EXPLORING COMPUTING APPLICATIONS WITH NON-VOLATILE MEMORY

SANDEEP KAUR KINGRA

DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY DELHI

MARCH 2023

EXPLORING COMPUTING APPLICATIONS WITH NON-VOLATILE MEMORY

by

SANDEEP KAUR KINGRA

Department of Electrical Engineering

Submitted

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

to the

INDIAN INSTITUTE OF TECHNOLOGY DELHI

MARCH 2023

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

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

Sri Raag Guru Amar Das

Jio Andhhaerai Dheepak Baaleeai Thio Gur Giaan Agiaan Thajaae //2//

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

Dedicated to the Almighty and all my Respected Teachers

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

Certificate

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

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

Prof. Manan Suri
Associate Professor,
Department of Electrical Engineering,
Indian Institute of Technology Delhi.
Date:

Acknowledgements

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

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

I would also like to thank the members of my student research committee (SRC), **Prof. Anuj Dhawan**, **Prof. Bhaskar Mitra** and **Prof. Preeti Ranjan Panda**. They provided me incredible technical and moral support and countless pieces of critical advice throughout my PhD journey.
I would also like to thank **Prof. Mamidala Jagadesh Kumar**, **Prof. Bodh Raj Mehta**, and **Dr. Kaushik Saha**, for teaching me valuable and relevant courses during that further strengthened the foundation for my research work. I am also grateful to our esteemed institute for providing me with the facilities to carry out my research.

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

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

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

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

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

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

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

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

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

Sandeep Kaur Kingra

Abstract

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

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

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

सार

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

Contents

List of Figures

List of Tables

List of Algorithms

Abbreviations

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

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

NV-SRAM	Non-Volatile Static Random Access Memory
OxRAM	Oxide based Random Access Memory
PCA	Principal Component Analysis
PCM	Phase Change Memory
PCMO	$Pr_{0.7}Ca_0.3MnO_3$
PDK	Process Development Kit
PE-ALD	Plasma-enhanced Atomic Layer Deposition
P-F	Poole-Frenkel
PIM	Processing-In-Memory
PMOS	P-channel Metal-Oxide Semiconductor
PMU	Pulse Measure Unit
PWM	Pulse-Width Modulation
QI	Query Input
QNN	Quantized Neural Network
RNN	Recurrent Neural Network
RRAM/ReRAM	Resistive Random Access Memory
S/H	Sample and Hold
SA	Sense Amplifier
SCA	Side-Channel Attack
SCLC	Space Charge Limited Conduction
SD	Stored Data
SEM	Scanning Electron Microscope
SHA-3	Secure Hash Algorithm-3
SHL	Shift Logical Left
SL	Select Line
SLC	Single Level Cell
SLIM	Simultaneous Logic-In-Memory
SMU	Source Measure Unit
SRAM	Static Random Access Memory
SSD	Solid State Drive
STT-MRAM	Spin Transfer Torque Magnetic Random Access Memory
TCAM	Ternary Content-Addressable Memory
TDMAHf	Tetrakis(dimethylamido)hafnium
TDMATi	Tetrakis(dimethylamido)titanium
TEM	Transmission Electron Microscopy
TIA	Trans-Impedance Amplifier
TNN	Ternary Neural Networks
t-SNE	t-Distributed Stochastic Neighbour Embedding
V_{TB}	Voltage applied at TE-Voltage applied at BE
VVM	Vector Vector Multiplication

VWW	Visual Wake Words
WL/WLB	Wordline/Wordline bar