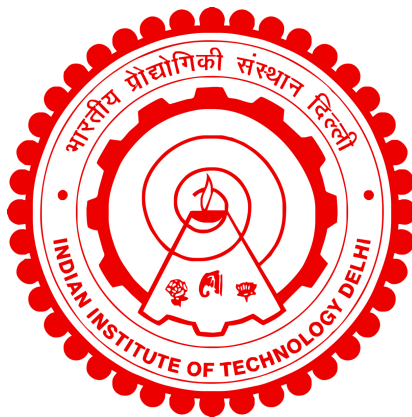


INVESTIGATION OF IN-MEMORY COMPUTING ARCHITECTURES FOR EDGE AI APPLICATIONS

CHITHAMBARA MOORTHII J



DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY DELHI

JULY 2026

© **Indian Institute of Technology Delhi (IITD), New Delhi, 2026**

Investigation of In-Memory Computing Architectures for Edge AI Applications

by

Chithambara Moorthii J

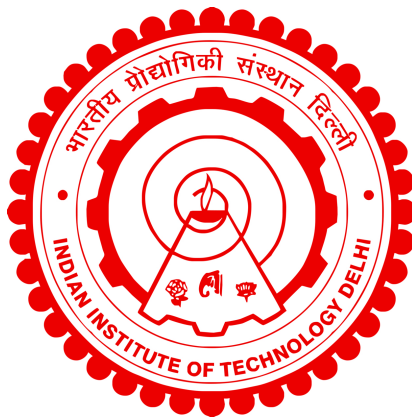
Department of Electrical Engineering

Submitted

in partial fulfillment of the requirements of the degree of

Doctor of Philosophy

to the



INDIAN INSTITUTE OF TECHNOLOGY DELHI

JULY 2026

*Dedicated to my family and loved ones, whose
unshakeable support carried me through every challenge.*

Certificate

This is to certify that the thesis entitled “**Investigation of In-Memory Computing Architectures for Edge AI Applications**”, submitted by **Chithambara Moorthii J** to the Indian Institute of Technology Delhi, for the award of the degree of **Doctor of Philosophy** in the Department of ELECTRICAL ENGINEERING, is a record of the original, bona fide 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

Professor,

Department of Electrical Engineering,

Indian Institute of Technology Delhi.

Acknowledgements

I would like to express my deepest gratitude to all those who have supported me throughout my PhD journey. This thesis would not have been possible without their guidance, encouragement, and unwavering faith in me.

First and foremost, I thank Prof. Manan Suri, my PhD advisor, for selecting me to work under his guidance and for supporting me from day one until the completion of this work. His mentorship, trust, and constant encouragement have shaped my academic and professional growth in profound ways.

I am forever grateful to my wife, Versha, whose love, understanding, and constant motivation have been the backbone of this journey. I also extend my heartfelt thanks to my parents, sister, and brother-in-law for their unconditional love, support, and belief in me.

I would like to acknowledge Deepak Verma and Richa Mishra, my juniors, for their extended support throughout the journey. Their enthusiasm and readiness to help made many challenging phases more manageable. I am equally grateful to my seniors, Supriya Chakraborty and Ahmed Shaban, whose guidance helped me navigate several hurdles during my research.

My sincere thanks to Bhanprakash Goswami, Greeshma, Yuvraj Chowhan, Anmol Singla, Avanthika, Vedang, Haridarshan, Sahil, Utsab, Ayan Ray, Digamber Kumar Pandey, and Shubham Panwar for their support at various stages of my work. Each of them contributed in meaningful ways, whether through collaboration, discussions, or simply being there during crucial moments.

A special note of appreciation goes to Merbin John, Vinayan H, and Hari, who have been like family to me. Their emotional, mental, and professional support helped me stay grounded and motivated throughout the highs and lows of this journey.

I am also thankful to Dwijay Bane, Sufyan, and Devendra Singh, who helped me immensely during the first year of my PhD. Their early guidance and companionship played a significant role in easing my transition into research life.

Lastly, I express my sincere gratitude to Himanshu, Yathish, Yathinder, Usha, and Prof. Anuj Dhawan for their continuous support with administrative paperwork and procedures. Their timely assistance ensured smooth progress throughout my PhD.

To all of you—thank you for being part of this journey and for helping me reach this milestone.

Abstract

The memory wall problem is a fundamental bottleneck in modern computer architecture, where the increasing speed of processors has far outpaced the rate at which data can be transferred to and from memory. To investigate these challenges, we present a heterogeneous simulation framework, which models the performance of heterogeneous memory subsystems with systolic arrays for the simulation of various memory configurations, including hybrid SRAM-NVM setups, for different neural network workloads. Upon examination of the memory subsystem’s impact on the accelerator, we present novel and energy-efficient In-Memory Computing (IMC) architectures designed to tackle the ”memory wall problem” by performing multi-bit multiply and accumulate (MAC) operations directly within an SRAM-based IMC accelerator. The proposed architecture features a unique input and weight mapping strategy that removes the need for power-hungry Digital-to-Analog Converters (DACs), which also incorporates an innovative analog carry computation method that computes the final multi-bit product directly within the IMC macro, eliminating the latency and power consumption of external digital circuitry.

The study applies these notions to RRAM (Resistive Random-Access Memory), a form of non-volatile memory. This encompasses the creation of a variable bit-precision vector co-processor that incorporates an RRAM-based IMC unit alongside a RISC-V core. The system employs an innovative data mapping technique for vector-matrix multiplication, enhancing energy efficiency and executing multi-bit MVM in a single cycle, therefore reducing latency. These RRAM-based designs seek to deliver efficient and high-performance solutions for deep learning inference on edge devices.

In addition to deep learning, the thesis briefly explores applications in genomics by exploiting RRAM-based IMC accelerator for DNA sequence analysis. This architecture addresses the high computational complexity of DNA sequencing by converting DNA into a binary format to enable in-memory bitwise XOR and AND operations, showcasing the versatility of IMC architectures beyond conventional machine learning tasks. This research presents a collection of specialized hardware accelerators and a simulation framework that jointly enhance the area of IMC by showcasing high-performance, energy-efficient solutions for intricate computational challenges.

सारांश

मेमोरी वॉल समस्या आधुनिक कंप्यूटर आर्किटेक्चर की एक मूलभूत बाधा है, जहाँ प्रोसेसर की गति में तीव्र वृद्धि की तुलना में मेमोरी से डेटा स्थानांतरण की गति उतनी तेज़ी से नहीं बढ़ सकी है। इन चुनौतियों का अध्ययन करने के लिए हम एक विषम (heterogeneous) सिमुलेशन फ्रेमवर्क प्रस्तुत करते हैं, जो विभिन्न न्यूरल नेटवर्क कार्यभारों के लिए हाइब्रिड SRAM-NVM विन्यास सहित कई प्रकार के मेमोरी कॉन्फ़िगरेशन का अनुकरण करते हुए, सिस्टोलिक एरे के साथ विषम मेमोरी सबसिस्टम के प्रदर्शन को मॉडल करता है। मेमोरी सबसिस्टम के एक्सेलेरेटर पर प्रभाव का विश्लेषण करने के बाद, हम नवीन और ऊर्जा-कुशल इन-मेमोरी कम्प्यूटिंग (IMC) आर्किटेक्चर प्रस्तुत करते हैं, जो SRAM-आधारित IMC एक्सेलेरेटर के भीतर ही मल्टी-बिट मल्टिप्लाई-ऐंड-अक्यूमुलेट (MAC) संचालन करते हुए “मेमोरी वॉल समस्या” को हल करने की दिशा में कार्य करते हैं।

प्रस्तावित आर्किटेक्चर में एक विशिष्ट इनपुट और वज़न-मैपिंग रणनीति शामिल है, जो ऊर्जा-गहन डिजिटल-टू-एनालॉग कन्वर्टर्स (DACs) की आवश्यकता को समाप्त करती है। इसके साथ ही, यह एक नवाचारपूर्ण एनालॉग कैरी कम्प्यूटेशन विधि को सम्मिलित करता है, जो IMC मैक्रो के भीतर ही अंतिम मल्टी-बिट उत्पाद की गणना करता है, जिससे बाहरी डिजिटल परिपथों की विलंबता और ऊर्जा खपत दोनों कम होती हैं।

अध्ययन इन अवधारणाओं को RRAM (Resistive Random-Access Memory) तक विस्तारित करता है, जो एक नॉन-वोलेटाइल मेमोरी है। इसमें एक चर बिट-प्रिसीजन वेक्टर को-प्रोसेसर का निर्माण शामिल है, जिसमें RRAM-आधारित IMC यूनिट और एक RISC-V कोर एकीकृत होते हैं। यह प्रणाली वेक्टर-मैट्रिक्स गुणन के लिए एक नई डेटा-मैपिंग तकनीक का उपयोग करती है, जो ऊर्जा दक्षता में सुधार लाती है तथा मल्टी-बिट MVM को एक ही चक्र में निष्पादित करती है, जिससे विलंबता कम होती है। ये RRAM-आधारित डिज़ाइन एज डिवाइसों पर डीप लर्निंग इन्फ़रेंस के लिए उच्च-दक्षता और उच्च-प्रदर्शन समाधान प्रदान करने का लक्ष्य रखते हैं।

डीप लर्निंग से परे, यह शोध जीनोमिक्स अनुप्रयोगों का भी संक्षिप्त अध्ययन करता है, जहाँ DNA अनुक्रम विश्लेषण के लिए RRAM-आधारित IMC एक्सेलेरेटर का उपयोग किया गया है। यह आर्किटेक्चर DNA को बाइनरी प्रारूप में परिवर्तित करके इन-मेमोरी बिटवाइज़ XOR और AND संचालन सक्षम करता है, जिससे पारंपरिक मशीन लर्निंग कार्यों से आगे IMC आर्किटेक्चर की बहुमुखी क्षमता प्रदर्शित होती है। यह शोध विभिन्न विशेषीकृत हार्डवेयर एक्सेलेरेटर और एक सिमुलेशन फ्रेमवर्क प्रस्तुत करता है, जो उच्च-प्रदर्शन, ऊर्जा-कुशल समाधान उपलब्ध कराते हुए IMC के क्षेत्र को आगे बढ़ाते हैं और जटिल कम्प्यूटेशनल चुनौतियों का समाधान प्रदान करते हैं।

Contents

Certificate	i
Acknowledgements	ii
Abstract	iv
सारांश	v
Contents	vi
List of Figures	xi
List of Tables	xv
Abbreviations	xvii
1 Introduction	1
1.1 Motivation	1
1.2 Genesis of non-Von Neumann architectures	3
1.3 Thesis Organization and Contributions	5
1.3.1 Objective of Thesis	5
1.3.2 Key Contribution of Thesis	5
1.3.3 Thesis Organization	7
2 Literature Review and Background basics	11
2.1 Breaking the Memory Wall	11
2.2 The Crossbar Architecture: An Analog Computing Engine	13
2.3 SRAM-Based In-Memory Computing	14
2.3.1 SRAM Cell Structures for IMC	15
2.3.2 Computing Mechanisms in SRAM	16

2.3.3	Advantages and Limitations	16
2.4	RRAM-Based In-Memory Computing	17
2.4.1	RRAM Switching Mechanisms	17
2.4.2	RRAM for Analog Computation	18
2.4.3	Advantages and Limitations	19
2.5	Conclusion	19
3	Simulation Framework for Emerging Memories with Systolic Array	21
3.1	Introduction	22
3.2	Proposed Simulation Framework (NVSystolic)	22
3.2.1	Modelling Systolic array and Dataflow	23
3.2.2	Modelling Topology	24
3.2.3	Modelling SRAM/NVM Memory Architecture	24
3.3	Analysis in Google TPU	27
3.3.1	Memory Configuration 1:	27
3.3.2	Memory Configuration 2:	30
3.4	Validation	32
3.5	Summary	32
4	SRAM IMC-based Multi-bit MVM with Analog Carry Computation	35
4.1	Introduction	36
4.2	IMC Module and Architecture	37
4.3	Proposed Multi-bit MVM using SRAM IMC	38
4.3.1	Multi-bit Input and Weight Mapping	39
4.3.2	Parallel Architecture for Matrix multiplication	40
4.4	Analog Carry Propagation Circuit	42
4.5	Workload Mapping of CNN	43
4.6	Simulation Results and Analysis	45
4.7	Summary	48
5	RRAM IMC based Analog Carry Propagation and Multi-bit MVM	49
5.1	Introduction	50
5.2	Proposed In-Memory Compute Module Architecture and Mapping	51
5.2.1	Analog Carry Propagation Methodology	52
5.2.2	Proposed mapping strategy for MVM	54
5.3	Results and Discussion	55
5.4	Summary	56
6	RRAM IMC based Vector Co-processor	57
6.1	Introduction	58
6.2	Proposed VPU-CIM	60

6.2.1	Compute-in-memory Macro	60
6.2.2	Data Mapping Strategy for vector-matrix multiplication	61
6.2.3	Proposed VPU-CIM Microarchitecture	62
6.3	Architecture of the Tiled Design	63
6.4	Results and discussion	65
6.4.1	Workload analysis and variability	66
6.4.2	Power Performance Area Analysis	67
6.4.3	Energy efficiency and comparison with related work	68
6.5	Summary	69
7	DNA Sequence Analysis using RRAM IMC based Accelerator	71
7.1	Introduction	72
7.2	Proposed DNA-CIM Architecture	75
7.2.1	Binary Encoding of the DNA Sequences	76
7.2.2	Design and Mapping for DNA Sequence analysis	77
7.2.3	Matching and Mutation Analysis of the DNA sequences	79
7.3	Results and Analysis	80
7.3.1	Analysis of PVT Variations	80
7.3.2	RRAM Variability Analysis	82
7.3.3	Performance Evaluation	83
7.4	Summary	84
8	Efficient Implementation of Gated Recurrent Unit using IMC	85
8.1	Introduction	86
8.2	Proposed Architecture	88
8.2.1	In Memory Computing Co-processor for MAC	89
8.2.2	Non-IMC function mapping	91
8.3	Pipelining and Hardware Reuse	91
8.4	Workload Analysis and Accuracy	93
8.5	Benchmark and results	95
8.6	Summary	96
9	Conclusion	97
	Scope for Future Work	99
	 Bibliography	 101
	List of Publications	123
	Curriculum Vitae	125

List of Figures

1.1	Development history of and new trends in MOSFETs with differing transistor structures and scaling stages [1])	2
1.2	(a) von Neumann architecture for computing systems, where the central processing unit and the memory unit are physically separated and connected through a data bus. (b) Near-memory computing architecture, where the processing unit is complemented with an eNVM unit to store commands and parameters. (c) SRAM-based in-memory computing architecture, where computation is performed directly within the SRAM unit via dedicated peripherals, while eNVM serves as storage for computational parameters. (d) eNVM-based in-memory computing architecture, where eNVM provides both the non-volatile storage of computational parameters and the computation. (e) Illustration of the memory hierarchy in traditional CMOS-based computing systems(Image adapted from [2], [3])	4
2.1	Different types of volatile CIM cells: (a) 6T SRAM cell, (b) 8T SRAM cell, (c) 8T-1C SRAM cell, (d) 9T-1C SRAM cell, (e) 10T-2C SRAM cell (Figure adopted from [4])	15
2.2	(a) Illustration of switching mechanism of RRAM (b) Energy efficiency versus dimension of MAC parallelism for NVM and SRAM IMC (Normalized to 1-bit MAC and 28-nm tech node), and (c) IMC Macro Latency versus dimension of MAC parallelism for NVM and SRAM IMC (Figure adopted from [4, 5])	17
3.1	Architecture of simulation framework. Memory Configuration 1: representing NVM (PCM, RRAM, and STTRAM) or SRAM choice for all three scratchpad memories. Memory Configuration 2: representing hybrid SRAM-NVM configuration. (R/W stands for Read /Write, DB - Double Buffered, SA - Systolic Array)	23
3.2	IFMAP Read Energy vs Latency for various types of memories for different workloads. All memories at 32nm technology node and size of 8Mb	26
3.3	Filter Read Energy vs Latency for various types of memories for different workloads. All memories at 32nm technology node and size of 8Mb	27

3.4	OFMAP Write Energy vs Latency for various types of memories for different workloads. All memories at 32nm technology node and a size of 2Mb	28
3.5	Read energy vs Latency vs Power for various NVM-SRAM ratios of IFMAP (8MB, 32nm tech node) for TinyYOLO workload	30
4.1	(a) Truth table for IMC bitcell. (b) Input data flow into IMC module. (c) Hierarchical representation of MVM operation. (d) 8T SRAM IMC Bitcell. (e) Detailed architecture of IMC module and mapping of weights in memory. (f) Elaborated view of single column of memory array with analog carry.	37
4.2	(a) Mathematical representation of multiplication of two 4-bit numbers expanded from the MAC operation. (b) Expansion of MVM operation. (c) Matrix-Vector representation of Neural network implementation.	39
4.3	CNN Network for MNIST image classification. The achieved accuracy is 98% with 4-bit precision.	40
4.4	CNN mapping in Hierarchical architecture: Weight Matrix (WM) of size $K \times K \times N$ is mapped in $K \times K$ IMC Modules, with each module having N columns. Each Output Feature Map (OFM) is calculated in 1 product cycle	41
4.5	(a) The bitline discharge voltage levels for MAC value of 0 to 7. (b),(c) Analysis of the impact of PVT conditions, Temperature variation and Process corners(SS, TT, FF) on MAC operations using the implemented SRAM-IMC, respectively. (d) Energy consumption for a single product cycle at different frequencies for each design block in the architecture. (e) Power breakdown of design blocks for complete architecture. (f) Variation of Energy efficiency with input-weight bit precision.	44
4.6	(a) Table shows the carry propagation and final pop count. (b) Bitline voltage plot over eight cycles of MAC operation and Final 8-bit output(red) result of 4b-4b IN-W multiplication(IN=1111, W=1111).	46
5.1	(a) Architecture of proposed IMC module (b) Mapping of multiplier in 1T1R array (c) Analog carry propagation methodology.	50
5.2	Representation of input vector multiplied with matrix, where each matrix element is 4-bits wide.	51
5.3	(a) Peak power distribution across MAC array, ADC and XOR logic (b) 1T1R bitcell and design specifications.	52
5.4	(a)Multiplicand and multiplier mapping in bitcell with corresponding pop count and carry generated (b)BL voltages for multiplication of 4b'1111x4b'1111 input combination (c)Output of the product obtained in single-cycle.	53

6.1	(a) RISC-V PicoRV32 architecture (b) Analog Mux (c) AND logic based 1T-1R RRAM CIM bitcell with truth table (d) Detailed CIM Architecture (e) Top-level architecture (f) VPU-CIM-v2: Tiled version of VPU-CIM with $N = \text{number of CIMs}$ (g) Data processing unit of VPU-CIM-v2 (i) VPU-CIM-v1: with $N=1$ (j) Data processing unit of VPU-CIM-v1	59
6.2	Data mapping strategy of multi-bit weight matrix and propagation of input matrix	61
6.3	VPU-CIM V2 or Design2 Data flow	64
6.4	Layout of VPU-CIM V1, Analog CIM Macro and Layout of VPU-CIM V2	65
6.5	Inputs, bitline voltage and ADC outputs for all possible input/output combinations for one column	66
6.6	Area distribution across different technology nodes.	67
6.7	Area distribution in analog CIM module	67
7.1	(a) Top-level block diagram of the proposed DNA sequence analysis CIM accelerator (b) Binary encoding scheme for the DNA sequences (c) DNA sequence matching using bit-wise XOR and AND logic operations mapped into different banks of CIM (d) Comparison of the different output sequences using digital logic.	75
7.2	(a) The proposed RRAM-based CIM Architecture (b) represents the mapping of the DNA sequences in the proposed Architecture (c) the 1T-1R RRAM bitcell (d) schematic of the proposed multi logic sense circuit (MLSC) (e) Truth table for all the possible combinations used to perform the comparison between pairwise DNA sequences.	77
7.3	(a) Monte Carlo simulation performed for 1000 points for different Bitline voltages (V_L , V_M , V_H) (b) Temperature and corner variation analysis for V_H level (c) Temperature and corner variation analysis for V_M level (d) Temperature and corner variation analysis for V_L level (e) RRAM variability analysis for each level.	81
7.4	Transient result output waveforms from MLSC for all possible input combinations, showing both XOR and AND logic outputs.	82
8.1	(a) Mapping of weights in the RRAM memory in a skewed manner in order to perform MAC operation in a single cycle, (b) 1T1R bitcell with the truth table of operation and design parameters, (c) Top-level block diagram of proposed architecture with RISC-V core and IMC co-processor, (d) Description of pipelining (Color coded) performed in IMC co-processor, (e) Tile:1 with M modules to store and perform MAC operation on W_{ir} , W_{in} and W_{iz} , (f) Tile:2 with K modules to store and perform MAC operation on W_{hr} , W_{hn} and W_{hz} , (g) Multi-bit digital adder for accumulation of partial products with bias. Shift logic will be enabled only for 8-bit MAC operations	88

8.2	Proposed GRU pipelining with distribution of operations across IMC and RISC-V over 10 stages	91
8.3	(a) Bit line voltages for various MAC values and reference voltages for sensing, (b) Analog IMC core block, (c) Digital output from multi-bit Digital adder block, (d) ADC Output of IMC core.	93

List of Tables

3.1	Layerwise topology of TinyYOLO network[6]	24
3.2	Validation results for Resnet18 workload. Shows the percentage error between the literature and the simulated results.	31
4.1	Benchmarking of this work with other literature. All comparisons are for In-memory mixed-signal implementations	43
5.1	Benchmark results of current work with existing literature.	56
6.1	Proposed vector extensions	62
6.2	Performance parameters: Comparison of proposed design with exist- ing work (RRAM CIMs)	69
7.1	Binary Encoding of DNA sequence	76
7.2	Outcomes for DNA Mutation Analysis	79
7.3	Benchmarking table comparing with other state-of-the-art works. . .	81
8.1	Summary of proposed IMC architecture	89
8.2	Benchmarking of GRU implementation with other works	95

Abbreviations

SRAM	Static Random Access Memory
NVM	Non Volatile Memory
RRAM	Resistive Random Access Memory
STT-RAM	Spin Transfer Torque Random Access Memory
DRAM	Dynamic Random Access Memory
PCM	Phase Change Memory
IFMAP	Input Feature MAP
OFMAP	Output Feature MAP
WS	Weight Stationary
IS	Input Stationary
OS	Output Stationary
SA	Systolic Array
nm	nano meter
TPU	Tensor Processing Unit
CNN	Convolutional Neural Network
SOTA	State Of The Art
DAC	Digital to Analog Converter
IMC	In Memory Computing
AI	Artificial Intelligence
MAC	Multiply and Accumulate
NMOS	N-channel Metal Oxide Semiconductor
MVM	Matrix Vector Multiplication
ADC	Analog to Digital Converter
VLSA	Voltage Latch Sense Amplifier
LSB	Least Significant Bit
MSB	Most Significant Bit

RWL	Read Word Line
GOPS	Giga Operations Per Second
TOPS/W	Tera Operations Per Second / Watt
FOM	Figure Of Merit
CMOS	Complementary Metal Oxide Semiconductor
PVT	Process Voltage Temperature
RBL	Read Bit Line
HRS	High Resistance State
LRS	Low Resistance State
1T-1R	1 Transistor - 1 Resistor
TC	Thermometric Code
WL	Word Line
BL	Bit Line
SL	Select Line
RISC-V	Reduced Instruction Set Computer - V
CIM	Compute In Memory
PDK	Process Design Kit
MUX	Multiplexer
ALU	Arithmetic Logic Unit
CLSA	Current Latch Sense Amplifier
nMPPU	near Memory Post Processing Unit
DPU	Data Processing Unit
VRF	Vector Register File
TSMC	Taiwan Semiconductor Manufacturing Company
ASAP	Arizona State Predictive
VPU	Vector Processing Unit
ML	Machine Learning
NGS	Next Generation Sequencing
DNA	Deoxyribonucleic acid
FPGA	Field Programmable Gate Array
CPU	Central Processing Unit
GPU	Graphics Processing Unit
CAM	Linearly Polarised
SOT-MRAM	Spin Orbit Torque - Magnetoresistive Random Access Memory
BWT	Burrows Wheeler Transform

MTJ	M agnetic T unnel J unction
MLSC	M ulti L ogic S ense C ircuit
ASIC	A pplication S pecific I ntegrated C ircuit
GCUPS	G iga C ell U dates P er S econd
TCUPS/W	T era C ell U dates P er S econd / W att
GRU	G ated R ecurrent U nit
RNN	R ecurrent N eural N etwork
LSTM	L ong S hort T erm M emory