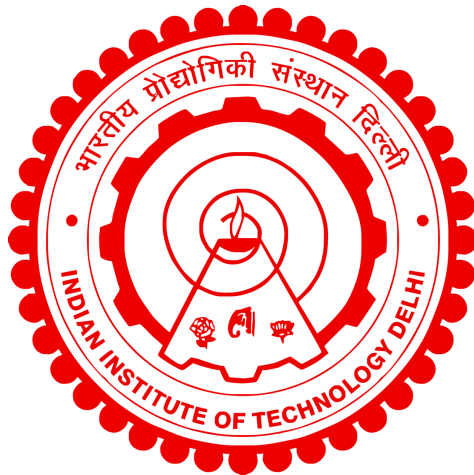


EXPLORING COMPUTING APPLICATIONS WITH NON-VOLATILE MEMORY

SANDEEP KAUR KINGRA



DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY DELHI

MARCH 2023

© Indian Institute of Technology Delhi (IITD), New Delhi, 2023

EXPLORING COMPUTING APPLICATIONS WITH NON-VOLATILE MEMORY

by

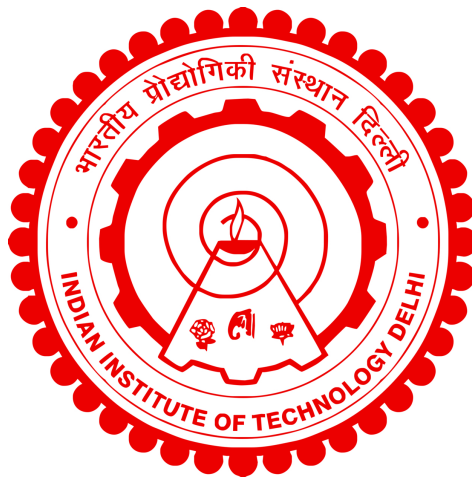
SANDEEP KAUR KINGRA

Department of Electrical Engineering

Submitted

in partial fulfillment of the requirements of the degree of Doctor of Philosophy

to the



INDIAN INSTITUTE OF TECHNOLOGY DELHI

MARCH 2023

ਜਿਉ ਅਧਿਰੈ ਦੀਪਕੁ ਬਾਲੀਐ ਤਿਉ ਗੁਰ ਗਿਆਨਿ
ਅਗਿਆਨੁ ਤਜਾਇ ॥੨॥

ਸਿਰੀਰਾਗੁ (ਮਃ੩) (੬੪) ੨:੪ - ਗੁਰੂ ਗ੍ਰੰਥ ਸਾਹਿਬ:
ਅੰਗ ੩੯ ਪੰ. ੧੦

Sri Raag Guru Amar Das

*Jio Andhhaerai Dheepak Baaleeai Thio Gur Giaan
Agiaan Thajaae ||2||*

*Like a lamp lit in the darkness, the spiritual
wisdom of the Guru dispels ignorance. ||2||*

Dedicated to the Almighty and all my Respected Teachers

*Dedicated to
my loving husband,
caring parents,
and supportive siblings*

Certificate

This is to certify that the thesis entitled “**EXPLORING COMPUTING APPLICATIONS WITH NON-VOLATILE MEMORY**”, submitted by **Sandeep Kaur Kingra** to the Indian Institute of Technology Delhi, for the award of the degree of **Doctor of Philosophy** in Department of ELECTRICAL ENGINEERING, is a record of the original, bona fide research work carried out by her under my supervision and guidance. The thesis has reached the standards fulfilling the requirements of the regulations related to the award of the degree.

The results contained in this thesis have not been submitted in part or in full to any other University or Institute for the award of any degree or diploma to the best of our knowledge.

Prof. Manan Suri

Associate Professor,

Department of Electrical Engineering,

Indian Institute of Technology Delhi.

Date:

Acknowledgements

The six years of my PhD journey was nothing less than a roller coaster ride. These years have been filled up with enriching, frustrating, fun, painful, insightful, colorful, and enlightening moments, but they are all truly memorable. As this journey is coming to its end, this is a humble attempt to thank everyone who were alongside me for their contribution and moral support.

First and foremost, I would like to express my deep and sincerest gratitude towards my advisor, **Prof. Manan Suri**, who believed in me, even when I did not. There is a long list of tasks/moments when I ended up saying that I am not capable of this but his counselling sessions and his words, “*Trust me, you can do it, just give it a try*” made everything possible. His unwavering support kept me motivated through this long and arduous journey. He generously provided me with invaluable guidance, exceptional opportunities, incredible resources, and more importantly, extraordinary freedom to carry out my research. I also thank him for teaching me how to think critically, write thoroughly, and perform impactful research. His influence in shaping me certainly goes above and beyond this thesis and extends to countless real-life lessons that I have learned from him. I thank him for all his help and support through the times where I needed it the most.

I would also like to thank the members of my student research committee (SRC), **Prof. Anuj Dhawan**, **Prof. Bhaskar Mitra** and **Prof. Preeti Ranjan Panda**. They provided me incredible technical and moral support and countless pieces of critical advice throughout my PhD journey.

I would also like to thank **Prof. Mamidala Jagadesh Kumar**, **Prof. Bodh Raj Mehta**, and **Dr. Kaushik Saha**, for teaching me valuable and relevant courses during that further strengthened the foundation for my research work. I am also grateful to our esteemed institute for providing me with the facilities to carry out my research.

My research journey would have been incomplete without the support of my international collaborators. I would like to thank **Prof. Tuo-Hung Hou** (Alex) (Distinguished Professor, National Yang Ming Chiao Tung University, Taiwan), **Dr. Boris Hudec** (currently Scientific Researcher in Slovak Academy of Sciences), **Che-Chia Chang** (PhD Student, National Yang Ming Chiao Tung University, Taiwan), **Dr. Amir Regev** (CTO Weebit Nano, Israel), **Dr. Giuseppe Piccolboni** and **Dr. Alessandro Bricalli** for the technical discussions and suggestions that helped me improve the experiments leading to some of the key results of this thesis. I am also grateful to **CEA-LETI CMP Memory Advanced Demonstrator (MAD)** service for fabricating our custom circuits designed and validated as part of this study. I am also thankful to

the team at **ASCENT PhD research accelerator programme** for selecting me for semiconductor characterization programme in 2018 at CEA-Leti, Grenoble. It gave me an exposure to state-of-the-art testing facilities that motivated me to design our own test-setups.

I would specially like to thank my friends **Vivek Parmar, Priya Vinayak, Shubham Negi** and **Salam Thoi Thoi Singh** (Research Engineer at National University of Singapore) that became my family. Apart from the brainstorming discussions regarding our research, they were always available in my thick and thin times. Our long walks around the campus discussing bottlenecks in our research, exploring Delhi markets, shopping trips and lively discussions over tea are some of the most memorable moments of my journey. I am further grateful to my juniors **Chithambara Moorthy, Sufyan Khan** (University of Wisconsin-Madison), **Tinish Bhattacharya** (PhD student at University of California Santa Barbara) and **Deepak Verma** who kept inspiring and encouraging me through their curiosity and willingness to learn and experiment.

I am thankful to all my fellow research group members from **NVM and Neuromorphic Research group** at IIT-D who made this long and at times tough journey easier for me. Specifically, I would like to mention and appreciate **Supriya Chakraborty, Manoj Kumar, Abhishek Gupta** and **Narayani Bhatia** for their co-operation and informal support. I am grateful to **Shubham Sachdeva** (CAD Engineer at Intel Corporation) who taught me layout designing, **Shubham Sahay** for fruitful discussions about nanodevices and memory technologies and **T.R. Ashish** for technical discussions related to VLSI tools and circuits. I would also like to thank **Mr. Devendra, Mr. Rakesh Kumar** (Staff, VDTT Lab) and **Ms. Usha Devi** (Staff, U.G. Electronics Lab) for simplifying our lives by helping us with the equipment. I would also like to mention the support from Electrical Engineering office staff (**Mr. Yatindra Mani Tripathi, Mr. Satish Sah**) right from the start of my journey till the very end. I would also like to express my gratitude for the technical support provided by **Raghu Sir, Abhaya Joshi, Ravichandra, Dev Prakash** and **V. Chandrashekar**.

I will never forget the support of my friends and roommates (**Sakshi, Ruchi, Vasudha, Neha Priyadarshini**) during my stay in the hostel. I would like to acknowledge the ultimate support for providing me a home away from home to **Nitin Bhai** and **his family**.

This journey would not have been possible without the continuing support, patience, and encouragement from my family members. I am most grateful to **my grandparents** and my parents **Lachhman Singh** and **Sukhdev Kaur** for enabling me to pursue my passion for learning, and for being by my side at each and every step of this journey. I thank my elder sister **Soni Di, Rajbeer jiju**, beloved niece (**Saanjh**) and nephew (**Arunveer**) for their immense love and

support which has been a sublime source of motivation for me. I also want to thank my younger brother **Garry** for always believing in me. I am also grateful to my in-laws (**Mummy Ji, Gugu Veerji, Jassu Bhabhi** and **Mehreen**) for understanding the trials and tribulations of the PhD and being patient with my journey.

I would like to give my heartfelt gratitude to my loving husband, **Jasjeet Singh**, for being my number one supporter, and my safe place. This journey would have been impossible without him by my side. His unconditional love and continued support have nourished and sustained me through the toughest moments of this journey.

And last, but not the least, I am profoundly beholden to “**ਵਾਹਿਗੁਰੂ**” for his blessings to help me raise my academic level to this stage.

Sandeep Kaur Kingra

Abstract

In conventional computing, intensive workloads dissipate most resources (time and energy) on data shuttling between isolated processing- and storage- blocks, leading to fundamental limitations such as the “Memory Wall”. Promising solutions to eliminate these bottlenecks utilize concepts such as “In-Memory Computing/Near Memory Computing” (IMC/NMC), where computations are performed either in-situ or near the actual storage.

In this thesis, we present an exhaustive study of new IMC techniques and IMC based application mapping based on emerging resistive non-volatile memory (NVM) devices. We propose a novel “Simultaneous Logic in-Memory” (SLIM) methodology wherein the bitcells are capable of implementing both Memory and Logic operations simultaneously in space (silicon) and time. We demonstrate novel non-stateful SLIM bitcells (1T-1R/2T-1R) and propose a detailed programming scheme, array level implementation, and controller architecture. To study the impact of the proposed SLIM approach for real-world implementations, we performed analysis for three applications: (i) Sobel Edge Detection, (ii) Binary Neural Networks-Multi layer Perceptron (BNN-MLP), and (iii) Keccak-f hash function. For edge-AI applications, we propose an IMC based low-precision deep neural network (DNN) implementation that utilizes oxide-based random access memory (OxRAM) devices.

We experimentally demonstrate a dual-configuration stateful 2T-2R XNOR IMC bitcell using fabricated 1T-1R OxRAM arrays and analyze the trade-off in terms of circuit overhead, energy, and latency. Additionally, using the proposed 2T-2R bitcell, we present a fully-binarized XOR based In-Memory Similarity Search (IMSS) operation. It enables simultaneous match operation across multiple stored data vectors by performing analog column-wise XOR operation and summation to compute Hamming Distance (HD). We also propose an efficient hardware mapping methodology for vector matrix multiplication (VMM) using analog OxRAM based IMC technique. For implementing Quantized Neural Networks (QNNs), two key building blocks (CMOS based neurons and OxRAM-synaptic blocks) are experimentally demonstrated. Our evaluations indicate that emerging NVM based IMC architectures can attain significant improvement in system-level performance compared to conventional von Neumann computing systems.

सार

पारंपरिक कंप्यूटिंग में, गहन कार्यभार विभिन्न 'कंप्यूट' और 'स्टोरेज' ब्लॉक के बीच डेटा स्थानांतरण करने पर अत्यधिक संसाधनों (समय और ऊर्जा) का उपयोग करता है, जिससे 'मेमोरी वॉल' जैसी मूलभूत समस्याएँ उत्पन्न होती हैं। इन बाधाओं को दूर करने के लिए नए प्रकार के समाधान जैसे 'इन-मेमोरी कंप्यूटिंग/नियर मेमोरी कंप्यूटिंग' (IMC/NMC) का प्रस्ताव कई शोधकर्ताओं ने किया है। इन समाधानों में गणना या तो विशेष रूप से स्टोरेज के भीतर या स्टोरेज के पास की जाती है। इस शोध में आधुनिक नॉन-वोलाटाइल मेमोरी (NVM) उपकरणों पर आधारित नई IMC तकनीकों और IMC आधारित अनुप्रयोग-मैपिंग का एक विस्तृत अध्ययन प्रस्तुत किया है। हम एक सृजनात्मक 'सायमलटेनियस लॉजिक इन-मेमोरी' (SLIM) तकनीक का प्रस्ताव करते हैं, जिसमें बिटसेल 'मेमोरी' और 'लॉजिक' दोनों ऑपरेशनों को एक साथ स्थानिक (सिलिकॉन) और समय में लागू करने में सक्षम हैं। सृजनात्मक 'नॉन-स्टेटफुल IMC बिटसेल' (1T-1R/2T-1R) को प्रदर्शित करने के साथ एक विस्तृत प्रोग्रामिंग योजना, सरणी स्तर कार्यान्वयन और नियंत्रक आर्किटेक्चर का भी प्रस्ताव किया गया है। SLIM तकनीक के प्रभाव का अध्ययन करने के लिए, तीन अनुप्रयोगों के लिए विश्लेषण किया गया: (1) 'सोबेल एज' डिटेक्शन, (2) 'बाइनरी न्यूरल नेटवर्क्स-मल्टी लेयर परसेप्ट्रॉन' (BNN-MLP), और (3) केचक (Keccak) -एफ हैश फंक्शन। एज-कृत्रिम बुद्धिमत्ता (Artificial Intelligence) अनुप्रयोगों के लिए, एक IMC आधारित लो प्रिसिशन 'डीप न्यूरल नेटवर्क' (DNN) कार्यान्वयन का प्रस्ताव किया गया है जो 'ऑक्साइड आधारित रैंडम एक्सेस मेमोरी' (OxRAM) उपकरणों का उपयोग करता है। 1T-1R ऑक्सराम उपकरणों का उपयोग करके प्रयोगात्मक रूप से 2T-2R XNOR IMC बिटसेल को प्रदर्शित किया गया है, और साथ ही सर्किट-व्यय, ऊर्जा और विलंबता के संदर्भ में ट्रेड-ऑफ का विश्लेषण भी किया गया है। इसके अतिरिक्त, प्रस्तावित 2T-2R बिटसेल का उपयोग करते हुए, पूरी तरह से बिनरीकृत XOR आधारित इन-मेमोरी समानता खोज (IMSS) ऑपरेशन को प्रस्तुत किया गया है। यह हैमिंग दूरी (HD) की गणना करने के लिए एनालॉग XOR ऑपरेशन का सक्षम प्रयोग करता है। इस शोध में एनालॉग OxRAM आधारित IMC तकनीक का उपयोग करके वेक्टर मैट्रिक्स गुणन (VMM) के लिए एक कुशल हार्डवेयर मैपिंग पद्धति का भी प्रस्ताव करते हैं। क्वांटाइज्ड न्यूरल नेटवर्क (QNN) को लागू करने के लिए, दो प्रमुख मूलभूत आधार (CMOS आधारित न्यूरॉन्स और OxRAM-सिनैप्टिक ब्लॉक) प्रयोगात्मक रूप से प्रदर्शित किए जाते हैं। इस शोध के मूल्यांकन और अध्ययन से संकेत मिलते हैं कि अत्याधुनिक NVM आधारित IMC आर्किटेक्चर पारंपरिक वॉन न्यूमैन कंप्यूटिंग सिस्टम की तुलना में महत्वपूर्ण सुधार प्राप्त कर सकते हैं।

Contents

Acknowledgements	i
Abstract	iv
Contents	vi
List of Figures	ix
List of Tables	xviii
Abbreviations	xx
1 Introduction and Motivation	1
1.1 Compute-Memory Bottleneck	1
1.2 IMC Status and Outlook	4
1.3 Thesis Organization and Contributions	6
1.3.1 Objective of Thesis	6
1.3.2 Key Contribution of Thesis	6
1.3.3 Thesis Organization	6
2 Background Basics and Concepts	8
2.1 Oxide-based Memory (OxRAM)	8
2.1.1 Filamentary OxRAM	9
2.1.2 Non-Filamentary/Interfacial OxRAM	9
2.2 Basics of von Neumann Computation	11
2.3 Memory-Centric Computing	12
2.3.1 Near-Memory Computing (NMC)	13
2.3.2 In-Memory Computing (IMC)	13
2.4 IMC Applications Relevant to this Thesis	18
2.4.1 Deep Neural Networks (DNNs)	18
2.4.2 Associative Memory for Similarity Search	21
2.4.3 Cybersecurity: Hash Primitives	22
3 SLIM exploiting Bilayer Analog OxRAM Devices	24
3.1 Background	25
3.2 Experimental Results: SLIM Bitcell	28
3.2.1 Interfacial OxRAM Device Fabrication and Characterization Test Setup	28
3.2.2 SLIM Bitcell	29

3.2.3	Concept of SLIM	31
3.2.4	Experimental Validation: Memory-Write Operation	33
3.2.5	Experimental Validation: Logic Operation	35
3.3	Array-Level Implementation of SLIM	38
3.4	SLIM Application Analysis	41
3.5	Impact of Device Characteristics on the Performance of SLIM Bitcell	46
3.6	Summary	48
4	IMC based Mapping of Keccak-f	49
4.1	IMC for Hashing Algorithm	49
4.2	Basics of Keccak-F	51
4.3	SLIM Operation Mapping	53
4.3.1	Proposed Keccak-f Mapping	53
4.4	Performance Benchmarking for Proposed SLIM-based Keccak-f	57
4.4.1	SLIM Implementation and NVM Device Reliability	59
4.4.2	Comparison with other LIM Methods	60
4.5	Summary	60
5	Dual-Configuration IMC Bitcells for BNNs	62
5.1	Limitations of Edge-AI hardware	62
5.2	XNOR-Net based Binary Neural Networks	64
5.3	Fabricated Filamentary SiO _x OxRAM Array	66
5.4	IMC XNOR bitcell configurations	68
5.4.1	XNOR _{row} implementation:	69
5.4.2	XNOR _{col} implementation	70
5.5	Experimental Demonstration of OxRAM XNOR IMC	72
5.6	Learning Performance and Energy Estimation	75
5.6.1	Network Training Methodology	76
5.6.2	IMC Energy Estimation	77
5.6.3	Performance Trends: Impact of MAT size	78
5.7	Performance comparison of XNOR _{row} and XNOR _{col}	78
5.7.1	BER and OxRAM Variability Analysis	79
5.8	Benchmarking with Literature NVM based XNOR IMC Implementations	81
5.9	Summary	81
6	Fully-Binarized, Parallel, IMC Primitive for Similarity Search	82
6.1	Introduction	82
6.2	Fabricated OxRAM Array	83
6.3	OxRAM IMSS Experiments and Working	84
6.3.1	SPICE Simulations	86
6.4	HSI Classification Task	89
6.5	Summary	92
7	VMM with Binary OxRAM Arrays	93
7.1	Introduction	93
7.2	Basics and Background	95
7.2.1	Vector Matrix Multiplication (VMM) in Hardware	95
7.2.2	ADALINE	96

7.3	Proposed Methodology for VMM-based BNN	96
7.3.1	Training for Binarized-ADALINE	96
7.3.2	Weight Mapping Strategy	97
7.4	Experimental Results	98
7.4.1	Testbench & Dataset	98
7.4.2	Fabricated OxRAM Crossbar Chip	99
7.4.3	BNN Results on OxRAM Crossbar	100
7.5	Summary	101
8	Time-Multiplexed IMC Scheme for Mapping QNNs	102
8.1	Quantized Neural Networks	102
8.2	Prior Art	104
8.3	Fabricated CMOS-OxRAM circuits	106
8.4	VMM operation mapping	108
8.5	Network Simulations	110
8.6	Summary	115
9	Conclusion and Future Work	116
9.1	Scope for Future Work	118
A	CV Characterization of Ni/HfO₂/ATO/TiN Bilayer OxRAM Device	120
	List of Publications	153
	References	153
	Curriculum Vitae	155

List of Figures

1.1	(a) Sources of computing performance have been challenged by the end of Dennard’s scaling in 2004 [1]. (b) Processor-memory performance gap grows at a rate of 50%/year [2]. (c) Trends in training compute of n=121 milestone ML systems between 1952 and 2022. [Source: OpenAI, The Economist]	2
1.2	(a) Energy costs for various operations in 45 nm 0.9 V highlighting processor-memory bottleneck [3]. (b) GPU profiling results for some common Image Processing Workloads [4].	3
1.3	(a) IMC target position on power for performance [5]. (b) Exponential increase in published manuscripts exploring IMC techniques. Source: Scopus	3
1.4	LPDDR5-PIM performance and energy [5].	4
1.5	Object Lookup performance using RRAM Crossbar-based IMC benchmarked against von Neumann architecture [6].	5
1.6	Potential eFlash/SRAM market cannibalization by emerging NVM[7].	5
2.1	(a) TEM (Transmission Electron Microscopy) cross-section of the advanced OxRAM demonstrator (MAD300) chip with 28 nm Fully Depleted Silicon-On-Insulator (FDSOI) transistor. OxRAM devices are integrated between M5 and M6 [8]. (b) EDX (Energy Dispersive X-Ray Analysis) cross-section close-up of the OxRAM device [8]. (c) DC I-V characteristics and potential device conduction mechanism of filamentary OxRAM device (N ⁺ -Si/SiO _x /P ⁺ -Si) [9]. Note the formation/dissolution of CF.	10
2.2	Schematic diagrams showing O ²⁻ migration in the V _O -rich (oxygen-deficient) layer near the Ta TE during bi-polar (a) SET and (b) RESET operations [10]. (c) Real-time TEM images (from 1 to 6) during analog resistance changes in TiN/PCMO/Pt junction device. The switching sequence and direction are numbered and indicated by arrows. TEM images showing the intermediate reaction layer at the TiN/PCMO interface after switching to the LRS. As schematically illustrated next to the TEM images, the thickness of the reaction layer increases as the device is switched from the LRS to the HRS [11].	11
2.3	(a) Schematic of the traditional von Neumann computer architecture. Here <i>A</i> denotes information stored in a memory location. To perform a computational operation, <i>f(A)</i> , and to store the result in the same memory location, data is shuttled back and forth between the memory and the computing unit [12]. (b) The energy efficiency of compute operations continues to improve at a slow pace, but the energy needed in terms of picojoules per bit per millimeter of data movement is not improving with each technology generation. The result is that data movement now consumes more on-chip energy than the compute operations performed on the data after it is moved [13].	12
2.4	Three conceptual approaches to computing: (a) conventional digital computing, (b) NMC, and (c) IMC [14].	13

2.5	Schematic of variants of SRAM bitcells that are proposed for IMC [15].	14
2.6	(a) Bit-wise AND/NOR logical operations using an SRAM array [16, 17]. BL and BLB are pre-charged to the supply voltage (V_{DD}), prior to the execution of the operation. When the two activated SRAM cells in a column are both 1 (0), V_{BL} (V_{BLB}) will be comparable to V_{DD} , whereas for the other bit combinations, both V_{BL} and V_{BLB} will be lower than V_{DD} . Hence, by sensing V_{BL} and V_{BLB} with a SA, AND and NOR operations are performed, respectively. (b) Schematic illustration of bit-wise AND/OR logical operations performed using three DRAM bitcells [18]. The image is adapted from [19].	15
2.7	OxRAM based digital logic gates: (a) V–R logic gate and corresponding truth table for material implication (IMP) operation. (b) V–V logic and the corresponding input/output characteristic for AND operation. Four configurations of input values can be linearly separated (dashed line) according to the weights G_j and the comparator threshold V_{thresh} , thus yielding a reconfigurable Boolean function. (c) Parallel R–R stateful logic gate for IMP operation and corresponding truth table. The output variable Y is the final resistance state of the OxRAM that changes state. (d) Serial R–R stateful logic gate for OR operation [20].	16
2.8	(a) Block diagram of the basic MAC unit, (b) Memory read/write for each MAC unit. (c) One-step VMM using memory arrays. (d) Schematic of convolution operation in an image. (e) Typical mapping method of 2D convolution to memory arrays. The image is adapted from [21].	18
2.9	IMC applications are grouped into three main categories based on the overall degree of computational precision (Low degree of precision, Computational precision, High degree of precision) that is required [19]	19
2.10	(a) Two efficient variations of L-layer CNN architectures: Binary-Nets (the weight filters contains binary values), and XNOR-Nets (both weigh and input have binary values). (b) This figure illustrates the procedure for approximating a convolution operation using binary operations. Note: I, W are a set of tensors, where each element $\mathbf{I} = I_{l(l=1, \dots, L)}$ is the input tensor for the l^{th} layer, $\mathbf{W} = W_{lk(k=1, \dots, K^l)}$ is the k^{th} weight filter in the l^{th} layer. K^l is the number of weight filters in the l^{th} layer of the CNN. * represents a convolutional operation with I and W as its operands	20
2.11	(a) Traditional RAM design: Input is the Address and the Output is the contents at this address. (b) CAM and TCAM: Input is the Search word (content) and the output is the address of a match. (c) Conventional 10T BCAM structure. (d) Conventional 16T TCAM structure [16].	21
2.12	(a) Overview of SHA-3 operation. (b) Representation of 1600 bit hash state. (c) Depiction of the first 4 steps of Keccak-f. This figure is adapted from [22]	22
3.1	(a) von Neumann architecture with separate processing and memory units, LIM: memory blocks can perform logic, and SLIM: all memory blocks capable to perform Storage and Logic operations simultaneously and non-destructively, (b) Ideal computing system with LIM/SLIM co-existing with von Neumann CPU, enabling computation at all levels of memory hierarchy.	25
3.2	(a, b) HR-TEM and DC IV curve of bilayer Ni/HfO ₂ /ATO/TiN OxRAM device, fabricated for this chapter (inset shows quite stable repeatable switching for 30 cycles).	28
3.3	Experimental setup used for SLIM characterization. It shows integrated 2T-1R/1T-1R SLIM bitcell, CMOS chip, OxRAM chip and parameter analyzer used for measurements.	29

- 3.4 Repeatable analog conductance tuning characteristics observed in the OxRAM device using identical SET and RESET pulses trains for (a) $V_{SET} = 3$ V (10 ms) , $V_{RESET} = -5.5$ V (10 ms) and $V_{READ} = -0.4$ V. Effect of (b) D2D variability (from 10 devices), $\sigma_{max} = 28$ nS (at mean= 57.3 nS, $C_V = 0.49$), (c) C2C variability (for 30 cycles) , $\sigma_{max} = 7.35$ nS (at mean= 79.6 nS, $C_V = 0.09$) has been observed. Similar characteristics for (d) $V_{SET} = 3$ V (1 ms), $V_{RESET} = -3$ V (5 ms) and $V_{READ} = -1.5$ V, along with (e) D2D variability (from 16 devices), $\sigma_{max} = 0.86$ μ S (at mean= 9.49 μ S, $C_V = 0.09$) and (f) C2C variability (from 30 cycles), $\sigma_{max} = 0.031$ μ S (at mean= 11.2 μ S, $C_V = 2.77e-3$) has been observed experimentally. 30
- 3.5 Experimentally measured NMOS characteristics: I_D - V_{DS} plot, I_D - V_{GS} plot and relationship between NMOS transistor enforced compliance current (Y1 axis) and ON resistance (Y2 axis) with gate voltage ($V_{DS} = 3$ V). 30
- 3.6 Proposed circuit schematic of (a) 1T-1R, (b) 2T-1R SLIM bitcell. 31
- 3.7 (a) SLIM state assignments for Logic and Memory operation. Note Memory LRS and HRS sense regions. Histograms show resistance distributions for 4 selected SLIM states on 1T-1R/2T-1R SLIM bitcell (>100 trials). (b) Resistance distribution for 4 selected SLIM states. (c) Endurance for states: ‘11’, ‘10’, ‘01’, ‘00’ for 200 cycles. (d) Proposed SLIM programming signals and Memory-Logic state transitions for 1T-1R/2T-1R SLIM bitcell. All further SLIM operations use P1/P2/P3 pulses at V_1/V_2 32
- 3.8 (a) Applied example signals for Memory Write ‘1’ and Memory Write ‘0’. Read conditions are specified in brackets.[All pulse duration = 7 ms] (b) Experimental measurements for Memory Write ‘1’ operation (program device to state ‘11’) with device’s initial state as: (i) ‘10’, (ii) ‘01’, and (iii) ‘00’. Memory Write ‘0’ operation (program device to state ‘01’) with device’s initial state as (iv) ‘11’, (v) ‘10’ and (vi) ‘00’. Blue line: transient current through OxRAM device. Black line: P1/P2/P3 (applied signals). Black square: Initial resistance state, Red circle: Final resistance state post SLIM operation. Please note in (iv, v), the transient current through OxRAM device falls due to gradual increase in non-volatile resistance with application of successive reset pulses. **Note:** *The current scale in (i-vi) is varied for clear demonstration of gradual switching in the OxRAM device.* 34
- 3.9 Four possible input operand combinations: (a) a = b = ‘0’; (b) a = ‘0’, b = ‘1’; (c) a = ‘1’, b = ‘0’; (d) a = b = ‘1’; corresponding to NAND truth table and proposed signal mapping for each case for the 1T-1R SLIM bitcell. [$V_{TB} = V_{TE} - V_{BE}$; operand a is mapped to $V_G = 10$ V (7 ms long); operand b is mapped to $V_2 = P3 = 5.5$ V (7 ms long)]. Experimental results for NAND logic implemented using 1T-1R SLIM bitcell with device’s initial state: ‘11’ (e-h), and ‘01’ (i-l). Among the four operand combinations, OxRAM device switches to Logic HRS state (‘10’ or ‘00’) only for a=b= ‘1’. Blue: transient current through OxRAM device. 36

- 3.10 Four possible input operand combinations: (a) $a = b = '0'$; (b) $a = '0'$, $b = '1'$; (c) $a = '1'$, $b = '0'$; (d) $a = b = '1'$; corresponding to NOR truth table and proposed signal mapping for each case for the 2T-1R SLIM bitcell. [$V_{TB} = V_{TE} - V_{BE}$; $V_G = 10$ V (7 ms long); $V_2 = P3 = 5.5$ V (7 ms long)]. Experimental results for NOR logic implemented using 2T-1R SLIM bitcell with device's initial state: '11' (e-h), and '01' (i-l). Among the four operand combinations, OxRAM device switches to Logic HRS state ('10' or '00') for $a = '0'$, $b = '1'$; $a = '1'$, $b = '0'$ and $a = b = '1'$. Blue: transient current through OxRAM device. Black line: P3 in all cases (applied signal). 37
- 3.11 (a) Flowcharts for SLIM: Memory Write operation and Logic operations. Intelligent read is performed in both operations. Refresh scheme is an internal part of Logic operation. (b) Optimized Refresh Scheme corresponding to multiple SLIM MATs (Matrices). Each SLIM MAT has 8×8 bits. 1 Tag register is allocated to each SLIM MAT to track row status. Within each Tag register, 1-bit corresponds to 1 row of 8×8 SLIM MAT. Tag Byte is initialized to zero once a fresh SLIM MAT is used. Once a row is used for Logic operation, tag bit corresponding to it, is set high. When the tag byte contain all '1's, the Refresh block is triggered and it sends instruction to refresh the contents of complete SLIM MAT. After Refresh operation, all the SLIM bitcells in given SLIM MAT will have absolute Memory states ('11'/'01'). 39
- 3.12 (a) Block diagram of SLIM processing unit. Refresh block forms an internal part of the Logic operation block. User operands are passed first to the SLIM control unit. The control unit with other blocks maps Logic operations on the 1T-1R array. (b,c) Proposed sensing mechanism used for reading the SLIM bitcell state. 40
- 3.13 1-bit Full Adder operation mapping on 1T-1R SLIM bitcell array (NAND logic) using SLIM Operation Compiler. Here A_n , B_n indicate inputs, E_n indicates output of Logic operation. Black circles indicate input/output for the adder. . . 41
- 3.14 (a) EDP comparison for performing 64-bit Logic operations using SLIM array w.r.t. conventional CPU architecture (fetching operands from DRAM (DDR3)), (b) Edge-detection output of 'CPU+DRAM' (center) and 'CPU+ SLIM bitcell array' (right) along with the original image (left). Image sources: (b) Original image shown on left is a resized and cropped version of "BW Rubik's Cube" by Gerwin Sturm, licensed under CC BY-SA 2.0. Images in center and right are generated by performing edge detection). 42
- 3.15 (a) Network topology with 784 input neurons, 100 hidden neurons and 10 output neurons, (b) MNIST dataset used for training the network in our chapter, and (c) Hierarchical *POPCOUNT* tree for neuron accumulation realized using RCA of increasing bit width. 44
- 3.16 (a) Mapping of operations (XNOR and *POPCOUNT*) over SLIM MATs with addition SRAM based *POPCOUNT* LUT modules. All Logic operations are mapped on independent bitcells, considering no bitcell is reused during one inference cycle. Mapping ensures proximity for bitcells performing the same type of operation in the same layer. (b) Flowchart of SLIM-BNN computation. Since binary weight used for XNOR computation are $[0,1]$ rather than $[-1,1]$, a fixed offset has been subtracted in order to compute the same dot product. 45
- 3.17 Endurance requirement analysis for performing a single BNN inference operation with varying SLIM array sizes. Significant decrease in OxRAM maximum write-hits per device is observed with increase in the array size. 47

4.1	CSH performed using (a) Conventional computing architecture, (b) Proposed SLIM based architecture.	50
4.2	Flowchart describing steps followed for workload mapping on a SLIM MAT and energy/latency estimation. Here pipeline stages refer to independent large functions extracted by dividing the workload into independent steps without external data dependency.	52
4.3	Illustration of XOR/NOT operation mapping on SLIM MAT with each SLIM bitcell performing NAND logic depending on the inputs. XOR operation takes three cycles of SLIM operations. Intermediate Logic outputs are read and stored in buffer memory (see Figure 4.1(a)) for further signal application. AND operation can be realized by using NAND gate followed by a NOT gate and thus requiring 2 SLIM bitcells.	54
4.4	(a) Variable mapping for Keccak-f function realized on SLIM MAT of size 64×64 . SLIM based Logic gate realization estimates of: (b) Energy, and (c) Latency. (d) Cycle-wise mapping of operations on SLIM MAT for a single round of Keccak-f function. Each time unit represents computation time for a XOR ($3 \times$ SLIM) operation.	55
4.5	Energy contribution breakdown based on operation type when mapping Keccak-f computation on SLIM bitcells using device stacks: (a) CBRAM (Ag/MoS ₂ /Ag), (b) OxRAM (Pt/HfAlO _x /TaN), (c) PCM (Doped-GST), (d) FeRAM (Pt/BTO/S-NTO).	58
5.1	Example of Binarized layer implementations with steps for (a) Binarized Convolutional, and (b) Binary fully connected layers.	64
5.2	Convolution building blocks used in (a) FracBNN [23] and (b) MobileNet-v1 [24] architecture. All parameters i.e. weights, inputs and outputs in the convolution layers are binarized. Computation blocks implemented on fabricated IMC cells are highlighted in blue. (c) Comparison of activation functions used for networks in the study. (d) Thermometric encoding [25] used for binarizing inputs compared to conventional fixed point representation.	65
5.3	(a) SEM cross-section of the SiO _x OxRAM cell integrated on top of the 130 nm CMOS, (b) IV characteristics showing electro-forming, SET and RESET operation highlighting D2D variability (20 devices), (c) C2C variability during SET/RESET distribution over 10 cycles.	66
5.4	(a) 10^6 endurance switching cycles. It is clearly visible that the resistive states are well separable for 10^6 cycles. (b) Statistical resistance state distribution for LRS and HRS. LRS ranges from 3 k Ω to 20 k Ω and HRS ranges from 60 k Ω to 1 M Ω	67
5.5	Schematic representation of the 8×8 OxRAM array.	68
5.6	(a) BNN computation mapping on XNOR _{row} bitcell and its corresponding <i>POPCOUNT</i> implementation. Output current from the 2T-2R bitcell is converted to voltage using a CSA followed by summation in the <i>POPCOUNT</i> block. The <i>POPCOUNT</i> is then compared to a pre-fixed threshold to obtain the output of VVM. (b) BNN computation mapping using XNOR _{col} configuration and <i>POPCOUNT</i> is implemented inherently over fabricated OxRAM array.	69

5.7	Schematic representations of input activations and weights with computation equations for: (a) XNOR_{row} (4×2 2T-2R bitcell array), (b) XNOR_{col} (2×4 2T-2R bitcell array) IMC implementations on a 4×4 1T1R array. (c) CSA block used for the study. Binary activations are mapped onto the differential SLs (in case of XNOR_{row}) and WLs (in case of XNOR_{col}). Binary weights are mapped onto the HRS/LRS values of XNOR-OxRAM cells.	71
5.8	Our custom designed experimental setup for XNOR IMC validation. Switch board helps to intelligently choose between multiple inputs. An interface board is used for routing control signals (from micro controller). Chip interface board and Breakout board are used for accessing OxRAM array test chip.	73
5.9	Schematic representation/operand mapping corresponding to possible combinations of input activations ('-1', '+1') and weights ('-1', '+1') are shown for (a,b) XNOR_{row} , and (d,e) XNOR_{col} . (c,f) Experimentally characterized bitcell output current of four possible operand combinations for XNOR_{row} and XNOR_{col} respectively.	74
5.10	Statistical distribution for VMM output variability for (a) XNOR_{row} and (b) XNOR_{col} configuration. The simulations are performed on 8×8 1T-1R array with all input combinations applied for ≥ 1000 trials. Energy trade-off analysis based on MAT sizes for CIFAR-10 workload in terms of XNOR operation energy and total energy (XNOR operations + CMOS periphery) for (c) XNOR_{row} , and (d) XNOR_{col}	76
5.11	Sample images from: (a) VWW dataset, and (b) CIFAR-10 dataset. (c,d) Training weight map (floating-point precision) from an intermediate layer for VWW and CIFAR-10 datasets respectively. (e,f) Inference weight map (binary precision) from an intermediate layer for VWW and CIFAR-10 datasets respectively.	77
5.12	Simulated performance trends of XNOR_{col} IMC bitcell based BNN implementation with varying MAT Sizes for: (a) VWW dataset, and (b) CIFAR-10 dataset.	78
5.13	(a) Impact of BER on BNN accuracy for VWW and CIFAR-10 workloads. For XNOR_{row} bitcell, the impact of $I_{sense,th}$ on (b) BER (for 1 million instances), and (c) BNN accuracy for VWW and CIFAR-10 workloads. (d) Impact of MW on BER and Inference accuracy (for VWW and CIFAR-10 workloads). All BNN accuracy simulations have been averaged over 10 trials and exhibits negligible variability ($\approx 1\%$).	79
6.1	(a) I-V characteristics showing electro-forming [Inset: SEM cross-section of the SiO_x OxRAM cell integrated on top of the 130 nm CMOS], (b) Statistical resistance state distribution for pre-forming and post-forming resistance state (64 devices), (c) I-V characteristics showing SET and RESET operation highlighting D2D variability, (d) Statistical resistance state distribution for LRS and HRS (64 devices).	84
6.2	Block diagram of the proposed binarized IMSS engine with periphery blocks showing mapping of inputs for 2T-2R XOR bitcell. WL decoder maps QI vector to differential encoding. Current integrated along the bitlines ($I_{BL,n}$) is translated to voltage using SA to compute HD between applied QI vector and corresponding SD vector.	85
6.3	(a) Truth table summarizing 2T-2R XOR bitcell operation, (b) Custom designed PCB for XOR IMC validation, (c) Experimental validation for all possible input combinations of 2T-2R XOR bitcell.	87

6.4	(a) Schematic of the simulated SA circuit using two-stage amplification. (b) SA transfer characteristic based on 130 nm Skywater PDK technology used for converting column-wise integrated current to voltage. (c) Simulations based SA outputs for different HD values using 4×8 2T-2R array. (d) Timing waveforms for QI vector (applied at Ws) and V_{out} for different BLs (storing binarized feature vectors) demonstrates successful match operation when QI vector matches with SD vector (extracted from SPICE simulations using 130 nm Skywater PDK).	88
6.5	HSI classification results using proposed IMSS engine. (a) RGB data, (b) Ground truth of classification, (c) Prediction results using proposed IMSS engine. Decision space representation based on t-SNE for: (d) Ground truth and (e) Predicted results.	90
7.1	(a) Basic VMM operation implemented using two-terminal resistive nanodevice crossbar. (b) Structure of generic ADALINE Network.	94
7.2	Proposed scheme for computation and weight mapping using two-terminal resistive crossbar. Every logical weight value (i.e. '+1' or '-1') is mapped on the crossbar using 2 paired devices from consecutive rows (i.e. rows W^+ and W^-) of the same column. The paired devices are always programmed to complementary states (LRS-HRS or vice-versa). In particular, to realize logical weight '+1', device from the first row is programmed to LRS and the paired device in the consecutive row is programmed to HRS. For realizing logical weight '-1', programming is inverted (i.e. first row device is in HRS while consecutive row device is in LRS). Eight logical weight values ('-1', '+1', '+1', '-1', '+1', '+1', '-1', '-1') are programmed using 16 (4×4) devices. The first four weights corresponding to class $k=0$ ('-1', '+1', '+1', '-1') are partitioned in rows 1 and 2, while the next four weights corresponding to class $k=1$ ('+1', '+1', '-1', '-1') are partitioned in rows 3 and 4. Note, input voltages are applied on columns and current integration occurs across rows. This is due to the fact that DAC units used in our experimental setup generate only positive voltages. Negative voltages are effectively realized by grounding device top-electrode and applying +ve DAC signal at device bottom electrode. Programming and read paths are isolated using CMOS switches for each channel.	95
7.3	Flowchart summarizing sequence of operations to perform binarized-ADALINE computation.	97
7.4	Custom-testbench of proposed OxRAM based VMM. Since our DAC units can only generate +ve voltage signals, -ve voltage across the RRAM device is realized by applying a +ve voltage to the bottom terminal while grounding the top terminal of the device.	98
7.5	(a) (i) Packaged 8×8 OxRAM crossbar IC used for testing. (ii) Image of the crossbar die acquired using an optical microscope. (iii) HR-TEM image of bilayer Ni/ HfO_2 /ATO/TiN OxRAM device fabricated for this chapter. (b) Overlaid DC I-V curves of 64 OxRAM devices in 8×8 crossbar indicating low D2D variability. ($V_{set} = 3.3$ V, $V_{reset} = -5.5$ V) (c) Resistance distribution of LRS and HRS states used ($V_{read} = -0.8$ V).	99
7.6	(a) Implemented 30×2 ADALINE network. (b) Proposed PWM based input encoding scheme for mapping analog inputs to crossbar columns. (c) Mapping strategy used in this chapter for realizing 30×2 ADALINE network on a 8×8 crossbar. Resistance state and distribution of 8×8 OxRAM crossbar before programming (d, e) and after programming (f, g). All resistance values are in $M\Omega$	100

8.1	(a) TEM cross section of the integrated $TiN/HfO_2/Ti/TiN$ OxRAM [26]. (b) 1T-1R bitcell with OxRAM device stack. (c) Test setup used in this chapter. (d) Layout of the fabricated 1T-1R bitcell using 130 nm CMOS technology.	103
8.2	Comparison of responses of activation functions used in this chapter.	104
8.3	Measured NMOS selector device characteristics: (a) I_D vs V_{DS} , and (b) I_D vs V_{GS}	105
8.4	(a) DC I-V characteristics of 1T-1R device highlighting Forming, SET and RESET switching action. (b) Device switching action observed during pulse test. $V_{read} = 0.2$ V, $V_{gate} = 1.2$ V, is used to read the OxRAM resistance state.	105
8.5	MLC characteristics exhibited by OxRAM device by varying: (a) V_{stop} during RESET programming and gaussian distribution of 3 resistive states selected in the HRS region (with C2C/D2D distribution) for realizing BNNs/TNNs in (b). (c) V_{gate} to tune I_{comp} during SET programming and gaussian distribution of 3 resistive states (with C2C/D2D distribution) selected from possible states shown in (c) for realizing BNNs/TNNs in (d).	107
8.6	(a) Circuit schematic of 6T CMOS sigmoid neuron. (b) Measured sigmoid transfer curve validating circuit functionality along with its layout (inset).	108
8.7	Vector multiplication for BNN/TNN using MLC OxRAM states and 2-step READ operations for applied input vector = ['-1', '0', '0', '+1']. Note sigmoid neuron based activation is used for class detection.	109
8.8	Simulated S/H-based time-multiplexed differentiation circuit for validation of proposed methodology. The circuit is simulated with resistance values based on experimental characterized 1T-1R data along with 130 nm CMOS technology.	110
8.9	Input and output waveforms based on SPICE simulations of time-multiplexed differential circuit validating proposed methodology. V_{out} is high in case output of S/H_2 exceeds output of S/H_1 . V_{in} and V_{clk} are pulses of amplitude 1.4V. Since clock signals are applied at the gate terminal of PMOS transistor, output is sampled on the input lines when V_{clk} is low. Similarly, final output voltage V_{out} is sampled when $V_{clk,diff}$ is low.	111
8.10	Simulated VMM output voltage for 4×4 matrix of 1T-1R devices using MLC states in HRS region for BNN and TNN in (a),(b) respectively; similarly using MLC states in LRS region, BNN and TNN Sense _{out} voltages in (c),(d) respectively. Note that x-axis has POPCOUNT values possible for BNN/TNN networks.	112
8.11	(a) Sample image from FMNIST dataset with binary images resulting from thermometric encoding (channels=8, resolution=32). (b) LeNet-based CNN [27] architecture used in this chapter. (c) Categorical cross-entropy loss evolution observed during training of BNN/TNN with QKeras [28] framework.	113
8.12	Weight distributions extracted using 3 resistive states from HRS region (as OxRAM device conductance) for simulated BNN and TNN in (a,b) respectively. Similarly, weight distributions extracted using 3 resistive states from LRS region (as OxRAM device conductance) for simulated BNN and TNN in (c,d) respectively.	113
8.13	Confusion matrices observed for simulated neural networks including OxRAM device conductance values for BNN/TNN weight realization using selected memory states from HRS region (a,b respectively) and LRS region (c,d respectively).	114
A.1	CV of (a) Pt/ATO/TiN OxRAM device, (b) Ta/HfO ₂ /TiN OxRAM device. Note that the HfO ₂ has less defects, therefore low f dispersion.	120
A.2	For Bilayer OxRAM devices, one oxide layer (HfO ₂ in our stack) is considered to have fixed R,C values and other oxide layer (ATO in our stack) has variable R,C values depending upon the amount of charge trapped [29].	121

A.3	(a) DC IV characteristics for Bilayer OxRAM device and its corresponding (b) CV measurements at different frequencies.	121
A.4	(a) C_{ATO} calculated from measured CV using two capacitors in series ($C_{HfO_2} = 590$ pF) (b) K-value of ATO calculated from C_{ATO}	122

List of Tables

1.1	Example of two generations of NVIDIA GPUs [30] to illustrate compute-memory gap.	3
3.1	Salient Features of different LIM techniques proposed in literature.	26
3.2	Two-input NAND logic Gate truth-table using our 1T-1R SLIM methodology	35
3.3	Two-input NOR logic Gate truth-table using our 2T-1R SLIM methodology	38
3.4	Program scheme for realizing Boolean operations using a SLIM bitcell.	41
3.5	SLIM bitcell count, Normalized Energy/Operation, Normalized Latency count for different operations using SLIM logic gate. (All values normalized w.r.t. SLIM 1T-1R NAND and provide worst-case estimates i.e. maximum device switching.)	42
3.6	Device parameters used for 1T-1R/2T-1R SLIM bitcell array based application analysis [31, 32].	43
3.7	Performance results for application benchmarks using conventional and SLIM based system configuration	46
4.1	Operation mapping of Keccak-f steps on SLIM MATs	53
4.2	Performance benchmarking of SLIM based Keccak-f for a single round using various device technologies from literature.	57
4.3	NVM Device Technology comparison in terms reliability parameters.	59
4.4	Comparison of proposed SLIM-based SHA-3 implementation with other works in literature based on NVM-based LIM	60
5.1	Truth table showcasing XNOR operation realized using IMC bitcell.	73
5.2	Performance of Trained BNN implemented using XNOR IMC bitcells. Performance parameters used for the simulation are MAT size: 256×256 ; T_{read} : $10 \mu s$; V_{read} : $0.2V$	75
5.3	Performance benchmarking with XNOR-based hardware architectures in literature on CIFAR-10 workload.	80
6.1	Estimated Power Dissipation for each circuit component with simulation parameters.	89
6.2	Comparison of IMSS realized using structures with 2T-2R circuits.	91
7.1	Classification accuracy for VMM-based binarized-ADALINE.	101
8.1	Performance benchmarking of the proposed BNN/TNN hardware.	112

List of Algorithms

- 1 Algorithm for $n \times n$ SLIM bitcell array Refresh 40
- 2 Keccak-f Round Algorithm 51
- 3 Proposed IMSS method based on bit-wise XOR. 91
- 4 Algorithm for training binarized-ADALINE 97

Abbreviations

1R	1 Resistor
1T-1R	1 Transistor-1 Resistor
ADALINE	Adaptive Linear
AI	Artificial Intelligence
ALU	Arithmetic-Logic Unit
ANN	Artificial Neural Network
ASIC	Application Specific Integrated Circuit
ATO	Al-doped- TiO_2
AVIRIS	Airborne Visible/Infrared Imaging Spectrometer
BCAM	Binary Content-Addressable Memory
BE/TE	Bottom Electrode/Top Electrode
BEOL	Back End of Line
BER	Bit Error Rate
BL/BLB	Bitline/Bitline bar
BNN	Binary Neural Network
BRS	Bipolar Resistive Switches/Switching
C2C/D2D	Cycle-to-Cycle/Device-to-Device
CAM	Content Addressable Memory
CBRAM	Conductive Bridge Random Access Memory
CF	Conductive Filament
CMOS	Complementary Metal Oxide Semiconductor
CNN	Convolutional Neural Networks
CPU	Central Processing Unit
CRS	Complementary Resistive Switches/Switching
CSA	Current Sense Amplifier
CSH	Cryptographically Secure Hash
CV	Capacitance Voltage
C_v	Coefficient of Variance
DRAM	Dynamic Random Access Memory
DSP	Digital Signal Processor
EDP	Energy Delay Product

EDP	Energy Delay Product
FA	Full Adder
FeFET	Ferroelectric Field Effect Transistor
FeRAM	Ferroelectric Random Access Memory
FLOP	Memory bandwidth per processor floating point operation
FP32	Floating point 32-bit
FPGA	Field Programmable Gate Array
GPU	Graphics Processing Unit
GXNOR-Net	Gated XNOR-Network
HBM	Homogeneous Barrier Modulation
HBM	High Bandwidth Memory
HD	Hamming Distance
HDD	Hard Disk Drive
HMC	Hybrid Memory Cube
HRS	High Resistance State
HSI	Hyper-Spectral Image
ICP	Inductively-coupled plasma
IMC	In-Memory Computing
IMSS	In-Memory Similarity Search
INT32	Integer 32-bit
IRDS	International Roadmap for Devices and Systems
ITRS	International Technology Roadmap for Semiconductors
IV	Current Voltage
LIM	Logic-In-Memory
LMS	Least Mean Square
LPDDR	Low Power Double Data Rate
LRS	Low Resistance State
LSB/MSB	Least Significant Bit/ Most Significant Bit
LUT	Look Up Table
MAC	Multiply-and-Accumulate
MAT	Matrix
MB/GB/TB	Mega Byte/Giga Byte/Tera Byte
MIM	Metal-Insulator-Metal
ML	Machine Learning
MLC	Multi Level Cell/Multi Level Capability
MLP	Multi-Layer Preceptron
MLP	Multi Layer Perception
MW	Memory Window
NMOS	N-channel Metal-Oxide Semiconductor
NVM	Non-Volatile Memory

NV-SRAM	Non-Volatile Static Random Access Memory
OxRAM	Oxide based Random Access Memory
PCA	Principal Component Analysis
PCM	Phase Change Memory
PCMO	$\text{Pr}_{0.7}\text{Ca}_{0.3}\text{MnO}_3$
PDK	Process Development Kit
PE-ALD	Plasma-enhanced Atomic Layer Deposition
P-F	Poole-Frenkel
PIM	Processing-In-Memory
PMOS	P-channel Metal-Oxide Semiconductor
PMU	Pulse Measure Unit
PWM	Pulse-Width Modulation
QI	Query Input
QNN	Quantized Neural Network
RNN	Recurrent Neural Network
RRAM/ReRAM	Resistive Random Access Memory
S/H	Sample and Hold
SA	Sense Amplifier
SCA	Side-Channel Attack
SCLC	Space Charge Limited Conduction
SD	Stored Data
SEM	Scanning Electron Microscope
SHA-3	Secure Hash Algorithm-3
SHL	Shift Logical Left
SL	Select Line
SLC	Single Level Cell
SLIM	Simultaneous Logic-In-Memory
SMU	Source Measure Unit
SRAM	Static Random Access Memory
SSD	Solid State Drive
STT-MRAM	Spin Transfer Torque Magnetic Random Access Memory
TCAM	Ternary Content-Addressable Memory
TDMAHf	Tetrakis(dimethylamido)hafnium
TDMATi	Tetrakis(dimethylamido)titanium
TEM	Transmission Electron Microscopy
TIA	Trans-Impedance Amplifier
TNN	Ternary Neural Networks
t-SNE	t-Distributed Stochastic Neighbour Embedding
V_{TB}	Voltage applied at TE-Voltage applied at BE
VVM	Vector Vector Multiplication

VWW	Visual Wake Words
WL/WLB	Wordline/Wordline bar