NON-VOLATILE MEMORY-CENTRIC COMPUTING ADVANCES

VIVEK KAMALKANT PARMAR

DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY DELHI

JULY 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 Penkovskyi** (currently AI R&D Project Mananger, 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↔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 (ArtificiaI 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

List of Figures

List of Tables

List of Algorithms

Abbreviations

1R	1 Resistor
1T-1R	1 Transistor-1 Resistor
AI	Artificial Intelligence
ALU	Arithmetic-Logic Unit
ANN	Artificial Neural Network
AR	Augmented Reality
ASIC	Application Specific Integrated Circuit
ATO	Al-doped-TiO_2
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
DBN	Deep Belief Network
DRAM	Dynamic Random Access Memory
DSP	Digital Signal Processor
EAI	Edge Artificial Intelligence
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
FSL	Few-Shot Learning
GPU	Graphics Processing Unit
HAR	Human Activity Recognition
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
IoT	Internet of Things
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 Perceptron
MRAM	Magneto-Resistive Random Access Memory
MW	Memory Window
NMOS	N-channel Metal-Oxide Semiconductor
NVIA	Non-Volatile AI Inference Accelerator
NVM	Non-Volatile 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
RBM	Restricted Boltzmann Machine
RNN	Recurrent Neural Network
RRAM/ReRAM	Resistive Random Access Memory
S/H	Sample and Hold
SA	Sense Amplifier
SCA	Side-Channel Attack
SDA	Stacked Denoising Autoencoder
SCLC	Space Charge Limited Conduction
SD	Stored Data
SEM	Scanning Electron Microscope
SHL	Shift Logical Left
SL	Select Line
SLC	Single Level Cell
SMU	Source Measure Unit
SOT-MRAM	Spin Orbit Torque Magnetic Random Access Memory
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
VGSOT-MRAM	Voltage-Gated Spin Orbit Torque Magnetic Random Access Memory
VR	Virtual Reality

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