

# NON-VOLATILE MEMORY-CENTRIC COMPUTING ADVANCES

VIVEK KAMALKANT PARMAR



DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY DELHI

JULY 2023

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

# NON-VOLATILE MEMORY-CENTRIC COMPUTING ADVANCES

by

**VIVEK KAMALKANT PARMAR**

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**

**JULY 2023**

योगस्थः कुरु कर्माणि सङ्गं त्यक्त्वा धनञ्जय ।  
सिद्ध्यसिद्ध्योः समो भूत्वा समत्वं योग उच्यते ॥  
२-४८

*Shrimad Bhagvad Gita*

*By being established in Yoga, O Dhananjaya, undertake actions, casting off attachment and remaining equipoised in success and failure. Equanimity is called Yoga. ||2||*

*Dedicated to the Almighty and all my Respected Teachers*

*Dedicated to  
my caring parents,  
and dear friends*

# Certificate

This is to certify that the thesis entitled “**NON-VOLATILE MEMORY-CENTRIC COMPUTING ADVANCES**”, submitted by **Vivek Kamalkant Parmar** 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*

My PhD journey has been nothing short of a roller coaster ride filled with enriching, frustrating, fun, painful, insightful, and enlightening moments that I will cherish forever. As this journey comes to an end, this is a humble attempt to thank everyone who accompanied me on this journey for their contributions and moral support.

First and foremost, I would like to express my deepest gratitude to my advisor Prof. Manan Suri for believing in me and supporting me throughout this journey. He generously provided me the guidance to hone my skills and counselled me during adversities and challenges. He supported me in my endeavours to venture into the unexplored and conquer new frontiers. I am most grateful for his patience in handling my impulsive outbursts and making me more humble and patient. His influence in shaping me to become a better person 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 am grateful to the members of my student research committee (SRC), **Prof. Anuj Dhawan**, **Prof. Shubhendu Bhasin** and **Prof. Ajeet Kumar** for their critical advice and continual moral support. 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. Kangho Lee** (currently Master, Samsung Electronics), **Dr. Vinayak Bharat Naik** (Senior Technologist, GLOBALFOUNDRIES), **Dr. Amir Regev** (CTO, Weebit Nano, Israel), **Dr. Giuseppe Piccolboni**, **Dr. Alessandro Bricalli**, **Dr. Damien Querlioz** (Research Scientist, CNRS, Université Paris-Saclay), **Dr. Bogdan Penkovskiy** (currently AI R&D Project Manager, Alysophil), **Dr. Christopher H. Bennett** (Member of Technical Staff, Sandia National Laboratories), **Dr. Sourav De** (Researcher, Fraunhofer IPMS), **Dr. Thomas Kämpfe** (Group Manager Integrated RF & AI, Fraunhofer IPMS), **Prof. Tian-Li Wu** (Professor, National Yang Ming Chiao Tung University, Taiwan) for the technical discussions and support as well as suggestions that helped me improve the experiments leading to some of the key results of this thesis. I would like to give a special mention to the support of Meta Reality Labs Research team (Dr. Syed Shakib Sarwar, Dr. Ziyun Li, Dr. Hsien-Hsin S. Lee, Dr. Barbara De Salvo) for their continued technical guidance, feedback and support over the last phase of my journey.

I would specially like to thank my friends **Sandeep Kaur Kingra**, **Shubham Negi** 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 research problems, exploring Delhi, and lively discussions over tea are some of the most memorable moments of my journey. I am further grateful to my juniors **Tinish Bhattacharya** (PhD student at University of California Santa Barbara), **Narayani Bhatia**, **Chithambara Moorthii**, **Sufyan Khan** (University of Wisconsin-Madison), **Deepak Verma**, **Digamber Pandey** and **Richa Mishra** 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 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 all my friends.

This journey would not have been possible without the continuing support, patience, and encouragement from my parents **Kamalkant H. Parmar** and **Jayshreeben K. Parmar**.

Vivek Kamalkant Parmar

# *Abstract*

Recent advances in the domain of Artificial Intelligence has led to a wide variety of solutions across multiple domains. Specialized hardware accelerators have been developed to facilitate high-speed computations for these data-intensive workloads. While dedicated AI hardware has been dominantly explored for cloud/enterprise applications, true benefits of AI can be realized by enabling low-power edge computing. For IoT (Internet of Things) devices with constrained area and power, performing high-precision computations becomes infeasible. Conventional DNNs implementations rely on networks with sizes in the order of MBs and compute capacity of the order of Tera FLOPs/sec. Such implementations require high-precision computing using floating-point computations, that escalates energy costs. Additionally, due to physical separation between the storage/memory unit and the processor, memory $\leftrightarrow$ compute bottleneck causes a further limitation.

This thesis focusses on different possible use-cases/benefits of exploiting emerging NVM technology for advanced computing applications. An exhaustive study of NMC as well as IMC techniques is presented focusing on utilizing NVM devices. As a physical realization of the NMC concept for edge-computing the NVIA (non-volatile inference accelerator) architecture based on 22nm-MRAM (magneto-resistive random access memory) has been explored both experimentally and through large scale simulations. Benefits of non-volatile inference with resilience to harsh operating conditions (800 Oe and 125 °C) has also been demonstrated using the MNIST dataset. On the application front, the concept has also been validated for the domain of Mixed-Reality workloads (such as eye segmentation and hand detection). Differential NVM in-memory-compute bitcells have been validated both experimentally and through large-scale simulations on multiple application workloads (Thermal Images, Fashion-MNIST, CIFAR-10, Visual Wake Words). In particular we show DM-FeFET (differential mode ferro-electric field effect transistor) realized using 28nm HKMG technology, which exhibits excellent BER (bit-error rate) tolerance of upto  $10^{-2}$  for both storage and IMC applications involving multi-bit precisions.

A novel methodology is proposed for realizing few-shot learning application with binary precision and on 130nm-RRAM based IMC hardware and validated over the miniImageNet and ORBIT datasets. Further, IMC based realization of stochastic BNN (binarized neural networks) is proposed exploiting device variability. Macro-level realization including Analog and Digital periphery circuits for RRAM IMC was demonstrated using the open-source Skywater 130nm PDK.

# सार

कृत्रिम बुद्धिमत्ता (Artificial Intelligence) के क्षेत्र में हाल के विकास ने डेटा-गहन वर्कलोड को जन्म दिया है जिसके लिए जीबी डेटा पर किए गए संचालन की आवश्यकता होती है। समानांतरवाद में सक्षम पारंपरिक कंप्यूटिंग आर्किटेक्चर ऐसी उच्च मेमोरी आवश्यकताओं के समर्थन में सीमाओं का सामना करते हैं जिन्हें "मेमोरी वॉल" के रूप में भी जाना जाता है। नियर-मेमोरी कंप्यूटिंग (NMC) और इन-मेमोरी कंप्यूटिंग (IMC) जैसी मेमोरी-सेंट्रिक कंप्यूटिंग अवधारणाएँ स्टोरेज या इन-सीटू के पास कंप्यूट प्रदर्शन करके आशाजनक समाधान के रूप में उभरी हैं, जिससे मेमोरी बैंडविड्थ की अड़चन दूर हो गई है।

इस थीसिस में एनएमसी के साथ-साथ आईएमसी तकनीकों का एक विस्तृत अध्ययन प्रस्तुत किया गया है जो उभरती हुई गैर-वाष्पशील मेमोरी (ईएनवीएम) उपकरणों के उपयोग पर ध्यान केंद्रित कर रहा है। एमआरएएम (मैग्नेटो-रेसिस्टिव रैंडम एक्सेस मेमोरी) पर आधारित NVIA (गैर-वाष्पशील निष्कर्ष त्वरक) आर्किटेक्चर के Edge-Computing के लिए NMC अवधारणा के भौतिक अहसास के रूप में प्रयोगात्मक रूप से और बड़े पैमाने पर सिमुलेशन के माध्यम से इसके लाभों को मान्य करते हुए विस्तार से पता लगाया गया है। -एआई जिसे MRAM का उपयोग करके कठोर परिचालन स्थितियों के लिए दुर्लभ संगणना और लचीलेपन की आवश्यकता होती है। अवधारणा को MR (मिश्रित-वास्तविकता) पहनने योग्य उपकरणों के उभरते क्षेत्र में भी लागू किया गया है। डिफरेंशियल बिटकल्स (प्रयोगात्मक रूप से और सिमुलेशन के माध्यम से) का विस्तार से पता लगाया गया है और DM-FeFET (डिफरेंशियल मोड फेरो-इलेक्ट्रिक फील्ड इफेक्ट ट्रांजिस्टर) प्रस्तावित है जो स्टोरेज और IMC दोनों अनुप्रयोगों के लिए उत्कृष्ट BER (बिट-एरर रेट) सहिष्णुता प्रदर्शित करता है जिसमें मल्टी- शामिल हैं। बिट परिशुद्धता। द्विआधारी परिशुद्धता के साथ कुछ-शॉट सीखने के साथ अनुकूली एआई को साकार करने के लिए एक उपन्यास पद्धति प्रस्तावित है और एल्गोरिदम और आईएमसी हार्डवेयर दोनों पर मान्य है। स्टोचैस्टिक BNN (बिनाराइज्ड न्यूरल नेटवर्क) का एक पूर्ण आईएमसी आधारित अहसास डिवाइस परिवर्तनशीलता के साथ-साथ आईएमसी क्षमताओं का दोहन करने का प्रस्ताव है। ओपन-सोर्स स्काईवाटर 130 nm पीडीके का उपयोग करके RRAM (रेसिस्टिव रैंडम एक्सेस मेमोरी) को साकार करने के लिए सभी एनालॉग और डिजिटल परिधि सहित एक पूर्ण-प्रणाली प्राप्ति डिजाइन की गई थी।

# Contents

Certificate

Acknowledgements i

Abstract iii

Contents v

List of Figures viii

List of Tables xvi

Abbreviations xix

**1 Introduction and Motivation 1**

1.1 Motivation . . . . . 1

1.2 Thesis Organization and Contributions . . . . . 4

1.2.1 Objective of Thesis . . . . . 4

1.2.2 Key Contribution of Thesis . . . . . 4

1.2.3 Thesis Organization . . . . . 5

**2 Accelerators exploiting NVM 7**

2.1 Introduction . . . . . 7

2.2 Performance Enhancement of Edge-AI-Inference Using Commodity MRAM . . . . . 11

2.2.1 Proposed NVIA Architecture and Test Setup . . . . . 11

2.2.2 NVIA Operation . . . . . 12

2.2.3 NN accelerator and AI Workloads: Network and Dataset . . . . . 13

2.2.4 Results and Discussion . . . . . 14

2.2.5 NVIA Power Analysis . . . . . 14

2.2.5.1 Impact of network parameters . . . . . 14

2.2.5.2 Impact of Memory Parameters . . . . . 16

2.3 BER-Resilient Edge-AI-Inference Accelerator With Quantized Neural Networks 17

2.3.1 Proposed NN Architecture and Test Setup . . . . . 17

2.3.2 Results and Discussion . . . . . 17

2.4 Memory-Oriented Design-Space Exploration of Edge-AI Hardware for XR Applications . . . . . 19

2.4.1 Analysis on Representative XR-AI Workloads . . . . . 19

2.4.1.1	Dataset Description	20
2.4.1.2	Network Training and Quantization	21
2.4.2	Implementation on Edge-AI Accelerators	22
2.4.3	Proposed NVM-based Enhancement	23
2.4.4	Results and Discussion	24
2.5	Summary	29
<b>3</b>	<b>Building error-resilient NVM-IMC primitives</b>	<b>30</b>
3.1	Introduction	30
3.2	VMM Computational Mapping on Crossbar for BNN	34
3.2.1	Results and Discussion	37
3.2.1.1	Device Characterization	37
3.2.1.2	BNN experiments	38
3.2.1.3	BNN Simulations with device non-idealities	41
3.3	Dual-Configuration IMC Bitcells for BNNs	44
3.3.1	Proposed IMC-XNOR bitcell configurations	44
3.3.1.1	XNOR <sub>row</sub> implementation:	46
3.3.1.2	XNOR <sub>col</sub> implementation	46
3.3.2	Results and Discussions	49
3.3.2.1	Device Characterization	49
3.3.2.2	Experimental Demonstration of OxRAM XNOR IMC	50
3.3.2.3	Learning Performance and Energy Estimation	54
3.3.2.4	Performance comparison of XNOR <sub>row</sub> and XNOR <sub>col</sub>	56
3.4	DMFeFET-Array based IMC-Macro for multi-precision NN applications	60
3.4.1	Proposed DM-FeFET IMC bitcell	61
3.4.2	Results and Discussions	62
3.4.2.1	Device Characterization	62
3.4.2.2	Binary MAC Operation Validation	64
3.4.2.3	Neural Network Simulations	65
3.4.2.4	Reliability factors for FeFET	66
3.5	Summary	68
<b>4</b>	<b>Advanced learning use-cases with IMC architectures</b>	<b>70</b>
4.1	Introduction	70
4.2	Fully-Binarized Distance Computation based On-device Few-Shot Learning	71
4.2.1	Proposed BinDC for FSL	72
4.2.2	Results and Discussion	73
4.2.2.1	Network Results	75
4.2.2.2	Benchmarking on Embedded Platforms	76
4.2.2.3	IMC-based optimizations	78
4.3	Hardware-Efficient Stochastic Binary CNN Architectures for Near-Sensor Computing	81
4.3.1	Proposed SBCNN Methodology	82
4.3.2	Hardware Implementation based on RRAM	84
4.3.3	Results and Discussions	86
4.3.3.1	Case study A : CIFAR-10	86
4.3.3.2	Case study B: Microscopy	87
4.3.3.3	Learning Performance	90

4.3.3.4	Performance analysis: Memory, Energy, Delay	92
4.4	Summary	93
<b>5</b>	<b>IMC Macro realizations using open-source EDA tools</b>	<b>94</b>
5.1	Introduction	94
5.2	Prior Art	95
5.3	Proposed RRAM-based IMC (Open-RIMC)	96
5.4	Results and discussions	99
5.5	Summary	101
<b>6</b>	<b>NVM-based stochastic neuron for exploratory applications</b>	<b>102</b>
6.1	Introduction	102
6.2	Basics of OxRAM and DGM Architectures	104
6.2.1	OxRAM Switching Mechanism	104
6.2.2	Deep Generative Architectures	104
6.2.3	Restricted Boltzmann Machines (RBM)	104
6.2.4	Stacked Denoising Autoencoder (SDA)	105
6.2.5	Deep Belief Network (DBN)	106
6.3	Implementation of Proposed Architectures	106
6.3.1	RBM Block	106
6.3.2	Synaptic Network	107
6.3.3	Stochastic Neuron Block	107
6.3.4	CD Weight Update Block	108
6.3.5	Output Normalization block	108
6.4	Deep Learning Simulations and Results	109
6.4.1	Stacked Denoising Autoencoder performance analysis	109
6.4.2	Deep Belief Network performance analysis	109
6.4.3	Benefits of weight initialization with Proposed DGM	110
6.5	Switching activity analysis for the Proposed DGM architectures	111
6.5.1	Impact of hidden layer size	113
6.6	Limitations	114
6.7	Summary	114
<b>7</b>	<b>Conclusion and Future Work</b>	<b>116</b>
7.1	Scope for Future Work	118
<b>A</b>	<b>Summary tables</b>	<b>121</b>
	<b>Bibliography</b>	<b>122</b>
	<b>List of Publications</b>	<b>147</b>
	<b>Curriculum Vitae</b>	<b>150</b>

# List of Figures

1.1	An overview of applications of Transformers sometimes called foundation models which are the current state-of-the-art for deep learning applications (Adapted from <a href="https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model/">https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model/</a> ).	2
1.2	Computation mapping flow for performing function $f$ using data $D$ stored in memory with three different architectures: (a) Conventional computing, (b) Near-memory computing (NMC), (c) In-memory computing (IMC). For a conventional computing system, data needs to be fetched from memory to processing unit leading significant costs in latency and energy. When using the NMC approach, computations are performed using peripheral circuits by reading data internally within the memory block and storing back the result. In case of IMC, computation is performed by exploiting the physical attributes of the devices. Both charge-based memory technologies, such as SRAM, DRAM and flash memory, and resistance-based memory technologies, such as RRAM, PCM and STT-MRAM, can serve as elements of such a computational memory unit. Adapted from [1].	3
2.1	MRAM devices considered in this study: (a) Field induced switching (toggle-MRAM), (b) Spin-transfer torque (STT) switching. TEM images of MRAM devices: (c) Everspin Toggle-MRAM, and [2] (d) GlobalFoundries 22nm eMRAM [3].	8
2.2	(a) Proposed NVIA architecture (b) Block diagram of FPGA based testbench for experimental validation.	10
2.3	(a) Photo of our FPGA-based NVIA testbench with integrated PCB for MRAM/SRAM chips and sensor, (b) HAR dataset classes (walking, sitting, walk upstairs, sleeping, walk downstairs, standing).	12
2.4	(a) Timing model of operation for proposed NVIA (b) Memory current usage profile for different memory types.	13
2.5	(a) Neural Network Architecture used for the study (b) Our implemented inference pipeline on FPGA.	13
2.6	Memory power dissipation for <i>HAR</i> workload for memory chips used in the testbench as a function of: (a) Event-frequency (Weight size = 17 kB), and (b) Weight size (Event-frequency = 5 Hz).	15
2.7	Impact of technology node on memory performance parameters w.r.t. event-frequency for <i>HAR</i> workload: (a) Power (b) Power-density. (c) Power and area scaling trend w.r.t. memory technology node (Event-frequency = 1 Hz). Device data based on [4, 5, 6, 7, 8].	16
2.8	(a) Fabricated 40 Mb STT-MRAM macro (Inset: TEM Image of MTJ) (b) Block diagram of FPGA testbench showing proposed Quantized Neural Network Inference IP interfaced with STT-MRAM. (c) Sources of External Magnetic Field [9, 10, 11].	17

2.9	(a) Sample images from MNIST dataset, (b) Network architectures used for AI accelerator performance estimations: (b) MLP (c) CNN. Inference Accuracy vs. BER for MNIST dataset: (d) MLP (e) CNN. . . . .	18
2.10	Experimentally measured STT-MRAM macro BER variation in presence of external magnetic field as a function of write voltage (b) BER vs $H_{ext}$ at room-temperature (25°C) and high-temperature (125°C). (c) Stand-by Magnetic Immunity vs Operating Temperature for target BER. (d) Contour plot showing impact of Temperature and $H_{ext}$ on Inference Accuracy for Binarized CNN (Color indicates accuracy level). . . . .	18
2.11	Normalized Write power vs Precision at different Magnetic field conditions during Write operation for (a) Quantized MLP (b) Quantized CNN. . . . .	19
2.12	Sample images from datasets: (a) FPHAB and (b) OpenEDS. (c) MobileNetV2 building block (Inverted Residual Bottleneck [12]). (d) DetNet. (e) EDSNet (UNet model + MobileNetV2 Backbone). (f) Training loss evolution. (g) DetNet evaluation, on sample image, with FP32 and INT8 precision. Red circle shows ground truth and purple shows predicted. (h) EDSNet evaluation, on sample image, with FP32 and INT8 precision. (i) Trained and quantized weight distributions for both networks. . . . .	20
2.13	Simulated: (a) CPU + Memory in QKeras, (b) Eyeriss PE [13] and (c) Simba [14] in Timeloop. (d) Specification of simulated architectures used in the study. Numbers mentioned in bracket indicate bus size. (e) Energy Breakdown of simulated architectures. (f) Estimated EDP for inference of DetNet and EDSNet. SRAM-only variant was estimated at 45 nm for CPU and 40 nm for Eyeriss/Simba. Scaling estimates on other nodes are based on [15, 16]. . . . .	22
2.14	(a) Operation breakdown for XR-AI accelerator (b) Memory activity profile during XR-AI workload execution: (i) SRAM (ii) NVM. (c) Breakdown of memory specific operations in the AI Inference mode. Proposed NVM introduction strategies: (ii) PO (SRAM+MRAM) and (iii) P1 (MRAM-only). (d) Single inference energy dissipation for 9 simulated architectural variants on DetNet and EDSNet. . . . .	23
2.15	Simulated energy breakdown in terms of memory and compute for NVM-based architectural variants for DetNet on: (a) CPU (b) Eyeriss (c) Simba, and EDSNet on: (d) CPU (e) Eyeriss (f) Simba. . . . .	25
2.16	Simulated memory power vs. IPS benchmarking for proposed architectural variants utilizing SRAM, STT, SOT, and VGSOT devices. (a,b,e,f) correspond to Simba while (c,d,g,h) correspond to Eyeriss. For each case, IPS cross-over points w.r.t SRAM and VGSOT are indicated in the plots. Any IPS value below cross-over point would lead to energy-savings while using NVM. Plots (a-d) (top row) are for P1 variants while (e-h) (bottom row) are for P0 variants. For P0 variants, cross-over points are limited based on maximum frequency supported by the memory architecture. . . . .	27
3.1	Example of Binarized layer implementations with steps for (a) Binarized Convolutional, and (b) Binary fully connected layers. . . . .	32
3.2	(a) Illustration of a typical BNN network highlighting weight and input vectors. To address the limitation of negative weight and negative input using physical mapping, we propose new VMM computation strategies using mapping as defined in: (b) BNN Input mapping, (c) BNN Weight mapping, (d) Components used for BNN computations. (e) Proposed VMM computation strategies. . . . .	34

3.3	Mapping of binary VMM operation on OxRAM Crossbar for a $6 \times 3$ weight matrix in differential scheme. Weights are represented by device conductance and inputs in form of voltages. VMM output is computed in the form of integrated at the columns of the crossbar with final outputs realized in form of output voltages at the last comparator stage. . . . .	35
3.4	(a) HR-TEM image of bilayer Ni/ $HfO_2$ /ATO/TiN OxRAM device fabricated for this study along with crossbar image. (b) Fabrication flow for the OxRAM device. (c) DC IV curves of 36 OxRAM devices in $6 \times 6$ crossbar indicating low D2D variability. (d) Analog conductance tuning characteristics observed in the OxRAM device using identical SET and RESET pulses trains. Statistical distributions of selected LRS (4-6) and HRS (1-3) states depicting different levels of overlap in state distribution: (e) PDF, (f) Box plot, (g) CDF. (h) Experimental setup used for characterization in the study depicting custom testbench built using ADCs, DACs, CMOS switches and OxRAM crossbar chip. (i) Single channel for device access and control. (j) Block diagram of programming and read path depicting components used in test bench. . . . .	36
3.5	(a) Sample images from Thermal camera based Rock-Paper-Scissor dataset with corresponding binarized representations used as network input. (b) Binarized MLP network used for Thermal RPS application. (c) Binarized weight map of output layer. (d) Differential representation of binarized weight map. (e) Programmed conductance on OxRAM crossbar for the output layer computations. (f) Comparison of learning performance of all computation strategies on ‘Training’ and ‘Test’ splits of the dataset. (g) Learning performance comparison of the computation strategies in terms of training accuracy estimated using theoretical model (ideal BNN), simulated (including device Conductance) model and experimental measurements. . . . .	38
3.6	Normalized neuron response for inputs of each class using VMM computation strategies: (a) Positive Weight, (b) Negative Weight, (c) Positive Input, (d) Negative Input, (e) XNOR, (f) Differential. . . . .	39
3.7	Confusion matrix on Training Set for Thermal RPS dataset using experimental measurements for: (a) Positive Weights (b) Negative Weights (c) Positive Input, (d) Negative Input, (e) XNOR, (f) Differential. Error observed in column-wise neuron currents between simulation and experiments w.r.t. active input rows with +ve Weights (g-i) and -ve Weights (j-l) for each class from Thermal RPS dataset. . . . .	40
3.8	(a) Sample images of two classes from Fashion-MNIST dataset with binary images resulting from thermometric encoding (channels=8, resolution=32). (b) LeNet based binarized CNN architecture used for validation of proposed VMM strategies. . . . .	41
3.9	(a-c) Experimentally measured device state distributions with varying memory windows. (d) Experimentally measured statistical distributions for the 6 conductance states. . . . .	42
3.10	Accuracy as a function of overlap in conductance state distributions using six different VMM computation strategies: (a) Positive weight (b) Negative weight (c) Positive input (d) Negative input (e) XNOR (f) Differential. Box plots represent experimentally characterized device distributions. . . . .	43
3.11	(a) Distributions of LRS and HRS conductance as a function of variance using mean values based on MW=3. Accuracy as a function of variation in conductance states using six different VMM computation strategies: (b) Positive weight (c) Negative weight (d) Positive input (e) Negative input (f) XNOR (g) Differential. . . . .	43

3.12	(a) BNN computation mapping on $\text{XNOR}_{row}$ 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 $\text{XNOR}_{col}$ configuration and <i>POPCOUNT</i> is implemented inherently over fabricated OxRAM array. . . . .	45
3.13	Schematic representations of input activations and weights with computation equations for: (a) $\text{XNOR}_{row}$ ( $4 \times 2$ 2T-2R bitcell array), (b) $\text{XNOR}_{col}$ ( $2 \times 4$ 2T-2R bitcell array) IMC implementations on a $4 \times 4$ 1T1R array. (c) CSA block used for the study. Binary activations are mapped onto the differential SLs (in case of $\text{XNOR}_{row}$ ) and WLs (in case of $\text{XNOR}_{col}$ ). Binary weights are mapped onto the HRS/LRS values of XNOR-OxRAM cells. . . . .	47
3.14	(a) SEM cross-section of the $\text{SiO}_x$ 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. . . . .	49
3.15	(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 $\Omega$ to 20 k $\Omega$ and HRS ranges from 60 k $\Omega$ to 1 M $\Omega$ . . . . .	50
3.16	Schematic representation of the $8 \times 8$ OxRAM array. . . . .	51
3.17	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. . . . .	52
3.18	Schematic representation/operand mapping corresponding to possible combinations of input activations ('-1', '+1') and weights ('-1', '+1') are shown for (a,b) $\text{XNOR}_{row}$ , and (d,e) $\text{XNOR}_{col}$ . (c,f) Experimentally characterized bitcell output current of four possible operand combinations for $\text{XNOR}_{row}$ and $\text{XNOR}_{col}$ respectively. . . . .	53
3.19	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. . . . .	54
3.20	Statistical distribution for VMM output variability for (a) $\text{XNOR}_{row}$ and (b) $\text{XNOR}_{col}$ configuration. The simulations are performed on $8 \times 8$ 1T-1R array with all input combinations applied for $\geq 1000$ 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) $\text{XNOR}_{row}$ , and (d) $\text{XNOR}_{col}$ . . . . .	56
3.21	Convolution building blocks used in (a) FracBNN [17] and (b) MobileNet-v1 [18] 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 [19] used for binarizing inputs compared to conventional fixed point representation. . . . .	57
3.22	Simulated performance trends of $\text{XNOR}_{col}$ IMC bitcell based BNN implementation with varying MAT Sizes for: (a) VWW dataset, and (b) CIFAR-10 dataset. . . . .	58

3.23	(a) Impact of BER on BNN accuracy for VWW and CIFAR-10 workloads. For XNOR <sub>row</sub> 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 ( $\approx 1\%$ ).	59
3.24	(a) Schematic of proposed DM-FeFET IMC bitcell. (b) MLC state mapping for proposed DM-FeFET bitcell for multiple precisions. Statistical distribution of $I_{read}$ for weight encoding schemes: (c,f) Conventional MLC; (d,g) Differential MLC. Inset on top left for conventional schemes shows overlap in $I_{read}$ distributions for the lower MLC states (S1-S3). State for 2-bit storage using: (e) Conventional, (h) Proposed DM-FeFET scheme.	61
3.25	(a) Error probability calibrated over 1000 samples with conventional and DM-FeFET programming schemes, showing improvement in BER in program-erase operation achieved by DM operation. (b) 2-bit image of SuperMario written over a 32x32 matrix proves this fact.	61
3.26	(a) Stack of the fabricated FeFET device. (b) SEM image of the device stack. (c) IV characterization showing impact of CL (current limiters). FeFETs were programmed and erased with 500 ns pulses. (d) Distribution of $I_{BL}$ states using SET/RESET pulses for LVT and HVT showcasing benefit of using CL.	62
3.27	(a) 3D illustration of fabricated IMC array. (b) Layout of the fabricated FeFET array with inset showing crossbar. (c) Schematic representation of DM-FeFET crossbar array demonstrates the operation of IMC bitcell.	63
3.28	(a) TVB-sensing based program and erase characteristics show presence of MLC. (b) CDF of MLC $V_{th}$ states obtained. (c) DCB-sensing based program and erase characteristics. (d,e) The distribution of $I_d$ reveals overlapping states and absence of MLC at wafer scale.	63
3.29	(a) $I_{BL}$ as a function of #Active lines with varying $V_{WL}$ . (b) $I_{BL}$ response for different values of $V_{WL}$ . (c) Measured CDF of $I_{BL}$ for 20 crossbar arrays with cells activated serially in a column. (d) Data retention of MAC output.	64
3.30	(a) Output voltage responses for MAC operations performed using 4x8 DM-FeFET arrays using 3 weight precisions (binary, ternary, 3-bit) and MLC based on program (top) and erase (bottom). (b) MAC operation mapping on 2x4 DM-FeFET array. (c) CNN architecture used for the study (VGG-8) along with weight mapping strategy. (d) Benchmarking of current work with state-of-the-art IMC macro.	65
3.31	(a) Training accuracy and (b) Loss evolution during training over 30 epochs for CIFAR-10 dataset. (c) Binary MAC output error as a function of MAT block size for all 6 combinations (3 weights, 2 MLC methods). (d) Impact of binary MAC output BER on accuracy performance shows BER tolerance of upto 1%.	65
3.32	(a) TDDDB characteristics and (b) Weibull plot for different bias values. Relatively high $\beta$ is obtained, indicating good uniformity. (c) Extrapolation of operating voltages for 3, 5 and 10-year lifetime of 1% failure. (d,e) The flicker noise characterization was conducted for optimizing the pulse width and amplitude value for READ operation. The gate was biased at an overdrive voltage of 500 mV.	66
4.1	t-SNE based distributions of embeddings derived using combination of ProtoNet with (a) ResNet12,(b) ResNet18 backbones and FEAT with (c) ResNet12,(d) ResNet18 backbones. Highlighted areas in red show class-wise confusion.	74

4.2	t-SNE based distributions of embeddings derived using proposed binarized embeddings with precisions of (a) 2-bit, (b) 4-bit, (c) 8-bit with normalization method 1. Corresponding results with normalization method 2 (d-f). Highlighted areas in red show class-wise confusion. . . . .	75
4.3	Example images from datasets used in the study: (a) miniImageNet[20], (b) ORBIT[21]. . . . .	76
4.4	(a) Prototype learning representation in 2-D space. (b) Feature extraction pipeline for image classification task. (c) Feature extraction and inference pipeline used for object classification with videos. . . . .	77
4.5	Measurement based results on Jetson Xavier NX platform (CPU-based) for the two workloads: (a,b) Inference latency, (c,d) Inference energy and (e,f) Peak power. Dashed black line shows the floating-point baseline utilizing cosine distance computations. . . . .	79
4.6	(a) I-V characteristics showing electro-forming [Inset: SEM cross-section of the $\text{SiO}_x$ 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). . . . .	80
4.7	(a) Schematic of IMC-array showing mapping of binarized support vector bits and application of query inputs at the WL of the array. (b) $I_{BL}$ as a function of computed HD. (c) Custom PCB used for performing experimental measurements with RRAM-based memory array for IMC applications. . . . .	81
4.8	(a) Stochastic sampling for multi-channel input images based on normal distribution (b) Modified AlexNet architecture used for CIFAR-10 based case study. (c) CIFAR-10 dataset samples for each class. . . . .	82
4.9	Variation in stochastic input representations based on $N_{pre}$ and for CIFAR-10 image sample by stochastic sampling using distributions: (a) Uniform (b) Normal. Variation in stochastic input representations based on $N_{pre}$ and for ALD-AIR image sample by stochastic sampling using distributions: (c) Uniform (d) Normal. . . . .	83
4.10	Histograms of pixel-value distribution over image samples from the datasets used in the study: (a)-(d) Original image, (e)-(h) Uniform distribution based stochastic sampling, (i-l) Normal distribution based stochastic sampling . . . . .	84
4.11	(a) Stochastic neuron circuit based on OxRAM device used for input sampling. 2T-2R in-memory XNOR circuit for BNN computation. (b) C2C variability observed in LRS state for the fabricated OxRAM device of [22]. (c) Memory array chip photograph (d) OxRAM cell, (e) Binarized Neural Network implementation highlighting connections to one specific neuron. (f) Implementation of Binarized Neural Network in the "parallel to sequential" configuration. (g) 2T-2R bitcell array (h) Schematic of 2T-2R bit-cell for XNOR operation computation based on PCSA [23] . . . . .	85
4.12	Variation of network performance metrics with $N_{pre}$ for inference using stochastic BNN (AlexNet model) for CIFAR-10 dataset : (a) Accuracy (b) Mean Average Precision (c) Receiver Operating Characteristics Area under Curve. The results have been averaged over 5 iterations. . . . .	87
4.13	Microscopy image samples from dataset [24] for diseases. (a) Malaria, (c) Tuberculosis (e) Intestinal Parasite. Sample slices used for training for classification network to detect pathogen: (b) Malaria (d) Tuberculosis (f) Intestinal Parasite. . . . .	88

4.14	Network precision and architecture analysis for microscopy diagnosis task. For all stochastic networks, training and inference is performed with $N_{pre}=32$ . . . . .	89
4.15	Heatmaps generated based on sliding window-based detections using Reduced Stochastic BNN classifiers: (a) Malaria (b) Tuberculosis (c) Intestinal Parasite. Detections on microscopy sample images: (d) Malaria (e) Tuberculosis (f) Intestinal Parasite. Red box indicates network based detection, white box indicated ground truth. The $N_{pre}$ used for training and inference are 32. . . . .	90
4.16	ROC curves for proposed SBCNN :(normal distribution and reduced layers): (a) Malaria (b) Tuberculosis (c) Intestinal Parasite. Precision recall curves for proposed SBCNN (normal distribution and reduced layers): (d) Malaria (e) Tuberculosis (f) Intestinal Parasite. . . . .	91
4.17	Variation of network performance metrics with $N_{pre}$ for SBCNN (reduced model) on microscopy diagnosis task : (a) Accuracy (b) mAP (c) ROC_AUC. Results have been averaged over 5 iterations. . . . .	91
5.1	(a) Chip Architecture including detailed block-diagram of implemented Open-RIMC. (b) AND-Gate based IMC bitcell with truth-table. (c) XNOR-Gate based IMC bitcell with truth-table. (d) Input signal mapping for the MUX used on all 3 terminals of the IMC array. . . . .	95
5.2	(a) Layout of the RRAM-based IMC macro implemented in Skywater-130nm PDK. Inset shows the integration of RRAM. (b) Layout of Open-RIMC including digital periphery. (c) Area breakdown of implemented IMC macro. (d) Area breakdown of digital periphery. . . . .	96
5.3	(a) Mapping of input, weights and outputs on the implemented RRAM-IMC macro depicting weight storage for both AND and XNOR-based computations. (b) SPICE simulations depicting MAC outputs generated using RRAM-IMC macro in form of ADC output and $I_{BL}$ as a response to applied inputs at WL. Distributions of $I_{BL}$ corresponding to ADC outputs when utilizing $4 \times 16$ weight matrix based multiplications for: (c) XNOR, (d) AND. . . . .	97
5.4	Heatmaps showing comparison of obtained ADC results against expected Popcount values for: (a) XNOR, (b) AND. . . . .	98
5.5	(a) BNN network used for evaluating performance of implemented XNOR-based MAC. (b) Sample of class images from MNIST dataset. (c) Impact of MAT block size on MAC output BER for binarized activations. (d) MNIST test accuracy as a function of MAC output BER. . . . .	100
6.1	Cycle-to-Cycle ON/OFF-state resistance distribution for $HfO_x$ device presented in [25]. ( $R_{ON}$ : $\mu = 3.47, \sigma = 0.11$ , $R_{OFF}$ : $\mu = 6.72, \sigma = 0.69$ ). Inset shows Filament state during HRS and LRS. . . . .	104
6.2	(a) Graphical representation of RBM hidden/visible nodes. (b) DBN architecture comprising of stacked RBMs. (c) Denoising noisy image using autoencoder. . . .	105
6.3	(a) Fully digital CD based weight update module. (b) Circuit schematic of stochastic sigmoid neuron.(c) Hybrid CMOS-OxRAM RBM layer architecture. (d) Cascaded RBM blocks for realizing the proposed deep autoencoder with shared weight update module Programmable normalization circuit: (e) Circuit Schematic, (f) Gain variation vs. OxRAM resistance state. . . . .	106
6.4	(a) Simulated response curve of sigmoid neuron circuit extracted in Cadence Spectre. (b) Simulated response of comparator using C2C distribution based variable threshold. (c) Comparator output voltage $V_{out}$ (d) Current passing through device M2, (e) Switching of M2 device. . . . .	107

6.5	(a) 3-layer deep SDA-1, (b) 5-layer deep SDA-2. Denoising results of 100 corrupted images from the MNIST dataset for: (c) SDA-1 and (d) SDA-2. . . . .	110
6.6	Comparison of DBN fine-tuning performance over epochs for weights initialized using proposed OxRAM DGM and initialization using random numbers for two different network architectures using MNIST dataset: (a) 784x512x10 (b) 784x512x1024x10. . . . .	111
6.7	SDA denoising performance after fine-tuning on MNIST dataset. (a) Original test images, (b) Images with added noise, (d) Denoising using deep SDA architecture (784x512x1024x512x784), (d) Denoising using shallow SDA architecture (784x512x784). . . . .	112
6.8	SDA performance analysis w.r.t hidden layer size. MSE with 5% noise added for network configurations: (a) $784 \times X \times 784$ , and (b) $784 \times 512 \times X \times 512 \times 784$ . Maximum device switching activity in the last layer for network configuration: (c) $784 \times X \times 784$ , and (d) $784 \times 512 \times X \times 512 \times 784$ . Spread shows variation obtained from 10 simulation runs each. . . . .	113
6.9	DBN performance analysis w.r.t hidden layer size. Learning performance in terms of test accuracy for network configurations: (a) $784 \times X \times 10$ (b) $784 \times 512 \times X \times 10$ . . Max. switching activity in the last layer for network configuration: (c) $784 \times X \times 10$ (d) $784 \times 512 \times X \times 10$ . Spread shows variation during 10 simulations performed for the same configuration. . . . .	113

# List of Tables

2.1	Projected specs of state-of-the-art XR devices [26]. . . . .	9
2.2	Learning Performance of the Proposed Network . . . . .	14
2.3	Memory Power Benchmarking for AI Accelerator Pipeline . . . . .	15
2.4	Benchmarking with respect to other reports in literature. . . . .	19
2.5	Estimation of Area Benefits on Systolic Accelerators using Proposed P0 and P1 variants at 7nm node. . . . .	26
2.6	IPS Analysis summary for proposed architectures using PE configuration v2 (64×64). 26	26
3.1	Learning performance using proposed VMM computations methods through simulated BNN utilizing multiple memory windows of fabricated device (averaged over 10 trials). . . . .	42
3.2	Truth table showcasing XNOR operation realized using IMC bitcell. . . . .	51
3.3	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. . . . .	53
3.4	Performance benchmarking with XNOR-based hardware architectures in literature on CIFAR-10 workload. . . . .	60
3.5	Comparison of current work with FeFET-based state-of-the-art synaptic bitcells. 61	61
3.6	Performance Benchmarking w.r.t. recent IMC implementations. . . . .	67
4.1	Description of datasets used in the study. . . . .	73
4.2	Impact of normalization technique, FSL model and backbone on performance of proposed BinDC method for miniImageNet. . . . .	77
4.3	Benchmarking learning results based on proposed method against cosine distance at floating point precision for two workloads. . . . .	78
4.4	Benchmarking of performance for BinDC-based FSL with 4-bit encoding using OxRAM-based IMC implementations. . . . .	80
4.5	Test Accuracy Benchmarking of different precision networks, simulated in this study, for CIFAR-10 dataset. . . . .	86
4.6	Dataset samples used for experiments. . . . .	88
4.7	Performance estimates of inference with multiple architectures for microscopy image analysis. . . . .	92
4.8	Benchmarking performance w.r.to other literature studies for implementation of binarized AlexNet. . . . .	92
5.1	Comparison of implemented logic gates for IMC-based MAC. . . . .	99
5.2	Summary of implemented IMC Macro. . . . .	101
6.1	Proposed SDA performance on MNIST dataset. . . . .	109
6.2	Proposed DBN Performance on MNIST dataset . . . . .	110

---

6.3	Proposed SDA performance on MNIST dataset post-finetuning. . . . .	111
6.4	Proposed DBN performance on MNIST dataset post-finetuning. . . . .	111
6.5	Review of prior approaches for realizing generative models based on memristive devices in hardware using CD. . . . .	112
6.6	Maximum device switching activity for 5 layer DBN (training) . . . . .	113
A.1	Summary of experimental device data used in the thesis. . . . .	121
A.2	Summary of chapter-wise contributions from the thesis. . . . .	121

# List of Algorithms

1	Proposed binarized method for FSL. . . . .	73
2	SBCNN inference algorithm . . . . .	85

# Abbreviations

<b>1R</b>	1 Resistor
<b>1T-1R</b>	1 Transistor-1 Resistor
<b>AI</b>	Artificial Intelligence
<b>ALU</b>	Arithmetic-Logic Unit
<b>ANN</b>	Artificial Neural Network
<b>AR</b>	Augmented Reality
<b>ASIC</b>	Application Specific Integrated Circuit
<b>ATO</b>	Al-doped- $TiO_2$
<b>BCAM</b>	Binary Content-Addressable Memory
<b>BE/TE</b>	Bottom Electrode/Top Electrode
<b>BEOL</b>	Back End of Line
<b>BER</b>	Bit Error Rate
<b>BL/BLB</b>	Bitline/Bitline bar
<b>BNN</b>	Binary Neural Network
<b>BRS</b>	Bipolar Resistive Switches/Switching
<b>C2C/D2D</b>	Cycle-to-Cycle/Device-to-Device
<b>CAM</b>	Content Addressable Memory
<b>CBRAM</b>	Conductive Bridge Random Access Memory
<b>CF</b>	Conductive Filament
<b>CMOS</b>	Complementary Metal Oxide Semiconductor
<b>CNN</b>	Convolutional Neural Networks
<b>CPU</b>	Central Processing Unit
<b>CRS</b>	Complementary Resistive Switches/Switching
<b>CSA</b>	Current Sense Amplifier
<b>CSH</b>	Cryptographically Secure Hash
<b>DBN</b>	Deep Belief Network
<b>DRAM</b>	Dynamic Random Access Memory
<b>DSP</b>	Digital Signal Processor
<b>EAI</b>	Edge Artificial Intelligence
<b>EDP</b>	Energy Delay Product
<b>FA</b>	Full Adder

---

<b>FeFET</b>	Ferroelectric Field Effect Transistor
<b>FeRAM</b>	Ferroelectric Random Access Memory
<b>FLOP</b>	Memory bandwidth per processor floating point operation
<b>FP32</b>	Floating point 32-bit
<b>FPGA</b>	Field Programmable Gate Array
<b>FSL</b>	Few-Shot Learning
<b>GPU</b>	Graphics Processing Unit
<b>HAR</b>	Human Activity Recognition
<b>HBM</b>	High Bandwidth Memory
<b>HD</b>	Hamming Distance
<b>HDD</b>	Hard Disk Drive
<b>HMC</b>	Hybrid Memory Cube
<b>HRS</b>	High Resistance State
<b>HSI</b>	Hyper-Spectral Image
<b>ICP</b>	Inductively-coupled plasma
<b>IMC</b>	In-Memory Computing
<b>IMSS</b>	In-Memory Similarity Search
<b>INT32</b>	Integer 32-bit
<b>IoT</b>	Internet of Things
<b>IRDS</b>	International Roadmap for Devices and Systems
<b>ITRS</b>	International Technology Roadmap for Semiconductors
<b>IV</b>	Current Voltage
<b>LIM</b>	Logic-In-Memory
<b>LMS</b>	Least Mean Square
<b>LPDDR</b>	Low Power Double Data Rate
<b>LRS</b>	Low Resistance State
<b>LSB/MSB</b>	Least Significant Bit/ Most Significant Bit
<b>LUT</b>	Look Up Table
<b>MAC</b>	Multiply-and-Accumulate
<b>MAT</b>	Matrix
<b>MB/GB/TB</b>	Mega Byte/Giga Byte/Tera Byte
<b>MIM</b>	Metal-Insulator-Metal
<b>ML</b>	Machine Learning
<b>MLC</b>	Multi Level Cell/Multi Level Capability
<b>MLP</b>	Multi-Layer Perceptron
<b>MRAM</b>	Magneto-Resistive Random Access Memory
<b>MW</b>	Memory Window
<b>NMOS</b>	N-channel Metal-Oxide Semiconductor
<b>NVIA</b>	Non-Volatile AI Inference Accelerator
<b>NVM</b>	Non-Volatile Memory

---

<b>OxRAM</b>	Oxide based Random Access Memory
<b>PCA</b>	Principal Component Analysis
<b>PCM</b>	Phase Change Memory
<b>PCMO</b>	$\text{Pr}_{0.7}\text{Ca}_{0.3}\text{MnO}_3$
<b>PDK</b>	Process Development Kit
<b>PE-ALD</b>	Plasma-enhanced Atomic Layer Deposition
<b>P-F</b>	Poole-Frenkel
<b>PIM</b>	Processing-In-Memory
<b>PMOS</b>	P-channel Metal-Oxide Semiconductor
<b>PMU</b>	Pulse Measure Unit
<b>PWM</b>	Pulse-Width Modulation
<b>QI</b>	Query Input
<b>QNN</b>	Quantized Neural Network
<b>RBM</b>	Restricted Boltzmann Machine
<b>RNN</b>	Recurrent Neural Network
<b>RRAM/ReRAM</b>	Resistive Random Access Memory
<b>S/H</b>	Sample and Hold
<b>SA</b>	Sense Amplifier
<b>SCA</b>	Side-Channel Attack
<b>SDA</b>	Stacked Denoising Autoencoder
<b>SCLC</b>	Space Charge Limited Conduction
<b>SD</b>	Stored Data
<b>SEM</b>	Scanning Electron Microscope
<b>SHL</b>	Shift Logical Left
<b>SL</b>	Select Line
<b>SLC</b>	Single Level Cell
<b>SMU</b>	Source Measure Unit
<b>SOT-MRAM</b>	Spin Orbit Torque Magnetic Random Access Memory
<b>SRAM</b>	Static Random Access Memory
<b>SSD</b>	Solid State Drive
<b>STT-MRAM</b>	Spin Transfer Torque Magnetic Random Access Memory
<b>TCAM</b>	Ternary Content-Addressable Memory
<b>TDMAHf</b>	Tetrakis(dimethylamido)hafnium
<b>TDMATi</b>	Tetrakis(dimethylamido)titanium
<b>TEM</b>	Transmission Electron Microscopy
<b>TIA</b>	Trans-Impedance Amplifier
<b>TNN</b>	Ternary Neural Networks
<b>t-SNE</b>	t-Distributed Stochastic Neighbour Embedding
<b>VGSOT-MRAM</b>	Voltage-Gated Spin Orbit Torque Magnetic Random Access Memory
<b>VR</b>	Virtual Reality

---

<b><math>V_{TB}</math></b>	Voltage applied at TE-Voltage applied at BE
<b>VVM</b>	Vector Vector Multiplication
<b>VWW</b>	Visual Wake Words
<b>WL/WLB</b>	Wordline/Wordline bar
<b>XR</b>	Extended Reality