

**EMERGING NON-VOLATILE MEMORY  
BASED DEVICE-CIRCUIT CO-DESIGN FOR  
NEUROMORPHIC COMPUTING**

**AHMED SHABAN**



**DEPARTMENT OF ELECTRICAL ENGINEERING  
INDIAN INSTITUTE OF TECHNOLOGY DELHI  
JUNE 2025**

© Indian Institute of Technology Delhi (IITD), New Delhi, 2025

**Emerging non-volatile memory based  
device-circuit co-design for neuromorphic  
computing**

by

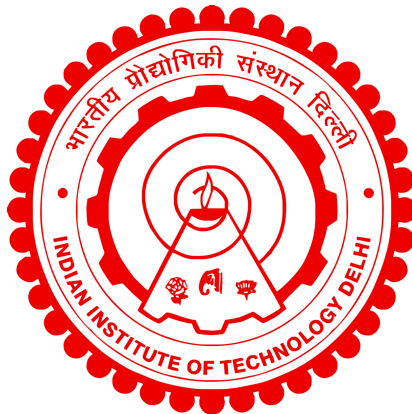
**AHMED SHABAN**

Department of Electrical Engineering

Submitted

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

to the



**INDIAN INSTITUTE OF TECHNOLOGY  
DELHI**

**June 2025**

*Dedicated to my parents*

# Certificate

This is to certify that the thesis entitled “**Emerging non-volatile memory based device-circuit co-design for neuromorphic computing**”, submitted by **Ahmed Shaban**, Research Scholar in the *Department of Electrical Engineering under IIT Delhi-NYCU Taiwan, Joint Doctoral Program, Indian Institute of Technology Delhi, New Delhi, India*, for the award of the joint degree of **Doctor of Philosophy** is a record of the original research work carried out by him under our 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**

Department of Electrical Engineering,  
Indian Institute of Technology Delhi,  
New Delhi, India.

**Prof. Tuo-Hung Hou**

Department of Electrical Engineering  
National Yang Ming Chiao Tung  
University, Hsinchu, Taiwan.

# *Acknowledgements*

This has been a long arduous journey during which I underwent lot of different emotions. However, today I can confidently say that I am proud of myself for undertaking this path of scientific research and innovation that helped me achieve a level of self-satisfaction. My PhD journey was also plagued with COVID-19 outbreak making it one of the toughest times. I am thankful to The Almighty for giving me strength, courage, health and steadfastness to complete my PhD.

I am extremely thankful to my supervisor **Prof. Manan Suri** for supervising me and showing confidence in me. His guidance, support, supervision and encouragement has been instrumental in what I have been able to achieve in my PhD. His attitude and dedication towards research has always been a source of motivation and inspiration to me. He always motivated us to do the best in our research and strive for excellence. He always provided us with the best lab facilities and whatever we needed to carry out our research seamlessly. His critical and constructive feedback always helped in bringing the best out of us and refining the results. I am also thankful to him for giving me the opportunity to enter the IITD-NYCU Joint Doctoral Program that further helped my research. I am also extremely thankful to **Prof. Tuo-Hung Hou (NYCU, Taiwan)** for agreeing to jointly co-supervise me as part of the IITD-NYCU Joint Doctoral Program. The time I spent at his lab - NanoSTLab at NYCU, Taiwan proved to be very beneficial for my PhD. It helped me expand my network, gain experience on working in an international multicultural environment with access to state-of-the-art lab facilities. I am thankful for his consistent support, supervision and guidance throughout my stay in Taiwan and providing the best facilities to carry out quality research.

I am also thankful to my Student Research Committee (SRC) members at IITD- **Prof. Shubendu Bhasin, Prof. Anuj Dhawan, Prof. Pintu Das** for their continuous constructive feedback.

A PhD is never complete without group members, friends and colleagues. I would like to thank all my colleagues at IIT Delhi NVM and neuromorphic hardware research group for being part of this journey. I am thankful to my seniors in the group-**Dr. Ashwani Kumar** (now post-doctoral researcher at UCSD), **Dr.**

**Swatilekha Majumdar** (now at IBM, Germany), **Dr. Sandeep Kaur Kingra** (now at CDIL semiconductors), **Dr. Supriyo Chakraborty** (now post-doctoral researcher at RWTH, Aachen) who provided me with necessary assistance and information I needed during the initial days. I would like to specially thank **Sai Sukruth Bezugam** (now pursuing PhD at UCSB) with whom I enjoyed carrying out significant research in my PhD. I can never forget the long telephonic conversations that we used to have during the COVID-19 lockdown to discuss technical research. I would also like to appreciate his attitude and dedication towards research and his ability to perform teamwork. I am also thankful to **Sufyan Khan** for his assistance in lab during his internship at IITD. I would also like to mention my juniors with whom I shared this journey – **Chithambara Moorthii, Shubham Negi, Narayani Bhatia, Tamoghno Das, Deepak Verma.**

I am also thankful to all my friends that I made in VDTT lab - **Imran Ahmad, Parvez Akhtar, Dr. Vikram Maharshi, Dr. Satish Verma, Amir Saud Khan** for all the stress-releasing talks, gossip and countless tea sessions. I am thankful to Imran bhai for always acting like an elder brother and always being available for any help I needed. I will also remember all the expert cricket analysis discussion that I used to have with Vikram and Amir. I am also thankful for the home cooked lunch that Imran bhai and Amir bhai always used to share with me. I also want to mention my college friends – **Imran Ali Khan, Isaar Ahmad, Dr. Tauheed Mian** with whom I started the PhD journey at IITD as roommates and hostel mates. I would also like to thank **T. R. Ashish** and **Iram Ali** for their assistance with queries related to VDTT server, PDKs and Cadence related issues.

I would also like to thank all my friends and colleagues at NanoST Lab, NYCU, Taiwan who supported and assisted me during my stay in Taiwan. Specifically, I would like to thank **Ming-Chun Hong** for his help and assistance in lab related matters and also useful technical discussions. I would also like to thank **Dr. Chih-Cheng Chang** (post-doctoral researcher at Academia Sinica, Taiwan) for the technical discussions and help with network simulations. I would also like to thank **Shreshtha Gothalyan** who worked jointly with me on a research project. I am also thankful to other colleagues- **Chen-Yi Cho, Dr. Ming-Hung Wu** (now at Macronix), **Kai-Jie Fan, Yi-shin** (now at MediaTek), **Hsin-Yu, Le-Chih Cho, Wei-Tien**

with whom I shared the lab. Further, I would like to acknowledge all my colleagues from IITD who were part of the IITD-NYCU JDP program-who supported each other during the tough times of living abroad away from home in Taiwan. I would like to thank **Shivendra Kumar** for his assistance on our arrival in Taiwan. I will also cherish the memories with **Taslim Khan**, my friend, roommate and cooking partner in Taiwan where we explored lots of new recipes and supported each other. I would also like to mention **Dr. Manu Garg** (now at Silicon Austria Labs) who was always ready with any help and assistance. I am also thankful to **Nadeem Ahamad**, **Iqbal Khan**, and **Aasif Bhutt** who helped and assisted me with my stay in Taiwan whenever needed.

I am also thankful to EE office staff-**Mr. Yatindra Tripathi**, **Mr. Satish Sah**; **Mr. Devendra Goswami**, **Mr. Rakesh Kumar** (Staff, VDTT Lab) for their help and co-ordination with the administrative work. I am also thankful to **Ms. Wei-Ching Kuo** (NanoST Lab admin), **Ms. Yi Ting Lin** (ICST staff), **Ms. Janet Chen** (OIA Staff, NYCU) and IITD coordinators of JDP program- **Dr. Archana Harendra Kumar Trivedi**, **Dr. Sheba Shadrach**, **Ms. Priyanka Kapoor** for their co-operation and support in the administrative work. I would also like to thank **Dr. Sankalp Singh** (now University Program Manager at Synopsis India) who was the coordinator of JDP program at IITD for his initial help and assistance regarding this program and travel.

Lastly, I would like to extend my heartfelt gratitude to my parents **Mr. Shabbir Ahmad** and **Mrs. Nasreen Ahmad** for all the love, support and prayers that made it possible for me to pursue and achieve my goals. I also want to thank my sisters – **Dr. Shamama Firdaus** and **Dr. Shazra Tasneem** for their constant encouragement and support that helped me keep going. I would like to thank each and everyone who was part of this journey and please forgive me if I forgot any name.

**Ahmed Shaban**

# *Abstract*

Efficient hardware implementation and deployment of modern day data-intensive artificial intelligent (AI) workloads has been challenged by the ‘memory wall’ in the conventional computing architecture and ‘power hungry’ acceleration units-graphics processing units (GPUs). This necessitates the exploration of bio-inspired algorithms and their efficient hardware implementation as well as newer architectures to enable energy efficient ubiquitous deployment of AI systems.

In this thesis, we propose to exploit the properties of emerging non-volatile memory (NVM) devices for realization of biologically-plausible spiking neural networks. We propose a double exponential adaptive threshold (DEXAT) spiking neuron model that improves the performance of neuromorphic Recurrent Spiking Neural Networks (RSNNs). We also present a hardware efficient methodology to realize the DEXAT neurons and experimentally demonstrate the DEXAT neuron block using oxide based non-filamentary resistive switching devices. Further, we demonstrate the performance improvements on two spatiotemporal tasks (a) Sequential MNIST (SMNIST) and (b) Speech recognition.

We also exploit the in-memory computing architecture to realize (a) area efficient high precision floating point neural network acceleration and (b) energy efficient and reliable binary neural network acceleration. For floating point neural network acceleration, we exploit a hardware array of filamentary resistive random access memory (RRAM) devices. We map the floating point mantissas of convolutional layer on 1T-1R hardware array, demonstrate the hardware inference and analyze the sources of errors. For binary neural network (BNNs) acceleration, we propose to use a hybrid CMOS-spin orbit torque magnetic random access memory (SOT-MRAM) device based circuit to realize the XNOR operation in BNNs. We also propose a pulse scheme for programming the voltage gated SOT-MRAM device to reduce the write error rate (WERs) thereby increasing the reliability of the accelerator. We also exploit another variant of SOT-MRAM i.e. spin transfer torque assisted SOT-MRAM for realizing an approximate non-volatile content addressable memory that provides high hamming distance tolerance for DNA classification application. Our thesis work demonstrates the potential of emerging NVM devices for enabling the next generation of neural networks and computing architectures.

# सार

आधुनिक डेटा-प्रधान कृत्रिम बुद्धिमत्ता (Artificial Intelligence – AI) कार्यभारों (workloads) के कुशल हार्डवेयर कार्यान्वयन (hardware implementation) और परिनियोजन (deployment) को पारंपरिक संगणन वास्तुकला (conventional computing architecture) में उपस्थित 'मेमोरी वॉल' (memory wall) तथा 'ऊर्जा-खपतकारी' (power hungry) त्वरण इकाइयों (acceleration units) जैसे ग्राफिक्स प्रोसेसिंग यूनिट्स (Graphics Processing Units – GPUs) की चुनौतियों का सामना करना पड़ता है। यह स्थिति जैव-प्रेरित कलन विधियों (bio-inspired algorithms) की खोज, उनके कुशल हार्डवेयर कार्यान्वयन तथा नवीन वास्तुकलाओं (architectures) के अन्वेषण की आवश्यकता को इंगित करती है, जिससे ऊर्जा-कुशल एवं सर्वव्यापी AI प्रणालियों (systems) का परिनियोजन संभव हो सके।

इस शोध प्रबंध (thesis) में, हम जैविक रूप से सुसंगत स्पाइकिंग न्यूरल नेटवर्क्स (spiking neural networks) को साकार करने के लिए उभरती हुई नॉन-वोलेटाइल मेमोरी (non-volatile memory – NVM) युक्तियों (devices) के गुणों का उपयोग करने का प्रस्ताव करते हैं। हम एक डबल एक्स-पोनेंशियल अडैप्टिव थ्रेशोल्ड (Double Exponential Adaptive Threshold – DEXAT) स्पाइकिंग न्यूरॉन मॉडल (spiking neuron model) प्रस्तुत करते हैं, जो न्यूरॉमॉर्फिक रिकरेंट स्पाइकिंग न्यूरल नेटवर्क्स (neuromorphic Recurrent Spiking Neural Networks – RSNNs) के प्रदर्शन में सुधार लाता है। हम DEXAT न्यूरॉनों को साकार करने की एक हार्डवेयर-कुशल विधि भी प्रस्तुत करते हैं और ऑक्साइड आधारित नॉन-फिलामेंटरी रेसिस्टिव स्विचिंग डिवाइसेज़ (oxide based non-filamentary resistive switching devices) का उपयोग करके DEXAT न्यूरॉन ब्लॉक का प्रायोगिक रूप से प्रदर्शन करते हैं। आगे, हम दो स्पेशियो-टेम्पोरल कार्यों (spatiotemporal tasks) पर प्रदर्शन सुधार को प्रदर्शित करते हैं: (क) सीक्वेन्शियल एमएनआईएसटी (Sequential MNIST–SMNIST) और (ख) वाक् पहचान (speech recognition)।

हम इन-मेमोरी कम्प्यूटिंग वास्तुकला (in-memory computing architecture) का उपयोग कर (क) क्षेत्र-कुशल उच्च-सटीकता फ्लोटिंग पॉइंट न्यूरल नेटवर्क त्वरण (area efficient high precision floating point neural network acceleration), तथा (ख) ऊर्जा-कुशल और विश्वसनीय बाइनरी न्यूरल नेटवर्क (binary neural network – BNN) त्वरण को साकार करते हैं। फ्लोटिंग पॉइंट न्यूरल नेटवर्क त्वरण के लिए, हम फिलामेंटरी रेसिस्टिव रैंडम एक्सेस मेमोरी (filamentary Resistive Random Access Memory–RRAM) युक्तियों की एक हार्डवेयर ऐरे (hardware array) का उपयोग करते हैं। हम कन्वोल्यूशनल लेयर (convolutional layer) के फ्लोटिंग पॉइंट मैण्टिसा (mantissas) को 1T–1R हार्डवेयर ऐरे पर मैप करते हैं, हार्डवेयर अनुकरण (hardware inference) का प्रदर्शन करते हैं और त्रुटि के स्रोतों का विश्लेषण करते हैं। बाइनरी न्यूरल नेटवर्क त्वरण के लिए, हम एक हाइब्रिड सीएमओएस-स्पिन ऑर्बिट टॉर्क मैग्नेटिक रैंडम एक्सेस मेमोरी (CMOS–spin orbit torque Magnetic Random

Access Memory –SOT–MRAM) आधारित परिपथ (circuit) का प्रस्ताव करते हैं, जो BNNs में XNOR क्रिया को साकार करता है। हम वोल्टेज–गेटेड SOT–MRAM डिवाइस को प्रोग्राम करने के लिए एक पल्स योजना (pulse scheme) भी प्रस्तावित करते हैं, जो लेखन त्रुटि दर (write error rate – WER) को कम करती है और इस प्रकार त्वरणक (accelerator) की विश्वसनीयता में वृद्धि करती है। हम SOT–MRAM के एक अन्य प्रकार — स्पिन ट्रांसफर टॉर्क असिस्टेड SOT–MRAM (spin transfer torque assisted SOT–MRAM) — का भी उपयोग करते हैं ताकि एक सन्निकट (approximate) नॉन–वोलेटाइल कंटेंट एड्रेसेबल मेमोरी (non–volatile content addressable memory – CAM) को साकार किया जा सके, जो डीएनए वर्गीकरण (DNA classification) अनुप्रयोग के लिए उच्च हैमिंग दूरी सहिष्णुता (high hamming distance tolerance) प्रदान करता है। हमारा शोधकार्य यह प्रदर्शित करता है कि उभरती हुई नॉन–वोलेटाइल मेमोरी युक्तियाँ (emerging NVM devices) अगली पीढ़ी के न्यूरल नेटवर्क्स और संगणन वास्तुकलाओं को सक्षम करने की दिशा में अत्यधिक संभावनाएं रखती हैं।

# Contents

<b>Certificate</b>	<b>i</b>
<b>Acknowledgements</b>	<b>ii</b>
<b>Abstract</b>	<b>v</b>
<b>Contents</b>	<b>vii</b>
<b>List of Figures</b>	<b>xii</b>
<b>List of Tables</b>	<b>xx</b>
<b>Abbreviations</b>	<b>xxi</b>
<b>Symbols</b>	<b>xxii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Beyond Von-Neumann Computing . . . . .	3
1.3 Thesis Organisation and Contributions . . . . .	6
1.3.1 Objective of the thesis . . . . .	6
1.3.2 Key contribution of the thesis . . . . .	6
1.3.3 Thesis Organization . . . . .	8
<b>2 Background concepts and literature</b>	<b>10</b>
2.1 Resistive Random Access Memory . . . . .	11
2.1.1 Filamentary RRAM . . . . .	11
2.1.2 Interfacial RRAM . . . . .	13
2.2 Spin-Orbit Torque MRAM device . . . . .	14
2.3 Applications . . . . .	16

---

2.3.1	Spiking Neural Network . . . . .	16
2.3.2	Artificial neural network acceleration . . . . .	17
2.3.3	Approximate content addressable memory . . . . .	20
<b>3</b>	<b>An Adaptive Threshold Neuron for Recurrent SNN with Nanode-</b>	
	<b>vice Hardware Implementation</b> . . . . .	<b>22</b>
3.1	Introduction . . . . .	23
3.2	Adaptive threshold neuron model . . . . .	25
3.3	Hardware Implementation . . . . .	30
3.3.1	Experimental setup for basic characterization . . . . .	30
3.3.2	Characterization and circuit testing . . . . .	32
3.4	LSNN learning with proposed DEXAT neuron . . . . .	40
3.4.1	Simulation Methods . . . . .	40
3.4.2	Experimental setup for speech recognition . . . . .	42
3.4.3	Resultant device variability extraction methodology . . . . .	44
3.4.4	Results . . . . .	45
3.5	Performance estimation of proposed DEXAT neurons . . . . .	52
3.6	Summary . . . . .	53
<b>4</b>	<b>Block Floating Point Neural Network Acceleration on RRAM Com-</b>	
	<b>pute in-Memory Hardware</b> . . . . .	<b>55</b>
4.1	Introduction . . . . .	56
4.2	Hardware mapping strategy for FP8 inference . . . . .	58
4.2.1	Background . . . . .	58
4.2.2	Weight mantissa hardware mapping . . . . .	60
4.3	RRAM array device characterization . . . . .	61
4.3.1	Multi Level Cell (MLC) retention . . . . .	62
4.3.2	D2D variations in MLC states . . . . .	62
4.3.3	Effect of read voltage on SLC CDFs . . . . .	64
4.4	FP8 inference . . . . .	65
4.4.1	Simulation methodology . . . . .	65
4.4.2	Hardware FP8 Inference . . . . .	66
4.4.2.1	Error analysis . . . . .	66
4.4.2.2	Sparsity analysis and word line parallelism . . . . .	69
4.4.2.3	Energy Analysis . . . . .	70
4.5	System level performance analysis . . . . .	71
4.6	Summary . . . . .	74
<b>5</b>	<b>SOT-MRAM Based Energy Efficient and Reliable Binary Neural</b>	
	<b>Network Acceleration</b> . . . . .	<b>75</b>
5.1	Introduction . . . . .	76
5.2	Proposed XNOR BNN cell . . . . .	79

5.2.1	4T-2R EB-VGSOT / STT-SOT based BNN cell . . . . .	79
5.2.2	Sensing of BNN-XNOR cell . . . . .	81
5.2.3	SOT device model characterization . . . . .	82
5.2.4	Functional verification of proposed 4T-2R cell . . . . .	84
5.3	Proposed pulse scheme for VGSOT-Write . . . . .	85
5.4	Performance analysis . . . . .	89
5.4.1	BER Analysis . . . . .	89
5.4.2	Energy consumption . . . . .	90
5.4.3	System level performance analysis . . . . .	92
5.5	Summary . . . . .	93
<b>6</b>	<b>SOT-MRAM based Approximate Content Addressable Memory for DNA Classification</b>	<b>94</b>
6.1	Introduction . . . . .	95
6.2	Proposed ANV-CAM . . . . .	97
6.2.1	Proposed 5T-1R CAM cell . . . . .	97
6.2.1.1	Data storage . . . . .	97
6.2.1.2	Data search . . . . .	97
6.2.2	Device choice for proposed cell . . . . .	99
6.3	Simulation results . . . . .	102
6.3.1	Functional simulation . . . . .	105
6.3.2	Performance analysis . . . . .	105
6.3.2.1	Effect of process variations . . . . .	106
6.3.2.2	Sensitivity analysis . . . . .	107
6.3.3	Design space for ANV-CAM . . . . .	110
6.4	Summary . . . . .	110
<b>7</b>	<b>Conclusion</b>	<b>112</b>
	<b>Scope for Future Work</b>	<b>114</b>
<b>A</b>	<b>Detailed explanation for better performance of DEXAT model</b>	<b>115</b>
<b>B</b>	<b>Fabrication details of interfacial OxRAM device</b>	<b>123</b>
<b>C</b>	<b>Endurance Analysis for OxRAM based hardware DEXAT neuron</b>	<b>125</b>
	<b>Bibliography</b>	<b>128</b>

---

<b>List of Publications</b>	<b>152</b>
<b>Curriculum Vitae</b>	<b>154</b>

# List of Figures

1.1	Different eras of scaling over the years beginning with geometric scaling followed by effective scaling and the futuristic hyper-scaling enabled by newer architectures, non-volatile memory devices and system level integration (Image adapted from [9]). . . . .	2
1.2	Increase in number of neural network parameters over the years to handle increasingly complex tasks (Image adapted from [11], [12]). . .	3
1.3	(a) Computing power demand has increased over the last four decades (shown in terms of peta FLOPS/days). (b) Performance (speed, power) of neural network accelerators on GPU, ASIC, FPGA platforms. (Image adapted from [13]). . . . .	4
1.4	(a) Diverse applications of spiking neural networks [14] and (b) Future roadmap to overcome the memory wall [15]. . . . .	5
2.1	Filamentary RRAM device. (a) Pristine or ‘As-formed’ state of device with no filament and high resistance. The device is first formed applying a high forming voltage bring the device in low resistance state [20]. After forming, normal SET and RESET operations can be carried out to switch the device from LRS to HRS and vice-versa. (b) DC I-V curve for filamentary device showing forming, SET and RESET operation [114]. . . . .	12
2.2	(a) SET and RESET process in interfacial RRAM device due to the movement of oxygen ions close to the Ta/ $TaO_x$ interface [27]. (b) DC I-V curve showing SET and RESET process for an interfacial RRAM device with Ni/ $HfO_2$ / Al doped $TiO_2/TiN$ stack. . . . .	13
2.3	(a) VCMA assisted SOT-MTJ or Voltage gated-SOT (VGSOT) device. A VCMA voltage is applied on the top terminal of MTJ to lower energy barrier and assist the SOT write process. (b) STT-assisted SOT-MTJ device with a STT voltage applied on top terminal. In this case, STT current through the MTJ is used to assist the SOT write current. (c) 2T-1R bitcell of a SOT-MTJ device with read and write access transistors. . . . .	15

2.4	Sketch of a biological neuron showing inter-neuron transmission of information via synapses. A SNN showing integration of incoming spike train weighted through synaptic connections. LIF behaviour can be seen. A spike is generated when the threshold is crossed [41].	17
2.5	(a) Deep neural networks consisting of multiple CONV layers to extract input image features followed by fully connected layers. (b) DNN inference acceleration performed by mapping individual layers on multiple NVM based crossbar arrays. Network weights are mapped on the array and inputs are applied in the form of voltage pulse. Multiplication between input and weight is performed across the column governed by physical Ohm's and Kirchoff's law [47], [48].	18
2.6	Illustration of BNN training and inference process [54]. . . . .	19
2.7	(a) A CAM/TCAM array showing search lines for each cell shared across a column and matchline shared across a row and connected to sense amplifier. Exact match CAM/TCAM cell can be built using conventional CMOS transistors utilizing two SRAM cells. Other alternative include exploiting non-volatile memory device like MTJs to build the cell. (b) Block diagram showing the basic steps involved on using approximate CAM for DNA classification application. . . . .	21
3.1	Qualitative description of neuron models and network architecture for LSNN tasks. (a) Model in which neuron's threshold decays exponentially with a single time constant and (b) Proposed DEXAT model used in this work where threshold decays with two time constants. (c) LSNN Network used for STORE-RECALL task and (d) Learning curves of LSNN with ALIF and DEXAT adaptive neurons on a STORE-RECALL task for a working memory requirement of 1200 ms. . . . .	26
3.2	Performance analysis of ALIF and DEXAT neurons. (a) Performance of LSNN on STORE-RECALL task with ALIF neurons and $\tau_a$ varied for different working memory requirements. Performance of LSNN on STORE-RECALL task with DEXAT neurons and (b) $\tau_{a1}$ fixed at 30 ms and $\tau_{a2}$ varied for different working memory requirements (c) $\tau_{a2}$ fixed at 600 ms and $\tau_{a1}$ varied for different working memory requirements (d) $\tau_{a1}=30$ ms, $\tau_{a2}=600$ ms and ratio $\beta_2/\beta_1$ varied for different working memory requirements (Shorter and yellower bars denote faster convergence and lower decision error, taller and redder bars denote non-convergence and higher decision error). (e) Radar plot showing design space for tuning parameters in DEXAT model for obtaining best performance (For optimum performance, the region formed by the three set of points should lie within the shaded region).	28

3.3	Experimental setup used for sample sequences (i.e. S1 and S2) testing to demonstrate DEXAT behaviour. Experimental setup showing neuron circuit on General purpose board, Digital Storage Oscilloscope (DSO) and external power supply. . . . .	31
3.4	DEXAT behaviour extraction from non-filamentary OxRAM devices. (a)-(d) Extracted LTP-LTD characteristics of Pt/PCMO/N-doped TiN/Pt, Pt/PCMO/ TiN/Pt, Mo/TiOx/TiN and Ni/HfO <sub>2</sub> /Al doped TiO <sub>2</sub> /TiN devices respectively. (e)-(h) Normalized and interpolated LTD conductance curves corresponding to (a)-(d) respectively fitted with DEXAT neuron equations. Extracted DEXAT neuron parameter values are indicated inside respective curves. . . . .	34
3.5	Proposed DEXAT neuron threshold modulator. (a) Functional block diagram of proposed adaptive neuron and associated control signals generated by pulse generator in a spike event and (b),(c),(d) SET, RESET and IDLE modes in our proposed 6T-1R threshold modulator circuit respectively. . . . .	36
3.6	DEXAT threshold modulator experiment. (a) Output spikes of the neuron denoting a sequence S1 of four firing events. (b) Input pulses applied to OxRAM device in threshold modulator circuit initiating threshold modulation in case of a spike event. A 3V, 30 ms SET pulse ( $V_{DD\_SET} = 3$ V) applied at top electrode (TE) is followed by a 3V, 50 ms RESET pulse ( $V_{DD\_RESET} = 3$ V) applied at bottom electrode (BE) after first spiking event. A voltage of 2 V ( $V_{DD\_IDLE}$ ) is applied on BE after threshold voltage saturates in absence of firing. (Voltages applied on BE are shown negative only for representation) (c) Experimentally observed threshold voltage of hardware neuron showing increase in threshold at each spike event and subsequent decay afterwards. . . . .	37
3.7	Hardware DEXAT neuron parameter extraction. (a),(c) Experimentally observed adaptive threshold behaviour for spike sequence S1 and S2 respectively showing increment in threshold voltage ( $\Delta$ ) at each spiking event and fitting using DEXAT mode. Extracted parameters for both sequences S1 and S2 are shown in inset of graphs. (b),(d) Variation in threshold voltage increment ( $\Delta$ ) for sequence S1 and S2 respectively at each spike event due to C2C device variability and decrease in threshold voltage increment ( $\Delta$ ) with subsequent spikes due to device behaviour. Error bars represent standard deviation. . . . .	39
3.8	Experimental setup for full end to end speech recognition task. Schematic of the designed experimental setup showing communication between different modules. . . . .	43
3.9	Real time experimental setup showing fabricated PCB for end to end speech recognition task. Fabricated PCBs showing parent board with required blocks like ADC, microcontroller, DAC and interface board with DEXAT neurons. . . . .	43

- 3.10 Experimental variability measurements for adaptive threshold voltage for sample sequence S1 on Ni/  $HfO_2$  /Al doped  $TiO_2$  / TiN device. (a) C2C variability, showing mean and standard deviation for each point. (b) C2C+D2D variability (obtained by cycling multiple devices multiple times), showing mean and standard deviation for each point. Solid blue dots represent the mean of each threshold voltage point and shaded grey region denotes the standard deviation. Variance for C2C+D2D case is higher than variance for C2C only case. 45
- 3.11 Simulated DEXAT neuron adaptive threshold cycles using resultant variability parameter. Each curve shows 10,000 simulated neuron traces for sequence S1 capturing effect of both C2C+D2D variability. 10,000 traces are representative of 100 neurons for 100 cycles (or X neurons for 10,000 / X cycles). (a)  $\eta_r = 10$  %, (b)  $\eta_r = 30$  % and (c)  $\eta_r = 40$  %. . . . . 46
- 3.12 Dynamics of LSNN network for sequential MNIST obtained after training the network. (a) Spike rasters from input neurons. (b), (c) Spike rasters of LIF and DEXAT neurons respectively. (d) Dynamics of the firing thresholds of adaptive neurons. (e) Activation of softmax readout neurons. (f) Input test image presented sequentially pixel by pixel row-wise. The network gradually infers the image over a duration of 840 ms. . . . . 47
- 3.13 Benchmarking of DEXAT based LSNN. (a) Test accuracy comparison for classifying SMNIST using different neuron cells. Numbers in brackets indicate number of cells. (b) Test accuracy comparison on SMNIST for parameters extracted from different device stacks and full DEXAT experiments on  $HfO_2$  /  $TiO_2$  device. (c) Test accuracy comparison on SMNIST with different resultant variabilities for  $HfO_2$  /  $TiO_2$  device. . . . . 48
- 3.14 LSNN classification performance on GSC dataset. (a) Test accuracy vs network dimension. DEXAT based LSNN achieves higher accuracy compared to ALIF based LSNN even for significantly fewer hidden layer neurons. (State-of-the-art-accuracy [94] is achieved even with 51% lesser neurons). All binary class simulations are performed after training the LSNN on a reduced 2-class GSC dataset. DEXAT binary class result for network size (10-10) corresponds to the end-to-end speech experiment as described in the simulation methods in the Chapter. Also, it can be seen that accuracy increases with increasing network size for both binary and 12-class GSC. (b) Test accuracy comparison for a 300-300 sized ALIF vs DEXAT LSNN. For all GSC simulations, DEXAT neurons are used based on the hardware extracted parameters on  $HfO_2/TiO_2$  device. . . . . 49

3.15	Real-time recognition of a spoken input speech sample using DEXAT based LSNN. (a) Normalized and pre-processed real-time input speech sample. (b) Input neurons spike rasters. (c) Hidden layer LIF and DEXAT neurons spike rasters. Spike rasters for neuron no. 1 to 10 shown in blue are for LIF neurons, spike rasters (11 to 12) in green and (13 to 20) in magenta are for hardware and software DEXAT neurons respectively. (d) Adaptive thresholds of software DEXAT neurons with variability included. (e) Adaptive thresholds of hardware DEXAT neurons. $V_{DD\_SET}$ of 3.5 V and $V_{DD\_RESET}$ of 3 V is taken in experiments. (f) Decision output plots for the two classes evolving with time and generating correct result corresponding to input sample at the end. . . . .	50
3.16	Energy consumption per inference task for SMNIST task. 60000 sample images are used for inference. Average DEXAT threshold modulation energy per neuron in a single inference is calculated and plotted for each inference task involving 100 DEXAT neurons in the hidden layer of LSNN. . . . .	53
4.1	(a) Shift-align scheme to obtain block floating point tensors. (b) SLC and MLC 1T-1R bitcell based mantissa storage and partial MAC calculation method. Weight mantissas are stored using 4-bits in E4M3 FP8 format. (c) CONV layer weight mantissa mapping strategy on the 1T-1R array. Weight mantissas are stored in differential pair method to obtain positive and negative weights. A $n \times n$ kernel with depth 'k' is mapped using $n \times n \times k$ rows on the array. . . . .	59
4.2	Hardware setup with 1T-1R RRAM CIM array with other on-chip peripherals and ADCs on board controlled using a host computer. . .	61
4.3	(a) Multiple conductance states obtained in our filamentary RRAM device using incremental $V_{WLSET}$ voltage. Conductance states obtained at low $V_{WLSET}$ are unstable and tend to decay to HRS state. $V_{WLRead}$ of 1 V is used that limits the maximum LRS state to 20 $\mu$ A. (b) Linearly separated MLC states obtained by extending the conductance range using a high $V_{WLSET}$ of 1.5 V. . . . .	63
4.4	(a) Cumulative distribution function (CDF) of MLC states obtained using write-verify scheme on 200 devices on the 1T-1R array. Lower conductance state S1 has a higher $\sigma/\mu$ and also shows relaxation effect post programming. state S2 and S3 are stable and a narrow distribution can be obtained. (b) CDFs of 144 devices in LRS state '1' at different $V_{WLRead}$ . CDF distribution becomes narrower (i.e. $\sigma/\mu$ reduces) as $V_{WLRead}$ is increased. . . . .	64

4.5	Bit-wise simulation framework for emulating hardware inference. (a) Graphical description showing the methodology for finding maximum exponents and performing MAC accordingly. (b) Input mantissa is presented bit-wise sequentially in multiple cycles and MAC is performed with weight mantissas. . . . .	66
4.6	RRAM hardware mapped layer-1 ( $CONV_1$ ) shift-aligned weight mantissas mapped using (a) SLC array of $27 \times 8$ (b) MLC array of $27 \times 4$	67
4.7	FP8 hardware inference error analysis showing MSE between ideal (software) and hardware obtained output maps for $CONV_1$ layer with $WL_{parallel} = 27$ for a) SLC mapping at $V_{WLRead} = 0.8$ V b) SLC mapping at $V_{WLRead} = 0.95$ V c) SLC mapping at $V_{WLRead} = 1.5$ V and d) MLC mapping at $V_{WLRead} = 1.5$ V . . . . .	68
4.8	(a) RRAM hardware mapped layer-2 ( $CONV_2$ ) shift-aligned weight mantissas mapped on $228 \times 8$ array using SLC devices. Partial MAC distribution during hardware inference on a CIFAR-10 image at $V_{WLRead}$ of 0.95 V for b) $WL_{parallel} = 72$ c) $WL_{parallel} = 144$ d) $WL_{parallel} = 288$ . Partial MAC distribution shows that a high amount of sparsity is present in input and weights. Mean square error (MSE) of output maps for each case are shown in each graph. . . . .	70
4.9	Energy per operation normalized w.r.t. (a) Energy at $V_{WLRead}$ of 0.8 V for each sparsity value (b) Energy at sparsity of 80 % at each $V_{WLRead}$ . . . . .	71
4.10	VGG-9, CIFAR-10 network analysis for INT4, INT8 and FP8 formats showing (a) $V_{WLRead}$ versus test accuracy analysis using SLC mapped inference dataset (b) D2D variation sweep versus test accuracy analysis.	72
4.11	D2D variation versus test accuracy analysis for MLC mapped inference using VGG-9 network and CIFAR-10 dataset for (a) FP8 format and (b) INT4 format and (c) INT8 format. . . . .	73
5.1	VGSOT based XNOR cells in existing literature (a) 5T-2R cell [129] (b) 2T-2R cell [125] (c) 11T-9R cell [124] and (d) Proposed EB-VGSOT / STT-SOT based 4T-2R XNOR cell. . . . .	77
5.2	(a) Proposed XNOR cell coupled with precharge sense amplifier circuit. (b)-(d) Steps showing the computation of XNOR in write based scheme in our design. (e) Current-encoded inputs generated through write driver, corresponding device states and final output of circuit implementing XNOR operation. . . . .	80
5.3	STT-SOT device calibrated switching curve against micromagnetic curve in [133] . . . . .	83
5.4	(a), (b) Switching in EB-VGSOT device for a VCMA voltage of 1.2 V and 0.8 V with 1 ns SOT pulse (c), (d) Switching in STT-SOT device using a SOT current pulse of 0.3 ns and a pair of STT voltages of 3ns.	84

5.5	Functional verification of the proposed 4T-2R XNOR cell with sense amplifier circuit for an input current ( $I_{SL}$ ) corresponding to (a) (0,0) (b) (0,1), (1,0) and (c) (1,1). . . . .	85
5.6	(a),(b) Switching probability ( $P_{SW}$ ) curve for AP to P switching at different post-SOT VCMA pulse duration using proposed pulse scheme as shown in the insets. 1000 MC runs are performed for each point including the effect of thermal fluctuations. . . . .	87
5.7	Monte Carlo simulations (1000 runs for each point) for AP to P switching using proposed pulse scheme at different values of $T_{rise}$ , $T_{fall}$ of VCMA and SOT current pulse for (a),(b) a post-SOT VCMA pulse of 1.2 V applied for 0.3 ns with an in-plane field of 75 Oe (c), (d) a post-SOT VCMA pulse of 1.2 V applied for 0.4 ns with an in-plane field of 100 Oe. For each measured point thermal noise is included along with process variations. . . . .	88
5.8	Monte Carlo simulations for calculating BER in (a) EB-VGSOT device and (b) STT-SOT device based XNOR cell for different process variation (PV) of oxide thickness $t_{ox}$ , free layer thickness $t_{sl}$ . Each point in the figure is obtained through 1000 MC runs by including thermal noise and PV. Moving across X-axis, the BER values obtained are representative of D2D variations. 0 % variation case considers only the effect of thermal noise. . . . .	90
5.9	(a) EB-VGSOT / STT-SOT based XNOR cell based computing array for write-based BNN architecture.(b) Asymmetric BER vs test accuracy of VGG-9 network. BER pair (x,y) in figure signifies x % BER for (0,1)/(1,0) combination and y % BER for (1,1) combination. . . . .	92
6.1	(a) Proposed 5T-1R ANV-CAM cell using STT-SOT MTJ device and CMOS transistors. Stacked transistors MN3 and MN4 (high $V_t$ ) are exploited for matchline discharge. (b) Search ‘1’ operation path and (c) Search ‘0’ operation path. . . . .	98
6.2	A 4x4 array using our proposed 5T-1R cell showing possible DNA encoding and storage of reference DNA base in CAM row for approximate search operation. . . . .	100
6.3	Write operation in STT-SOT device for (a) Anti-parallel (AP) to Parallel (P) switching and (b) Parallel (P) to Anti-Parallel (AP) switching. A small STT current is needed to support the switching. . . . .	102
6.4	Functional simulation of proposed ANV-CAM cell. (a) A 16 bit word topology used for the simulation is shown. (b) Precharge signal that is initially kept low to precharge the matchline. (c) Search lines that are driven to perform a Search ‘1’ operation and (d) Matchline showing the discharge in case of different bit mismatches / hamming distance. As long as the ML voltage is above the $V_{evalthr}$ , an approximate match can be obtained. . . . .	104

6.5	Distribution of node voltage ‘X’ after performing 1000 MC runs for (a) Search 1 and (b) Search 0 operation. . . . .	107
6.6	Sensitivity analysis of the proposed ANV-CAM cell using a 16 bit word. 1000 MC runs are performed including local variations for CMOS and SOT-MTJ device for each value of hamming distance at different global process corners (a) TT corner (b) SS corner and (c) FF corner. . . . .	107
6.7	Sensitivity analysis performed using a 256 bit ANV-CAM word for different $V_{evalthr}$ of (a) 0.1 V (b) 0.2 V (c) 0.3 V. 1000 MC simulations are performed including local variations for CMOS and SOT-MTJ device for each hamming distance value. A large hamming distance tolerance can be obtained by choosing suitable parameters. . . . .	108
A.1	. Pseudo-derivative behaviour comparison for DEXAT and ALIF neurons. A constant current of 50 mA is injected as input to all the three neurons, ALIF1 (with a small time constant) and ALIF2 (with a large time constant) and a DEXAT neuron (with a small and a large time constant). Neuron parameters are listed inside the graph. (a) Output spikes corresponding to input current (b) Membrane potential evolution (c) Adaptive threshold behaviour and (d) Pseudo-derivative magnitude behaviour for the three neurons. . . . .	119
A.2	Comparison of synaptic recurrent weight matrix evolution between ALIF and DEXAT based LSNN for SMNIST application. Recurrent weight distribution for ALIF based LSNN during training: (a) Initial random, (b) After 15000 iterations, and (c) after 25000 iterations. Synaptic recurrent weight distribution in DEXAT based LSNN during training: (d) Initial random, (e) After 15000 iterations, and (f) after 25000 iterations. . . . .	120
C.1	Dynamic modeling of degradation of adaptive threshold decay behaviour. Device $R_{on} / R_{off}$ ratio degrades with an increasing number of programming cycles. Scaling of $R_{on}/R_{off}$ ratio is based on characterization shown in [152]. Note overall window squeezes with cycling. . . . .	126
C.2	Post cycling DEXAT behaviour for multiple spike events. Simulated DEXAT neuron behaviour for multiple spike events after injecting dynamic cycle-lifetime based threshold decay curves for respective hidden layer neurons in the LSNN simulations. After a large number of cycles almost no DEXAT action is observed. . . . .	127

# List of Tables

3.1	Comparison with other implementations of adaptive neuron models. .	49
3.2	Estimated energy benchmarking with other relevant adaptive neurons.	51
5.1	EB-VGSOT and STT-SOT device parameters. . . . .	82
5.2	Comparison with other SOT-MRAM based XNOR designs . . . . .	91
6.1	Search operation . . . . .	98
6.2	STT-SOT MTJ device parameters . . . . .	103
6.3	Benchmarking with other approximate CAMs . . . . .	109
C.1	Estimation of programming cycle requirement/ per DEXAT neuron depending on network size and application. Table also shows the approximate number of inferences before device breakdown. . . . .	127
C.2	Test accuracy degradation with cycling for SMNIST task (each run corresponds to inference on 10000 test images using LSNN with 120 LIF and 100 DEXAT neurons). . . . .	127

# Abbreviations

<b>MOSFET</b>	<b>Metal Oxide Semiconductor Field Effect Transistor</b>
<b>RRAM</b>	<b>Resistive Random Access Memory</b>
<b>SOT</b>	<b>Spin Orbit Torque</b>
<b>STT</b>	<b>Spin Transfer Torque</b>
<b>MRAM</b>	<b>Magnetic Random Access Memory</b>
<b>SNN</b>	<b>Spiking Neural Network</b>
<b>ANN</b>	<b>Artificial Neural Network</b>
<b>DNN</b>	<b>Deep Neural Network</b>
<b>CNN</b>	<b>Convolutional Neural Network</b>
<b>BNN</b>	<b>Binary Neural Network</b>
<b>LIF</b>	<b>Leaky Integrate Fire</b>
<b>CAM</b>	<b>Content Addressable Memory</b>
<b>IMC</b>	<b>In Memory Computing</b>
<b>SLC</b>	<b>Single Level Cell</b>
<b>MLC</b>	<b>Multi Level Cell</b>
<b>LTP</b>	<b>Long Term Potentiation</b>
<b>LTD</b>	<b>Long Term Depression</b>
<b>VCMA</b>	<b>Voltage Controlled Magnetic Anisotropy</b>

# Symbols

$\tau_a$	adaptation time constant of spiking neuron
$\beta$	scaling factor of spiking neuron
$\alpha$	gilbert damping constant
$\gamma$	gyromagnetic ratio
$\Delta$	thermal energy barrier