previous sections.

As number of cascading stages increases, it is reasonable to believe that NM in each segment can be improved since $I_{discharge}$ in each stage decreases. As shown in Figure 4.13, increasing I_{charge} by increasing W_{charge} becomes less effective as reducing t_{search} cost more E_{search}, especially when number of stages increases. This is because the amount of time of intermediate searching result propagation between stages becomes dominant in t_{search}.

A performance comparison is conducted as summarized in Table 4.5 with other SRAM-based and RRAM-based TCAM designs implemented in 65nm technology and word size > 64-bit. For this comparison, the 2-stage TCAM design with SRC and SCPC is used (W_{charge} = 300nm). TDEs used in this 2-stage TCAM design can be implemented with RRAM-based TDE shown in Figure 3.17. High leakage current is observed from V_{PROG}=2.1V when using the ASU RRAM model fitted to the IMEC RRAM device [72, 73] by default. Therefore, for implementation of the TDE as part of the TCAM design, another set of the parameters of ASU RRAM model from [76] are used, which allows RRAMs to be programmed with V_{PROG}=1.5V. Transistor of N-type Metal-oxide-semiconductor (NMOS) of inverters are set to 120nm with width ratio of 1.5 between P-type Metal-oxide-semiconductor (PMOS) and NMOS. Programming transistors are set to minimum width required to properly program the RRAMs. With this setup, the TDE can provide a range of delay between ∼0.1ns to ∼0.5ns. Energy consumption due to leakage from V_{PROG} is limited to <3% of overall energy consumed by the TDE circuit during normal operation. One TDE can be potentially shared by multiple ML's connected to different CR-MLSA, lowering energy consumed by each CR-MLSA.

As indicated by results from Table 4.5, t_{search} of the proposed design is much lower than SRAM-based designs in [79] and [80]. Once a design is fabricated, it becomes difficult to measure t_{search}. Instead, f_{clk} becomes a more reliable metric for evaluating speed of the system. Maximum f_{clk} of the proposed design achieves up to 500MHz based on simulation data. As for E_{search}, simulated result from the proposed design is the same as SRAM-based design reported by [81] and much lower than the other SRAM-based design. Another RRAM-based TCAM design [62] is also included for comparison. Due to the difference in word size, direct comparison of t_{search} is not meaningful. However, the proposed design has advantage of lower E_{search}. Overall, [81] has the lowest t_{search} and E_{search} among all designs. However, it also has an enormous size of SA circuit with heavily cascaded structure of 6 stages per ML. In comparison, area of the proposed design is significantly smaller: the SA circuit in this work is only <10% area of that in [81]; and each TCAM cell is only 2T2R in this work compared with 12T in the SRAM cell in [81]. Overall, the proposed design is a better-rounded design in terms of energy, search speed and area than the other TCAM designs.

Table 4.5: Performance comparison with TCAM design using 65nm technology. (ML word size >64-bit)

	[81]	[79]	[80]	[62]	This work
Technology	65nm (SRAM)	65nm (SRAM)	65nm (SRAM)	65nm + RRAM (2.5T1R)	65nm + RRAM (2T2R)
ML Word Size (n-bit)	144	72	128	256	128
V_{DD} (V)	1	1	1.2	1	0.7
t_{search} (ns)	0.38	1.83	1.76	1	0.62
f_{clk} (MHz)	400	250	330	N/A	500
Avg. E_{search} (fJ/bit/search)	0.165	1.98	0.41	0.495	0.176 (TCAM) + 0.040 (TDE)
SA Circuit size (μm^2 per ML)	$2.8 \times 3.77 \times 24$	N/A	N/A	N/A	$2.14 \times 3.2 \times 2$

Another analysis is performed on how activation rate of each stage affects the expected E_{search} for the cascaded design. Here three activation scenarios are considered as listed in Table 4.6, where the first case is shown in Figure 4.13. As activation rate of later stages (r_{subsq}) increase, average E_{search} of all designs increases. In some cases, average E_{search} exceeds the 1-stage design, as shown in Figure 4.14. Therefore, it is important to study expected data pattern and reduce r_{subsq} strategically by arranging stored data bits.

Table 4.6: Three activation cases for evaluating average E_{search} in simulation

Case	Activation rate (Match rate)		
	1-stage	2-stage	4-stage
1	100% (5%)	100%→50% (5%)	100%→75%→50%→25% (5%)
2	100% (5%)	100%→60% (5%)	100%→80%→60%→40% (5%)
3	100% (5%)	100%→80% (5%)	100%→90%→80%→70% (5%)

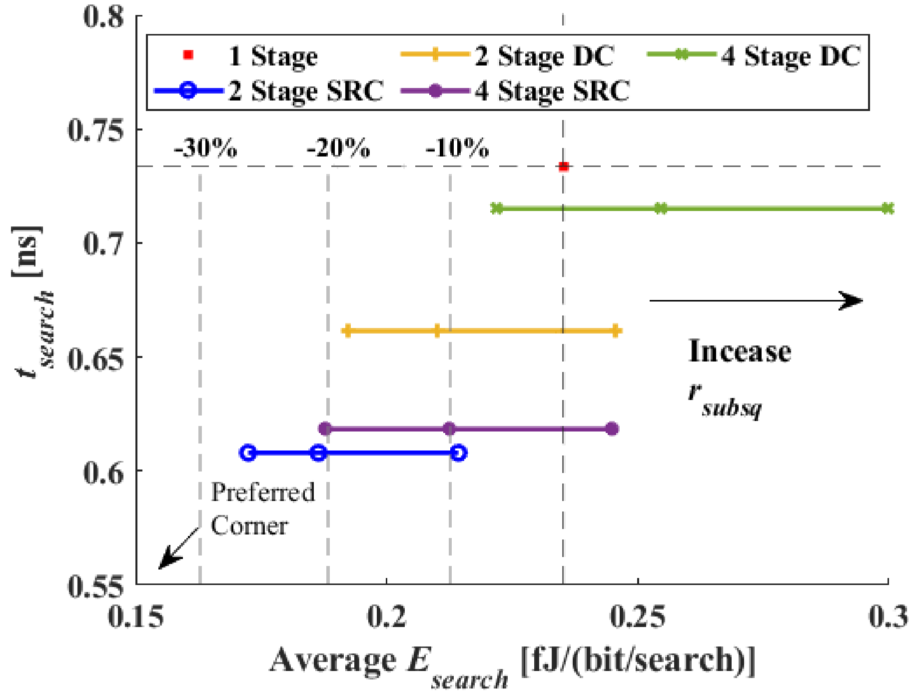

Figure 4.14: Average E_{search} and t_{search} of 2-stage and 4-stage vs 1-stage 128-bit TCAM design. With different activation rate scenarios (increasing r_{subseq}) and W_{charge} = 300nm, With high enough r_{subseq}, average E_{search} of multi-stage TCAM design can exceed E_{search} of the 1-stage design.

4.8 Summary

With the foundation of CR-MLSA design from the previous chapter, different cascading techniques have been explored in this chapter. A DC and a SRC structure are proposed, explained and compared in performance through simulation. The SRC structure shows lower t_{search} and E_{search} than the DC structure with the cost of lowering NM. Therefore, the SRC structure is preferable in a design with multiple cascading stages because segmentation improves NM in each ML segment because of reduced $I_{discharge}$. Due to the pre-discharge mechanism, CR-MLSA allows SCPC in cascading structure (both DC and SRC). Through verification from simulation, the proposed 2T2R CR-based TCAM design is illustrated to be a well-rounded design compared to other reported eNVM- and SRAM-based TCAMs, with comparable/better search time and speed of operation at lower

energy consumption. Moreover, the proposed design has the extra advantage of compact size. With all advantages combined, it is suitable for current and future large scale memory applications including network routers, image processing and neural network acceleration. Attention should also be paid to the stored data patterns since increasing r_{subsq} in multi-stage design also causes E_{search} to increase, diminishing the E_{search} reduction advantage in multi-stage TCAM design.

Chapter 5

Resistive Switching with Innovative Materials and Potential Neuromorphic Applications

5.1 Introduction

Capacitive effects of RRAM has been reported in studies [82, 83] and analyzed in depth by [84]. Such entangled effect can normally be discovered from their non-pinched [85] or "non-zero-crossing" [86] I–V hysteresis loop. Device of such kind can have different applications. One straight forward application is tunable RC filter [87], where the programmable resistance states can be utilized in modifying filter behavior during run-time, improving flexibility of circuit design. Another innovative application is LIF neurons that can be used to construct SNN. Over the years, many studies have been proposed to use circuit components such as transistors to build LIF neurons [88, 89]. However, external capacitors are usually required in the design to fill in the role of accumulating charges. In this chapter, a capacitive coupled RRAM device is presented with capacitve and resistive switching effect decoupled to determine equivalent capacitance. This is accomplished by using the GA to perform calculation on measurement data from physically fabricated device. The decoupled device is then used for designing LIF neuron circuit in Cadence Virtuoso providing post fabrication adjustable leakage rate. Simulation is done via Spectre to verify LIF neuron circuit performance. This is a collaboration study with PhD student Tao Guo from Department of Mechanical and Mechatronics Engineering, who is mainly responsible with fabrication of the device and decoupling the entangled effect of the device in section 5.2.

The main contributions of this thesis are: (a) creating circuit model of the device and (b) designing the LIF neuron circuit and verifying circuit functionalities through simulations, as described in section 5.3.

During the last decade, studies of photo-conductivity in resistive switching device has becoming a growing topic. Light illumination has shown to have impact on conductivity of RRAMs, both OxRAM and CBRAM, built based on different materials such as ZnO and SiOx [90, 91, 92, 93]. Meanwhile, Kesterite ($Cu_2ZnSnSe_4$) (CZTSe) is a well-studied material that can be used in thin films solar cell application [94, 95, 96]. Its resistive switching behavior as an RRAM device is also reported by [97]. Yet, there is a lack of study of concurrent resistive switching and light-activated conducting behavior of this kind of devices. In this chapter, CZSe, which has a similar composition to CZTSe, is used to fabricate RRAM device. Study on its resistive switching and light-activated conducting behaviors are established through measurement of the device switching characteristics. Its synaptic behavior is also evaluated by being used in simulation of ANN implemented in Python. This is also a collaboration study with PhD student Tao Guo, who is mainly responsible of carrying out ANN simulation in section 5.5. The main contributions of this thesis are (a) fabrication of the device and (b) carrying out measurements of the device performance in section 5.4.

5.2 Capacitive Coupled Effect in Al_2O_3-based RRAM Device

For this study, Ag/Al_2O_3/Al devices were fabricated with the following procedures. The bottom Al electrode is deposited onto a clean glass substrate using direct current sputtering. The active layer of Al_2O_3 was deposited using reactive sputtering with Al and a mixture of gas of O_2/Ar. (16% concentration of O_2) As for the top Ag electrode, it is deposited by directly applying silver paint on top of a mask attached to the deposited Al_2O_3 material to form dot electrode of diameter 100 μm.

Basic electrical characteristics of the fabricated device is shown in Figure 5.1. Low switching voltage of V_{SET}/V_{RESET} of 2/-2V is not enough to switch the device between LRS and HRS until they are raised to 4/-4V. The non-zero crossing I-V curve shown in Figure 5.1(b) indicates that it is a C1M1 type of capacitive-coupled memristor indicated by [84]. The device switching behaviour can also be confirmed with temporal measurement waveform shown in Figure 5.1(c). On the other hand, if the Ag top electrode is replaced with Al, the device loses the resistive switching behaviour and regress to a pure capacitor.

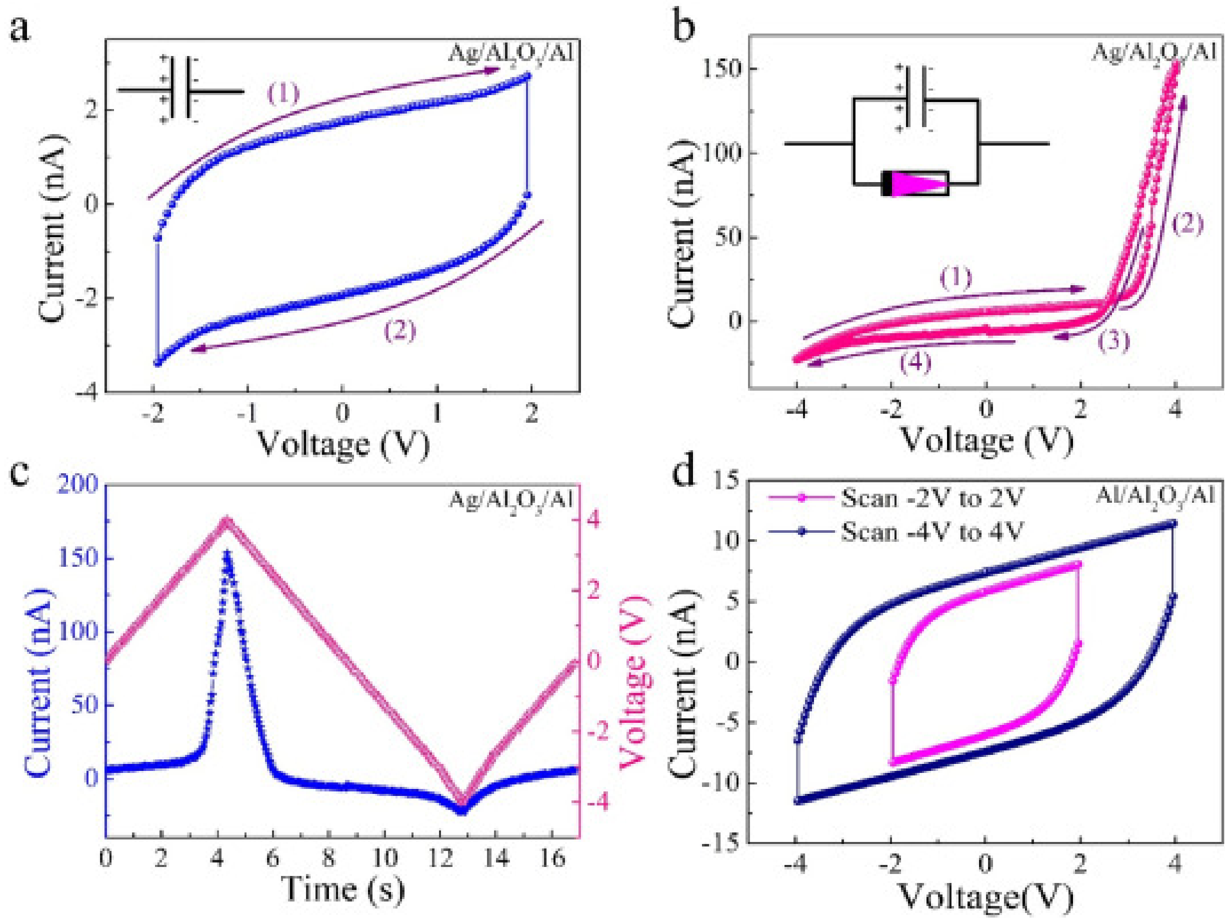

Figure 5.1: Measurement results of (a) I-V curve of Ag/Al2O3/Al device with V = -2 to 2V; (b) I-V curve of Ag/Al2O3/Al device with V = -4 to 4V; (c) V-t and corresponding I-t curves of Ag/Al2O3/Al device with V = -4 to 4V; (d) I-V curve of Al/Al2O3/Al device.

Resistive switching behavior of the fabricated device can be explained by electrochemical metallization. When a device in HRS is under forward bias with a positive electrical potential applied on the top Ag electrode, Ag atoms gets oxidized. Under the effect of applied external electrical field, the Ag ions migrate toward the bottom Al electrode, where they are reduced back to Ag atoms [98, 99]. Accumulation of Ag atoms eventually lead to formation of conductive channel connecting between the two electrodes. At this stage, the device enters LRS. When a device in LRS is reverse biased, the formed Ag conductive path is dissolved.

In order to utilize data from the fabricated device, the entangled device needs to be decoupled in order to obtain equivalent capacitance and range of resistance. Device effect decoupling is achieved by utilizing the GA, which is one of the most popular and powerful calculation tools [100]. Flowchart of this algorithm is presented as shown in Figure 5.2.

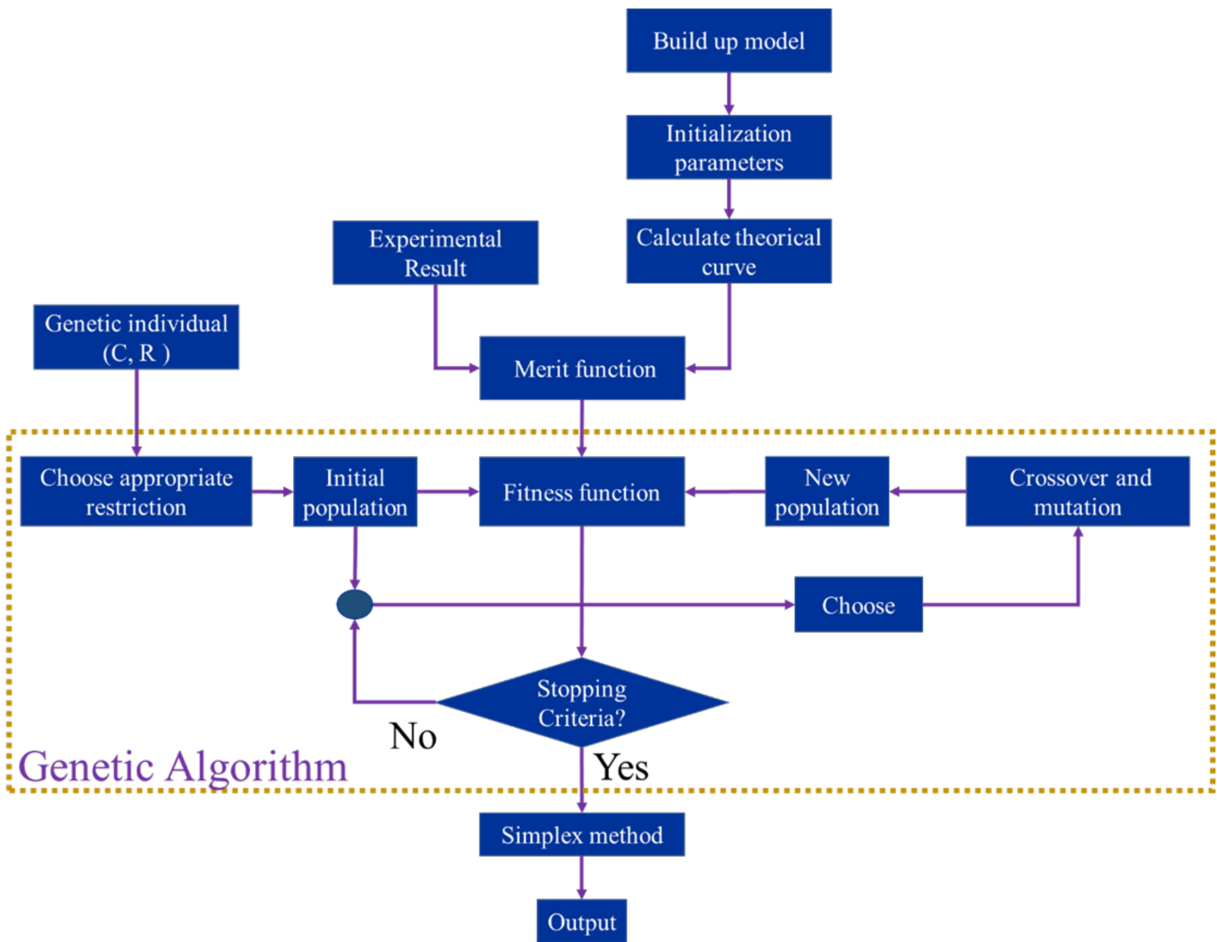

Figure 5.2: GA calculation process in a flow chart

The goal of applying this algorithm is to determine the unknown parameters in Equations (5.1) to (5.3). Equation (5.1) and (5.2) describes the I-V relationship contributed by the entangled capacitor and memristor correspondingly provided an external voltage of V. Meanwhile, Equation (5.3) describe how the Ag conductive path evolves over time with the external voltage of V applied. Equation (5.2) is developed based on the assumption that

conduction of the memristor device is mainly caused by Schottky Barrier Emission (first term) and tunneling current (second term) [101]. This is applicable because Schottky can be formed and observed between metal and Al_2O_3 thin film [102]. Variable ω in Equation (5.3) is an area index used to describe width of the Ag CF. The rest of parameters $(\alpha, \beta, \gamma, \delta, \lambda, \eta_1, \eta_2)$ are fitting parameters that correspond to material properties such as height of Schottky barrier [101].

$$I_C(V) = \frac{\mathrm{d}Q}{\mathrm{d}t} = C\frac{\mathrm{d}V}{\mathrm{d}t} \tag{5.1}$$

$$I_M(V) = (1-\omega)\alpha(1-\exp(-\beta V)) + \omega\gamma\sinh(\delta V) \tag{5.2}$$

$$\frac{\mathrm{d}\omega}{\mathrm{d}t}(V) = \lambda(\exp(\eta_1 V) - \exp(-\eta_2 V) \tag{5.3}$$

The GA is searching heuristic inspired by natural evolution in essence. By proposing new sets of unknown parameters, the resultant I-V data are then compared to the experimentally measured data, shown in Figure 5.1 (b). The sets that can more closely approximate the actual data than the others are considered better solutions. These solutions are then given subsequent opportunities to reproduce a new set of parameters that can be potentially more accurate than before [103]. With iterations of the above process, a set of essential parameters can be obtained for usage in circuit design as shown in Table 5.1.

Table 5.1: Decoupled memristor and capacitor parameter of the fabricated Al_2O_3-based RRAM device

LRS resistance	HRS resistance	Capacitance
337 MΩ	3.38 GΩ	7.34 nF

5.3 LIF Neuron Circuit Design with Al_2O_3-based RRAM Device

A LIF neuron circuit design using the Ag/Al_2O_3/Al device is shown in Figure 5.3, which is one of the contributions to this thesis. Due to the capacitive coupled effect, device $M1$ is essentially equivalent to a capacitor and resistor connected in parallel. The equivalent

capacitor stores the incoming signal pulse as charge accumulates. Meanwhile, the resistive portion of the circuit produce charge leakage effect. As V_{MEM} increases and exceeds threshold voltage of a switching circuit, V_{OUT} is set high and trigger a quick discharging path to be closed as soon as it happens. As the quick discharge happens, V_{MEM} rapidly drops, bringing V_{OUT} to 0V. This results in a short output pulse of V_{OUT} that can propagate to other neurons in the network. As shown in Figure 5.3, the switching mechanism is implemented based on a 8-transistor Schimitt Trigger based Amplifier (STBA). However, it can also be replaced by other switching circuit that has a simple structure, such as a TS device [104].

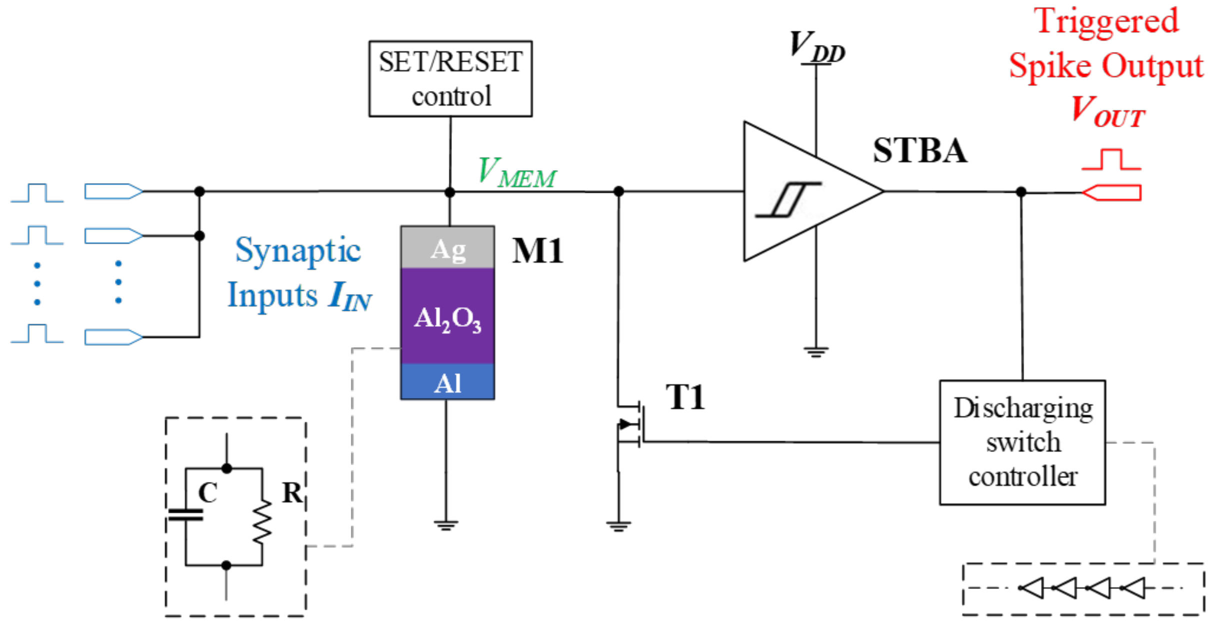

Figure 5.3: LIF neuron circuit implemented using capacitive coupled memory device

Due to the fact that $M1$ is non-volatile, its resistance can be adjusted after fabrication with peripheral circuits built with transistors controlling the its resistive programming process. This result in different equivalent leaking time constant of the LIF neuron circuit based on Equation (5.4) in the range defined by R_{HRS} and R_{LRS}. Thus, it can provide flexibility in building SNN that other components cannot provide.

$$\tau = RC \tag{5.4}$$

In order to demonstrate flexible performance that this LIF neuron can provide, simulations of circuit in Figure 5.3 with $M1$ in both HRS and LRS are shown in Figure 5.4. The circuit setup is done in Cadence Virtuoso and using TSMC 65nm PDK for implementation of transistors. Circuit simulation is carried out with Spectre. One simulation is done with same input signal amplitude but varying input signal frequency. The other simulation is done with same input signal frequency and different input signal amplitude. For each simulation, the corresponding V_{MEM} and V_{OUT} waveforms are shown. V_{MEM} increases due to the accumulation of charge injected by input pulse signal. V_{MEM} also constantly decreases due to leakage effect corresponding to resistance states of $M1$. Once V_{MEM} reaches the threshold value of STBA shown in Figure 5.3, which is around 0.75 to 0.8V, an output pulse is generated at the output correspondingly to propagate to other neurons. Meanwhile, V_{MEM} is discharged to 0V.

When $M1$ is in LRS, leakage effect of the LIF neuron is more severe compared to HRS. Therefore, with the same simulation conditions, LIF neuron with $M1$ in LRS requires higher input signal amplitude or frequency in order to generate active output and propagate to other neurons in the network. Even when output pulses are generated, output signal frequency of neuron with $M1$ in LRS is expected to be lower than neuron with $M1$ in HRS under the same input signal condition in terms of amplitude and frequency.

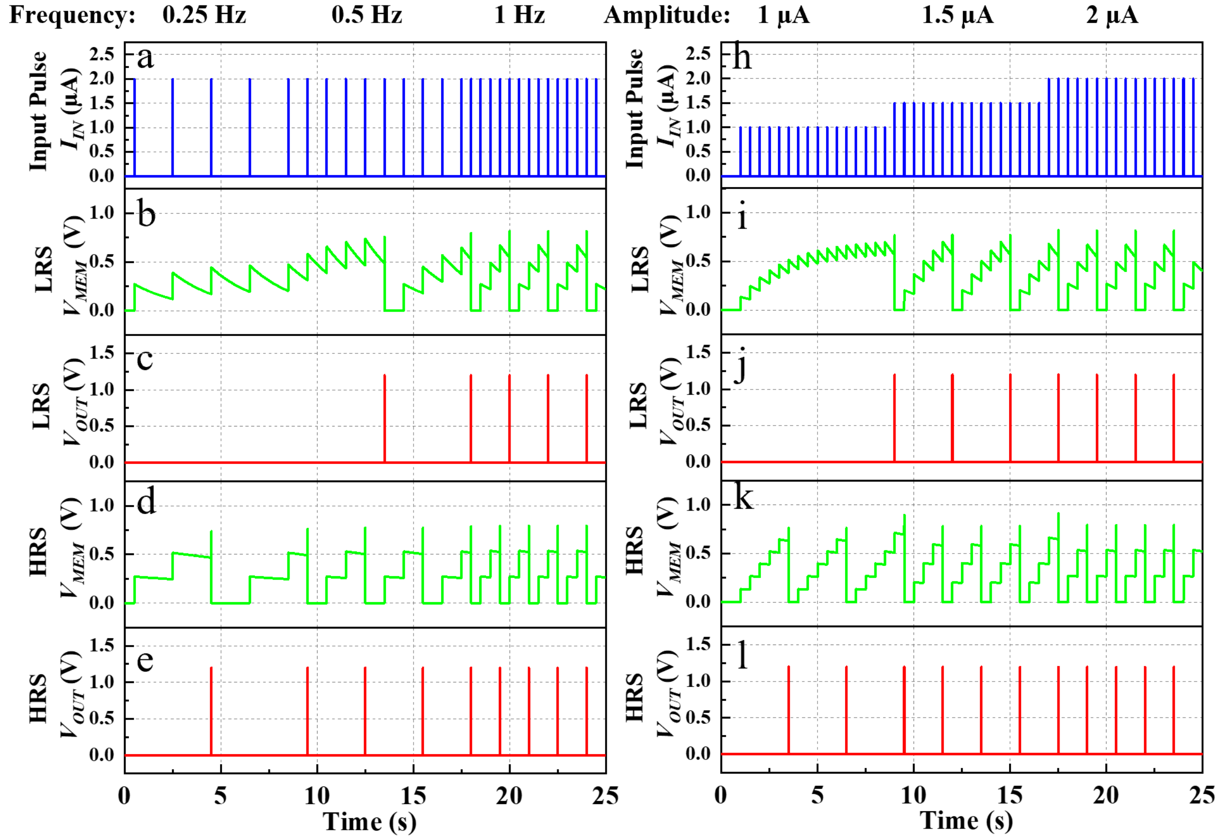

Figure 5.4: Simulated LIF neuron circuit performance: (a) Input pulse signal with increasing frequency; (b) V_{MEM} with $M1$ in LRS; (c) V_{OUT} with $M1$ in LRS; (d) V_{MEM} with $M1$ in HRS; (e) V_{OUT} with $M1$ in HRS; (h) Input pulse signal with increasing amplitude; (i) V_{MEM} with $M1$ in LRS; (j) V_{OUT} with $M1$ in LRS; (k) V_{MEM} with $M1$ in HRS; (l) V_{OUT} with $M1$ in HRS. Due to higher leakage rate of $M1$ in LRS than HRS, the neuron circuit with $M1$ in LRS requires higher input pulse signal amplitude and/or frequency to generate output spike signals.

5.4 Effect of Illumination in CZSe-based RRAM Device

For this study with CZSe, ITO/Ag/CZSe/Mo RRAM devices were fabricated and investigated in device performance, as one of the contribution to this thesis. The fabrication

process is as follow. The bottom Mo electrode was deposited onto a clean glass substrate using direct current sputtering. It is then used as a working electrode to deposit the active layer of CZSe using electro-deposition with the equipment of CHI 660E electro-chemical station and a three-electrode configuration. Meanwhile, a carbon electrode is used as the counter electrode and a saturated calomel electrode was used for the reference electrode. Composition of solution used for electro-deposition is shown in Table 5.2. The deposition was carried out with applied potential of -0.85V for 10 minutes.

Table 5.2: Decomposition of solution used for electro-deposition of CZSe

Material	Cu_2So_4	$ZnSO_4 \cdot 7H_2O$	H_2SeO_3	$C_6H_5Na_3O_7 \cdot 2H_2O$
Concentration	0.28M	0.01M	0.06M	0.2M

The samples were then separated into two batches to investigate photo-conductivity and resistive switching properties respectively. For the first batch of samples, the top electrode is deposited in a bi-layer structure using Indium Tin Oxide (ITO) and Ag using sputtering. Ag was first deposited with a very thin layer of less than 10nm. Afterward, ITO was deposited to form a transparent conductive layer on top of the Ag. Both layers of materials were deposited using sputtering with applying a mask attached to the deposited CZSe material to form dot electrode with diameter of $100\mu m$.

The first batch of device demonstrates resistive switching behavior with low V_{SET} (0.8V) and V_{RESET} (-0.7V)in measurement data as shown in Figure 5.5(a). Resistive switching behavior has been reported in devices with a similar chemical composition, CZTSe, in the active switching layer [97, 105]. When a positive potential is applied at the top electrode, Cu ions in the active layer of CZSe migrates to the bottom electrode under the effect of external electric field. They are then reduced to Cu atoms at the bottom electrode. Accumulation of the Cu atoms leads to formation of Cu CF, causing the device to enter LRS [97]. When the device is under reverse bias, the CF would be ruptured due to Joule heating and enter HRS [105].

Additionally, effect of illumination on conductance of the fabricated device is also investigated as shown in Figure 5.5(b). While a low external voltage of 0.2V is applied to the device, it can be observed that the conductance is changing over time. With full spectrum white light directly shining on the sample device, the conductance of the device rises to >12mS. It then drops back to a very low value of <1mS after the light source is removed and the sample is placed in a dark environment. It is believed that the light being shined on the CZSe layer generate electron-hole pairs in the materials due to photoelectric effect, similarly to general solar cell conduction mechanism. With the electric field generated by

the external voltage applied, electrons flow to the bottom Mo electrode and form a current flow.

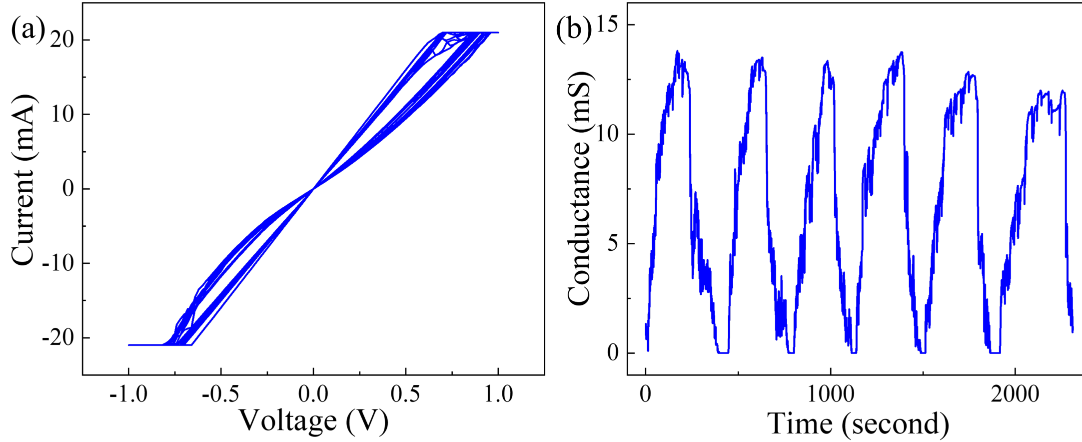

Figure 5.5: ITO/Ag/CZSe/Mo device measured performance (a) I-V curve of (b) Effect of light on low-voltage conductance

The impact of light on the switching characteristic of the ITO/Ag/CZSe/Mo device is shown in Figure 5.6. When light is shined on the device in a previously dark environment, it can be observed that the both halves of the I-V curve is lifted upward slightly, indicating an increase of current flow during through the device during the switching process. On the other hand, when light is removed and the environment becomes dark again, both halves of the I-V curve shift downward, indicating reduction of current flowing through the device. However, throughout the process, the resistive switching behavior of the device can always be observed. As mentioned in the previous section, resistive switching of CZSe is likely caused by $Cu2^{+}$ ions in the switching layer. On the other hand, the increase in conductance of CZSe under light illumination is believed to be caused by generated electron-hole pair. Therefore, the two mechanisms happen in parallel with no obvious evidence of interference with each other. This is different from resistive switching activities in OxRAM devices that are influenced by light illuminations. As pointed out in [90], ultraviolet (UV) illumination causes increase of oxygen vacancy in the ZnO material, which increases number of CFs during RRAM SET process.

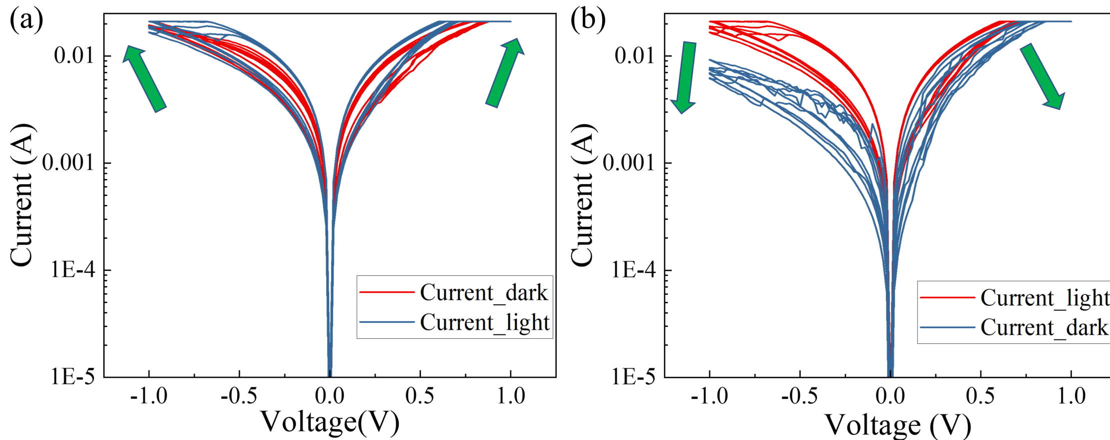

Figure 5.6: Effect of light on ITO/Ag/CZSe/Mo device switching behavior observed in resistive switching measurement: (a) from dark to shining light (b) from shining light to dark

As can be seen from both plots in both Figure 5.5 and Figure 5.6, the device is showing instability between voltage sweep cycles. This is likely caused by the inconsistency in the crystal structure due to the electro-deposition process. It is plausible that deposition method such as sputtering can yield monoclinic crystal cell structure and therefore more consistent resistive switching behaviour [96, 97] .

5.5 Synaptic Behavior in CZSe-based RRAM Device

In order to improve the device switching stability and evaluate its performance for NC applications, a batch of the CZSe/Mo samples underwent additional procedure and different switching measurement, as a part of contributions from this thesis. The work of using device measurement data to carry out ANN simulation in Python is contributed by PhD student Tao Guo. This batch of samples went through a Rapid Thermal Annealing (RTA) process with excess Se particles at atmospheric pressure. The samples are sent into a chamber pre-heated to 500 °C from room temperature for 10 minutes. Then the samples are quickly removed from the heated chamber and returned to room temperature. Throughout the RTA process, the chamber is filled with Ar gas to prevent reaction between the deposited material and oxygen in the air. After the RTA process, a layer of Ag

is deposited using sputtering to form the top electrode. Again, a mask attached to the deposited CZSe material to form dot electrode of diameter 100 μm.

This batch of samples demonstrate much more stable resistive switching characteristics compared to the first batch of samples. As shown in Figure 5.7, the device shows a similar I-V profile in terms of V_{SET}, V_{RESET} and resistance in both states compared to the first batch of devices. After 100 cycles of voltage switching scan, the resistive switching profile of the device still remains very stable. This proves that RTA is helpful for enhancing crystallinity of deposited CZSe, as suggested by [95] and [106].

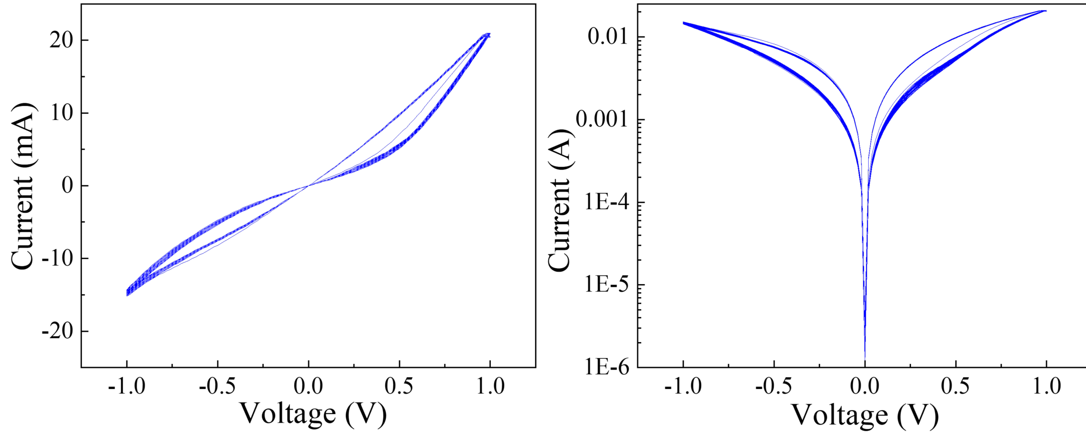

Figure 5.7: Ag/CZSe/Mo (Annealed) I-V curve of 100 sweeping cycles in with y-axis (current) in (a) linear scale and (b) log scale.

Synaptic behavior of this device is investigated through measurement by applying pulse programming and monitoring conductance of the device. For evaluating Long-term Potentiation (LTP), a train of 50 consecutive pulse with 0.5V amplitude and 100msec pulse width. For evaluating long-term Depression (LTD), the amplitude of the pulse signal is changed to -0.5V with number of pulses and pulse width unchanged. The conductance is monitored after each programming pulse by applying a small read voltage of 0.1V. The conductance of the device is equivalent to synaptic weight connected between neurons when using the device as synapse to construct neural networks. The performance of how conductance of the device changes with each pulse applied is shown in Figure 5.8(a).

This LTD and LTP data profile is then used for simulation of a three-layer feedfoward ANN, implemented in Python. As shown in Figure 5.8(b), this neural network has 64

neurons in the input layer, 36 neurons in the hidden layer and 10 neurons in the output layer. The network is trained using the backpropagation algorithm commonly used in ANN training. Images with 8×8 resolutions from [107] are used as data set for training and testing of this ANN. The accuracy of this ANN with respect to number of training epochs is shown in Figure 5.8(b). Accuracy of this network can reach 94% with 5 training epochs. Therefore, this device demonstrates good synaptic behavior to be used in constructing neural network.

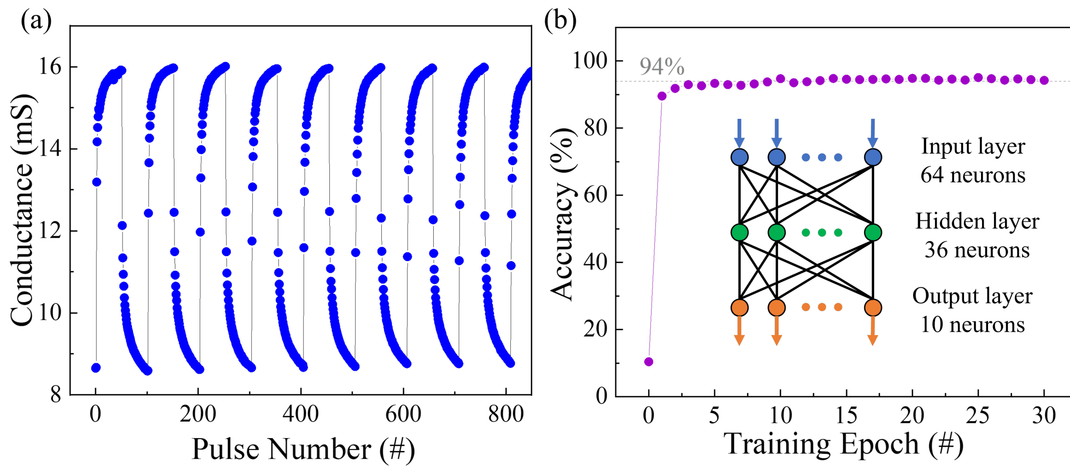

Figure 5.8: (a) Synaptic behavior (LTP and LTD) of Ag/CZSe/Mo device shown through measurement by applying sequence of SET and RESET voltage pulse to program resistance. (b) Simulation result of a 3-layer feedforward ANN constructed in Python using the Ag/CZSe/Mo device as synapses.

5.6 Summary

In this chapter, a Al_2O_3-based and a CZSe-based RRAM devices are shown and evaluated in performance. The Al_2O_3-based RRAM device shows capacitor-coupled switching behavior in measurement, which can be decoupled using the GA. A LIF neuron circuit is proposed based on this device and its performance is verified through simulation. The LIF neuron circuit can provide after-fabrication leakage rate tunability, increasing flexibility in neural network hardware design. On the other hand, the CZSe-based RRAM device demonstrates both resistive switching and light-activated conducting behavior at

the same time through measurement. Yet, instability can be observed during testing and measurement of the device. After an RTA process, the switching characteristic of this device improves in terms of stability. It is then demonstrated through simulation that this device is suitable be used in constructing ANN with high accuracy after training.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this study, an innovative CR-based TCAM circuit architecture is proposed using 2T2R TCAM cell configuration with RRAM. With the introduction of a compact ML booster, the proposed design is capable to provide high search speed and low energy consumption compared to previous eNVM-based TCAM design. It also provides tolerance to RRAM switching variation, for which there is still a lack of studies in eNVM based TCAM circuit design. This foundation of work is then extended to build TCAM cascading circuit structure to account for TCAM applications with large word size. Two cascading approaches are presented and compared: DC and SRC. SRC structure provides fast searching speed and low energy consumption while DC cascading provides high noise margin. Meanwhile, a SCPC approach, which is enabled by CR-based MLSA but not the commonly used PnE MLSA. This SCPC approach maintains the energy saving feature in cascading structure, while reducing the output latency from conventional cascading structures. With verification through simulation, the proposed multi-stage TCAM design is illustrated to be a well-rounded design compared to other reported eNVM- and SRAM-based TCAMs, with comparable/better search speed of operation at lower energy consumption and a compact TCAM cell structure, With all advantages combined, it is suitable for current and future large scale memory applications including network routers, image processing and neural network acceleration.

In addition, an RRAM-based TDE circuit is also proposed. It can be used in TCAM circuit design for tuning delay of reference signal sent to each TCAM MLSA to ensure correct functionality and efficiency of the TCAM circuit system. An improved TDE circuit

with two RRAM placed in parallel configuration is also proposed and evaluated. It shows great potential in reducing impact of RRAM switching variation and improving tunable delay resolutions, as indicated by its simulated performance.

In the third part of this study, two innovative RRAM device, Al_2O_3-based and CZSe-based, are presented and evaluated. The Al_2O_3-based RRAM device demonstrates capacitor coupled resistive switching behavior. After decoupling the two effects, a simplified model is used for designing a LIF neuron circuit. The simulated LIF neuron circuit can provide adjustable leakage rate after fabrication because of the non-volatile resistive switching behavior from the RRAM device. This can provide another flexibility when designing SNN with LIF neurons. On the other hand, the CZSe-based RRAM device demonstrates both resistive switching and light-activated conducting behavior at the same time. After a RTA process, the device switching stability is improved such that it can be used as synapses in constructing ANN with high output accuracy, which is verified through simulation.

6.2 Future Work

The work and data presented in this study are mostly based on simulation of circuit design. Devices fabricated for the studies used methods that have constrictions on device size and process optimization. From this standpoint, the proposed future work are listed as follow:

1. A more detailed and established design and layout of the TCAM circuit system proposed in this thesis is required. This also includes the detailed design integration of a RSG system, TDE circuit and RRAM cell programming circuit. A fabricated prototype is required for the next step of evaluating overall TCAM system design performance through measurement.

2. The TCAM design proposed in this thesis does not cover any SL encoding mechanism that can potentially reduce impact from RRAM switching variation to system performance. Yet, such a mechanism will require extra hardware and software to be implemented. A future study in this direction can be established to evaluate the trade-off between circuit performance and overall circuit area.

3. The proposed RRAM-based TDE design is utilizing quasi-analog switching behavior of RRAM. So far, the simulation of circuit performance has only been verified in simulation with the ASU RRAM model. Further study should be established with preferably fabricated prototype circuit to generate enough experimental data. Such

data can then be used to compare with predictions from simulation. This will also be helpful for further study of RRAM device switching variation and improve RRAM-based TDE design.

4. The Al_2O_3-based and CZSe-based RRAM devices presented in this thesis are fabricated with less-than-optimized fabrication process. Other fabrication methods should be investigated in attempt to reduce device dimension and preferably integrate the device with CMOS transistor to form complete circuit systems. Only then can the devices be more thoroughly examined and tested for their performance in neuromorphic computation circuit design and/or other circuit system design.

References

[1] S. Thompson *et al.*, "A 90-nm logic technology featuring strained-silicon," *IEEE Transactions on Electron Devices*, vol. 51, no. 11, pp. 1790–1797, 2004.

[2] K. Mistry *et al.*, "A 45nm logic technology with high-k+metal gate transistors, strained silicon, 9 cu interconnect layers, 193nm dry patterning, and 100% pb-free packaging," in *2007 IEEE International Electron Devices Meeting*, pp. 247–250, 2007.

[3] L. Chang *et al.*, "Extremely scaled silicon nano-CMOS devices," *Proceedings of the IEEE*, vol. 91, no. 11, pp. 1860–1873, 2003.

[4] Y. Taur *et al.*, "CMOS scaling into the nanometer regime," *Proceedings of the IEEE*, vol. 85, no. 4, pp. 486–504, 1997.

[5] S. Nassif, "Modeling and analysis of manufacturing variations," in *Proceedings of the IEEE 2001 Custom Integrated Circuits Conference (Cat. No.01CH37169)*, pp. 223–228, 2001.

[6] K. Rupp, "microprocessor-trend-data." https://github.com/karlrupp/microprocessor-trend-data, 2022.

[7] H.-S. P. Wong, C.-S. Lee, J. Luo, and C.-H. Wang, "CMOS technology scaling trend." https://nano.stanford.edu/cmos-technology-scaling-trend. Accessed: 2010-09-30.

[8] W. Haensch *et al.*, "Silicon CMOS devices beyond scaling," *IBM Journal of Research and Development*, vol. 50, no. 4.5, pp. 339–361, 2006.

[9] H.-S. P. Wong, "Beyond the conventional transistor," *IBM Journal of Research and Development*, vol. 46, no. 2.3, pp. 133–168, 2002.

[10] D. Frank *et al.*, "Device scaling limits of Si MOSFETs and their application dependencies," *Proceedings of the IEEE*, vol. 89, no. 3, pp. 259–288, 2001.

[11] T. N. Theis and H.-S. P. Wong, "The end of Moore's Law: A new beginning for information technology," *Computing in Science Engineering*, vol. 19, no. 2, pp. 41–50, 2017.

[12] S. Yu and P.-Y. Chen, "Emerging memory technologies: Recent trends and prospects," *IEEE Solid-State Circuits Magazine*, vol. 8, no. 2, pp. 43–56, 2016.

[13] B. J. Choi *et al.*, "High-speed and low-energy nitride memristors," *Advanced Functional Materials*, vol. 26, no. 29, pp. 5290–5296, 2016.

[14] C. Pan *et al.*, "Coexistence of grain-boundaries-assisted bipolar and threshold resistive switching in multilayer hexagonal boron nitride," *Advanced functional materials*, vol. 27, no. 10, p. 1604811, 2017.

[15] D. Patterson *et al.*, "A case for intelligent RAM," *IEEE Micro*, vol. 17, no. 2, pp. 34–44, 1997.

[16] M. V. DeBole *et al.*, "TrueNorth: Accelerating from zero to 64 million neurons in 10 years," *Computer*, vol. 52, no. 5, pp. 20–29, 2019.

[17] M. Davies *et al.*, "Loihi: A neuromorphic manycore processor with on-chip learning," *IEEE Micro*, vol. 38, no. 1, pp. 82–99, 2018.

[18] N. P. Jouppi *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *Proceedings of the 44th annual international symposium on computer architecture*, pp. 1–12, 2017.

[19] S. Park *et al.*, "RRAM-based synapse for neuromorphic system with pattern recognition function," in *2012 International Electron Devices Meeting*, pp. 10.2.1–10.2.4, 2012.

[20] S. Park *et al.*, "Nanoscale RRAM-based synaptic electronics: toward a neuromorphic computing device," *nanotechnology*, vol. 24, no. 38, p. 384009, 2013.

[21] J. Park *et al.*, "TiOx-Based RRAM synapse with 64-levels of conductance and symmetric conductance change by adopting a hybrid pulse scheme for neuromorphic computing," *IEEE Electron Device Letters*, vol. 37, no. 12, pp. 1559–1562, 2016.

[22] L. Chua, "Memristor-the missing circuit element," *IEEE Transactions on Circuit Theory*, vol. 18, no. 5, pp. 507–519, 1971.

[23] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, "The missing memristor found," *Nature*, vol. 453, pp. 80–83, May 2008.

[24] J.-G. Zhu, "Magnetoresistive random access memory: The path to competitiveness and scalability," *Proceedings of the IEEE*, vol. 96, no. 11, pp. 1786–1798, 2008.

[25] H.-S. P. Wong *et al.*, "Phase change memory," *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2201–2227, 2010.

[26] H.-Y. Chen *et al.*, "Resistive random access memory (rram) technology: From material, device, selector, 3d integration to bottom-up fabrication," *Journal of Electroceramics*, vol. 39, no. 1, pp. 21–38, 2017.

[27] X. Guan, S. Yu, and H.-S. P. Wong, "On the switching parameter variation of Metal-Oxide RRAM—Part I: Physical modeling and simulation methodology," *IEEE Transactions on Electron Devices*, vol. 59, no. 4, pp. 1172–1182, 2012.

[28] C.-W. Hsu *et al.*, "Self-rectifying bipolar tao_x/tio_2 rram with superior endurance over 10^{12} cycles for 3d high-density storage-class memory," in *2013 Symposium on VLSI Technology*, pp. T166–T167, 2013.

[29] Y. Huang *et al.*, "Amorphous zno based resistive random access memory," *RSC advances*, vol. 6, no. 22, pp. 17867–17872, 2016.

[30] M. Kund *et al.*, "Conductive bridging ram (CBRAM): an emerging non-volatile memory technology scalable to sub 20nm," in *IEEE InternationalElectron Devices Meeting, 2005. IEDM Technical Digest.*, pp. 754–757, 2005.

[31] P.-Y. Chen and S. Yu, "Compact modeling of rram devices and its applications in 1T1R and 1S1R array design," *IEEE Transactions on Electron Devices*, vol. 62, no. 12, pp. 4022–4028, 2015.

[32] S.-S. Sheu *et al.*, "A 5ns fast write multi-level non-volatile 1 k bits rram memory with advance write scheme," in *2009 Symposium on VLSI Circuits*, pp. 82–83, 2009.

[33] S. R. Lee *et al.*, "Multi-level switching of triple-layered TaOx RRAM with excellent reliability for storage class memory," in *2012 Symposium on VLSI Technology (VLSIT)*, pp. 71–72, 2012.

[34] S. Kvatinsky, A. Kolodny, U. C. Weiser, and E. G. Friedman, “Memristor-based IMPLY logic design procedure,” in *2011 IEEE 29th International Conference on Computer Design (ICCD)*, pp. 142–147, 2011.

[35] S. Kvatinsky *et al.*, “MAGIC—memristor-aided logic,” *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 61, no. 11, pp. 895–899, 2014.

[36] S. Kvatinsky *et al.*, “MRL — memristor ratioed logic,” in *2012 13th International Workshop on Cellular Nanoscale Networks and their Applications*, pp. 1–6, 2012.

[37] X.-Y. Wang *et al.*, “High-density Memristor-CMOS ternary logic family,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 1, pp. 264–274, 2021.

[38] P. Yao *et al.*, “Fully hardware-implemented memristor convolutional neural network,” *Nature*, vol. 577, no. 7792, pp. 641–646, 2020.

[39] S. Lashkare *et al.*, “PCMO RRAM for integrate-and-fire neuron in spiking neural networks,” *IEEE Electron Device Letters*, vol. 39, no. 4, pp. 484–487, 2018.

[40] J. Woo, D. Lee, Y. Koo, and H. Hwang, “Dual functionality of threshold and multi-level resistive switching characteristics in nanoscale HfO2-based rram devices for artificial neuron and synapse elements,” *Microelectronic Engineering*, vol. 182, pp. 42–45, 2017.

[41] R. Degraeve *et al.*, “Causes and consequences of the stochastic aspect of filamentary RRAM,” *Microelectronic Engineering*, vol. 147, pp. 171–175, 2015.

[42] A. Grossi *et al.*, “Impact of intercell and intracell variability on forming and switching parameters in RRAM arrays,” *IEEE Transactions on Electron Devices*, vol. 62, no. 8, pp. 2502–2509, 2015.

[43] M. Trapatseli *et al.*, “Engineering the switching dynamics of TiOx-based RRAM with Al doping,” *Journal of Applied Physics*, vol. 120, no. 2, p. 025108, 2016.

[44] S. H. Misha *et al.*, “Effect of nitrogen doping on variability of TaOx-RRAM for low-power 3-bit MLC applications,” *ECS Solid State Letters*, vol. 4, no. 3, p. P25, 2015.

[45] H.-Y. Chen *et al.*, “Resistive random access memory (RRAM) technology: From material, device, selector, 3D integration to bottom-up fabrication,” *Journal of Electroceramics*, vol. 39, no. 1, pp. 21–38, 2017.

[46] M. Yu *et al.*, "Novel vertical 3d structure of TaOx-based RRAM with self-localized switching region by sidewall electrode oxidation," *Scientific reports*, vol. 6, no. 1, pp. 1–10, 2016.

[47] M.-F. Chang *et al.*, "A sub-0.3V area-efficient L-Shaped 7T SRAM with read bit-line swing expansion schemes based on boosted read-bitline, asymmetric-V_{TH} read-port, and offset cell VDD biasing techniques," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 10, pp. 2558–2569, 2013.

[48] P. Jain *et al.*, "13.2A 3.6Mb 10.1Mb/mm^2 embedded non-volatile ReRAM Macro in 22nm FinFET technology with adaptive forming/set/reset schemes yielding down to 0.5V with sensing time of 5ns at 0.7V," in *2019 IEEE International Solid- State Circuits Conference - (ISSCC)*, pp. 212–214, 2019.

[49] Y. Y. Chen *et al.*, "Improvement of data retention in HfO_2/Hf 1T1R RRAM cell under low operating current," in *2013 IEEE International Electron Devices Meeting*, pp. 10.1.1–10.1.4, 2013.

[50] T.-B. Pei and C. Zukowski, "VLSI implementation of routing tables: tries and CAMs," in *IEEE INFCOM '91. The conference on Computer Communications. Tenth Annual Joint Comference of the IEEE Computer and Communications Societies Proceedings*, pp. 515–524 vol.2, 1991.

[51] H. Chao, "Next generation routers," *Proceedings of the IEEE*, vol. 90, no. 9, pp. 1518–1558, 2002.

[52] A. McAuley and P. Francis, "Fast routing table lookup using CAMs," in *IEEE INFOCOM '93 The Conference on Computer Communications, Proceedings*, pp. 1382–1391 vol.3, 1993.

[53] S. Panchanathan and M. Goldberg, "A content-addressable memory architecture for image coding using vector quantization," *IEEE Transactions on Signal Processing*, vol. 39, no. 9, pp. 2066–2078, 1991.

[54] M. Imani, D. Peroni, Y. Kim, A. Rahimi, and T. Rosing, "Efficient neural network acceleration on GPGPU using content addressable memory," in *Design, Automation Test in Europe Conference Exhibition (DATE), 2017*, pp. 1026–1031, 2017.

[55] N. Mohan, W. Fung, D. Wright, and M. Sachdev, "Design techniques and test methodology for low-power TCAMs," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 14, no. 6, pp. 573–586, 2006.

[56] N. Mohan, "Low-power high-performance ternary content addressable memory circuits," 2006.

[57] K. Pagiamtzis and A. Sheikholeslami, "Content-addressable memory (CAM) circuits and architectures: a tutorial and survey," *IEEE Journal of Solid-State Circuits*, vol. 41, no. 3, pp. 712–727, 2006.

[58] J. Li, R. K. Montoye, M. Ishii, and L. Chang, "1 Mb 0.41 µm² 2T-2R cell nonvolatile TCAM with two-bit encoding and clocked self-referenced sensing," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 4, pp. 896–907, 2014.

[59] X. Wang *et al.*, "A 4T2R rram bit cell for highly parallel ternary content addressable memory," *IEEE Transactions on Electron Devices*, vol. 68, no. 10, pp. 4933–4937, 2021.

[60] M.-F. Chang *et al.*, "A 3T1R nonvolatile TCAM using MLC ReRAM for frequent-off instant-on filters in IoT and big-data processing," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 6, pp. 1664–1679, 2017.

[61] L. Zheng, S. Shin, and S.-M. S. Kang, "Memristors-based ternary content addressable memory (mTCAM)," in *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 2253–2256, 2014.

[62] C. Lin *et al.*, "A 256b-wordlength ReRAM-based TCAM with 1ns search-time and 14× improvement in wordlength-energyefficiency-density product using 2.5T1R cell," in *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 136–137, 2016.

[63] A. M. S. Tosson, M. Anis, and L. Wei, "RRAM refresh circuit: A proposed solution to resolve the soft-error failures for HfO2/Hf 1T1R RRAM memory cell," in *2016 International Great Lakes Symposium on VLSI (GLSVLSI)*, pp. 227–232, 2016.

[64] A. M. Tosson *et al.*, "Analysis of RRAM reliability soft-errors on the performance of rram-based neuromorphic systems," in *2017 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, pp. 62–67, 2017.

[65] A. Rezaeian, G. Ardeshir, and M. Gholami, "A low-power and high-frequency phase frequency detector for a 3.33-GHz delay locked loop," *Circuits Syst Signal Process*, vol. 39, p. 1735–1750, 2019.

[66] J. Morales, Chierchie, M. F, and E. PS, Paolini, "A high-resolution all-digital pulse-width modulator architecture with a tunable delay element in CMOS," *Int J Circ Theor Appl.*, vol. 48, no. 8, p. 1329–1345, 2020.

[67] J. Gu and J. Li, "Exploration of self-healing circuits for timing resilient design using emerging memristor devices," in *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1458–1461, 2015.

[68] S. M. A. B. Mokhtar and W. F. H. W. Abdullah, "Memristor based delay element using current starved inverter," in *RSM 2013 IEEE Regional Symposium on Micro and Nanoelectronics*, pp. 81–84, 2013.

[69] T. Bunnam, A. Soltan, D. Sokolov, and A. Yakovlev, "Pulse controlled memristor-based delay element," in *2017 27th International Symposium on Power and Timing Modeling, Optimization and Simulation (PATMOS)*, pp. 1–8, 2017.

[70] I. Arsovski, T. Chandler, and A. Sheikholeslami, "A ternary content-addressable memory (TCAM) based on 4T static storage and including a current-race sensing scheme," *IEEE Journal of Solid-State Circuits*, vol. 38, no. 1, pp. 155–158, 2003.

[71] X. Guan, S. Yu, and H.-S. P. Wong, "A SPICE compact model of metal oxide resistive switching memory with variations," *IEEE Electron Device Letters*, vol. 33, no. 10, pp. 1405–1407, 2012.

[72] Y. Y. Chen *et al.*, "Balancing SET/RESET pulse for $> 10^{10}$ endurance in HfO_2/Hf 1T1R bipolar rram," *IEEE Transactions on Electron Devices*, vol. 59, no. 12, pp. 3243–3249, 2012.

[73] Y. Y. Chen *et al.*, "Improvement of data retention in HfO_2/Hf 1T1R rram cell under low operating current," in *2013 IEEE International Electron Devices Meeting*, pp. 10.1.1–10.1.4, 2013.

[74] N. Mohan, W. Fung, D. Wright, and M. Sachdev, "A low-power ternary CAM with positive-feedback match-line sense amplifiers," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 56, no. 3, pp. 566–573, 2009.

[75] W. W. Fung, "Low power circuits for multiple match resolution and detection in ternary CAMs," 2004.

[76] E. R. Berikaa, A. Khalil, H. Hossam, M. El-Dessouky, and H. Mostafa, "Multi-bit RRAM transient modelling and analysis," in *2018 30th International Conference on Microelectronics (ICM)*, pp. 232–235, 2018.

[77] J. Rajendran, R. Karri, and G. S. Rose, "Improving tolerance to variations in memristor-based applications using parallel memristors," *IEEE Transactions on Computers*, vol. 64, no. 3, pp. 733–746, 2015.

[78] C.-X. Xue *et al.*, "A 28-nm 320-Kb TCAM macro using split-controlled single-load 14T cell and triple-margin voltage sense amplifier," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 10, pp. 2743–2753, 2019.

[79] I. Hayashi *et al.*, "A 250-MHz 18-Mb full ternary CAM with low-voltage matchline sensing scheme in 65-nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 11, pp. 2671–2680, 2013.

[80] K.-C. Woo and B.-D. Yang, "Low-area TCAM using a don't care reduction scheme," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 8, pp. 2427–2433, 2018.

[81] P.-T. Huang and W. Hwang, "A 65 nm 0.165 fJ/Bit/Search 256×144 TCAM macro design for IPv6 lookup tables," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 2, pp. 507–519, 2011.

[82] J. Song *et al.*, "Effects of reset current overshoot and resistance state on reliability of RRAM," *IEEE Electron Device Letters*, vol. 35, no. 6, pp. 636–638, 2014.

[83] S. Ambrogio, V. Milo, Z. Wang, S. Balatti, and D. Ielmini, "Analytical modeling of current overshoot in oxide-based resistive switching memory (RRAM)," *IEEE Electron Device Letters*, vol. 37, no. 10, pp. 1268–1271, 2016.

[84] B. Sun *et al.*, "A unified capacitive-coupled memristive model for the nonpinched current–voltage hysteresis loop," *Nano Letters*, vol. 19, no. 9, pp. 6461–6465, 2019.

[85] S. Sarma, B. M. Mothudi, and M. S. Dhlamini, "Observed coexistence of memristive, memcapacitive and meminductive characteristics in polyvinyl alcohol/cadmium sulphide nanocomposites," *Journal of Materials Science: Materials in Electronics*, vol. 27, no. 5, pp. 4551–4558, 2016.

[86] F. Messerschmitt, M. Kubicek, and J. L. Rupp, "How does moisture affect the physical property of memristance for anionic–electronic resistive switching memories?," *Advanced Functional Materials*, vol. 25, no. 32, pp. 5117–5125, 2015.

[87] S. Ranjan *et al.*, "Passive filters for nonvolatile storage based on capacitive-coupled memristive effects in nanolayered organic–inorganic heterojunction devices," *ACS Applied Nano Materials*, vol. 3, no. 6, pp. 5045–5052, 2020.

[88] X. Wu, V. Saxena, K. Zhu, and S. Balagopal, "A CMOS spiking neuron for brain-inspired neural networks with resistive synapses and *InSitu* learning," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 62, no. 11, pp. 1088–1092, 2015.

[89] S. A. Aamir, P. Müller, A. Hartel, J. Schemmel, and K. Meier, "A highly tunable 65-nm CMOS LIF neuron for a large scale neuromorphic system," in *ESSCIRC Conference 2016: 42nd European Solid-State Circuits Conference*, pp. 71–74, 2016.

[90] P. Russo, M. Xiao, R. Liang, and N. Y. Zhou, "UV-Induced multilevel current amplification memory effect in zinc oxide rods resistive switching devices," *Advanced Functional Materials*, vol. 28, no. 13, p. 1706230, 2018.

[91] C.-C. Shih *et al.*, "Resistive switching modification by ultraviolet illumination in transparent electrode resistive random access memory," *IEEE Electron Device Letters*, vol. 35, no. 6, pp. 633–635, 2014.

[92] A. Mehonic, T. Gerard, and A. Kenyon, "Light-activated resistance switching in SiOx rram devices," *Applied Physics Letters*, vol. 111, no. 23, p. 233502, 2017.

[93] D. Jana, S. Chakrabarti, S. Z. Rahaman, and S. Maikap, "Resistive and new optical switching memory characteristics using thermally grown Ge 0.2 Se 0.8 film in Cu/GeSe x/W structure," *Nanoscale research letters*, vol. 10, no. 1, pp. 1–8, 2015.

[94] D. Colombara *et al.*, "Electrodeposition of kesterite thin films for photovoltaic applications: Quo vadis?," *physica status solidi (a)*, vol. 212, no. 1, pp. 88–102, 2015.

[95] J.-O. Jeon *et al.*, "Highly efficient copper–zinc–tin–selenide (CZTSe) solar cells by electrodeposition," *ChemSusChem*, vol. 7, no. 4, pp. 1073–1077, 2014.

[96] P. Zheng, B. Sun, Y. Zhao, and Z. Yu, "Tunneling of carrier at the interface barrier induced nonvolatile resistive switching memory behaviors," *Materials Today Communications*, vol. 16, pp. 164–168, 2018.

[97] T. Guo *et al.*, "Overwhelming coexistence of negative differential resistance effect and RRAM," *Physical Chemistry Chemical Physics*, vol. 20, no. 31, pp. 20635–20640, 2018.

[98] Y. Yang *et al.*, "Electrochemical dynamics of nanoscale metallic inclusions in dielectrics," *Nature Communications*, vol. 5, p. 4232, Jun 2014.

[99] T. Guo *et al.*, "Effect of electrode materials on nonvolatile resistive switching memory behaviors of Metal/In2S3/Mo/Glass devices," *Journal of Electronic Materials*, vol. 47, pp. 5417–5421, Sep 2018.

[100] L. Brezočnik, I. Fister, and V. Podgorelec, "Swarm intelligence algorithms for feature selection: A review," *Applied Sciences*, vol. 8, no. 9, 2018.

[101] T. Chang *et al.*, "Synaptic behaviors and modeling of a metal oxide memristive device," *Applied Physics A*, vol. 102, pp. 857–863, Mar 2011.

[102] L. Dissado and S. Le Roy, "The effect of contact charge upon the injection current at an electrode-insulator interface," in *2007 IEEE International Conference on Solid Dielectrics*, pp. 31–34, 2007.

[103] D. Whitley, "A genetic algorithm tutorial," *Statistics and Computing*, vol. 4, pp. 65–85, Jun 1994.

[104] P. Stoliar *et al.*, "A leaky-integrate-and-fire neuron analog realized with a Mott insulator," *Advanced Functional Materials*, vol. 27, no. 11, p. 1604740, 2017.

[105] S. Gao, C. Song, C. Chen, F. Zeng, and F. Pan, "Dynamic processes of resistive switching in metallic filament-based organic memory devices," *The Journal of Physical Chemistry C*, vol. 116, no. 33, pp. 17955–17959, 2012.

[106] S. Pawar *et al.*, "Single step electrosynthesis of Cu2ZnSnS4 (CZTS) thin films for solar cell application," *Electrochimica Acta*, vol. 55, no. 12, pp. 4057–4061, 2010.

[107] D. Dua and C. Graff, "UCI machine learning repository," 2017.

APPENDICES

Appendix A

List of Publications

Here is a list of published papers from this work. They correspond to work presented in Sections 3.8, 3.9, 5.2 and 5.3.

- K. Pan, A. M. S. Tosson, N. Y. Zhou and L. Wei, "A Novel Programmable Variation-Tolerant RRAM-based Delay Element Circuit," 2021 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH), 2021, pp. 1-2.
- T. Guo, K. Pan, B. Sun, L. Wei, Y. Yan, Y. Zhou, and Y. Wu, "Adjustable leaky-integrate-and-fire neurons based on memristor-coupled capacitors," Materials Today Advances, vol. 12, 2021, 100192.