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

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. Shubhendu 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 ग्राफ़िक्स 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 न्यूरोमॉर्फिक रिकरेंट स्पाइकिंग न्यूरल नेटवर्क्स (neuromorphic Recurrent Spiking Neural Networks – RSNNs) के प्रदर्शन में सु–धार 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 ने–टवर्क 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

List of Figures

List of Tables

Abbreviations

MOSFET	**M**etal **O**xide **S**emiconductor **F**ield **E**ffect **T**ransistor
RRAM	**R**esistive **R**andom **A**ccess **M**emory
SOT	**S**pin **O**rbit **T**orque
STT	**S**pin **T**ransfer **T**orque
MRAM	**M**agnetic **R**andom **A**ccess **M**emory
SNN	**S**piking **N**eural **N**etwork
ANN	**A**rtificial **N**eural **N**etwork
DNN	**D**eep **N**eural **N**etwork
CNN	**C**onvolutional **N**eural **N**etwork
BNN	**B**inary **N**eural **N**etwork
LIF	**L**eaky **I**ntegrate **F**ire
CAM	**C**ontent **A**ddressable **M**emory
IMC	**I**n **M**emory **C**omputing
SLC	**S**ingle **L**evel **C**ell
MLC	**M**ulti **L**evel **C**ell
LTP	**L**ong **T**erm **P**otentiation
LTD	**L**ong **T**erm **D**epression
VCMA	**V**oltage **C**ontrolled **M**agnetic **A**nisotropy

Symbols

τ_a	adaptation time constant of spiking neuron
β	scaling factor of spiking neuron
α	gilbert damping constant
γ	gyromagnetic ratio
Δ	thermal energy barrier