Powering Next-Generation Artificial Intelligence by Designing Three-dimensional High-Performance Neuromorphic Computing System with Memristors

Hongyu An

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

Electrical Engineering

Yang Yi, Chair

Dong. S. Ha

Marius K. Orlowski

Lingjia Liu

Andrew J. Kurdila

Zhen Zhou

July 27, 2020

Blacksburg, VA

Keywords: Neuromorphic Computing, Memristors, Three-dimensional Integrated Circuits, Associative Memory Learning

Powering Next-Generation Artificial Intelligence by Designing Three-dimensional High-Performance Neuromorphic Computing System with Memristors

Hongyu An

ABSTRACT

Human brains can complete numerous intelligent tasks, such as pattern recognition, reasoning, control and movement, with remarkable energy efficiency (20 W). In contrast, a typical computer only recognizes 1,000 different objects but consumes about 250 W power [1]. This performance significant differences stem from the intrinsic different structures of human brains and digital computers. The latest discoveries in neuroscience indicate the capabilities of human brains are attributed to three unique features: (1) neural network structure; (2) spike-based signal representation; (3) synaptic plasticity and associative memory learning [1, 2].

In this dissertation, the next-generation platform of artificial intelligence is explored by utilizing memristors to design a three-dimensional high-performance neuromorphic computing system. The low-variation memristors (fabricated by Virginia Tech) reduce the learning accuracy of the system significantly through adding heat dissipation layers. Moreover, three emerging neuromorphic architectures are proposed showing a path to realizing the next-generation platform of artificial intelligence with self-learning capability and high energy efficiency. At last, an Associative Memory Learning System is exhibited to reproduce an associative memory learning that remembers and correlates two concurrent events (pronunciation and shape of digits) together.

Powering Next-Generation Artificial Intelligence by Designing Three-dimensional High-Performance Neuromorphic Computing System with Memristors

Hongyu An

GENERAL AUDIENCE ABSTRACT

Human brains can complete numerous intelligent tasks with low power consumption (20 W), high accuracy, and fast speed, such as handwriting recognition. However, the digital computer either cannot achieve these sophisticated missions totally or has no capability of accomplishing them with the satisfying efficiency on power, speed, and accuracy. Usually, machines need an excessive amount of time and energy for calculation in specific algorithms.

This research aims to is to address these challenges in the field of artificial intelligence through rebuilding and mimicking the structure of the human brain physically using circuits. This concept is so-called neuromorphic computing or brain-inspired computing. Moreover, a self-learning method is exhibited to remembers and correlates two concurrent events.

Unlike the conventional path for realizing artificial intelligence with complicated mathematical concepts and equations, the approach is more straightforward. The significance of rebuilding the human brain is not only to reveal a way of designing a brain-like self-learning intelligence system but also to explore a method of comprehending the learning mechanism of our brains.

Acknowledgment

Artificial Intelligence is always one of the most challenging and exciting scientific mission in human history. I am deeply honored and feel lucky to have an opportunity of exploring this area in my Ph.D. period. Here, first, I would like to express my deep gratitude to my advisor Prof. Yang Yi for introducing this exciting research field and dedicatedly guiding and supporting me. I also want to show gratitude to all my Ph.D. advisory committee members: Dr. Dong. S. Ha, Dr. Marius K. Orlowski, Dr. Lingjia Liu, Dr. Andrew J. Kurdila, and Dr. Zhen Zhou (Intel Labs). Without their years of supports, my research and this dissertation cannot be accomplished. I also want to thank all my lab mates and friends in Multifunctional Integrated Circuits and Systems Group (MICS). I learned a lot from them. Specifically, I would like to show my special appreciation to my friends Mohammad Shah Al-Mamun (Intel Labs) for the discussion on memristors and Kang Jun Bai for the analog IC design counsel.

More important, I would like to express my genuine gratitude to my wife Yan Zhang for her selfless dedication to our family. Since 2017, after I joined Virginia Tech, she has been taking care of our sons at Kansas City for three years, balancing and struggling between the family and her academic career. It is extremely challenging, especially she is such an excellent scientist in her field. I deeply appreciate her sacrifice and devotion to our family.

Additionally, I would like to thank my parents and parents in law: Yonghui Liang, He An, Guizhen Li, Xueqing Zhang. During these years, they were laboriously traveling back and forth between China and the United States for supporting my wife and me to taking our children. Without their help, there will no scientific contribution I can make. I would like deeply to appreciate their devotion and support.

At last, I hope all my experience and what I learned in my doctoral period will inspire my children, William and Adam, with strength, determination, optimism, endurance, and wisdom on their paths later, and hope they will make valuable contributions to society and be a good man.

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

1.1 Motivation

Human brains can complete numerous intelligent tasks, such as pattern recognition, reasoning, control and movement, in remarkable energy efficiency (20 W). In contrast, a typical computer only recognizes 1,000 different objects but consumes about 250 W power [1]. This performance significant differences stem from the intrinsic different structures of human brains and digital computers. The latest discoveries in neuroscience indicate the capabilities of human brains are attributed to three unique features: (1) neural network structure; (2) spike-based signal representation; (3) synaptic plasticity and associative memory learning [1, 2]. Firstly, the neural network structure has demonstrated its capability of handling cognition tasks in deep learning [1, 2]. Secondly, the low firing rate of spiking signals enables the brains to operate with high energy efficiency. Thirdly, the synaptic plasticity and associative memory learning have been proved to be highly related to the memory mechanism and enable the brains to learn from the surroundings [3].

Despite deep learning with neural network structure demonstrates the capability of Artificial Neural Networks (ANNs) in solving complicated cognition tasks, the computing platforms built upon von Neumann architecture restrict the performance and efficiency of ANNs [4-6]. The von Neumann architecture is designed for efficient Boolean calculation and rather for neural network-based learning. Moreover, the inevitable demanding requirements on computational resources and large datasets restrict the deployment of ANNs on resource-constraint platforms, like cell phones, unmanned aerial vehicles, autonomous cars, and spacecraft [7-10]. Thereby, the next-generation platform of Artificial Intelligence should aim to rebuild a brain-like neuromorphic system physically to fully take advantage of human brains for overcoming all these challenges. Specifically, the novel neuromorphic system should include the following features:

- Spike-based information representation with a low firing rate (both on forwarding inference and backward learning parts) to achieve brain-comparable energy efficiency;

- Organ-like sensory system (for example, eyes and ears) to capture signals from the external world and transform them into spiking signals;
- Human-like learning methodology (associative memory) that enable the neuromorphic system to learn from the surroundings by themselves rather than from cumbersome datasets.

These unique features can be achieved by reversing engineering of human brains with emerging technologies at all levels of the architecture, algorithm, circuit, and device. In this dissertation, I preliminary design and analyze a high-performance neuromorphic system with memristors and associative memory learning.

The next-generation hardware platform of artificial intelligence cannot be simply designed with one discipline alone. A successful high-performance of neuromorphic computing system requires a deep understanding of neural science, mathematics, hardware design, and software. Figure 1-1 illustrates how these disciplines influence and support each other.

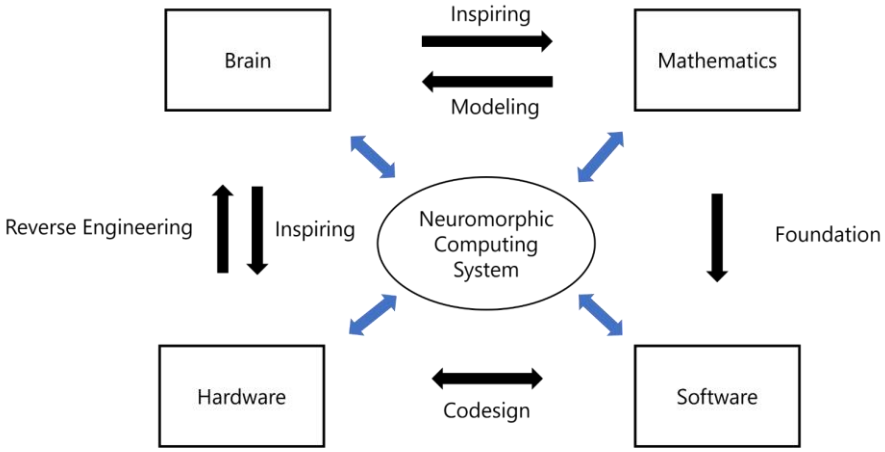


Figure 1-1: Methodology of realizing high-performance next-generation platform of artificial intelligence

Neuromorphic Computing, or so-called brain-inspired computing, requires deeply comprehensively understanding and study on brain functions. The features and operating mechanism should be future precisely defined and described by solid mathematic formulas, just like the calculus built by Leibniz and Newton in the 1600s to explain the physical world, particularly in the motion of objects. Despite deep learning demonstrates a massive success in many applications, its underlying mathematical causality is still unclear. A more comprehensive

mathematic theory needs to be built for explaining the functions of the neural system in the future. Moreover, the development of mathematical theory helps the design of software and hardware co-design methodology. Eventually, the hardware built by the solid mathematic theory which precisely describing the activity of the neural system will realize an artificial intelligence system with full unbelievable capabilities of human brains.

In my work [11], I first introduced and proposed three emerging neuromorphic architectures: Distributive Neuromorphic Computing Architecture (DNCA); Cluster Neuromorphic Computing Architecture (CNCA), and Associative Neuromorphic Computing Architecture (ANCA), which are illustrated in Figure 1-2.

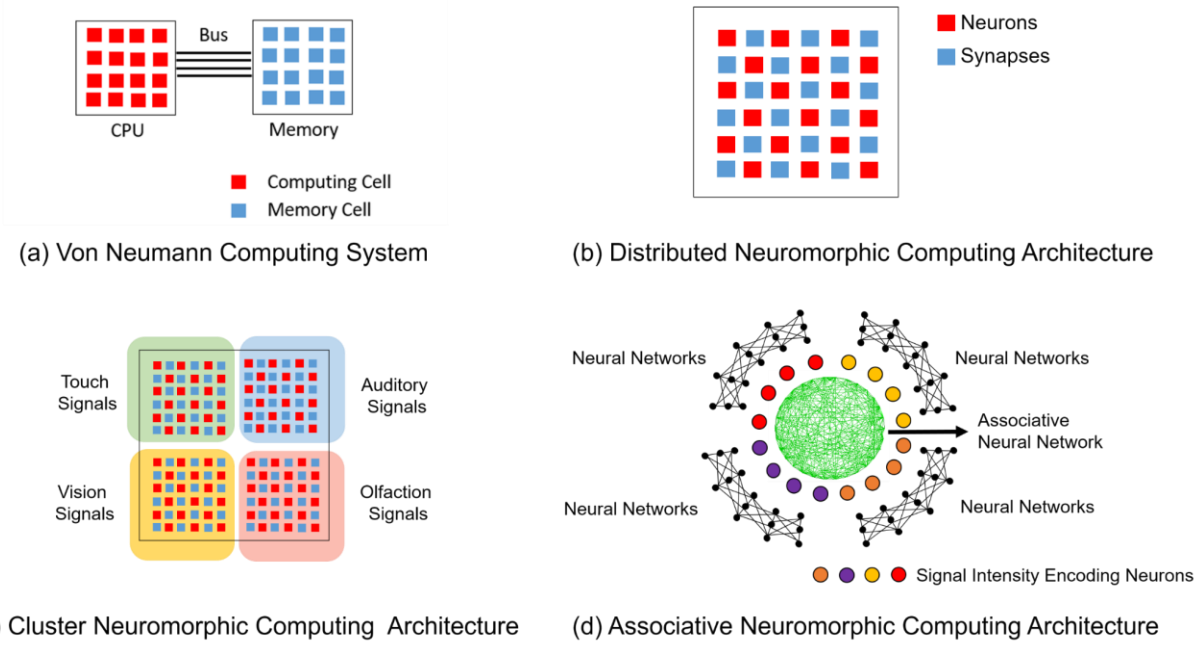


Figure 1-2: Emerging neuromorphic computing architectures: (a) traditional von Neumann Computing System (b) Distributive Neuromorphic Computing Architecture (c) Cluster Neuromorphic Computing Architecture (d) Associative Neuromorphic Computing Architecture.

In DNCA, the neurons and synapses are placed in a distributed neural network structure minimizing the distance between the computing units (neurons) and the memory units (synapse). As a result, the computing of neural networks can be performed between physically adjacent neurons and synapses reducing the energy spent on signal propagation. The utilization of threshold

neurons and the spike-based training methods will further reduce the power consumption of the DNCA-based system to a brain-comparable level.

Next, the CNCA, which is built upon the DNCA, divides the whole large neural network into multiple regions (Figure Figure 1-2 (b)). Each of them is responsible for a specific type of signal, such as visual and auditory signals. These signals are separately captured by organ-like sensors. In CNCA, different signals will be processed at separated neural networks, enabling a parallel information processing ability.

At last, the ANCA correlates the outputs of CNCA together realizing a behavior level associative memory learning as illustrated in Figure 1-2(c). The ANCA-based system will have the brain-like self-learning capability and learn directly from dynamically changing surroundings.

1.2 Contribution

In my Ph.D. study, I preliminarily explored the next-generation platform of artificial intelligence by utilizing low-variation memristors to a high-performance neuromorphic computing system. The contributions can be summarized as:

- Propose three emerging architectures of next-generation neuromorphic systems, showing a possible path to realizing the next-generation platform of artificial intelligence with self-learning capability and high energy efficiency;
- Apply emerging device *memristor* and three-dimensional integration technology increases the performance of the neuromorphic computing system;
- Design an Associative Memory Learning System that mimics the learning mechanism of neural system that remembers and correlates two concurrent events together.

In the remainder of this dissertation, Chapter 2 introduces backgrounds of the biological neural system, neuromorphic computing, and memristors, Chapter 3 exhibits the design of high-performance neuromorphic computing system with memristors, Chapter 4 presents how to implement associative memory learning, Chapter 5 shows the three-dimensional neuromorphic computing system with VT two-layer memristive synapses, Chapter 6 concludes the contributions and future work.

Chapter 2. Backgrounds of Neural System, Neuromorphic Computing and Memristors

In this chapter, the brief background of the biological neural system, neuromorphic computing, and memristors are introduced first at the preface of the dissertation.

2.1 Brain Structure and Organ

Human brains are built upon the neural networks that consist of neurons and synapse as the core organs. A brief introduction to the neurons and synapses is conducive to revealing a path of building a high-performance artificial intelligence system through mimicking their functions. In this section, the history of the discovery of a neural system will be introduced. Furthermore, the functions of neuron and synapses, the mechanism of memory at the cellular level, and how the signals propagate among them are discussed.

Spanish anatomist Santiago Cajal first identified and determined that neurons are the basic building organs and signal processing units in a nervous system in the 1890s [3, 12]. Figure 2-1 (a) illustrates his hand drawing of a group of neurons, which was published in 1899 [13]. The individual neuron under the microscope is depicted in Figure 2-1 (b). In a nervous system, neurons are connected with each other in a high-degree network configuration. Generally, each neuron connects with hundreds or even thousands of other neurons. As a result, a neuron can simultaneously communicate with thousands of other neurons using a sequence of low-rate spiking signals. The spiking signals propagating among neurons were monitored by Hodgkin and Huxley in 1939 (see Figure 2-1 (c)). Unlike the high-speed modern computer, the main frequency of the spiking signals in the nervous system is as low as ~ kilohertz level (1-10 millisecond duration) with millivolt-level magnitudes as illustrated in Figure 2-1 (c) [14, 15].

Figure 2-1 (e) depicts an abstract illustration of a typical neuron. Four critical functional parts of a typical neuron had been identified by the neurologists, which are dendrites, soma, axon, and synapses as illustrated in Figure 2-1(d, e) [3, 16, 17]. The function of the tree-like dendrites is to receive the spiking signals from other neurons, like a receiver.

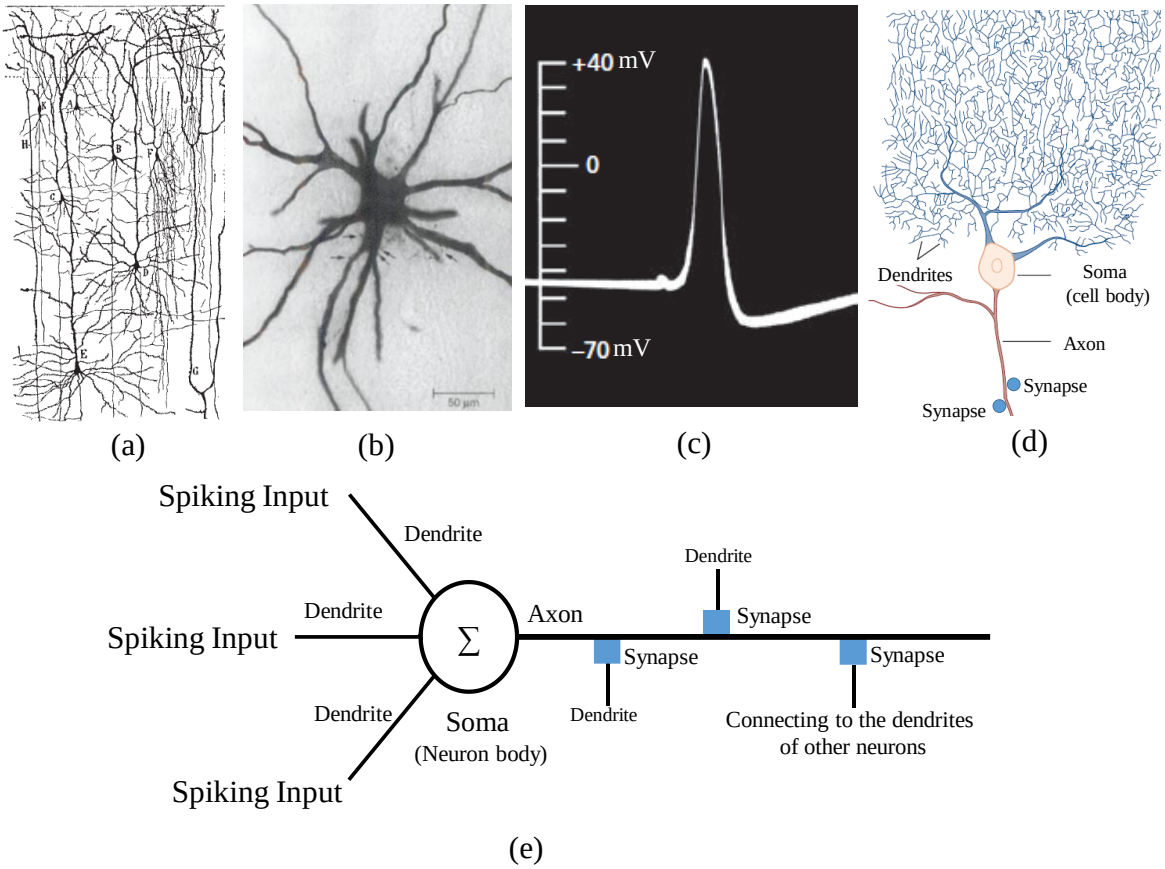


Figure 2-1: Neural network and the detailed neuron structure. (a) The hand drawing of the neural network by Santiago Cajal in 1899 [13, 18]. (b) Image of a motor neuron [3]. (c) The first measured membrane potential (Spikes) of a neuron in 1939 recorded by Hodgkin and Huxley (The interval is 2 ms) [3] (d) The structure of a typical neuron including four critical parts: soma, dendrites, axon, and synapse [3]. (e) Abstract illustration of a typical neuron

Then, the received spiking signals would be integrated together at the soma, and if the magnitude of the integrated value exceeds a specific threshold voltage, the soma generates and launches a sequence of spiking signals to the axon. Therefore, the function of an axon is similar to a transmitter sending the signals out of the neuron. At last, the axon sends the spiking signals to the dendrites of other neurons through synapses. Through the synapses, the magnitude of the signals propagated from the presynaptic neuron to the postsynaptic neuron can be either attenuated or amplified. This feature of synapses is referred to as plasticity.

The length of the axons of neurons is at the range of 0.1 mm to 2 m. The spiking signal (membrane potential) propagation within a neuron is a fundamentally different mechanism with

current/voltage signal traveling in conductive metal. Action potentials are generated by a sudden flow of Na ions (Na+) between the interior side and exterior side of the neuron. The flow of Na+ is generated through the opening and closing of the channels in the cell membrane. With an input signal exceeds the threshold of a neuron, the local high voltage in the membrane of the neuron stimulates the behavior of opening of channels allowing Na+ to flow between the interior and exterior of the membrane of the neurons. The sequence opening and closing behavior along with the membrane of the neuron body lead to a spiking signal propagation [3].

A synapse acting as a connecting organ between neurons in a nervous system [3]. At average, there are 10^{11} neurons in human brains with synaptic connections in a range of 1000 to 10,000, leading an amount of 10^{14} to 10^{15} synapse existing in the human brains. The signals transferring between neurons rely on the synaptic transmission.

The synapse is firstly introduced by Ramon Cajal in the late 19^{th} century using simple light microscopy. The size of the synapse is at the range of 4 nm to 40 nm. In the chemical synapse, there is no real physical connection between pre- and post-synaptic neurons. Thus, the signal is conveyed from presynaptic neurons to postsynaptic neurons through chemical neurotransmitters.

While a presynaptic action potential arrives at the end of the presynaptic neuron as shown in Figure 2-2, the voltage-gated (Ca_{2+}) channels open, triggering a biochemical reaction that releases neurotransmitter into the synapse region [3]. Then, the neurotransmitter diffuses from presynaptic ends to the postsynaptic end and further bind to their receptors on the postsynaptic cell. More specifically, the synaptic transmission can be summarized into two phases: release phase and bonding phases. In the releasing phase, a chemical messenger containing neurotransmitter is released from the terminals of the presynaptic cell. In the bonding phases, the released neurotransmitter binds to and stimulates the receptor molecules in the postsynaptic cell as illustrated in Figure 2-2. Moreover, the neurotransmitter crosses a small distance (4nm to 40nm) to arrive at the postsynaptic neuron, a time delay will be observed between the membrane potentials (spikes) from presynaptic neuron and the postsynaptic neuron. Through this way, the signals from presynaptic neurons are conveyed to the postsynaptic neuron through the neurotransmitters.

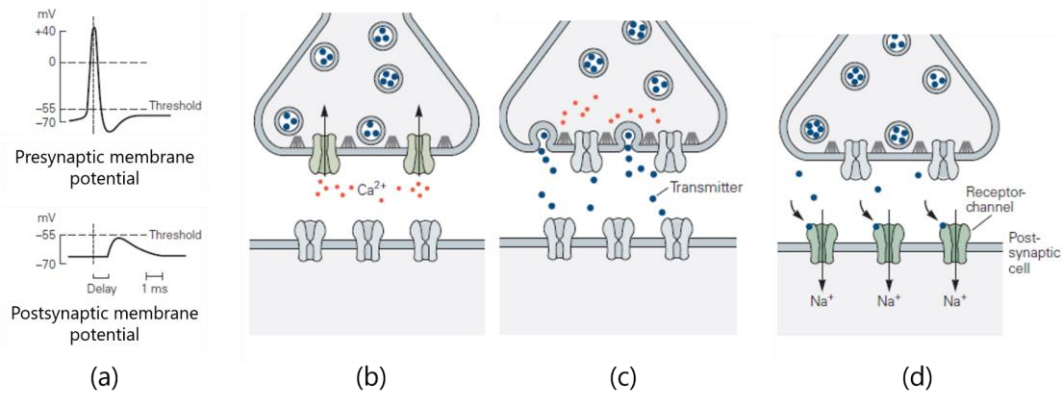


Figure 2-2: Illustration of synaptic transmission between neurons: (a) Presynaptic membrane potentials (spike) and the stimulated postsynaptic potentials (spike) with a time delay. (b) Membrane potentials (spikes) arrives at the terminal of a presynaptic neuron stimulating voltage-gated channels open. (c) The opened channel produces neurotransmitters to diffuse from the presynaptic neuron. (d) The released neurotransmitter molecules diffuse across the synaptic region and eventually bind the receptors on the postsynaptic neuron [3].

During the synaptic transmission, the spiking signal delivered to the postsynaptic neuron can be either attenuated or amplified which is determined by the connecting strength, which is referred to as the plasticity of a synapse. The plasticity of synapses has been widely believed as a critical feature for memory and other functions of human brains [3].

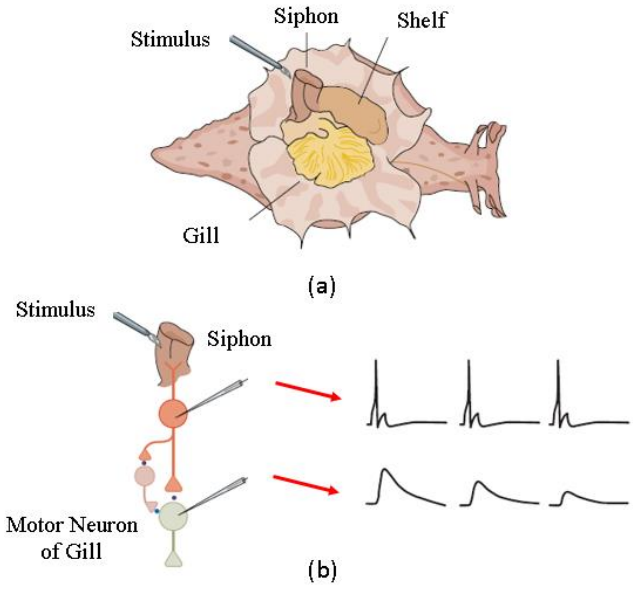


Figure 2-3: Connection strength is gradually strong by repeated stimulus signals [3]: (a) The experiment setup; (b) The signals are monitored at both sensory and responsive neurons.

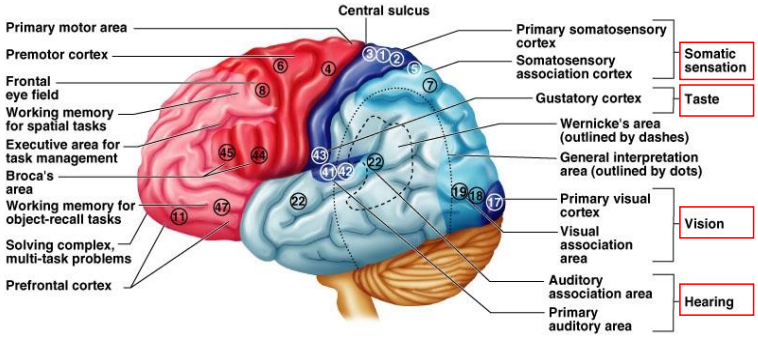
Habituation is one form of an implicit memory that can be interpreted and explained by a modification of synaptic plasticity. The habituation is described when an animal learns to ignore or present less active to a repetitive exterior stimulus, like the loud and noisy sound. The phenomena of habituation are prevalently existing in high-level animals. But their nervous system is quite complex. Here a sea slug, called *Aplysia California,* is selected for introducing the mechanism of synapse in habituation due to its relatively simple nervous system. The nervous system of Aplysia only contains 20,000 central neurons [3].

As shown in Figure 2-3, Tactile stimulus signals were repeatedly applied to the siphon of sea slug; then monitored at both sensory and response neurons. According to the number of stimuli applied to the siphon, the signal magnitude of the response neuron (in this case the motor neuron of the gill) reduced gradually, indicating a diminution in the strength of the connection between the sensory and responsive neuron. The experimental results are as depicted in Figure 2-3, wherein the amplitude of the response waveform is observed to decay over time. In the experiments, the stimulus was repeatedly applied to the Aplysia's sensory neurons as shown in Figure 2-3. This experiment indicates that the sea slug presents a smaller and smaller response (shrink of gill) to the stimulus from the siphon when the stimulus repeats many times, according to the definition of habituation. Meanwhile, the smaller response comes from the smaller amount of released neurotransmitters under repetitive stimulus signals. This phenomenon is widely considered to be a memory mechanism operating at the cellular level.

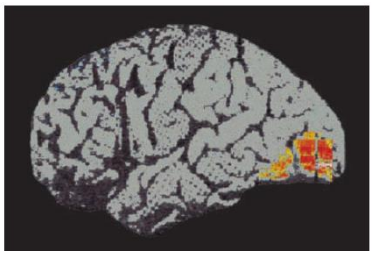
During studying different neurons in brains, one interesting fact captures people's attention and curiosity that the shapes and durations of the spiking signals in a nervous system are almost the same (spikes), whatever the signals generated by the sensation of light, of touch, or hearing.

This fact raises several reasonable questions that if the spiking signals are stereotyped reflecting no properties of the stimulus, how do the neural signals carry and convey specific behavioral information? How does our brain distinguish the spiking signals from seeing a flying butterfly, or smelling a flower? After decades of research on the human brain, the neurologists reveal that the signals of distinct sensations are routed and processed in different regions of the brain and the signals are distinguished by the signal pathway rather than their particular magnitudes or shapes [3].

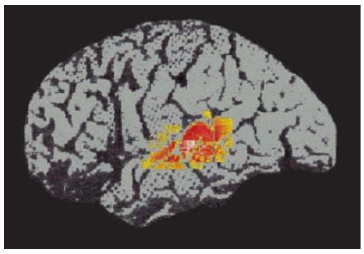
Figure 2-4 illustrates the regions of the human brain identified with Positron Emission Tomography (PET) images [3]. The PET is an imaging technique for visualizing the cerebral blood flow and metabolism accompanying mental activities [3]. As illustrated in Figure 2-4 (b, c), the visual pathways activated by receptor cells in the retina responding to light are completely different to the auditory sensory pathways activated by sensory cells in the ears that respond to sound. Thus, the function of the signal is determined not by the signal itself but by the pathway and the processing regions in the brain.



(a) The brain regions



(b) Positron Emission Tomography (PET) image of looking at words



(c) Positron Emission Tomography (PET) image of listening to words

Figure 2-4: Distinct signals are processed in different regions of the cortex. (a) The regions of the human brain. (b) Positron Emission Tomography (PET) image of recognizing a written word (c) Positron Emission Tomography (PET) image of recognizing a spoken word. The PET images are from the left side of the cerebrum, which represents the averaged brain activity. The red and yellow colors represent high activity, and blue and gray colors indicate low activity.

2.2 Neuromorphic Computing

Neuromorphic Computing is a concept developed by Carver Mead [19] in the 1980s. The concept of neuromorphic computing expounds an idea of realizing artificial intelligence by

physically rebuilding the organs and structure of human brains. Specifically, Prof. Carver Mead proposes to utilize the very-large-scale integration (VLSI) technology to mimic neuro-biological architectures in the nervous system.

Neuromorphic Computing [11, 20-59] exhibits a path of designing an untraditional non-von Neumann architecture system that potentially has the capability of achieving a high-performance artificial intelligence system. As we know, the human brain can process real-time signals with remarkable low-power consumption (about 20 Watts). Furthermore, the animals can adjust their behaviors according to the changes in surrounding environments and memorize and learn from their experiences. Neurologists believe these advantages are attributed from the network-based structure of neural system processing signals in parallel, and the neurons performing computing in low frequency. These unique signal processing characteristics are fundamentally different from the modern digital computers built upon von Neumann architecture computing Boolean and arithmetic operations.

Unlike the network configuration of the nervous system, the electronic computer is proposed and designed for a superior computing capability compared to mechanical computers. The underlying computing methodology of electronic computer is depicted in Figure 2-5 (a). The target problem is first abstracted by mathematic formula and then be calculated with the algorithms. The corresponding algorithms and calculations are performed within the hardware constituted with electronic devices. The capability of von Neumann computers relies on the speed of operations of algorithms. In order to efficiently perform these formulae, Von Neumann proposed the architecture in the 1940s [60, 61].

Since the digital computer operates Boolean and arithmetic computing, the information of the real world needs to be encoded into binary format through an analog-to-digital converter and a digital-to-analog converter. The binary data further would be used for executing the calculations in the arithmetic logic units (ALU) [62-64] as the critical computing units in central processing units (CPUs). In von Neumann architecture, the computing units and memory units are located separately connecting with a system bus as shown in Figure 2-5 (b). The transmission of data between the CPUs and the memory relies on the bus and the high-frequency signals carried on it. However, as the density of data continuously escalates, transferring valuable information back and

forth between CPUs and memory becomes computationally expensive. Specifically, as the rising of the data-driven artificial intelligence (deep learning), a quantity of data is dramatically increasing these years. The traditional von Neumann architecture based on a digital computer is not practical and applicable anymore since the data transferring congestion and dramatically large power consumption spent on data transferring [65]. The computer scientists and the neurologists share the same opinion that the performance difference between the digital computer and the human brain mainly comes from their distinctions of computing methodology and structure as listed in Table 2-1.

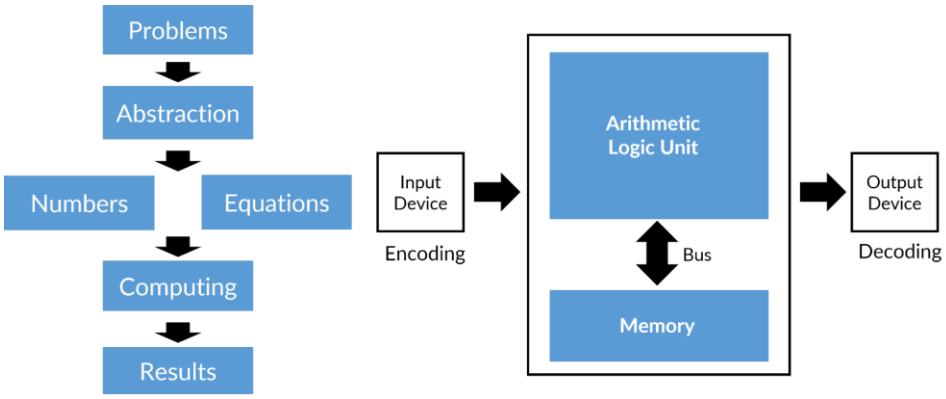


Figure 2-5: von Neumann computing system: (a) the underlying computing scheme of von Neumann computing architecture; (b) the hardware structure of von Neumann architecture

Table 2-1: Comparison between the computer and human brain

	Computer	**Brain**
Processing elements	The execution modules in CPU	Soma
Computing	Arithmetic & Boolean	Integration (summation)
Memory device	SRAM, DRAM, etc.	Synapses
	Storing the information in a binary format with two states	The information is stored as the connecting strength (weight) in analog format with multiple levels

Signal transmission/ communication device	Bus (Transmission lines)	Dendrites and Axons
Signal Format	Binary signals	Spiking signals
Learning/programming methods	Manually-coding programs	Self-learning
	Computer languages, like Java, PHP, Pythons, etc.	Associative memory
Communication scheme	One-to-one (CPU-to-memory)	Networks
System-level complexity	Low	High
	CPU connecting with memory through a communicating bus	20 billion neurons in a complex neural network topology
Operating frequency	High	Low
	~Gigahertz Level	~ Kilohertz Level
Power Consumption	High	Low
	~Kilowatts	~20Watts

On the contract, Neuromorphic Computing attempts to mimic the real-time power-efficient working mechanism of the nervous system through rebuilding three critical components of the nervous system: (1) neuron, (2) synapse, and (3) neural network architecture. Figure 2-6 illustrates the main difference between the human brain and the von Neumann architecture from the device to the algorithm levels. In a brain-like neuromorphic computing system, the building devices (computing units and memory units) need to be replaced from traditional CPUs and SRAMs to the electronic neurons and synapses. This is the first step for mimicking the brain at the device level. Unlike the computing units in the CPUs which perform binary code-based computing, the data in the electronic neurons and synapses need to be represented in a spiking sequence format for

generating the brain-like signals [3]. Then, these electronic neurons and synapse are interconnected with each other in a brain-like neural network configuration at the architecture level, which is demonstrated in Figure 2-6. Table 2-2 compares the contemporary state-of-the-art fabricated neuromorphic chips with a neural system of human brains.

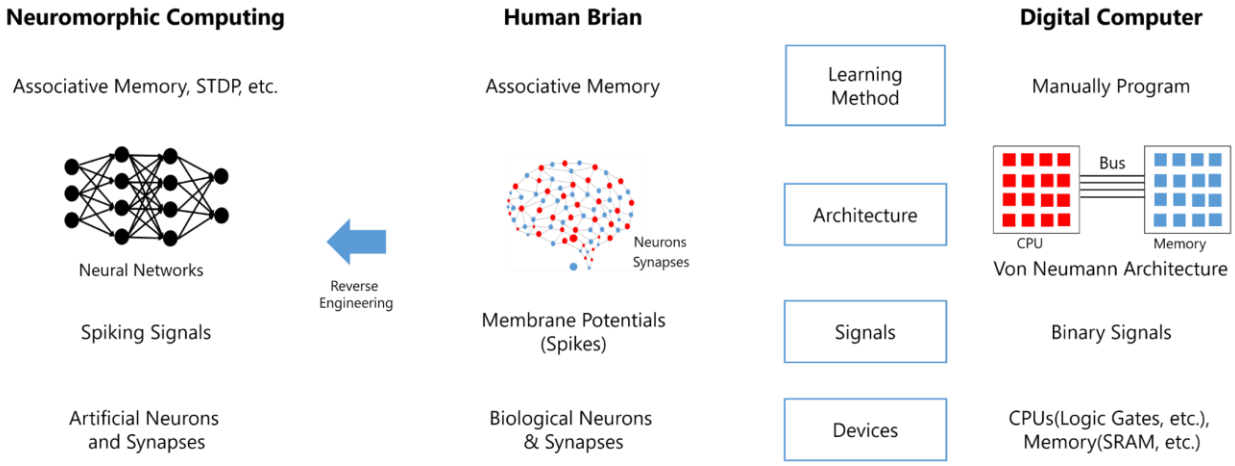


Figure 2-6: Comparison between brain computing architecture, von Neumann computing architecture, and neuromorphic computing architecture.

Table 2-2 Comparison of Power Density, Neuron Density, Synapse Density, and Neuron Connection Degree

	TrueNorth [30, 66]	**Neurogrid [30, 35]**	**BrainScaleS [30, 67]**	**SpiNNaker [30, 47, 48]**	**Brain [30, 68]**
Neurons	1,048,576	65,535	511	20,833	20 Billion
Synapses	256 million	N/A	113,636	20,833,333	200 Trillion
Area/volume	430mm^2	168mm^2	50mm^2	102 mm^2	1130 cm^3
Neuron Density	2438.55 per mm^2	390 per mm^2	10 per mm^2	204 per mm^2	17,699 per mm^3
Synapses Density	0.595 million per mm^2	N/A	2272 per mm^2	204,248 per mm^2	177 million per mm^3

Ratio of synapses to neurons	244	N/A	222	1,000	10,000
Power density	0.15 mW/mm^2	18 mW/mm^2	57 mW/mm^2	0.012 mW/mm^2	0.0177 mW/mm^3

2.3 Memristor

Four decades ago, Professor Leon Chua mathematically postulated the concept of memristor describing the relations symmetrizing the four fundamental circuit variables. However, the physical device of memristor has not been found. Although the phenomenon of nonlinear switching of the resistor had been observed and studied for many years, no one connects this particular phenomenon of resistors to the concept of memristor until HP Labs unintentionally and successfully found a connection between their results and the concept of memristor in 2008. In this section, a short review of memristor history is introduced, and after that, the characteristics of memristors are discussed in detail. The history of memristor exploration is summarized in Table 2-3.

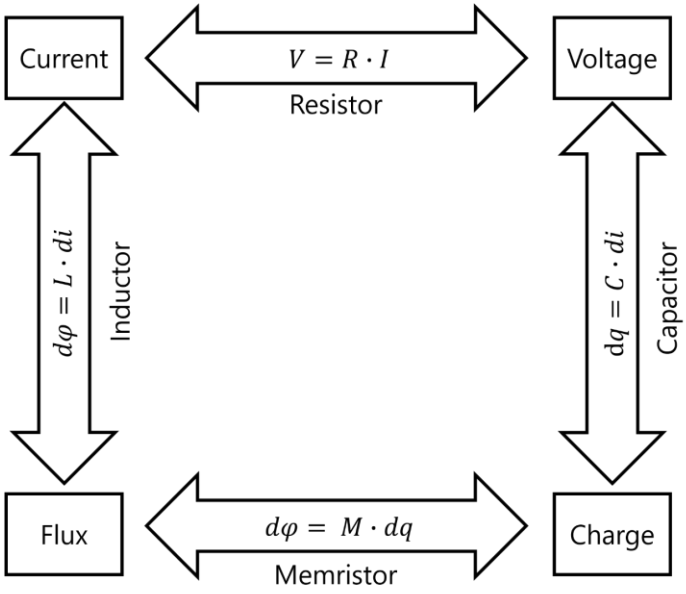


Figure 2-7: Relationships between the four basic circuit variables

In electrical engineering, there are four basic circuit variables, which are Current *i*, Voltage *v*, Charge *q*, and Flux *φ*. Correspondingly, three fundamental circuit components reveal the relationship between them. Figure 2-7 shows these three relationships. Resistors reveal the relationship between the voltage and current, while the relationship between *voltage and charge*, and *current and flux* are defined by a capacitor and an inductor, respectively.

Table 2-3: History of Memristor Exploration

Year	History of Memristor Exploration
1967	J. G. Simmons & R. R. Verderber described a hysteretic resistance switching phenomenon in a silicon oxide thin film with gold ions injected [69].
1968	F. Argall recorded a resistance switching phenomenon in a metal oxide thin film [70].
1971	Leon Chua mathematically postulated the concept of memristor describing the relations symmetrizing the four fundamental circuit variables, which is a similar mathematical attempt by Constantine A. Balanis in electromagnetics [71]. In his paper, he states that there should exist another circuit element that relates the charge and flux [72].
1998	Bhagwat Swaroop, William West, Gregory Martinez, Michael Kozicki & Lex Akers demonstrated an approach of using an ionic programmable resistance device for minimizing the complexity of an artificial synapse. [73]
2008	Dmitri Strukov, Gregory Snider, Duncan Stewart & Stan Williams at HP Labs published an article in Nature introducing a relationship between the two-terminal resistance switching characteristic of TaOx [74, 75].
2008	Leon Chua, Stan Williams, Greg Snider, Wolfgang Porod, Massimiliano Di Ventra, Rainer Waser, and Blaise Mouttet provided a discussion at the Symposium on Memristors and Memristive Systems talking the theoretical foundations of utilizing memristor for RRAM and neuromorphic architectures [76].

Initially, the missing part in Figure 2-7 is the relationship between the flux and electric charge. Thus Prof. Chua predicted that there should be another basic circuit device representing the relationship between the flux and electric charge based on symmetry. He called this hypothetical element as *memristor*. After almost four decades, finally, the HP labs build the memristor in physical form and further publish the paper introducing their discovery in Nature in 2008 that connects Chua's memristor and the resistance switching characteristic of the nanoscale device fabricated with $Pt/TaO_x/Pt$ as illustrated in Figure 2-8.

One of the main characteristics of a memristor is having a nonlinear and butterfly-shaped current-voltage curve [72, 74-76] as illustrated in Figure 2-9. With the increasing frequency, the loop of the current-voltage curve is going to shrink until it becomes a line at a large frequency.

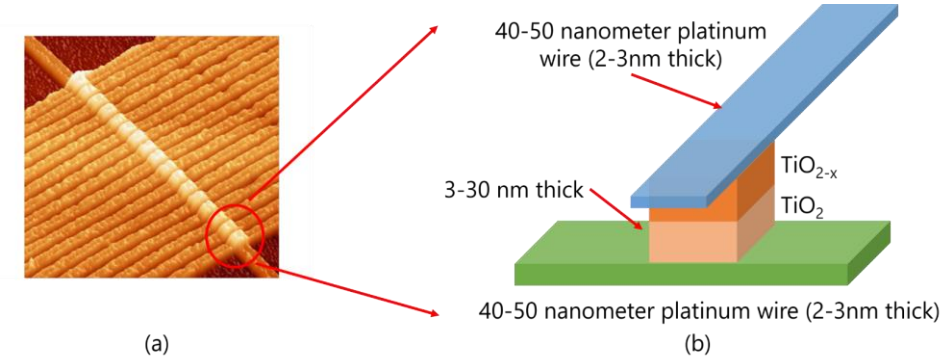


Figure 2-8: The memristors from HP labs: (a)The scanning tunneling microscope image of a memristor fabricated by HP Labs (b) the memristor cell is located at the cross-point of the crossbar structure with a 40-nanometer cube of titanium dioxide (TiO_2) in two layers. The lower layer is traditional of titanium dioxide with a 2:1 oxygen-to-titanium ratio.

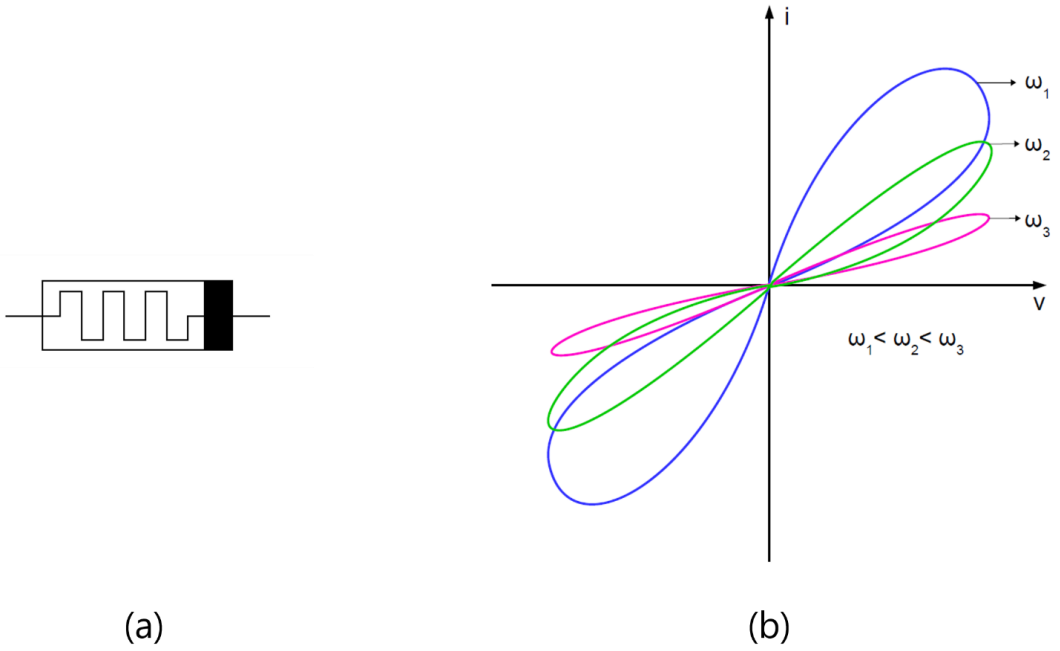


Figure 2-9: (a) Symbol of a memristor; (b) Current-voltage characteristic of the memristor with loops at different frequencies where $\omega_1 < \omega_2 < \omega_3$[76].

Chapter 3. High-Performance Neuromorphic Computing System with Low-variation Memristive Synapses

3.1 Introduction

Deep Neural Networks (DNNs) inspired by the high-degree structure of neural networks in mammalian brains have accomplished remarkable success in many applications, such as image recognition, natural language processing, machine neural translation [6], etc. A pristine DNN with random synaptic weights has no remarkable capability until its weights are trained by tremendous data. The larger sizes of the datasets and the neural networks lead to a higher inference accuracy [7, 8]. Thereby, the demand for excessively large datasets and neural networks is becoming inevitable.

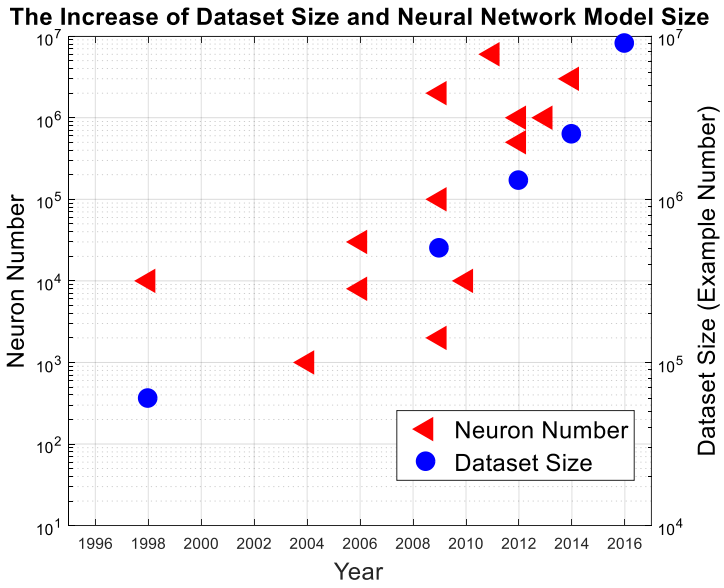


Figure 3-1: Increase trend of the dataset and neural network sizes [8]

As illustrated in Figure 3-1, the size of datasets is almost linearly increasing over the years, while the neural networks double their size roughly every two years [8, 10]. Accompanying the

growth of the scale of hypermeters, the capacity of the GPU memory has only increased by a factor of three [7, 8]. Hence, there is an urgent need for novel and reliable devices with higher capacity and lower power consumption, fulfilling the tremendous data storage demand for deep learning.

Nowadays, memristors are widely considered as one of the most promising candidates for next-generation memory because of its high density and low power consumption [6]. However, its wide distribution of resistance variation restricts its feasibility in deep learning as weight storing devices [77, 78], since the weight variation significantly reduces the inference accuracy [78-83]. Several methods involving circuit and algorithm optimizations have been proposed to mitigate this shortcoming. However, these methods entail inevitable drawbacks, like the large latency and circuit design overhead [84-86].

In this dissertation, the switching mechanism of memristors is studied to reveal the heat accumulated in the cell during the switching leads to a substantial metal atom diffusion effect. The metallic atoms diffusion at the tip ends of the conductive filaments (CFs) influences the gap size among of the filament in the off-regime when the filaments are ruptured [82]. As a result, the resistance variation increases significantly when the heat is accumulated interiorly [83, 87, 88]. In order to mitigate the resistance variation, a novel configuration of a memristor is designed and fabricated with an additional heat dissipation layer integrated into the cell's electrodes alleviating the heat-related switching variation by more than 30%. Unlike using low thermal conductivity material for subduing heat transfer between layers [89], the proposed approach dissipates the accumulated heat both on the metal and insulator layers. The candidates of the heat dissipation layer need to satisfy several requirements, such as high thermal conductivity, low cost, fabrication compatibility, electrochemistry stability at high temperature, etc. Several materials (Rh, Cr, Pt, Ti, Cu) have been tested for heat dissipation efficiency. It turned out that the Ti glue layer used for the adhesion of the inert electrode had to be supplanted by Cr with the most thermal conductivity to render the Joules heating effects less severe.

Furthermore, an experimentally verified memristor model capturing the electrical characteristics has been built. This memristor model is incorporated in the deep delay-feed-back reservoir computing (Deep-DFR) model for evaluation. The Deep-DFR is established by the system-level simulation platforms comprising PyTorch and *NeuroSIM* [81]. The parameters of VT

memristors in NeuroSIM are extracted from the measurement data. Through the proposed Deep-DFR model, the impact of reducing the switching variations of the memristor on a deep learning system is analyzed. The simulation results demonstrate that the accuracy has been increased by ~30% accompanying the reduction of the resistance variation of the memristor. The accuracy improvement, power consumption, design area, and latency reduction are evaluated with CIFAR-10 and CIFAR-100 datasets. The contributions are summarized as follows:

- A novel memristive device configuration with higher immunity to degradation induced by thermal effects has been fabricated and evaluated. The experiment results demonstrate a ~30% reduction in switching variation;
- The competent material for heat dissipation layer is determined;
- The accuracy improvement (~30%) on classification tasks is demonstrated through the Deep-DFR model, which deploys the proposed memristor model;
- The hardware performance improvement, e.g., power efficiency and design area reduction, is evaluated and analyzed through a co-simulation paradigm with PyTorch and the macro-circuit simulator *NeuroSIM* [81].

3.2 Memristors as Synapses

The plasticity of a synapse can be implemented as a new non-volatile device *Memristor*, which is also widely referred to as Resistive RAM (RRAM) [74, 90-101]. A typical memristor is constructed in a metal-insulator-metal (MIM) configuration, as illustrated in Figure 3-2 (a). Figure 3-2 (b-f) demonstrates an overall view of the VT memristive devices fabricated at the Micro & Nano Fabrication Laboratory at Virginia Tech (http://www.micron.ece.vt.edu/) [102]. The memristors are fabricated in a crossbar configuration on a thermally oxidized silicon wafer (730 nm thick) shown in Figure 3-2 (b) and (c). as illustrated in Figure 3-2(f), the memristor is fabricated in a 5 by 5 array, containing 25 devices. Figure 3-2 (e) demonstrates the details of each memristor, which is located at the cross-point of two accessing nanowires crossbar.

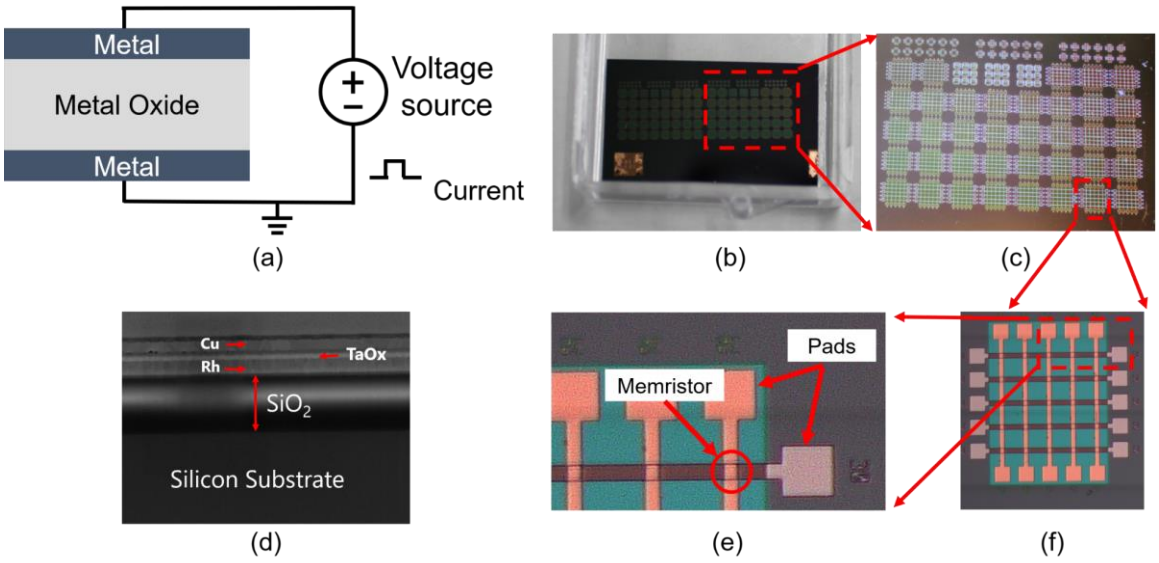


Figure 3-2: Memristor structure: (a) Schematic of metal-insulator-metal (MIM) with a voltage source stimulus. (b) VT memristor arrays at the wafer. (c) VT memristor arrays (d) Focused Ion Beam (FIB) cross-section image of a VT memristor. Cu/TaOx/Rh forms a one-layer memristor (e) zoom-in view of a memristor at the cross-point of the array. (f) a typical two-layer memristor cell.

At the ends of the nanowires, the pads are fabricated to place a testing probe. Figure 3-2 (d) depicts the Focused Ion Beam (FIB) cross-section image of a VT memristor forming with Cu/TaOx/Rh. The metals from the top and bottom contacts and the insulator is usually a resistive switching material [103, 104]. Its resistance is changeable through the construction and deconstruction of the conductive filaments in the oxide layer between two metal layers. There are four critical phases of the resistive switching process of a memristor. As illustrated in Figure 3-3, initially, the atomic structure of the memristor at its pristine state is intact, which is referred to the pristine stage. At this stage, the bonding between oxygen ions and metal atoms is strong. However, this bonding between oxygen ions and metal atoms is not unbreakable. Under the high electric field established by the applied voltage at the metal terminals of the memristor, some oxygen ions in the metal oxide would escape from the constraint of the bonding force [78]. Thus, the deficiency of oxygen ions leaves the oxygen vacancies or metal precipitates, consequently constructing the conductive filaments as depicted in Figure 3-3 [78, 105, 106]. The conductive filaments would provide an alternative current pathway resulting to decrease the resistance of the memristor. This transition process of the resistance from high to low is defined as a set process. On the opposite, the oxygen ions at the interface migrate back into the oxide to refill the oxygen vacancy or re-

oxidize the metal precipitates in the reset process. After this reset process, the resistance of the memristor would restore to its high resistance value.

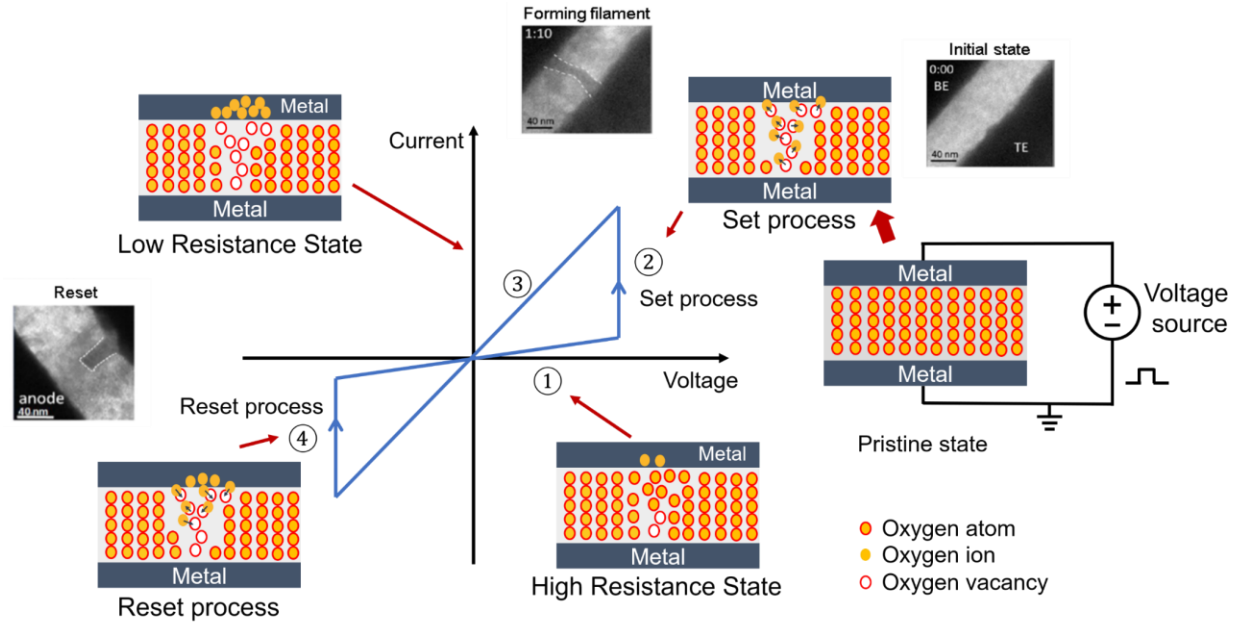


Figure 3-3: Illustration of the switching mechanism of a memristor. The memristor has two states (HRS and LRS) marked as ① and ③, and two transition states (set and reset processes) marked as ② and ④, respectively. Note that this paper would mainly focus on modeling the set process indicated as a remembering process instead of a biological disremembering process. TEM images of the dynamic evolution of conductive filaments [10].

As illustrated in Figure 3-3, when the voltage/current stimulus is applied on the terminals of memristors and it exceeds a specific value, the resistance of memristors will gradually change between its high resistance state (HRS) and low resistance state (LRS). The decrease in resistance of the switching material is due to the formation of the conductive filaments (CFs). As shown in Figure 3-3. (b). The phenomenon is called a soft breakdown in material science. This breakdown of the switching material can be recovered by applying a reversed stimulus at the terminals, which consequently resets the memristor to its high resistance state as shown in Figure 3-3.

3.3 Robust Deep Reservoir Computing through Reliable Memristors with improved Heat Dissipation Capability

As one of the most promising candidates of next-generation memory, memristive devices suffer a critical issue of low reliability, which diminishes its practicability for massive deployment [77, 78]. The low reliability of a memristor stems from the high variation on its on-state resistance (R_{on}) value [83]. Through the comprehensive study of the switching mechanism of a memristor [107, 108], we have discovered that the heat-related metal atom diffusion of conductive filaments (CFs) increases the resistive switching variation [109]. In order to address this issue, we designed and fabricated a novel configuration of a memristor, which can effectively mitigate the heat-related resistive switching variation.

A memristor is typically fabricated using a metallic oxide layer as a solid electrolyte sandwiched between an oxidizable active anode electrode and an inert cathode electrode. As illustrated in Figure 3-4, there are four resistively switching phases of a memristor. Initially, the atomic structure of the metallic oxide layer is intact.

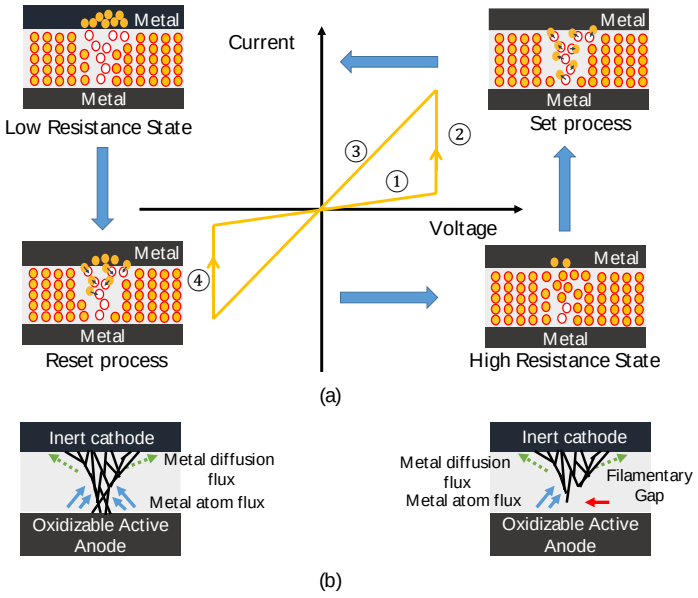


Figure 3-4: (a) Four typical switching phases of a memristor; (b) Formation mechanism of conductive filaments. The variation of the on-state resistance of a memristor results from a competition between the constructive metal atom flux and destructive metal atoms diffusion flux [83, 107].

At this stage, the bonding between oxygen ions and metal atoms of the metallic oxide is strong. However, under the high electric field established by the applied voltage to the cell's electrodes, the oxygen ions in the metallic oxide could be dislodged from the constraint of the bonding force and migrate to one of the terminals of the memristor. Consequently, the removal of oxygen ions leaves the oxygen vacancies behind leading to a build-up of conductive filament connecting the two electrodes. In another mode, the atoms of the active electrode are ionized and under the applied electric field migrate to the inert electrode where they are stopped and electrically reduced. Over time the active electrode metal atoms pile up on each other leading to a formation of metallic filament connecting the two electrodes. When this happens, the cell is in an on-state characterized by an on-resistance $R_{on.}$ Otherwise, the cell is in the off-state characterized by the off-resistance R_{off}. The ratio between R_{off} and R_{on} is large and exceeds in many cases 10^3. The switching process of the resistance from R_{off} to R_{on} is referred to as a set process. In contrast, the transition from R_{on} to R_{off} is called the reset process.

As illustrated in Figure 3-4 (a), the switching capability of memristors attributes to the construction and rupture of the conductive filaments. The shape and the size of the filaments could significantly influence the switching characteristic of a memristor.

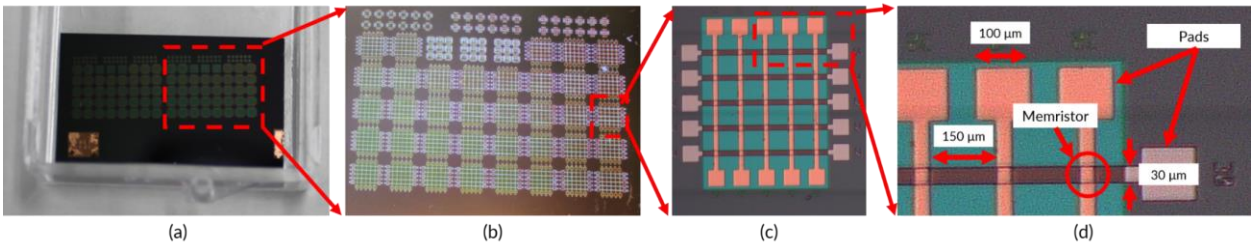


Figure 3-5: VT fabricated memristor die: (a) The overview of VT memristor die; (b) The zoom-in view of VT memristor; (c) The five by five crossbar structure of VT memristor; (d) The memristor located at the cross-point of the crossbar.

Table 3-1: Comparison of the memristor resistance switching variation

Device	Size (nm)	Thermal Conductivity	R_{on} (I_{cc} = 5 uA)	Target Value	Variation[1]	R_{on} (I_{cc} = 50 uA)	Target Value	Variation

Cu/TaOx /Rh/Cr	150/25 /50/20	Rh:150 Cr: 94	2.5 ± 0.1 KΩ	2.4 KΩ	~4 %	500 ± 5 Ω	500 Ω	~1 %
Cu/TaOx /Rh/Ti	150/25 /50/20	Rh:150 Ti: 20	2.3 ± 0.12 KΩ	2.4 KΩ	~5 %	225 – 750 Ω	500 Ω	~35 %
Cu/TaOx /Pt/Cr	150/25 /50/20	Pt: 72 Cr: 94	2.1 ± 0.1 KΩ	2.1 KΩ	~4.7 %	331 – 1000 Ω	400 Ω	~33.4 %
Cu/TaOx /Pt/Ti	150/25 /50/20	Pt: 72 Cr: 20	2.1 ± 0.9 KΩ	2.1 KΩ	~42.8 %	230 – 1000 Ω	400 Ω	~61.5 %

[1] The variation is the cycle-to-cycle variation that is measured by percent deviation.

During the set and reset switching processes, the considerable current flows through the CFs generally leads to a significant Joules heat dissipation. The temperature of the memristor cell is governed by the Joules heating and the rate of heat removal, which is determined by the thermal conductivity of the surrounding metallic oxide and the thermal conductivities of the electrodes. If the surrounding metallic oxide or the two electrodes cannot dissipate the heat fast enough, the temperature of the filament is bound to increase. Eventually, the high temperature of the CFs triggers a substantial metal diffusion. The metallic atoms of the filament, particularly at the tip of the cone-shaped CFs, diffuse out of the CFs consequently determining the size in the filament [82]. Macroscopically, the on-state resistance variation increases significantly [83, 87, 88, 110]. This phenomenon is even more severe during the rupturing process as the reset is dominated by a thermal dissolution effect [109]. When current flows through the memristive cell, Joule heat is deposited in the conductive filament. As a result, the temperature in the narrowest part (highest resistance) of the filament can reach 1000 ℃ [111, 112]. Such a high temperature triggers Cu atom diffusion from the constriction of filaments.

In order to address this issue, we proposed and investigated a solution of adding an extra metallic layer for facilitating heat dissipation. The copper (Cu) is selected as an oxidizable active anode due to its medium activation energy-yielding ions readily [113] Cu ←→ Cu++e-. The rhodium (Rh) is used for inert cathode since it is compatible with the back-end-of-line (BEOL)

integration technique and potentially can be integrated on the top of the metal-oxide-semiconductor field-effect transistors (MOSFETs) for a three-dimensional structure [114]. Furthermore, the Rh-Cu material configuration demonstrates a negligible solid solubility between two elements, rendering Rh an ideal inert electrode for Cu ions (Cu+). Besides, the Rh is 45 times less expensive than Pt with similar characteristics [109].

The oxygen-deficient tantalum oxide (TaOx) is used as the metallic oxide. In this work, the memristor Cu/TaOx/Pt is used as a benchmark device. VT memristive devices have been fabricated in a crossbar configuration on a thermally oxidized silicon wafer. The metal electrodes and solid electrolytes are deposited through e-beam evaporation. The TaOx layer was deposited by evaporating the Ta_2O_5 pellets with no oxygen injection at the evaporation chamber. A thin Ti layer was added between Pt and SiO_2 to improve the adhesion of Pt to the substrate. All the layers (Cu, TaO_x, Pt) are deposited by e-beam PVD in a Kurt Lesker PVD-250 chamber. The fabricated memristor die and the detailed geometry are illustrated in Figure 3-5. The range of the high resistance state (HRS) is ~1-900 MΩ, yielding a ratio of $R_{off}/R_{on} \approx 10^3–10^7$, which effectively avoids the negative effect caused by the sneak path.

The reliability of the memristive devices with different inert cathodes is evaluated by the variation of their on-state resistance. The testing results are summarized in Table 3-1. In Table 3-1, the cycle-to-cycle variation is measured by percent deviation. The precise temperature control is not practical in real measurement setups. Thus, we distinguish different temperatures (high and low) by applying different compliance currents during the set operation; they are I_{cc} = 5 uA and 50 uA respectively. The heat generated by the different currents, assuming constant current in the time interval t, is governed by $w = I^2 R_{on} t$. Table 3-1 demonstrates that the memristive device exhibit a higher spread of on-state resistance (R_{on}) values with higher temperatures (larger compliance current). For example, the on-state resistances of the Rh/Ti configuration are at the range of 225 Ω to 750 Ω for I_{cc} = 50 uA. This instability phenomenon comes from the competition between the constructive Cu^+ electro-migration flux and the destructive Cu diffusion flux, illustrated in Figure 3-4 (b). The measurement results demonstrate an effective metal dissipation layer (Cr) could effectively suppress the heat-related metal atom diffusion phenomenon, resulting in a significant reduction of switching variation (by ~30%).

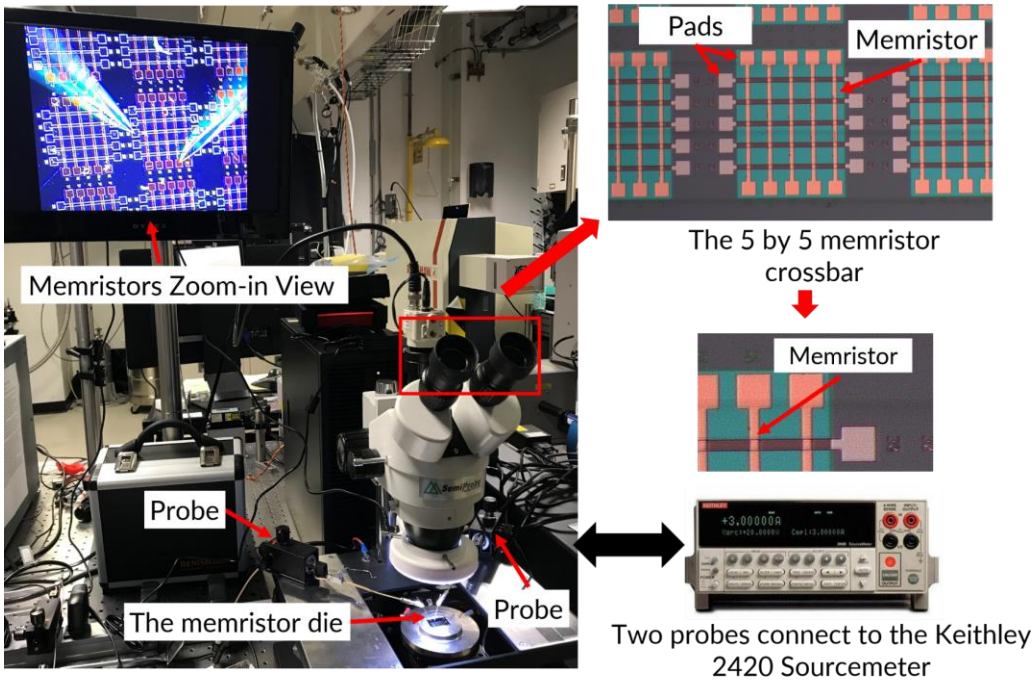


Figure 3-6: The testing setup of the memristor at Oak Ridge National Laboratory. The measurement was conducted at the Center for Nanophase Materials Sciences (CNMS) at Oak Ridge National Laboratory, which is a Department of Energy Office of Science User Facility.

The measurement is performed by applying a positive voltage to the electrode of the device and the voltage is swept at a constant voltage ramp rate (0.2V/s). Initially, the current remains small until the set voltage of the memristor is reached. The current switching is caused by the conductive filaments (CFs) formation when the applied voltage exceeds the set voltage of the memristor. The measurement usually performed more than 100 times. The variation is measured by the percent deviation from average, which shows the average percentage that a data point differs from the mean value.

The endurance of the devices depends on the compliance current (I_{cc}). For the I_{cc} is at the range of 10 μA and 5mA, the device can be switched more than 150 times. For smaller compliance current, like 1 μA, the endurance of the VT memristor device can be more than 1000 times switching. During the measurement, no incorrect switching of the unselect and adjacent memristor cells was detected which potentially caused by the sneak path issue. The high ratio of off-state and

on-state resistances of the VT memristor device (more than 10^3) effectively avoids the negative impact of the sneak path issue.

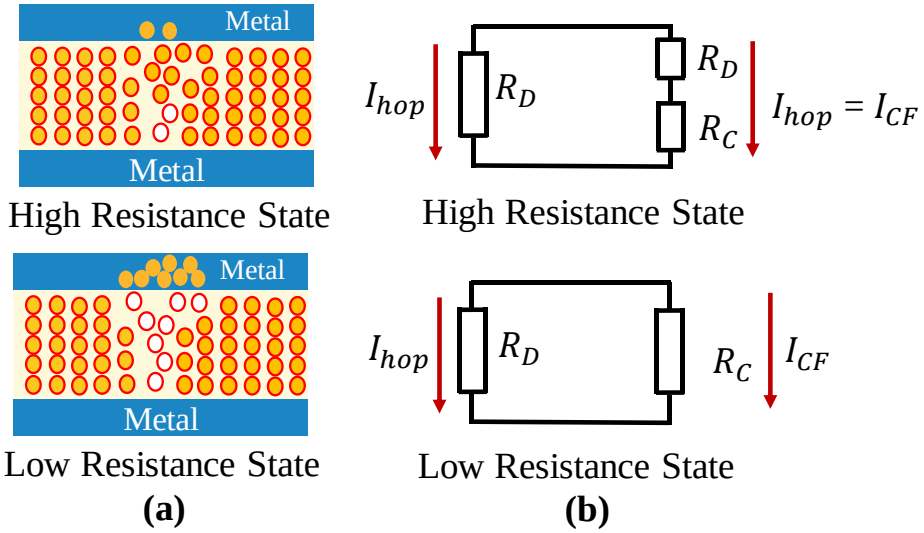


Figure 3-7: Current paths of the memristor at the HRS and LRS

Furthermore, to analyze the effect of resistance variation reduction of VT memristor on deep learning at a system level, a corresponding Verilog-A memristor model is built upon the filament growing method [115, 116]. The memristive synapse is modeled with the filament growing method [116] [117]. The resistance switching between HRS and LRS comes from the construction/deconstruction of the CFs in the metallic oxide. The CFs in the oxide provide an alternative current path with lower resistance. By modeling two current paths with different resistances, notated with R_D (dielectric resistance) and R_C (resistance of CFs), the memristor models at HRS and LRS are illustrated in Figure 3-7.

The growth of the conductive filaments determines the values of R_{on} and R_{off}. Hence, the conductive filaments of our memristor are modeled as a cylinder with parameters w for the diameter and x for the instantaneous height . The interval of x is at $0<x<H$, where H is the thickness of the metal oxide layer. The tantalum dioxide (TaO_x) is used as the metal oxide of our memristor. In the filament growing modeling method, the conductive filament is modeled as a cylinder with adjustable w and x, which represent an adjustable conductive filament width and length respectively. The nonlinear hopping current (I_{hop}) and current flowing in CFs (I_{CF}) with parameters of gap distance (x) and CF width (w) can be calculated [115, 116, 118] as:

$$I_{hop} = I_0 \left(\frac{\pi w^2}{4}\right) exp\left(-\frac{x}{x_T}\right) sinh\left(\frac{V_{gap}}{V_T}\right) \quad (3\text{-}1)$$

$$I_{CF} = \frac{\pi w^2 V_{CF}}{4\rho(x_0 - x)}, \quad (3\text{-}2)$$

where x_0 is the initial gap distance, x_T and V_T are the characteristic length and voltage in hopping, respectively. V_{gap} and V_{CF} are the voltage over the gap region and the conductive filament region, respectively. w contributes the Joules heat dissipated in the filament. In the reset process, the w and x grow with set voltage (V_{set}) by the Eq. (3-3) and (3-4), while the reset process is described by Eq. (3-5) to Eq. (3-7) [115, 116, 118]:

$$\frac{dx}{dt} = afex\,p\left(-\frac{E_a - \alpha_a ZeE}{k_B T}\right) \quad (3\text{-}3)$$

$$\frac{dw}{dt} = \left(\Delta w + \frac{\Delta w^2}{2w}\right) fex\,p\left(-\frac{E_a - \alpha_a ZeE}{k_B T}\right) \quad (3\text{-}4)$$

$$\frac{dx}{dt} = afex\,p\left(-\frac{E_i - \gamma ZeV}{k_B T}\right), \quad (3\text{-}5)$$

$$\frac{dx}{dt} = afex\,p\left(-\frac{E_h}{k_B T}\right) \sinh\left(\frac{\alpha_h ZeE}{k_B T}\right), \quad (3\text{-}6)$$

$$\frac{dx}{dt} = afex\,p\left(-\frac{\Delta E_r}{k_B T}\right). \quad (3\text{-}7)$$

The parameters from Eq. (3-3) to Eq. (3-7) are listed in Table 3-2 in details.

Table 3-2: Parameters of the Memristor Model

Parameter	Descriptions
I_0	Hopping current density in the gap region
ρ	Resistivity of the CF
a	Distance between adjacent oxygen vacancy

f	Vibration frequency of oxygen atom
x_T	Characteristic length in hopping region
V_T	Characteristic voltage in hopping
w_0	Initial CF width
R_H	High Resistance State
R_L	Low Resistance State
E_a	Average active energy
α_a	Enhancement factor
Z & e	Charge number & unit charge
k_B	Thermal resistance

Figure 3-8 illustrates the V-I characteristic curve comparison of the memristor model and the measurement data of VT memristors.

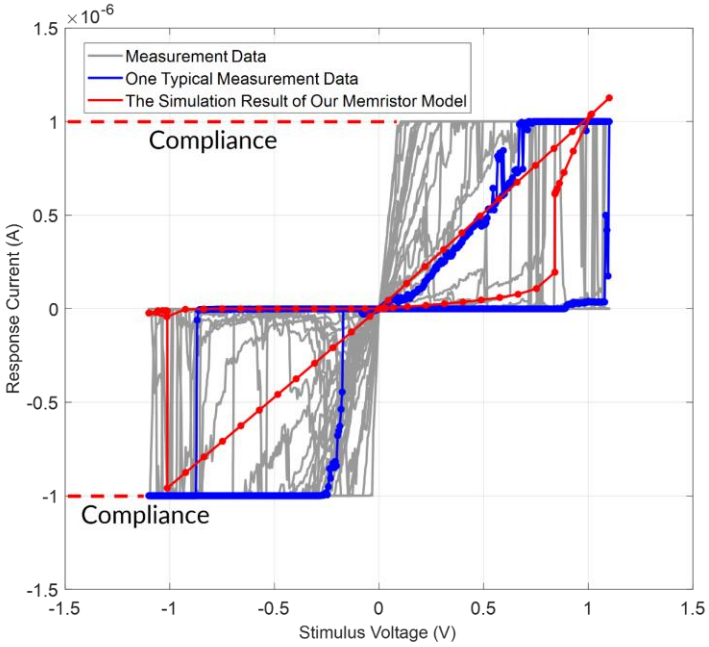


Figure 3-8: V-I switching characteristics of VT memristor (Cu/TaOx/Rh/Cr): The gray lines represent the measurement data, the blue line shows one typical measurement data, and the red line depicts VT memristor. Note: the compliance current is 1uA in this case.

As depicted in Figure 3-8, the resistance of the memristor model switches from ~1 MΩ to ~940 MΩ at V_{set} ~0.8V, which matches the measurement data. The sudden current cut-off at ± 1 μA in Figure 3-8 comes from the compliance current setting. The inconsistency of on-state resistance in Figure 5 and Table 1 comes from the different compliance current [78]. The relationship between R_{on} (low resistance state) and compliment current can be estimated by the equation:

$$R_{on} = \frac{K}{I_{cc}^{n}}, \tag{3-8}$$

where n and K are fitting parameters and I_{cc} is the compliance current [109]. Eq. (3-8) indicates the negative correlation between the compliance current and R_{on}.

3.4 Performance Evaluation of the memristors on Deep Delay Feedback Reservoir Computing

The emerging Deep-DFR demonstrates a strong capability of processing spatiotemporal data due to its recurrent loop and multiple layer structure [119, 120]. This specific structure allows the system to have more remarkable performance compared to other conventional reservoir computing system. Deep-DFR models demonstrate more than 50% better performance than the typical leaky echo state network (ESN) model [21, 121-125]. Furthermore, the Delay Feedback Reservoir (DFR) has a simplified structure, which merely consists of one nonlinear neuron in the reservoir [126, 127]. On the contrary, the traditional reservoir system requires numerous nonlinear neurons that demand more hardware resources increasing the hardware design challenge [128-132].

In this work, VT Deep-DFR model (Figure 3-9) is used for evaluating the impact of resistance variation reduction (cycle-to-cycle) of VT memristor on inference accuracy. In order to focus on studying the cause-and-effect between the resistance variation and the inference accuracy, other nonideal parameters of memristors that may influence the inference accuracy, e.g. device-to-device variation, are excluded (keeping constant) in this work. At last, The hardware performance improvement, e.g., power efficiency, latency, and design area, is evaluated through a co-simulation paradigm with PyTorch and *NeuroSIM* [81].

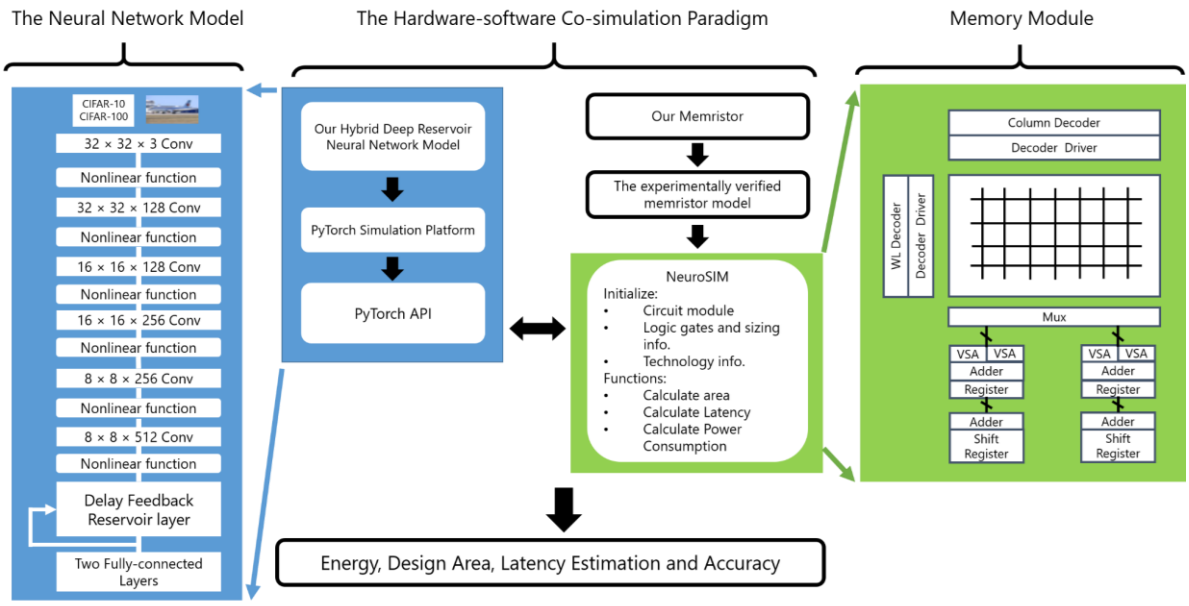


Figure 3-9: The diagram of the hardware-software co-simulation paradigm with NeuroSIM and PyTorch.

In this section, the crossbar configuration of the memristor as a memory array is introduced. Next, VT Deep-DFR model is introduced in detail. At last, the hybrid simulation paradigm is presented, combining the experimentally verified memristor model and the Python-based Deep-DFR model.

3.4.1 Weight Storage in Memristor Crossbar

Memristors typically are fabricated in a crossbar structure massively. As illustrated in Figure 3-10, the nanowires built with the inert cathodes and oxidizable active anodes are placed at the top and bottom of the crossbar, respectively. The metallic oxide layer is located at the cross points of the top and bottom nanowires. This crossbar structure is similar to the conventional memory array. As illustrated in Figure 3-10 (b), each memory cell of the memory array connects to a *wordline* and a *bitline*.

For example, the DRAM (Dynamic Random-access Memory) uses a capacitor for each memory cell, and the SRAM (Static Random-access Memory) generally has six transistors as one memory cell. The stored information is represented by the voltage states at the terminals of capacitor or transistor.

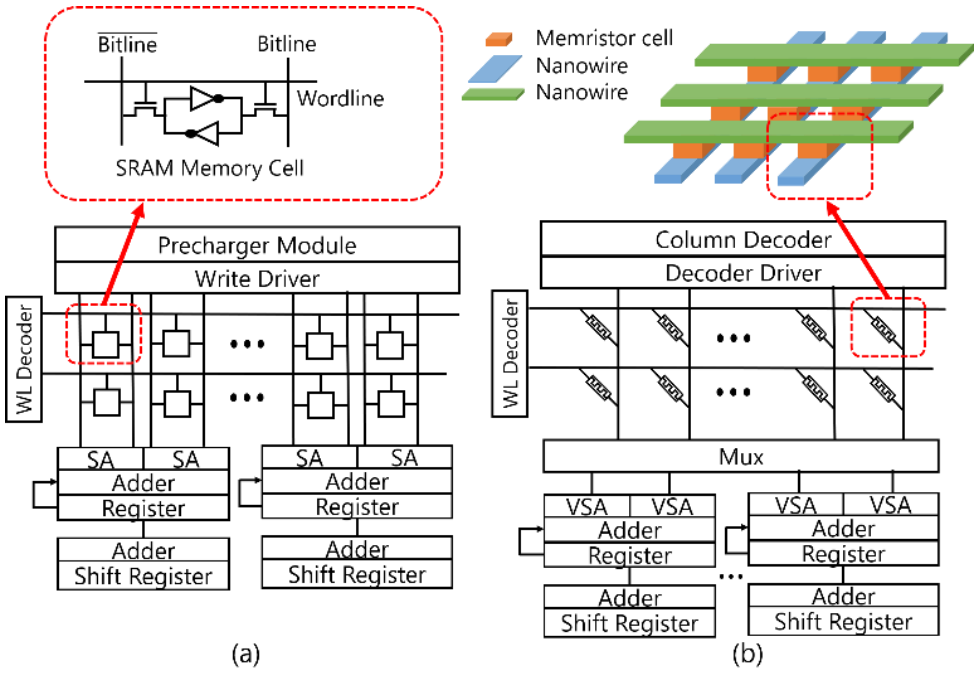


Figure 3-10: Configuration comparison between the memristive crossbar and the memory array with SRAM memory cells in NeuroSIM [81]: (a) the traditional memory array with SRAM (6T); (b) the structure of the memristive crossbar.

For memristor, the values are encoded in the resistance of a memristor and the nanowires serve as the bitline and wordline for accessing the memristive memory cells. In the writing phase, a voltage pulse, larger than set voltage, is applied to the nanowire of the crossbar structure and modifies the resistance value of the memristor. In the reading stage, the applied voltage is much smaller than the set voltage in order to preserve the resistance of the cell unaltered. The resistance value of the selected memristor equals the applied voltage divided by the measured current at the end of the nanowire. The weight matrices are mapped to the passive memristor crossbar with the memory cell selection devices, such as transistor or diode. The decoder of the system uses the *wordline* and *bitline* to access to every single memory cell. As illustrated in Figure 3-10 (a), the operations of weight sum and update in NeuroSIM are row-by-row-based write and reading [81]. The row selection is activated through the WL decoder. Then the BLs are precharged to each cell access. The memory data are captured by the sense amplifier (S/A). After that, the adder and register are used to sum the weight values in a row-by-row style. By replacing the SRAM core memory with the memristors, the architecture is not significantly modified. But the size of the memory cell reduces due to the intrinsic nanoscale of memristors. The weighted sum operation in

the memristor-based synaptic core is also a row-by-row style expect the use of multiplexers (Mux) [81].

3.4.2 Deep Reservoir Neural Network

Nowadays, hardware-friendly DFR demonstrates an impressive capability of processing temporal information [119, 120]. In this work, several convolutional layers are added for constructing a deep DFR structure. Figure 3-9 illustrates the details of the Deep-DFR structure. The six convolutional layers serve as feature extractor, which is followed with a delay-feed-back layer extracts the one-dimensional time series characteristics. Two fully connected layers are used for reducing the output dimensional serving as a classifier. The number of time delay reservoir layers matches the output of the convolutional layer. Initially, the weights in the reservoir ($\boldsymbol{W}^{res}$) layer is assigned as zeros. During the training process, the updating equation of the reservoir state is expressed as:

$$Res(t) = \alpha \times Res(t-1) + f_{nonlinear}(\boldsymbol{H}^{in}(t)), \tag{3-9}$$

where t is the time step, Res(t) is the reservoir state, α is the decay factor, $f_{nonlinear}$ is the nonlinear activation function, and $\boldsymbol{H}^{in}$ is the hidden layer. This equation reveals that the current state of the reservoir is not only determined by the current input but also highly related to the last time step.

To evaluate VT memristor performance, e.g., design area, accuracy, power consumption, a hardware-software co-simulation is established with *PyTorch* and *NeuroSIM* [81], as illustrated in Figure 3-9. The model is built as follows steps:

Firstly, the Deep-DFR model is built of six convolutional layers for extracting features, followed by a Delay Feedback Reservoir Layer, and two full-connected layers. There are no weights within the delay feedback loop [126]. The Deep-DFR model is trained on the PyTorch platform with CIFAR-10 and CIFAR-100 datasets. During the training progress, the weights and neural network configuration are monitored and stored.

Secondly, the experimentally verified memristor model is incorporated into the micro-architecture simulator *NeuroSIM* [81] including the set voltage, on-state resistance, off-state resistance. The resistance variation with different levels is incorporated in the memristor model in

NeuroSIM. To reveal intently the cause-and-effect relationship between resistance variation (cycle-to-cycle) and inference accuracy, other nonideal parameters of memristors are not included for eliminating the interference from them.

Algorithm 3.1: Performance Estimation

Initialize: The configuration of the Deep-DFR and the corresponding weights $\boldsymbol{W}_{i,j}$ with small random numbers

Initialize: $\mathbf{W}^{\mathrm{res}}$ of the reservoir as all zeros

Initialize: Memory cell configuration

Initialize: Peripheral circuits configuration

1 For epoch = 1, M do

2 **While** batch in dataset do

4 $\boldsymbol{y}_{conv}^{out} \leftarrow$ six convolutional layers to batch (input)

3 $\boldsymbol{h}_{res\_1} = \boldsymbol{W}_{res}^{in} \times \boldsymbol{y}_{conv}^{out} + bias$

4 $\boldsymbol{W}^{res} = \alpha \times \boldsymbol{W}^{res} + nonlinear(\boldsymbol{h}_{res\_1})$

5 $\boldsymbol{h}_{res\_2} = \boldsymbol{W}_{res}^{out} \times \boldsymbol{W}^{res} + bias$

6 $\boldsymbol{y}_{res}^{out} = f_{nonlinear}(\boldsymbol{h}_{res\_2})$

7 $\boldsymbol{y}_{classifier}^{out} \leftarrow$ full-connected layer as classifier to $\boldsymbol{y}_{res}$

8 $\hat{\boldsymbol{y}} = softmax\ (\boldsymbol{y}_{classifier}^{out})$

7 $loss = cross\_entropy\ (\hat{\boldsymbol{y}}, \boldsymbol{y})$

8 *Minimize(loss)*

9 **End While**

10 End For

11 Store weights and neural network configuration

12 Calculate Area of Peripheral circuits based on their configuration

13 Calculate total area = memristor memory array area + Σ area of the peripheral circuits

14 Recall Stored weights

15 For number of the weight index = 1, N do

16 Calculate latency of Peripheral circuits with RC as load parameters

17 Total latency = Σ (latency) of peripheral circuits in each operation

18 Total energy = array dynamic/static energy + Σ (dynamic energy) of peripheral circuits in each operation

19 End For

Thirdly, the Python API deploys the saved weights and configurations of the Deep-DFR to the *NeuroSIM* for hardware performance inference. The deployment method evaluates the performance of the neural network system on an offline training environment which demands a local computation. Compared to Online Learning, Offline Learning training keeps the trained neural network at the client-side and perform all prediction computation locally [133], due to the limited energy and space budget at the client-side.

At last, the performance improvements of VT memristor on energy, design area, execution latency, and accuracy are estimated through the co-simulation paradigm. The pseudocode of the hardware-software co-simulation paradigm is introduced in Algorithm 3.1.

3.4.3 Performance Evaluation

Using the co-simulation paradigm introduced in the previous subsection, the performance improvement of VT memristor on deep learning at the system level is evaluated and estimated. The inference accuracy degrades significantly while the resistance variation of the memristor increase [78, 82, 83]. Figure 3-11 presents a correlation analysis between the variation of the weights and the inference accuracy of the Deep-DFR model. The Deep-DFR models are trained with the CIFAR-10 and CIFAR-100 datasets in 150 epochs. The model structure details are depicted in Figure 3-9. The simulation results demonstrate a strong negative correlation between the testing accuracy and the variation of the weights. For example, in Figure 3-11 (a), the testing accuracy significantly reduces while the variation of the weight increases, specifically in the range from 0.2 to 0.6. After the weight variations reach the range larger than 0.6, the testing accuracies tend to be stable and are at low levels (lower than 13%). The testing accuracies with different memristive devices, associating with their variations, are marked in the testing accuracy curve. VT memristive device (Cu/TaOx/Rh/Cr) reaches the highest testing accuracy (~90%) due to its lower variation compared to other devices. The simulation results using the CIFAR-100 dataset (Figure 3-11 (b)) illustrates a similar degradation trend of the testing accuracy. The difference is the testing accuracy on CIFAR-100 reduces faster than CIFAR-10 and reaches its stable range on 0.4 weight variation.

The simulation results with CIFAR-10 and CIFAR-100 both demonstrate the accuracies of the Deep-DFR models constituted of VT memristor (1% variation) outperform the other state-of-

the-art memristors, and other material configurations we explored (listed in Table 3-1). The main advantage of storing weights of the neural networks in memristors is to enhance hardware performance. In this work, VT memristor is compared with SRAM and other state-of-the-art memristor reported, which are implemented with other materials, such as Ag:SiGe [134] and AlOx/HfO_2 [135].

The hardware performance enhancement with different memory techniques in the design area, power consumption, and computing latency are inferred and compared using *NeuroSIM* [81]. The settings of the model are summarized in Table 3-3. The SRAM is implemented in the typical six-transistor cell (6T) with 32 nm technology. The weights are stored in memristors in digital format since the analog memristive synapse degrades the learning accuracy [9]. The weights are stored in 4-bit precision. The feature size of the memristor is assigned at 40 nm because the current industry technology of integrating memristors and the transistors is at the range of 40 nm to 28 nm [9]. The simulation calculates all the latency, design area, and power consumption from different function modules, including the main memory module (SRAM and memristors) and the periphery circuits. The breakdown results of each module are listed in Table 3-4, which uses CIFAR-10 dataset.

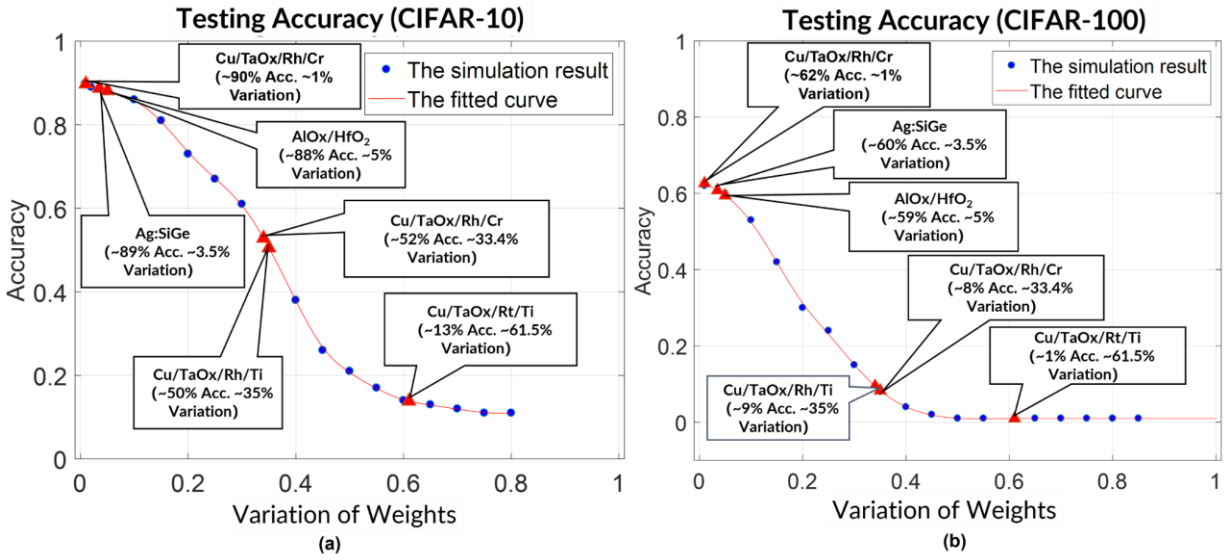


Figure 3-11: The reduction in the accuracy accompanying the increase of the weight variation: (a) CIFAR-10 and (b) CIFAR-100. The neural network model is the Deep-DFR. The blue cycles indicate the simulation results and the red line represents the fitted curve. The memristive device of Ag:SiGe and AlOx/HfO_2 are from [134] and [135] respectively.

Table 3-3: Simulation Setting of NeuroSim Model

Device	SRAM	Memristors
Frequency	1 GHz	1 GHz
Temperature	301 K	301 K
Subarray size	64 × 64	64 × 64
Read Voltage	1.1 V	0.5 V
Read Pulse Width	N/A	10 ns
Structure	6T	1R
Technology	32 nm	40 nm

Table 3-4: Simulation Result Breakdown of Chip Performance

Device	[34]	[35]	SRAM	VT Memristor
Chip Area (mm^2)	98.05	138.83	166.17	85.97
IC Area on chip (mm^2)	2.90	3.50	4.24	2.70
ADC Area on chip (mm^2)	14.03	14.03	42.68	14.03
Periphery circuits (mm^2)	47.50	84.66	52.89	39.05
Chip total Read Latency (us)	423.34	1082.33	803.38	264.97
Chip total Read Dynamic Energy (uJ)	44.1533	55.48	70.88	41.51
Chip total Leakage Energy (nJ)	223.37	699.91	966.03	108.90
Chip total Leakage Power (uW)	791.09	791.09	3074.87	791.09
Chip buffer Read Latency (us)	12.36	12.36	12.36	12.36
Chip buffer read Dynamic Energy (uJ)	4.16	4.16	5.87	4.16

Chip IC Read Latency (us)	36.22	49.40	28.40	32.77
Chip IC Read Dynamic Energy (uJ)	24.39	34.38	25.94	21.92
ADC Read Latency (us)	39.93	39.34	81.66	42.25
Periphery circuits read Latency (us)	214.10	873.62	93.78	53.48
ADC Read Dynamic Energy (uJ)	3.77	3.39	13.12	4.01
Periphery circuits read Dynamic Energy (uJ)	30.60	42.30	35.01	27.71

Figure 3-12 demonstrates that VT memristor reduces chip area, power consumption and latency reduce by ~48%, ~42%, and ~67% with respect to SRAM, respectively. Furthermore, the performance is improved at various degrees compared to other state-of-the-art memristors [134, 135]. The improvements show similar levels with the datasets of CIFAR-10 and CIFAR-100 in Figure 3-12 (a) and Figure 3-12 (b). This phenomenon probably stems from the same neural network model (Deep-DFR) and a similar value range of data (CIFAR-10 and CIFAR-100), which leads to a similar number and values of the weights.

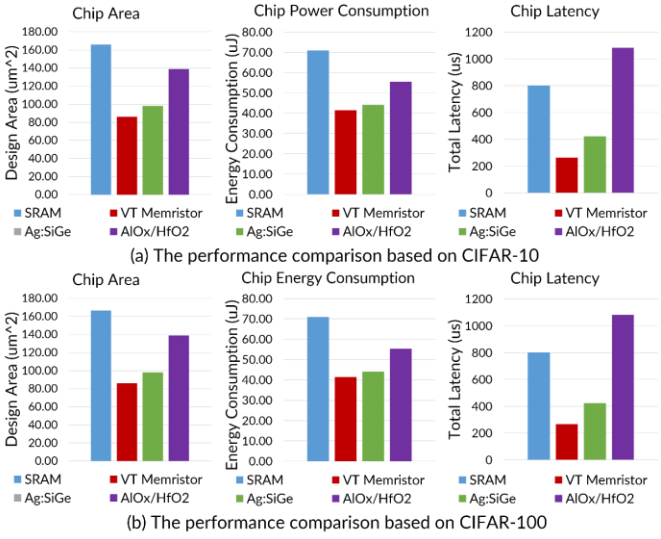


Figure 3-12: Performance evaluation on the different memory techniques: (a) CIFAR-10 and (b) CIFAR-100. The memristive device of Ag:SiGe and AlOx/HfO_2 are from [134] and [135] respectively

The area difference of memristors in Figure 3-12 mainly comes from the periphery circuits. The larger area of periphery circuits of memristors of Ag:SiGe and AlOx/HfO_2 [134, 135] stem from their small on-state resistance [134-136]. The small on-state resistance requires the larger size (W/L) of transistors in peripheral circuits, e.g., Mux or switch matrixes, to avoid the significant current drop and impedance mismatch [136]. Accordingly, the latency of periphery circuits also increases due to the large size of the transistors, which needs a longer time for charging and discharging.

As a non-volatile device, the memristors store the data in their resistances. Unlike SRAM, the non-volatile memory cores do not need a power supply to retain the data in memory cells thus their leakage power is much smaller than a typical SRAM. The energy reduction of other state-of-the-art memristors (Ag:SiGe and AlOx/HfO_2 [134, 135]) is much less than VT memristors because of their smaller on-state resistance (R_{on}). The small on-state resistance leads the array static energy (consumed by cells) dominates rather than the dynamic energy in the system. The static energy consumes more energy in the system, which leads VT memristor is much energy efficient compared to Ag:SiGe and AlOx/HfO_2 [134, 135].

In this work, a novel memristor configuration with the enhanced heat dissipation feature is designed and fabricated. The measurement data demonstrate VT memristor has higher immunity to degradation induced by the thermal effect. The on and off resistance variations are reduced correspondingly, leading to an increase of the testing accuracy within the same range. The Deep-DFR model is used for evaluating VT memristor as the weight storing devices. The datasets CIFAR-10 and CIFAR-100 are used for training the Deep-DFR model. The design area, power consumption, and latency of the system using VT memristor are reduced by ~48%, ~42%, and ~67% compared to conventional SRAM memory technique. At last, these hardware parameters are also improved at various degrees (~13%-73%) compared to other state-of-the-art memristors [134, 135].

3.5 Discussion

In this work, a novel memristor configuration with the enhanced heat dissipation feature is designed and fabricated. The measurement data demonstrate VT memristor has higher immunity

to degradation induced by the thermal effect. The on and off resistance variations are reduced correspondingly, leading to an increase of the testing accuracy within the same range. The Deep-DFR model is used for evaluating VT memristor as the weight storing devices. The datasets CIFAR-10 and CIFAR-100 are used for training the Deep-DFR model. The design area, power consumption, and latency of the system using VT memristor are reduced by ~48%, ~42%, and ~67% compared to conventional SRAM memory technique. At last, these hardware parameters are also improved at various degrees (~13%-73%) compared to other state-of-the-art memristors [134, 135].

Chapter 4. Three-dimensional Neuromorphic Computing System with Two-layer Memristive Synapses

Three-dimensional Integrated Circuits (3D-ICs) is a cutting-edge design methodology of placing the circuitry vertically to providing high speed, low power consumption, and small design area. In this chapter, a novel 3D neuromorphic system constituted with monolithic 3D integration and VT two-layer memristive synapse is proposed and analyzed. The simulation results demonstrate VT fabricated two-layer memristors outperform the one-layer configuration on design area, power consumption, and latency with the factors of 2, 1.48, and 2.58, respectively. The proposed 3D low-variation memristive synapse shows the significant improvement (10% to 66%) on design area, power consumption, and latency, compared with the SRAM (Static Random-access Memory) and other state-of-the-art memristive synapses. The performance of the neuromorphic system with our memristors is evaluated through the benchmark datasets (MINST and CIFAR-10) and the system-level simulator NeuroSIM.

4.1 Introduction

Human Brains demonstrates a remarkable energy-efficiency on numerous cognition tasks [3]. The low power consumption of the neural system in human brains stems from the unique threshold neurons, discretely spiking signal representation, and neural network configuration. Neuromorphic Computing System (NCS) is an approach of achieving a power-efficient artificial intelligence system through mimicking human brains with low-precision spiking communication of threshold neurons (activation functions) [137]. The significantly high energy budget makes NCS more suitable to the resource-constraint applications, such as self-driving vehicles, unmanned aerial vehicles, and smartphones [138, 139]. The outputs of the threshold functions are either zero or one, which requires less hardware utilization and communication energy since the zero or one output can be represented as one single pulse (Spike) [137]. Several neuromorphic chips have been demonstrated the capability of NCS, such as the Loihi chip developed by Intel and TrueNorth built

by IBM [20, 140]. However, the non-differentiability of the threshold activation functions (neurons) of NCS deprives of the learning capability that relies on the gradient descent algorithm. Nowadays, a so-called *Whetstone* method overcoming this issue has been proposed [137]. In this particular method, different activation functions are used at the training and inference stages [137]. Initially, the neural networks are trained through the conventional backpropagation algorithms with differentiable activation functions, such as ReLU, sigmoid, etc. Nevertheless, these activation function will gradually transform into a threshold function during the training process. This activation transformation process is referred to as a sharpening procedure [137]. Through the sharpening procedure, the NCS can be trained through the conventional backpropagation algorithms. In this chapter, a hardware design with VT low-variation and two-layer memristors are discussed and analyzed for the *Whetstone*-based neuromorphic computing system.

Typically, the memristors are fabricated in a one-layer crossbar structure. However, this one-layer structure suffers the larger signal propagation delay, chip area, and power consumption [11, 58]. In this chapter, the aforementioned VT low-variation memristors (see Figure 3-5) are scaled vertically into a three-dimensional structure (two layers) for further increase the device density. As a result, VT two-layer memristors structure demonstrates a reduction on the design area (2×), power consumption (1.48×), and latency (2.58×), compared to the traditional one-layer structure. Furthermore, the VT 3D low-variation memristive synapse outperforming the SRAM-based synapses and other state-of-the-art memristive synapses [134, 141] on design area, power consumption, and latency at various degrees (10% to 66%).

4.2 Monolithic Three-dimensional Integration with Memristors

Modern computers upon the von Neumann architecture was designed for Boolean algebra and arithmetic calculation. To perform these calculations efficiently, the central processing units (CPUs) in computers extract data from memory. The transmission of data between the CPUs and the memory relies on the bus and the high-frequency signals carried on it. However, as the rising of the data-driven artificial intelligence (deep learning), a quantity of data is dramatically increasing these years. Thereby, the unprecedented demand for computational resources becomes the most critical challenge.

The latest researches reveal a strong correlation between the inference accuracy of deep learning and data quantity. Thus, the more data is used for learning, the higher inference accuracy can be expected [7, 8]. The strong correlation between the data quantity and learning accuracy inevitably leads to an endless pursuit of large datasets. The massive amounts of data also accompany the successive data transferring between CPU and memory that will significantly increase the power consumption. The computational and power demands for the data-driven learning methodologies, such as deep learning, cannot be resolved by merely improving current transistors and memory technologies alone [93]. Hence, there is an urgent need for a novel integrated circuit (IC) structure to overcome these fundamental limitations of the scaling issue of modern computers. The high memory device density and more compacted IC structure can be achieved through the utilization of the so-called monolithic 3D integration technology and a nanoscale memory device memristor [74]. Unlike traditional 3D integration technology using Through silicon Vias (TSVs) and interposers at micrometer scales, monolithic 3D integration can connect the modules at different layers through nanoscale monolithic inter-tier vias (MIVs) [142-153] that further reduce the design area. The size among a TSV, an inter-tier via, and a memristor is illustrated in Figure 4-1.

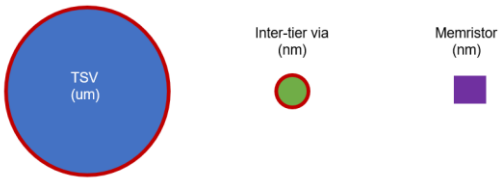


Figure 4-1: Size comparison among a TSV, an inter-tier via, and a memristor

Moreover, the fabrication processes of the monolithic 3D integration technology are compatible with 3D memristor fabrication processes. A typical transistor level monolithic 3D integration structure is illustrated in Figure 4-2.

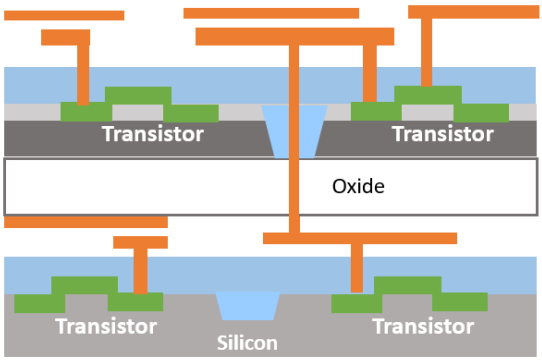


Figure 4-2: Monolithic 3D integration at transistor level

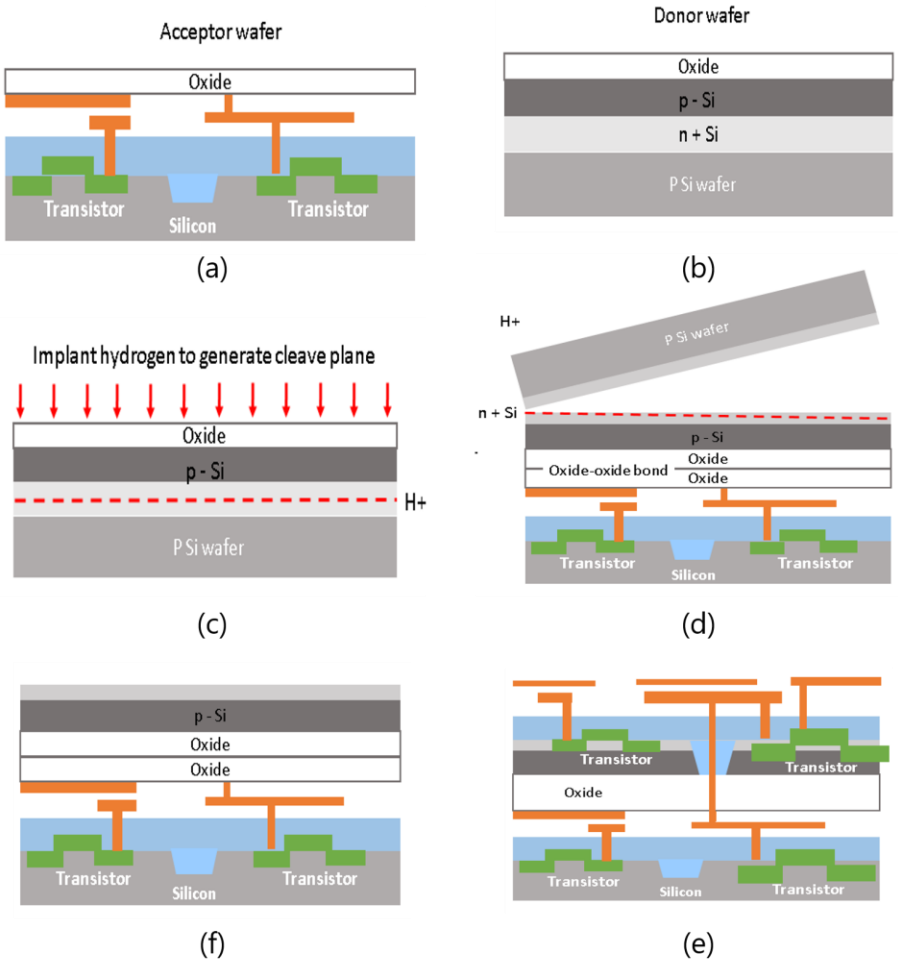


Figure 4-3: (a) Acceptor wafer fabrication; (b)Donor wafer fabrication; (c)Implant hydrogen to generate cleave plane; (d) Combining the donor wafer to the acceptor wafer and perform Ion-Cut cleave; (e) Remove the donor wafer and complete the Ion-Cut; (f) Fabricate another layer of transistors at low temperature

Unlike conventional TSV-based 3D-ICs technology, which fabricates discrete layers separately, the monolithic 3D integration technology fabricates the layers sequentially at low temperature, while concurrently interconnecting the layers with nanoscale monolithic inter-tier vias (MIVs). Figure 4-3 illustrates the details of Pulsed-laser and ion-cut fabrication procedures [154, 155].

For such fabrication process, circuits in an acceptor wafer, which include both transistors and interconnections, are first built conventionally. Meanwhile, the donor wafer containing both p – Si and n + Si layers is constructed on a new wafer; the dopants are activated applying the normal high temperature techniques. Then the hydrogen is implanted into the donor wafer with the p – Si and the n + Si regions, and a cutting plane is created. After the two wafers are formed, the donor wafer is inverted and bonded to the acceptor wafer of 3D-ICs. Then the donor wafer is cleaved along a plane created using either a low temperature anneal, or an applied sideways mechanical force. Then a recessed channel transistor (RCAT) architecture is utilized to create transistors in this thin layer under low temperature; without damaging the connections in the acceptor wafer.

The monolithic technology is a three-dimensional integration technology that can combine the memristors with traditional silicon-based transistors to further reduce the power consumption and design area. Through monolithic technology and low-temperature fabrication method, the memristors can be stacked on a silicon wafer as shown in Figure 4-4.

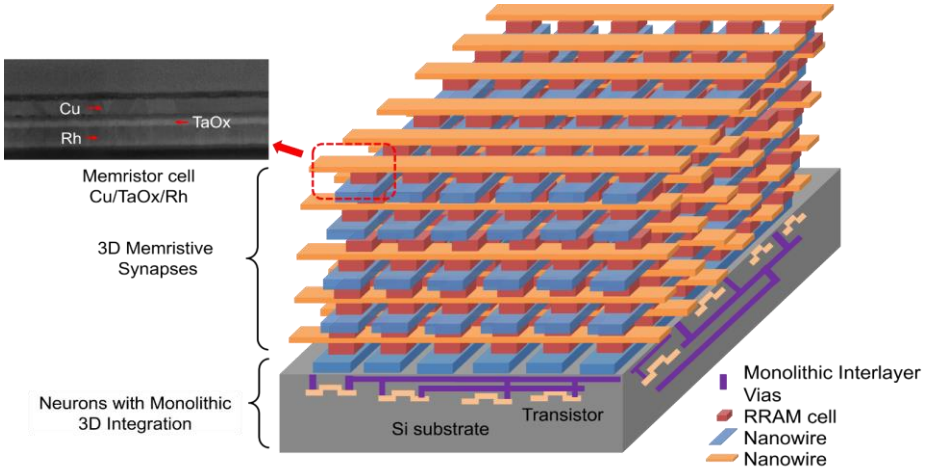


Figure 4-4: Monolithic 3D neuromorphic architecture constituted with memristive synapse and neurons. The neurons and memristive synapses are integrated through monolithic 3D integration technology.

The low temperature of fabrication protects the prior fabricated transistors at the bottom layer as shown in Figure 4-4. The fabrication temperature of the memristor is generally can be as low as 300°C [93]. The access of memristors which are at different layers is realized through nanoscale monolithic interlayer vias. With the Monolithic 3D technology, the vertical routing paths, the critical path lengths diminish by a factor of three, the power consumption can be decreased by 50%, and the design areas reduces by 35% [156].

4.3 Two-layer Memristor Fabrication and Evaluation

This section introduces the fabrication procedure of our three-dimensional (two layers) memristor. Our three-dimensional memristors are fabricated on a silicon wafer with a two-layer crossbar structure. Additionally. the cycle-to-cycle switching variation of our two-layer memristors reduces by more than 30% through heat dissipation layers. Furthermore, the corresponding three-dimensional model of the two-layer structure is built. The simulation indicates our 3D (two-layer) memristor structure reduces the design area, power consumption, and latency by 2×, 1.48×, and 2.58×, respectively. Figure 4-5 (a) shows that our memristors are fabricated in a crossbar configuration on a thermally oxidized silicon wafer (730 nm thick). The magnified figure of crossbars on the silicon wafer is shown in Figure 4-5 (a). Each crossbar structure contains 25 devices at a 5 by 5 array as illustrated in Figure 4-5 (c).

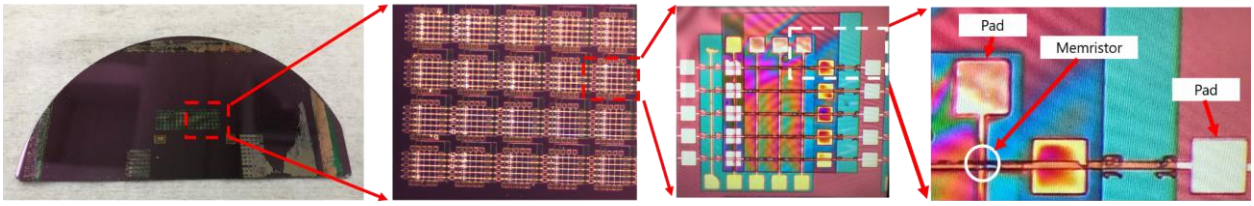


Figure 4-5: VT two-layer memristors: (a) two-layer memristor arrays at the wafer; (b) the memristor arrays at the wafer; (c) a typical two-layer memristor cell; (d) zoom-in view of a memristor at the cross-point of the array.

Each memristor cell is located at the cross-point of two accessing nanowires crossbar as illustrated in Figure 4-5 (d). At the ends of the nanowires, the pads are fabricated for placing a testing probe. Both metal electrodes and oxide (solid electrolyte) are deposited by electron beam evaporation and patterned by lift-off technology. A thin layer (20 nm) of chromium (Cr) is

deployed between Rh and SiO_2 to improve the adhesion of the rhodium (Rh) layer and enhance the capability of the heat dissipation. The processes are performed twice forming two layers of memristors. Moreover, the additional heat dissipation layers are added at each layer to reduce the cycle-to-cycle switching variations. The cross-section view of our two-layer memristors with heat dissipation layers is illustrated in Figure 4-6 through Focused Ion Beam (FIB) technology. The material Cr is used for the heat dissipation layer in our work.

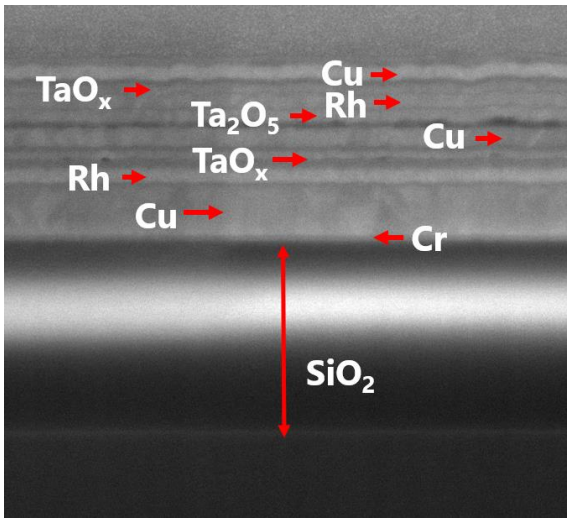


Figure 4-6: Focused Ion Beam (FIB) cross-section image of the two-layer memristor

The metals Cr and Rh have been selected as electronics due to their superior thermal conductivity properties among the traditional Ti, Pt, W, or TN metals. The oxygen-deficient TaO_x ($x \approx 1.9$) is deposited in the PVD-250 chamber through evaporating Ta_2O_5 pellets without O_2 injection into the evaporation chamber. The layer of TaO_x deposited on our device is 25 nm. Its thickness (TaO_x) is controlled during the deposition process through a calibrated quartz crystal. Our Kurt J. Lesker PVD-250 e-beam evaporation system possesses a built-in quartz crystal microbalance oscillator, which has the capability of precisely measuring deposited thin film thickness by the material and chamber geometry parameters, such as density, Z-ratio, and tooling factor. The width of nanoscale metal lines is 5μm resulting in rectangular device areas of the device at 25 μm^2. All metal layers (Cu, Rh, Cr) are deposited by Physical Vapor Deposition (PVD) in a Kurt Lesker e-beam PVD-250 chamber.

Before fabricating the second (top) layer of memristor, a layer (70nm) of TaO_x is deposited by evaporating Ta_2O_5 pellets into the e-beam evaporation system with oxygen injection into the

chamber. The purpose of oxygen injection is to improve the stoichiometry of the TaO_x layer. This stoichiometric Ta_2O_x (x = 2.5) layer is less defective and provides electrical isolation between two memristor layers. This step is followed by the deposition of the Cr adhesion layer. The Rh bottom electrode layer is fabricated through electron beam evaporation and patterned with lift-off technology. Then the oxygen-deficient TaO_x (x≈1.9) layer and Cu top electrode layer are deposited and patterned with lift-off technology in a similar fabrication process of the first layer. The detailed geometry of the layers is illustrated in Figure 4-7.

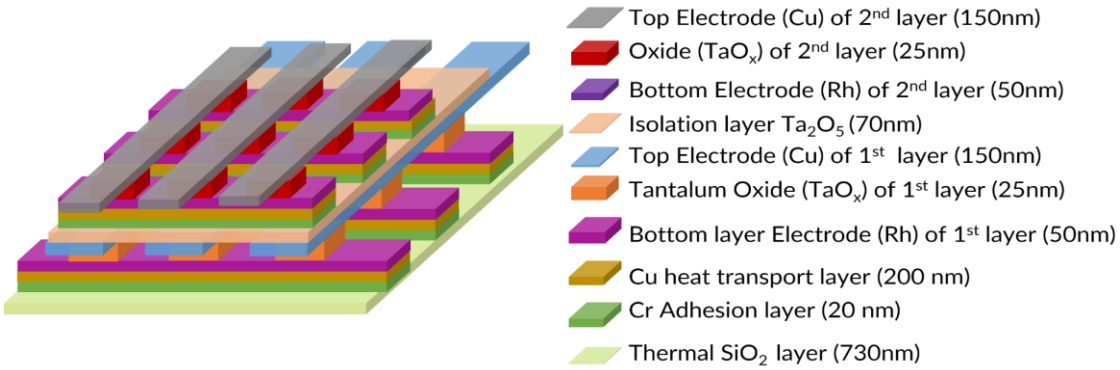


Figure 4-7: Structure of the two-layer memristor crossbar

During the switching of memristors, the active metal ion electromigration, which completes the conductive filaments (CFs) at the tip of the growing CFs, significantly influences the size of the filament and the value of the on-resistance R_{on} [82]. Because of the stochastic nature of ion migration and atom diffusion, the variation of the on-resistance R_{on} is rather high, especially in high temperature [82]. The temperature will increase gradually during the switching since the movement of oxygen atoms and ions in the metal oxide accumulates the heat inside of the memristor. The high-temperature interior device further enhances the metal diffusion effect. The latest research indicates the heat diffusion effect of metal will increase the resistive variation [83]. Thus, a reasonable attempt of mitigating the temperature-related resistance variation of memristors is to dissipate the accumulated heat inside of memristors. The high temperature can be reduced naturally by using an additional heat dissipation layer. Thereby, another heat dissipation layer with high heat conductivities is added to enhance the heat dissipation capability of memristors. Due to the high conductivity of the additional layer, the accumulated energy will be removed from the cells of memristors. In the measurement, the compliance current is set in 50 uA denoted as I_{cc}. Figure 4-9 illustrates the resistance variation of the memristors fabricated with different materials.

A positive correlation between the cycle-to-cycle resistance variation (R_{on}) and the thermal conductivity of the heat dissipation layer can be observed clearly in Figure 4-9. Note that the value of conductivity is the sum of the heat dissipation layer, such as Rh and Cr layers. Figure 4-9 illustrates the combination of Titanium (Ti) and Chromium (Cr) as the dissipation layer demonstrates the most effective mitigation of resistance variation, which is as small as ~1%.

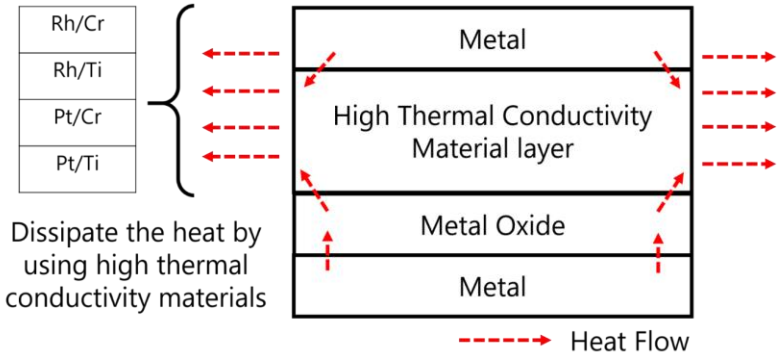


Figure 4-8: Mechanism of heat dissipation layer.

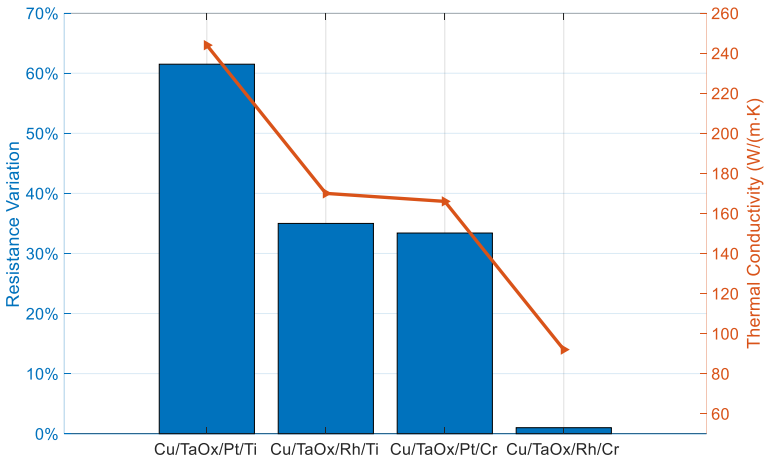


Figure 4-9: Resistance variation of the switching process of the memristors fabricated with different materials and the correlation between the memristor variation and thermal conductivity of heat dissipation layers

4.4 Three-dimensional Memristive Neuromorphic Computing System

In this section, the three-dimensional memristive neuromorphic computing system is designed using the circuit model of our fabricated memristor array and the model of the three-dimensional

structure of the memristors. The memristor model is built based on the filament growing method [115], capturing the main characteristics of our two-layer low-variation memristors, such as set voltage, on-state resistance (R_{on}), and off-state resistance (R_{off}). The three-dimensional structure of the memristors is modeled with the corresponding resistances and capacitances of the two-layer crossbar structure.

Based on the equations from Eq. (3-1) to (3-7), we built a Verilog-A model for our two-layer memristors. The model parameters are summarized in Table 4-1. Figure 4-10 illustrates the V-I characteristic curve of our model and the measurement data. The 1 μA of compliance current is applied for the protection purpose since the large current flowing through the devices will permanently damage the memristors. Furthermore, in the measurement, the characteristics of the memristors from the top and bottom layers have no significant differences. This probably due to the identical materials, size, and fabrication processes of those two memristor layers.

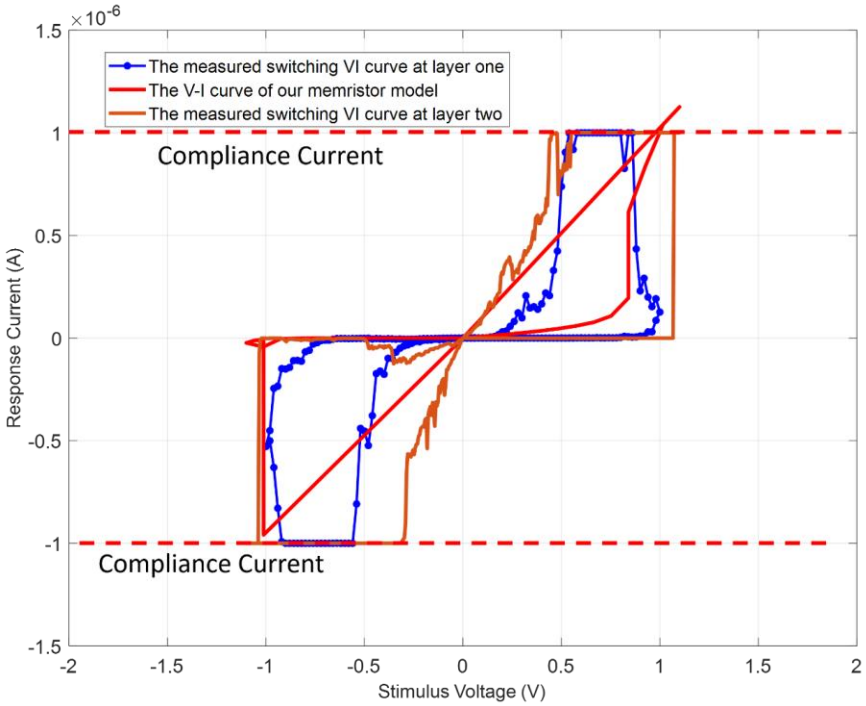


Figure 4-10: V-I curve of our two-layer memristor (Cu/TaOx/Rh/Cr): The blue line and brown lines represent two typical measurement data of the memristor at the top and bottom layers, respectively. The red line depicts our model capturing the critical switching characteristics.

Table 4-1: Parameters of the Memristor Model

Parameter	Descriptions	values
I_0	Hopping current density in the gap region	1e13
ρ	Resistivity of the CF	1.9635e-5
a	Distance between adjacent oxygen vacancy	0.2e-9
f	Vibration frequency of oxygen atom	1e13
x_T	Characteristic length in hopping region	0.4e-9
V_T	Characteristic voltage in hopping	0.4
w_0	Initial CF width	5e-9
E_a	Average active energy	0.75
α_a	Enhancement factor	0.75e-9
Z & e	Charge number & unit charge	1 & e
k_B	Thermal resistance	8.61733e-5

To evaluate the two-layer structure of our memristors, a full-wave model is built upon the resistances and capacitances of the 3D crossbar structure. A SPICE model of our two-layer memristor crossbar structure is developed that incorporates the memristor Verilog-A model and the resistance and capacitance list in Table 4-2. The improvement of 3D structure comes from the much less resistance and capacitance of the two-layer structure than the one-layer structure. Figure 4-11 demonstrates the geometry details of our models. The models are built using industry-standard simulator ANSYS Q3D extractor [157].

These two models have the same number of memristors, but the different structures. For the one-layer memristor (Figure 4-11 (a-c)), the memristor is modeled in 5 by 10 array. For the two-layer memristors (Figure 4-11 (d-f)), the same quantity of memristors is placed in two layers. Thus, the 25 memristors are located in a 5 by 5 array at every single layer. The resistances of the connecting nanowires (top and bottom electrodes) and the parasitic capacitance between them are extracted from ANSYS Q3D extractor. Table 4-2 summarizes the resistance and capacitance of one-layer and two-layer memristors. As a non-volatile device, memristors retain their resistances

(HRS and LRS) with no extra exterior voltage, as a result, saving the static power consumption. The main power consumption comes from the resistance of the accessing nanowires of the crossbar structure. Three critical parameters are used for evaluating the hardware performance of our two-layer design, which are power consumption, design area, and latency. The design area can be directed extracted from the device geometry. The power consumption caused by the resistance of nanowires is calculated with the equation $P = I^2R$, where R is the resistance and I is the current. The latency is calculated by the signal latency equation [158]:

$$t_{latency} = 0.7\ RC, \tag{4-1}$$

where the R and C are the resistance and capacitance of the nanowires listed in Table 4-2. Figure 4-12 illustrates the comparison between one-layer and two-layer memristors on these three critical parameters: design area, power consumption, and the latency. Figure 4-12 demonstrates that these three parameters have been reduced by a factor of 2, 1.48, and 2.58, respectively, through stacking the memristors from one layer into two layers.

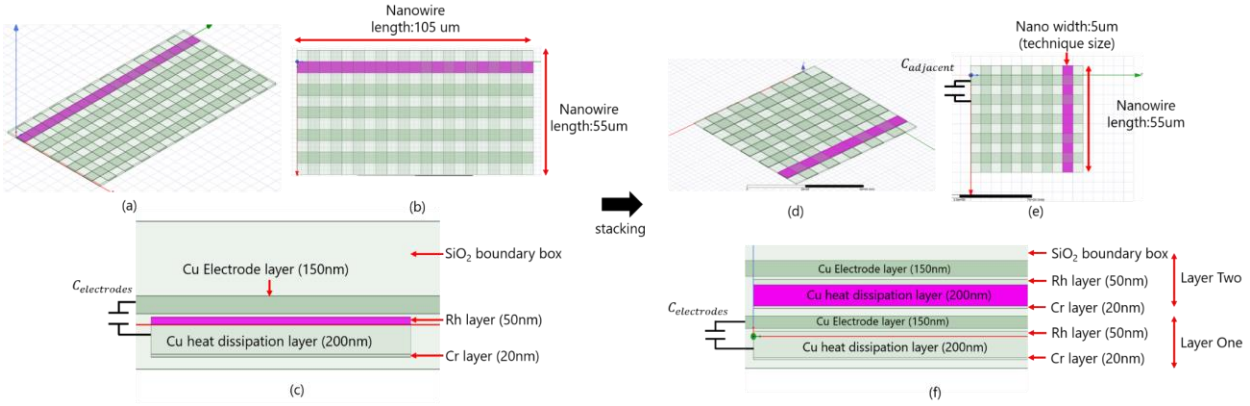


Figure 4-11: Full-wave model of memristor structure: (a) One-layer memristor (5×10×1); (b) Top-view of one-layer memristor; (c) Side-view of the of one-layer memristor; (d) Two-layer memristor (5×5×2); (e) Top-view of two-layer memristor; (f) Side-view of two-layer memristor.

Table 4-2 Parameters of our two-layer memristor structure

Characteristics	Descriptions	Two Layer Memristor Characteristics Values	One Layer Memristor Characteristics Values
$R_{Top\ Electrode}$	The resistance of the top electrode (150nm copper layer)	1.582 Ω	3.0172 Ω
$R_{Bottom\ Electrode}$	The resistance of the bottom electrode (50 nm Rh layer, 200 nm Copper heat dissipation layer, and 20 nm Cr layer)	0.855 Ω	0.855 Ω
$C_{Electrode}$	The capacitance between the top and bottom metal electrodes of a single memristor	0.0337 pF	0.035 pF
$C_{adjacent}$	The capacitance between the adjacent nanowires (Rh layer to Rh layer)	0.00016 pF	0.00025 pF
$C_{adjacent\_top}$	The capacitance between the adjacent nanowires (Copper to copper)	7.9e-5 pF	0.0007 pF

At last, our two-layer memristor is incorporated into a Whetstone-based neuromorphic system for a system-level evaluation [137].

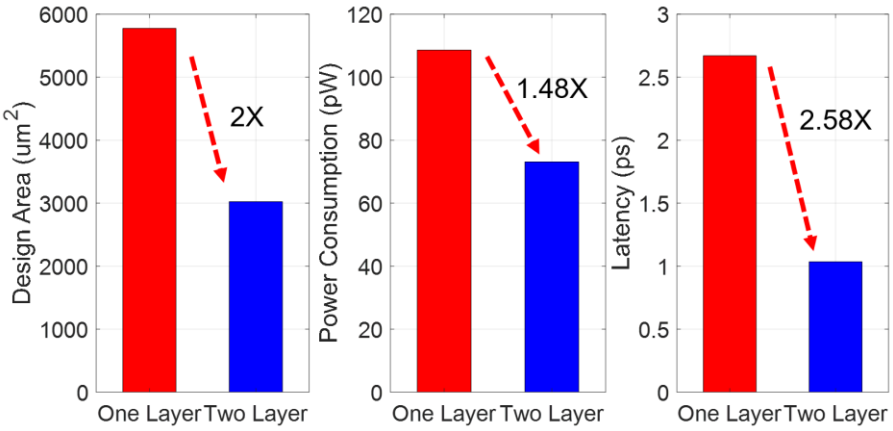


Figure 4-12 Comparison between one-layer and two-layer memristors on the critical performance parameters of the design area, power consumption, and latency.

In this paper, our two-layer is incorporated into a system-level simulator NeuroSIM [81] for hardware performance evaluation. Three critical hardware parameters, design area, power consumption, and latency, are compared with the SRAM-based synapse and other state-of-the-art

memristive synapses, such as Ag:SiGe [134] and AlOx/HfO_2 [141]. Figure 4-13 illustrates the methodology of evolution and the diagram of NeuroSIM. First, conventional neural networks are trained through *Whetstone* [137], which is an emerging technology that transforms a traditional neural network to a spiking neural network with threshold activations. in this work, we trained the MLP and CNN with two benchmark datasets, MNIST and CIFAR-10. These neural networks (MLP and CNN) initially utilize the traditional bounded ReLU activation.

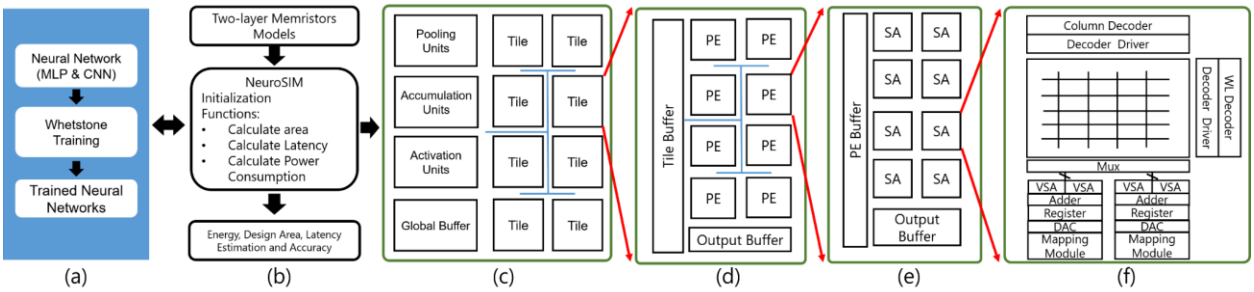


Figure 4-13: Diagram of our hardware-software co-simulation paradigm with *NeuroSIM* and *Whetstone*. (a) The neural networks are trained through *Whetstone*. (b) The weights are evaluated through *NeuroSIM*. (c) The architecture of the *NeuroSIM*. (d) The architecture of each Tile. (e) The structure of each Processing Element (PE). (f) The architecture details of each Synaptic Array (SA).

During the training, the bounded ReLU activation functions are gradually transformed into the threshold functions. Since the transition of activation function gradually occurs during the training, the derivatives of the activation function vanish rather suddenly than gradually.

Algorithm 4.1: Performance Estimation

Initialize: The configuration of the artificial neural network

Initialize: Memristive synapse configuration

Initialize: Peripheral circuits configuration

1 For epoch = 1, M do

2 **While** batch in dataset do

3 **For number of the layers = 1, N do**

5 sharpening the activation function (BReLU) through $\alpha \rightarrow 0.5$, $h \rightarrow 1$

6 **End For**

7 End While

8 End For

9 Store weights and neural network configuration

10 Calculate Area of Peripheral circuits based on their configuration

11 Calculate total area = memristor memory array area + Σ area of the peripheral circuits

12 Recall Stored weights

13 For number of the weight index = 1, N do

14 Calculate latency of Peripheral circuits with RC as load parameters

15 Total latency = Σ (latency) of peripheral circuits in each operation

16 Total energy = array dynamic/static energy + Σ (dynamic energy) of peripheral circuits in each operation

17 End For

Eq. (4-2) illustrates the bounded reified linear unit (bRELU) [137] with the assert of $|\beta - 0.5| = |\alpha - 0.5|$.

$$h_{\alpha,\beta} = \begin{cases} 1, & \text{if } x_i \geq \beta \\ \dfrac{x_i - \alpha}{\beta - \alpha}, & \text{if } \alpha \leq x_i \leq \beta. \\ 0 & \text{if } x_i \leq \alpha \end{cases} \quad (4\text{-}2)$$

As shown in Eq. (4-2), with $\alpha = 0$ and $\beta = 1$, $h_{\alpha,\beta}$ is a generic bound ReLU function. As $\alpha \to 0.5$, $h \to 1$, the function $h_{\alpha,\beta}$ is gradually sharpened into a threshold function. Thus, the final trained neural networks maintain a satisfying accuracy (~1% accuracy loss for MNIST and ~5% accuracy loss for CIFAR-10). During the training, the weights and neural network configurations are stored and extracted. Then, the stored weights are further imported into a macro-architecture simulator NeuroSIM. The pseudocode of this evaluation methodology is demonstrated in Algorithm 4.1.

Figure 4-13 illustrates the NeuroSIM framework showing its critical modules hierarchically, which contains multiple tiles, global buffer, accumulation units, and activation units. Each tile has several processing elements (PEs). The communication among Tiles is under a tree network

structure as illustrated in Figure 4-13 (c). The responsibility of the tile buffer is to load activations, accumulation modules for adding sums from PEs and transferring the result to the output buffers. Each PE module constructs with a group of synaptic arrays (SAs). The synapses in SA can be assigned with multiple memory devices, such as SRAM, memristors. The memristor crossbar structure is incorporated in the SA as shown in Figure 4-13 (e). Each memory cell of the memory array connects to a *wordline* and a *bitline*. For memristor, the values are encoded in the resistance of a memristor and the nanowires serve as the bitline and wordline for accessing the memristive memory cells. The resistance value of the selected memristor equals the applied voltage divided by the measured current at the end of the nanowire. The decoder of the system uses the *wordline* and *bitline* to access to every single memory cell. The operations of weight sum and update in NeuroSIM are row-by-row-based write and reading. The row selection is activated through the WL decoder. Then the BLs are precharged to each cell access. The memory data are captured by the sense amplifier (S/A). After that, the adder and register are used to sum the weight values in a row-by-row style.

In this work, the performance among our two-layer memristor-based neuromorphic system with SRAM and other state-of-the-art memristor [134, 141] are compared. The MLP includes three full-connected layers with the training accuracy at 98%; while the CNN contains the six convolutional layers with the training accuracy at 83%. The neural networks utilize the dataset benchmarks of MNIST and CIFAR-10, respectively. Figure 4-14 summaries performance comparisons. VT two-layer memristor shows a significant reduction in the design area (50%), power consumption (60%), and latency (66%). As shown in Figure 4-14, the design area, power consumption, and latency of the threshold-based neuromorphic system are reduced by 10%, 36%, and 23% compared with the system using Ag:SiGe [8], respectively; while the design area, power consumption, and latency are reduced by 10%, 58%, and 78.1% respect to the system with the electronic memristor AlOx/HfO_2 [9].

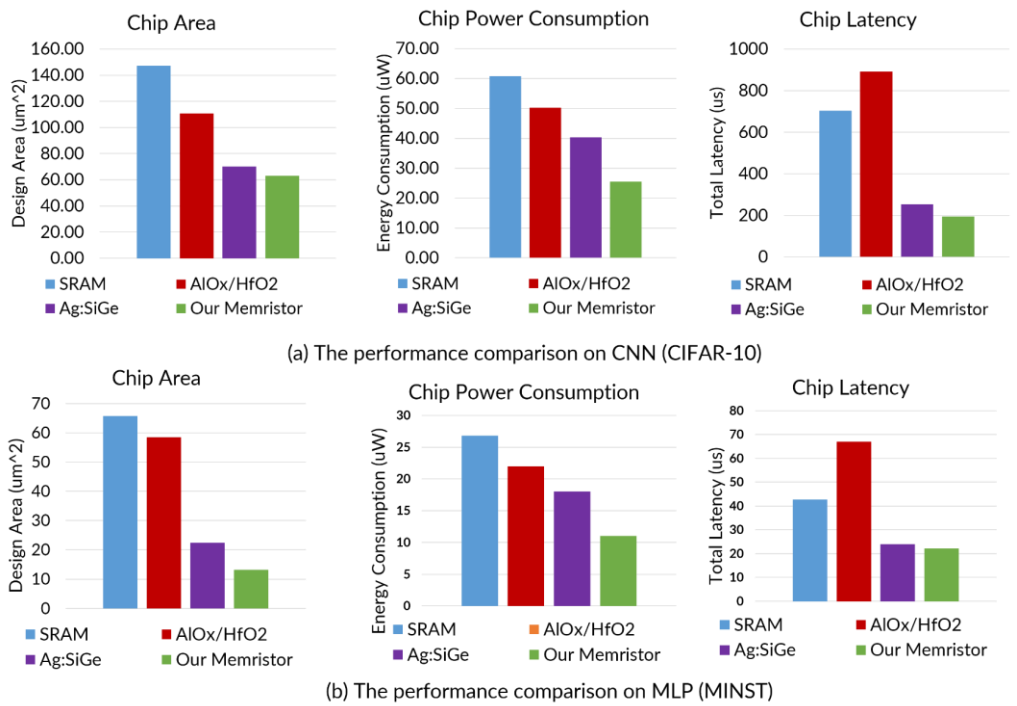


Figure 4-14: Performance evaluation of the different techniques

4.5 Discussion

In this work, VT two-layer memristive electronic synapses, constituting a Three-dimensional (3D) neuromorphic system. The 3D structure shows a significant improvement in the design area, power consumption, and latency with the factors of 2, 1.48, and 2.58, respectively. The improvement stems from the much less resistance and capacitance of the two-layer structure. At last, the performance of the neuromorphic system with the 3D memristors is evaluated through the benchmark datasets (MINST and CIFAR-10). These simulation results demonstrate an improvement in the design area, power consumption, and latency at various degrees compared to other state-of-the-art memristors (10% to 66%) [134, 141].

The integration of memristors and three-dimensional technology propels a neuromorphic system to a new stage of high performance on design area, power consumption, and latency. The improvements inherently come from the shorter signal propagation distance. Furthermore, the vertical connections among different layers are implemented by the nanoscale MIVs, which further reduce the vias from microscale (TSVs) to nanoscale (MIVs). Many approaches have been conducted to building a three-dimensional neuromorphic system with memristors [93, 159-164].

In [159], a three-dimensional system with eight layers of memristors is presented for CNNs specifically. The memristors in [159] are fabricated using materials of Pt/HfO2/Ta in an eight-layer three-dimensional structure. Similar to our multiple layer crossbar structure, the memristors in [159] are fabricated in a horizontal 3D structure with the feature sizes of 300 nm and 4 μm. The high state resistance and low state resistance are at 10KΩ and 1KΩ, respectively. These values are relatively low and the resistance ratio is small. Based on the discussion and analysis in Chapter 3, lower resistance may require a larger size of the transistors at peripheral circuitry, leading to a large design area and latency. Moreover, the R_{OFF}/R_{ON} ratio (the ratio of off-state resistance to on-state resistance) is relatively small, which potentially triggers the undesired switching due to the sneaking path. On the contrary, the R_{OFF}/R_{ON} ratio is 1000 that efficiently avoids the impact of sneaking path. The detailed comparison between VT 3D memristors and 3D memristors in [159] are listed in Table 4-3. Due to the scale limitation of fabricated memristors, the authors in [159] utilize a software (MATLAB) and hardware co-simulation method for evaluation that is similar to my approach. The system is evaluated with MNIST dataset reaching a satisfying inference accuracy (97.91%).

The work in [159] purposely designs a multiple-layer memristive structure to enhance the kernel calculation of CNNs that compromises the generality of the system. On the contrary, the typical VT two-layer memristors have a much simple structure, which is more friendly to the fabrication process. Moreover, the switching variation of memristors is still observed in the measurement data of their work. Based on the analysis in Chapter 3 (see Figure 3-11), the high variation significantly decreases the inference accuracy. At last, the neural network of our approach is spiking neural networks with threshold function, which is a more biology-plausible design compared to the other nonlinear activation functions used in [159].

Table 4-3: Comparison between VT 3D memristors and the 3D memristors of University of Massachusetts at Amherst

Specifications	VT 3D Memristors	[159]
Number of Layers	2	8

3D structure	Horizontal Crossbar Structure	Horizontal Crossbar Structure
Neural Network Model	Spiking Neural Networks	Convolutional Neural Networks
Materials	Cu/TaOx/Rh/Cr	Pt/HfO2/Ta
Set Voltage	0.7 V – 1.2 V	0.5 V - 1 V
Feature Size	5 µm to 35 µm	300 nm to 4 µm
Low Resistance Value	1 MΩ	1 KΩ
High Resistance Value	1 GΩ	10 KΩ
R_{OFF}/R_{ON} Ratio	1000	10

Chapter 5. Neuromorphic System with Associative Memory Learning

5.1 Introduction

As the core mission of artifical intelligence, the self-learning capability of brains attracts the scientists for a long time [19]. Self-learning capability of organisms stems from associative memory learning [3]. This particular real-time learning method enables the organisms to correlate two concurrent events together [3, 165-168]. Thus, dogs can learn the sound of whistle as a sign of food [3, 165]. Similarly, the children remember the words representing the objects while the parents repeatedly teach them. Thereby the investigation on the associative memory not only potentially disclose a path of designing a self-learning artificial intelligence system but also provide a method of comprehending the learning mechanism of organisms.

The nervous system is constructed with basic neurons connecting through synapse in a network configuration. The neural science scientists indicate that the modification of synaptic connecting strengths, also referred to synaptic weights, play a critical role in the associative memory learning [3]. The weight of a synapse represents the connection strength between two neurons. The connecting strength is realized by the amount of the chemical neurotransmitters as aforementioned in Chapter 2 (see Figure 2-2). This synaptic weight between neurons can be implemented by an emerging device memristor. The memristors are the nanoscale two terminal devices with an adjustable resistance under the exterior stimulus, which is similar with the function of synapses [78]. Several pioneers have applied memristor as electronic synapses on small scale associative memory learning recently [99, 167, 169-175]. But, these works only correlate several simple signals together with a few of neurons [99, 167, 170, 172-175]. Additionally, the information carried by simple two channels of signals is restricted [3, 176]. However, one of the most important steps to implement a self-learning artificial intelligence system is to make the system have a capability of correlating several pieces of sophisticated information together [176].

In the neural system of human brain, the different kinds of signals, such as auditory and visual signals, are processed at different regions [3] with distinct neural networks (see Figure 2-4).

Through this way, the information of auditory and visual signals captured by the eyes and ears are extracted by these neural networks using the spiking signals as the outputs. Similarly, artificial neural networks (ANNs) processes different types of signals independently [8, 177-179], and abstracts the input information into the probabilistic grades. For instance, the convolutional neural networks (CNNs) are used for processing image signals [178], while recurrent neural networks (RNNs) are more suitable for processing series signals [179]. A set of scores as the outputs of these neural networks that represent the probabilities of the input belonging to a particular category. Inspired by ANNs and distributed signal processing of brains, a behavior-level associative neuromorphic system is proposed and introduced in this chapter. Instead of relating simple signals together, the proposed system associates the outputs of the multiple ANNs together by correlating these outputs together. The correlation is realized by first transforming the output grades into corresponding spiking signals and then adding them together. The adding spiking signals are used for modification of synaptic weight. The association of spiking signals are implemented with one more layer of the neural network, referred to the associative memory network (AMN).

Specifically, the proposed system transforms the outputs (probabilistic scores) of the ANNs into the corresponding spiking signals with different frequencies and magnitudes. This function is implemented by a new neuron named Signal Intensity Encoding Neurons (SIENs). Then the spiking signals are imported into the AMN for an analog-based association. Through this way, the information preprocessed by the ANNs is associated together. The detailed contributions can be summarized as:

1) Proposed a large-scale associative memory learning system for correlating several ANNs together;
2) Associated the auditory signal and visual signal together;
3) Designed the corresponding circuitry modules: signal intensity encoding neurons that transforms the input into the magnitude and frequency of a spiking signal, a model of the vertical memristive synapse array;
4) Implemented a 3D large-scale associative memory learning system including 20 neurons and 100 memristive synapses.

5.2 Associative Memory in Biology

Associative memory is a ubiquitous learning mechanism in nature world that first investigated by Ivan Pavlov through a series of experiments on dogs [3] as illustrated in Figure 5-1. Initially, dogs have no salivation response to the sound of bells, while they salivate if the food is presented in front of them. Thereby, this experiment indicates the sound of bells does not evoke the salivation.

Next, Pavlov sounded the bell and presented the food to the dog simultaneously [3] and repeat this behavior several times. Then he observed that the dog started to salivate when the bell sounded around him even no sight of food. By studying this phenomenon, Pavlov indicated that salivation, normally from a visual sight of food, can also be stimulated from a normally different signal perception pathway.

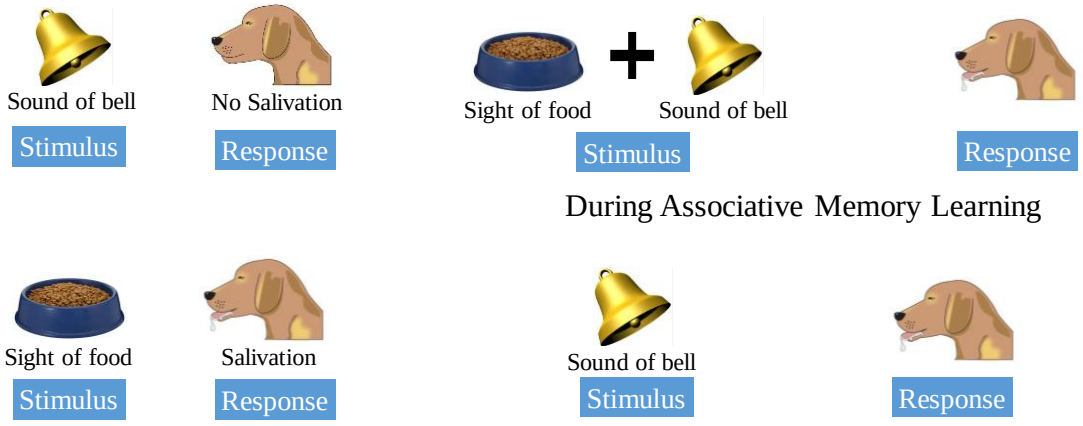


Figure 5-1: Pavlov's experience of associative memory learning on dogs

In this study, the visual perception (seeing food) is referred to as an unconditional stimulus (US) signal pathway since it unconditionally stimulates the salivation. Meanwhile, the auditory perception (hearing ringing the bell) is defined as a conditional stimulus (CS) because it needs the learning process for evoking the salivation. This research exhibits that the stimulus signals from two concurrent events can be correlated with each other if they occur repeatedly [3].

How do these associative memory learning between two initially unrelated events happens? As previously discussed in Chapter 2 (see Figure 2-4), the captured visual and auditory signals travel along two distinct signal pathways and are processed in different brain regions. With this major premise, a reasonable hypothesis of the incapability of bell's sound (conditional stimulus) on evoking the salivation reflection is the impassible signal pathway between the sensory neurons of the sound of bells and the response neurons of salivation reflection. Meanwhile, the visual signal of the presence of food determinedly (unconditionally) travels to the group of neurons stimulating the salivation refection. Thus, an underlying possible reason the phenomena of salivation reflection of dogs evoked by the conditional stimulus, the sound of the bell, after the associative memory learning process is a signal pathway modification process, which means the signal pathway of the sound of the bell is modified so that it learned the capability of stimulating the reflection of salivation. The associative memory is a pervasive learning mechanism in organisms, including primates, invertebrates, etc. Figure 5-2 illustrates the associative memory learning of mice correlating the tone and shock signals together [3].

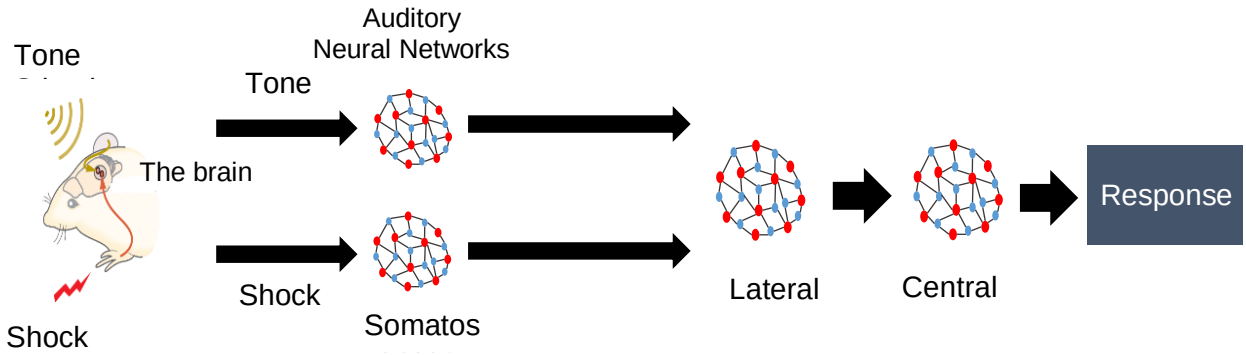


Figure 5-2: Associative memory in mouse.

However, how is this signal pathway modification achieved? The investigation of cellular level associative memory on Aplysia reveals the mechanism of signal pathway modification. The associative memory learning conducted by Dr. Kandel's research on Aplysia (2000 Nobel Prize) reveals the relationship between the synaptic weight and the associative memory learning [3]. The Aplysia was selected as the research object due to the simplicity of its neural system. The associative memory learning in Aplysia includes two signal pathways connecting the sensory and response neurons marked in blue and red respectively in Figure 5-3.

Normally, the gill motor is unresponsive to the siphon stimulation of the siphon before learning. However, by performing a training experiment which consisted of applying a shock to the tail (US) and touching the siphon (CS) simultaneously and repeatedly, the gill motor neuron became more responsive to inputs from the siphon sensory neuron (CS). As depicted in Figure 5-3 (b), the stimulus from the US and CS are paired and overlapped with each other in the time that is considered as a trigger condition of associative memory learning at the cellular level [3]. The increased magnitude of the gill motor response results from a stronger synaptic connection induced or imprinted between the sensory neuron of the siphon and the motor neuron of the gill during the associative learning process. This cellular association learning behavior comes from the increment connection strength between the sensory neuron and response neuron due to the repeatedly and simultaneously US and CS.

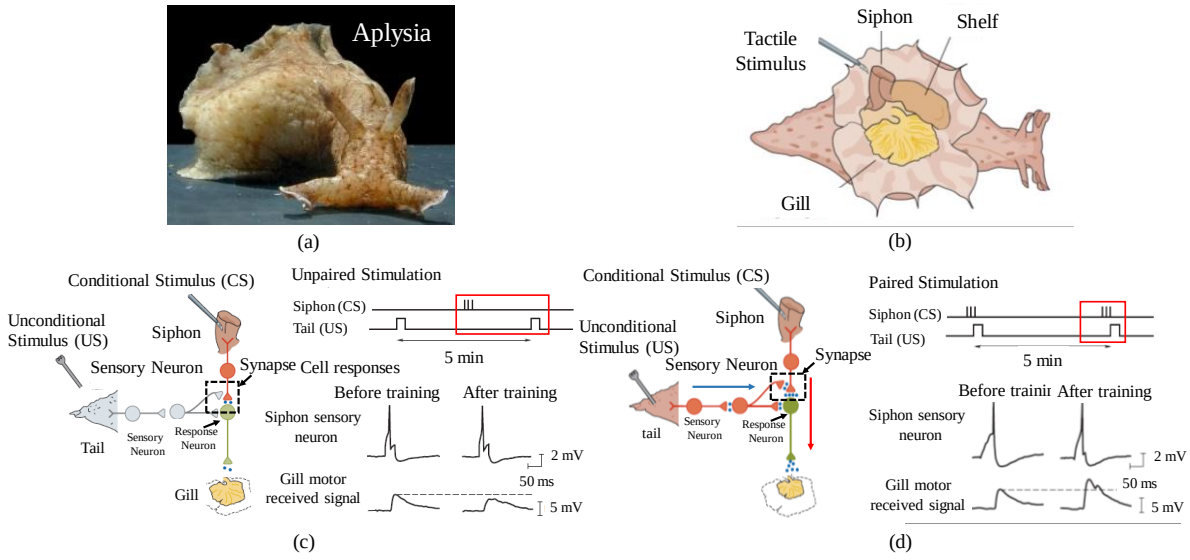


Figure 5-3: (a) Aplysia (b) The experimental setup (c) The siphon of aplysia and tail are stimulated by touching and shocking respectively. The received signal of response neuron (Gill motor neuron) stay almost same before and after training under the unpaired stimulation (d) Under the paired simulation, a larger magnitude of the received signal at gill motor neuron under is monitored [3].

The investigations on associative memory at the cellular level reveal that the changes in synaptic weight play a critical role in the associative memory [3]. The weight of a synapse, the amount of the chemical neurotransmitters, represents the connection strength between two

neurons. With the increase of the connecting strength between neurons, the relationship between two concurrent stimuli is memorized [3].

5.3 Realizing Associative Memory Learning with Memristive Synapses

In this subchapter, the methodologies of realizing associative memory learning are introduced and discussed. Figure 5-4 exhibits a neural network model mimicking associative memory learning in Aplysia (Figure 5-3) that consists of two main signal pathways: conditional and unconditional pathway. The unconditional pathway directly connects the sensory neuron A1 (US) to the response neuron, while the conditional pathway connects sensory neuron B1 (CS) to the response neuron through a memristive synapse. On the conditional signal pathway, an analog summation device is used to couple conditional stimulus (from neuron B1) and an unconditional stimulus (from neuron A1) together.

Initially, the stimulus signal from B1 to response neuron is small due to the attenuation effect of the high resistance of the memristive synapse. Furthermore, the magnitude of the spiking signals generated by A1 and B1 are both smaller than the set voltage of the memristive synapse, meaning the signals from A1 and B1 cannot modify the resistance of memristive synapse alone. Consequently, the associative memory learning cannot be accomplished. However, when the neuron A1 and B1 fire simultaneously generating overlapped spiking signals in time, the outputs spiking signals from them will couple together. Then the coupled signals potentially excee the set voltage of the memristive synapse resulting in decreasing its resistance. Thus, the magnitude of the signal arriving at response neuron is increased due to the reduced resistance of memristive synapses. This model (Figure 5-4) perfectly reproduces the cellular level associative memory learning of Aplysia.

The cellular level associative memory model (Figure 5-4) only associates signals that carries less sophisticated information. These two signal pathways merely mimic the simple exterior stimulus, like the tough from the tail and the cut from siphon (see Figure 5-3). The simplicity of the model restricts the capability of the system from learning more complex information, such as auditory or visual signals.

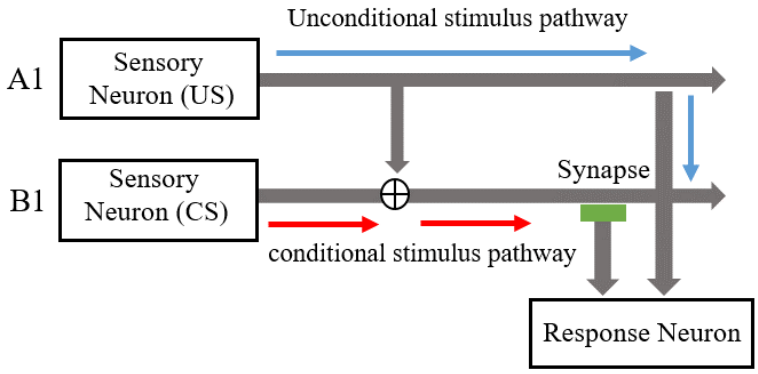


Figure 5-4: Cellular level associative memory model with two signal pathways and memristive synapse

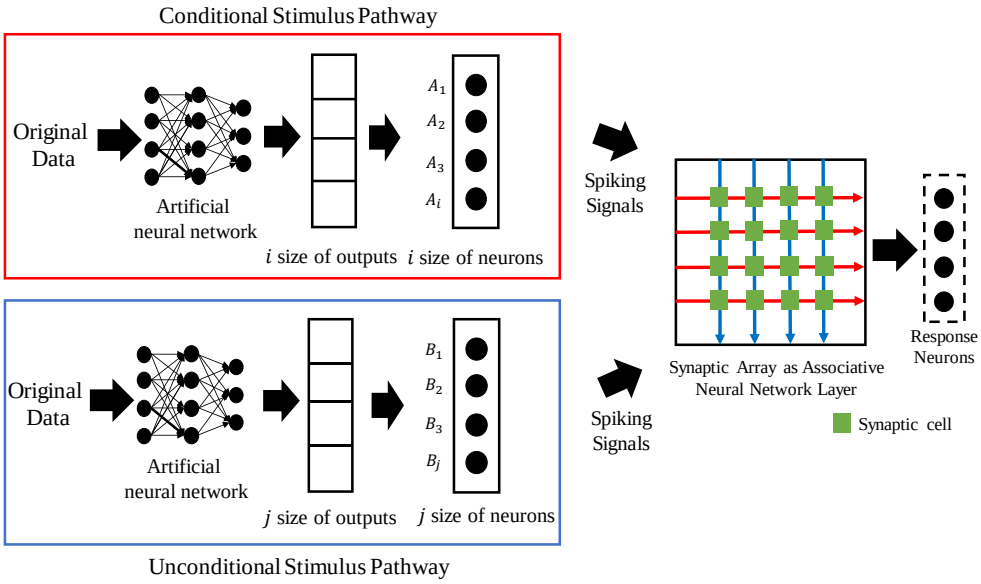


Figure 5-5: Large-scale associative neuromorphic system associating two ANNs together.

In order to resolve this limitation, a large-scale model of associative memory learning is proposed (Figure 5-5). In this model, the pieces of sophisticated information are preprocessed by various ANNs mimicking the distributed signal processing method in the brains (see Figure 2-4). The outputs of an ANN are usually a probabilistic number (score) between "0" to "1", representing a degree of prediction accuracy. The score indicates the probability of the original import data, e.g., image, voice, belonging to a specific category. In this way, the information carried by these input images, voices, etc., is transformed and embedded into a series of probabilistic scores.

Therefore, if these scores can be associated together, the information carried by these scores theoretically will be also correlated together. Generally, the neural networks are separated into training and operating phases [3]. In the operating phase neural network topology and synaptic weight are constant, whereas the synaptic weights are changeable. In the proposed associative architecture, the ANNs are in the operating phase, while the associative memory network is in the training/learning phase.

5.4 Signal Intensity Encoding Neuron

The main functionalities of a neuron as a computing unit can be summarized as: (1) to integrate/sum spiking signals from other neurons; (2) to generate a spiking signal sequence under the condition of the integrated signal exceeding some specific threshold voltage. The intrinsic behavior of the neuron is defined as the firing behavior. In the past century, many successful neuron models, as listed within Table 5-1, have been proposed to implement the two characteristics of the neuron.

Table 5-1: State-of-the-art Neuron Models

Neuron Model	Year	Reference
Integrate and Fire	1907	[180]
Hodgkin-Huxley	1952	[181]
Leaky integrate-and-fire	1965	[182]
Izhikevich	2003	[183]

However, all these neuron models focus on the spiking waveform characteristics (i.e. shapes, magnitude, rising/falling time, etc.) while ignoring the firing rate of the neurons which correspond to the stimulus signal intensity, which is a universal neuron firing feature in real biological systems. For example, the skin cold receptor neurons firing rate is corresponding to the temperature[184].

In our novel neuron design, we emphasize that actual biological neurons possess this capability In order to model this input intensity-dependent firing characteristic of the neurons, we designed a Signal Intensity Encoding Neuron (SIENs) using the Integrate and Fire neuron model [180] as the core spiking signal generating module. In the proposed associative memory learning

models (see Figure 5-4 and Figure 5-5), SIENs are used for encoding the analog input signals into spiking signals with different the frequencies and magnitudes [185]. The specific frequency and magnitude of spiking outputs of the SIEN depends on its input. A bigger input generates a spiking output signals with higher frequency and larger magnitude.

In the neural system. the amplitude and frequency of spiking signals is proportional to the input stimulus. For instance, the amplitude and duration of a muscle neuron, depend on the intensity of the muscle stretch [3]. The more intensive stretch stimulates a spiking signal with higher firing frequency [3]. Although these features widely exist in biological neurons [3], other state-of-the-art neuron designs [103, 183, 186-202] lack the realization of these features. The associative memory learning is realized through updating the synaptic weight with a concurrent firing behavior of the sensory neurons at US and CS pathways. The weight updating behavior occurrence depends on whether the magnitude of the coupling signal from the sensory neurons exceeds the set voltage of the memristive electronic synapse. Thus, the SIENs, as the sensory neurons, are specifically designed to generate a spiking signal, whose magnitude and firing frequency is proportional to the input stimulus. The model of SIEN is simulated by TSMC 180nm technology. The schematic of SIEN is illustrated in Figure 5-6.

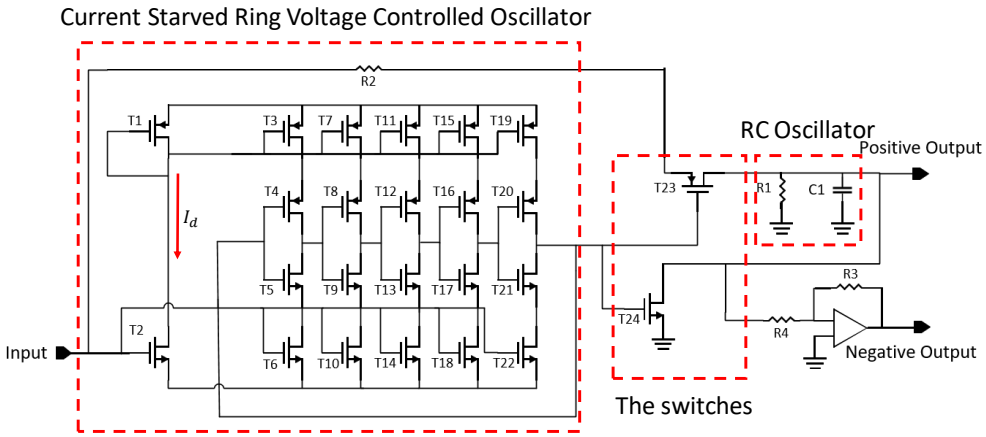


Figure 5-6: Signal Intensity Encoding Neuron (SIEN) schematic

As a result, the external stimulus signal with lower magnitude generates the spiking signals with smaller magnitude accordingly, which thus can not trigger the associative memory learning. The coupled spiking signal from neurons is responsive to updating the weight of memristive

synapse. The higher main frequency (smaller intervals between spikes) of the spiking signal would increase the opportunity of superposition of two spiking signals.

As depicted in Figure 5-6, there are three central parts of a SIEN: Current Starved Ring Voltage Controlled Oscillator (VCO), a switch pair, and a resistor-capacitor (RC) oscillator. The analog input signal would firstly be imported into the Current Starved VCO to generate an oscillating signal, and its frequency is proportional to the input signal magnitude. Next, this oscillating signal controls a switch pair constructed with a PMOS (positive channel metal oxide semiconductor) transistor and an NMOS (negative channel metal oxide semiconductor) transistor. By controlling the oscillating signal, the switch pair would be charging and recharging the RC oscillator to generate a spiking signal sequence. The frequency of the generated spiking signal sequence by RC oscillator would be proportional to the magnitude of the input analog signal due to the Current Starved VOC controlling the "on" and "off" switching frequency of the switch pair. The neuron firing frequency is determined by the Current Starved VOC with the governing equation [203]:

$$f_{fire} = \frac{I_d}{NC_{total}V_{DD}}, \tag{5-1}$$

where N is the number of inverter stage, C_{total} is total charging and discharging capacitance of one stage inverter in Current Starved VOC, and V_{DD} is the power supply voltage. The firing frequency is determined by the current I_d, controlled by the input stimulus as illustrated in Figure 5-6.

Moreover, the source terminal of the PMOS transistor in the switch pair is connected to the input signal serving as a charge provider to control the magnitude of the output spiking signal. The effective switching resistances of the PMOS and the NMOS are denoted as R_p and R_n, respectively.

The governing equations of the charging and discharging processes are listed as:

$$V_{charge} = \frac{R_1}{R_1 + R_c} \times V_{input}(1 - e^{-\tau t}), \tag{5-2}$$

$$V_{discharge} = \frac{R_1}{R_1 + R_p + R_2} e^{-t/(R_d C_1)} \times V_{input}, \tag{5-3}$$

where R_c equals $R_2 + R_p$, τ is $(R_1 + R_c)/R_1R_cC_1$, R_d represents $R_1R_n/(R_1 + R_n)$. The steady-state voltage value of the output is governed by the equation:

$$V_{output} = \frac{R_1}{R_1 + R_p + R_2} V_{input}. \quad (5\text{-}4)$$

Moreover, the SIENs could also generate positive and negative signals simultaneously, which is critical for the novel memristive synapse updating method. Figure 5-7 demonstrates the positive and negative output spiking signals of a SIEN with 700mV square waveform as the stimulus input. The firing response frequency and magnitude corresponding to the different input voltages is illustrated in Figure 5-8(a). As aforementioned, the behavior associative memory learning will correlate two type of information together (see Figure 5-5). In this Chapter, the pronunciations (auditory signal) and images (visual signal) of digits are associated together to perform a behavior level associative memory learning. The SIENs need to map the scores into the frequency and the magnitude of outputs. As depicted in Figure 5-8(b), the scores mainly distribute within the intervals [0 0.05] and [0.95 1], indicating the lowest and highest scores respectively. This means the input of SIENs will be within two separated ranges, below 0.05 V and above 0.7 V, accordingly, which are marked in Figure 5-8(a).

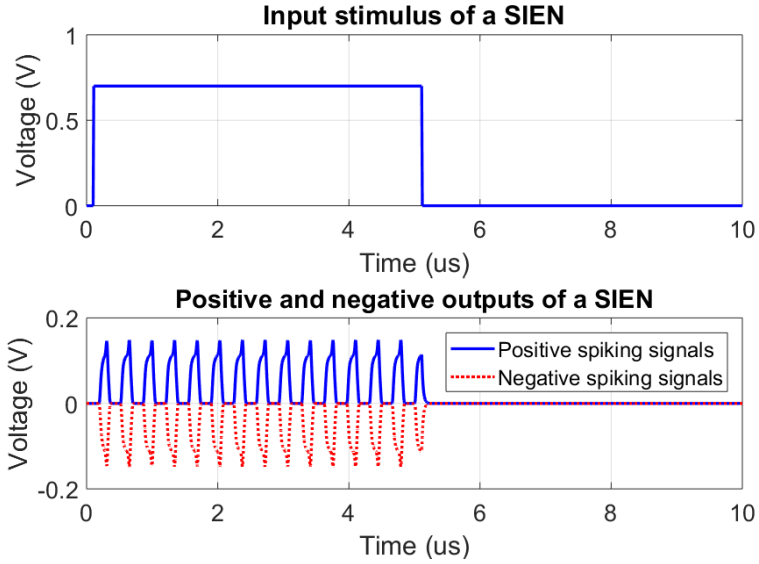
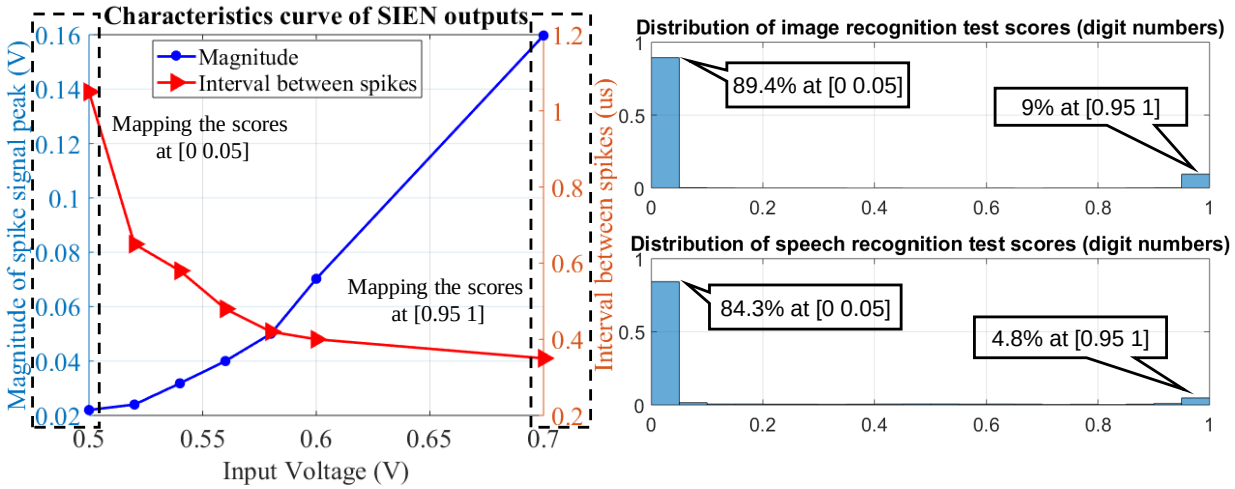


Figure 5-7: Positive and negative output spiking signals of a SIEN with 700 mV square wave signal as an input stimulus.

The scores in Figure 5-8(b) are generated by using the datasets of Modified National Institute of Standards and Technology database (MNIST) for digit image recognition [204], and Spoken Digit Commands Dataset (SDCD) for digit speech recognition. SDCD is a subset of the Speech Commands Dataset from Google containing 10,000 training and 1,000 test recordings corresponding to spoken digits from 0 to 9 [205].



(a) Characteristics curve of SIEN outputs

(b) Distribution of image and speech recognition scores (digit number)

Figure 5-8: (a) Characteristics curve of SIEN outputs (b) Distribution of image and speech recognition scores on digits using the datasets: MNIST and Spoken Digit Commands Dataset

5.5 Modeling of Vertical Three-dimensional Memristive Synapse

The memristive device, also referred to Resistive Random-access Memory (RRAM), is widely applied as an ideal electronic synapse candidate due to its programmable resistance [99]. The resistance of a memristor is modified with the applied voltage on its terminals excesses a specific value, called as its set voltage. The resistance modification from the high resistance state (HRS) to the low resistance state (LRS) is defined as a set process. Typically, the memristor is constructed by the metal-insulator-metal configuration. The decrease of the resistance is caused by the formation of the conductive filament in its insulator layer. The increase strength of synaptic connection, indicating a successful associative memory learning behavior [3], can be realized by programming the resistance of the memristor from its HRS into LRS. Consequently, the received

voltage/current of the postsynaptic response neuron increases demonstrating the accomplishment of the associative memory learning [3].

In the metal oxide, the bonding between oxygen ions and metal atoms is breakable. Under the high electric field (>10MV/cm) stimulated by the applied voltage, some oxygen ions in the metallic oxide would escape from the constraint of the bonding force and drift toward the anode side of a memristor [78]. The deficiency of oxygen ions leaves the oxygen vacancies or metal precipitates, which would further construct the conductive filaments (CFs) [105, 106, 206]. As a result, two current paths exist in its LRS. One is through the original oxide and the other is through CFs. These two paths in the parallel lead to the decline of the memristor resistance. In the reset process, the oxygen ions at the interface migrate back into the oxide to refill the oxygen vacancy or re-oxidize the metal precipitates to update the resistance of the memristor back to its HRS.

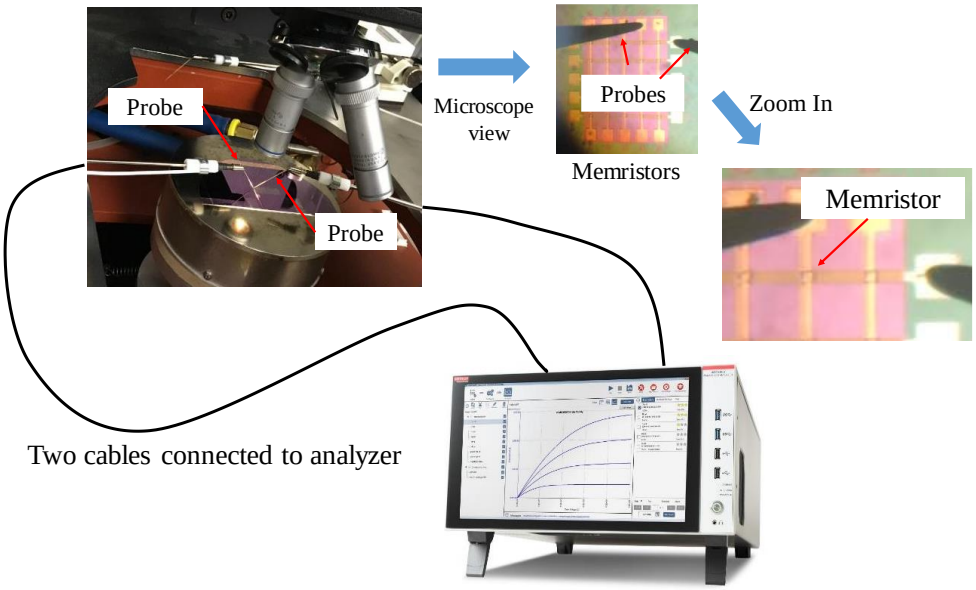


Figure 5-9: The memristor array and the experiment setup with semiconductor parameter analyzer from Micro & Nano Fabrication Laboratory at Virginia Tech (http://www.micron.ece.vt.edu/)

The memristive synapse in this paper is used for demonstrating a biological-like associative memory mechanism (Figure 5-3) indicating the synaptic connection strengthening between neurons as the associative learning accomplishment. This strengthening behavior is modeled as

the memristor resistance switching from HRS to LRS. Therefore, this paper would mainly focus on modeling the set process of the memristor without discussing the reset process, which reduces the connection strength between neurons and is considered as a biological disremembering phenomenon [3].

Based on the conductive filament evolution concept, we develop a memristor model for the memristive synapse array simulation in the large-scale associative memory learning system. Figure 5-10 illustrates the V-I characteristic curve comparison in the set process of VT memristor model and the measurement data. The measurement setup of the memristor is illustrated in Figure 5-9. As depicted in Figure 5-10, the resistance of the memristor model would switches from its HRS (1.6 MΩ) to LRS (64 KΩ) at ~3.2 V. the current is at ~50 uA, which matches the measurement data. The current response mismatch above 50 μA comes from the activated current-compliance for protecting the device on the measurement setting. The detailed parameters of the memristor model are listed in Table 5-2.

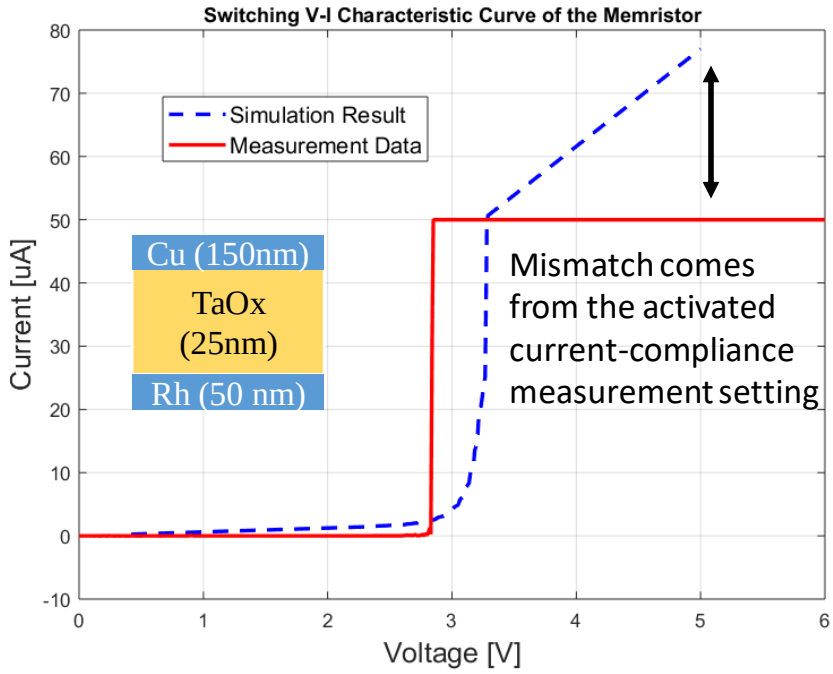


Figure 5-10: Switching V-I characteristic curve of the memristor. The current response mismatch above 50 μA comes from the activated current-compliance for protecting the device on the measurement setting.

The measurement data in Figure 5-10 come from the memristive device (Cu/TaOx/Rh) fabricated at the Micro and Nanofabrication Laboratory at Virginia Tech [108]. In the memristor, Copper (Cu) serves as a top metal electrode, oxygen-deficient tantalum oxide (TaOx) as solid electrolyte and Rhodium (Rh) as a bottom electrode. The device has been characterized by monitoring the forming voltage (Vform) when conductive filaments (CFs) are being formed initially. The reset voltage (Vreset), the set voltage (Vset), and the resistance switching characteristic with the applied ramp-shape stimulus having a rate of 2.0V/s. Table 5-3 lists the characteristic parameters of the fabricated memristor. For this device, the set voltage is 2.85V and the reset voltage is -3V.

Table 5-2: Parameters of the Memristor Model

Parameter	Descriptions	Values
I_0	Hopping current density in the gap region	1E13 A/m^2
ρ	Resistivity of the CF	2.5E-4 Ω/m^2
a	Distance between adjacent oxygen vacancy	0.25 nm
f	Vibration frequency of oxygen atom	1E13
x_0	Initial length of the memristor	5E-9
x_T	Characteristic length in hopping region	0.4Ee-9
V_T	Characteristic voltage in hopping	0.4
w_0	Initial CF width	1E-9
R_H	High Resistance State	1.6 MΩ
R_L	Low Resistance State	64 KΩ
E_a	Average active energy	1.2eV

α_a	Enhancement factor	0.75 nm
Z & e	Charge number & unit charge	1 & e
k_B	Thermal resistance	0.86177e-5
T	Temperature	300K

Table 5-3: Measurement results of the Memristor

Parameters	**Value**
V_{form}	4 V
V_{set}	2.85 V
V_{reset}	-3 V
Thickness of Cu layer	150 nm
Thickness of T_aO_x layer	25 nm
Thickness of Rh layer	50 nm
Set voltage ramp rate	2.0 V/s

The traditional large-scale memristor array is fabricated in a 2D crossbar configuration which suffers the large design area, power consumption, etc. Therefore, in this paper, we use a vertical memristor structure to offer the following promising benefits, the design area, and power consumption would be reduced by 50% [6] and 35% [156], respectively. Furthermore, we use a plane as the layer access port due to the large resistance attenuation effect of the narrow nanowire on accessing multiple memristors [114].

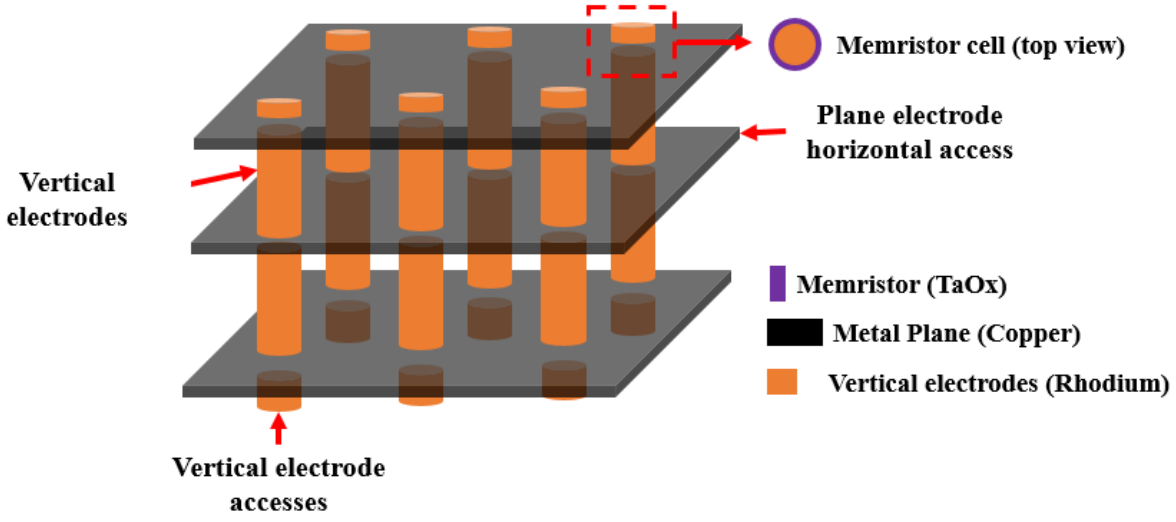


Figure 5-11: 3D vertical memristor structure

Figure 5-11 illustrates the vertical 3D memristive synapse array structure. The geometry of the structure is illustrated in Figure 5-12 and Figure 5-13. This structure uses vertical planes and monolithic inter-tier vias (MIVs) serving as horizontal and vertical access ports. The MIVs electrode and the plane materials were modeled as copper and rhodium, respectively. The TaOx is used as memristor material sandwiched at the intersection region between the horizontal plane and the vertical MIVs. The 3D vertical memristor structure can be modeled with an array configuration illustrated in Figure 5-14. Since the memristor at each layer are connected with each other with a plane metal physically, the port denoted as $Port\_P_i$, can access each memristor with the plane resistance denoted as R_{plane}. The resistances of the MIVs is denoted R_v. The values of the parasitic capacitance between the planes (C_{p_p}), the plane to the via (C_{p_v}), and the MIV to the MIV (C_{v_v}) are listed in Table 5-4. These values are extracted by the ANSYS Q3D Extractor, an industry standard tool for capacitance and resistance computation. The detailed geometry of the 3D vertical memristive synapse structure is listed in Table 5-5. Due to the extremely small parasitic capacitance (~fF), the effect of parasitic capacitance in the design is negligible.

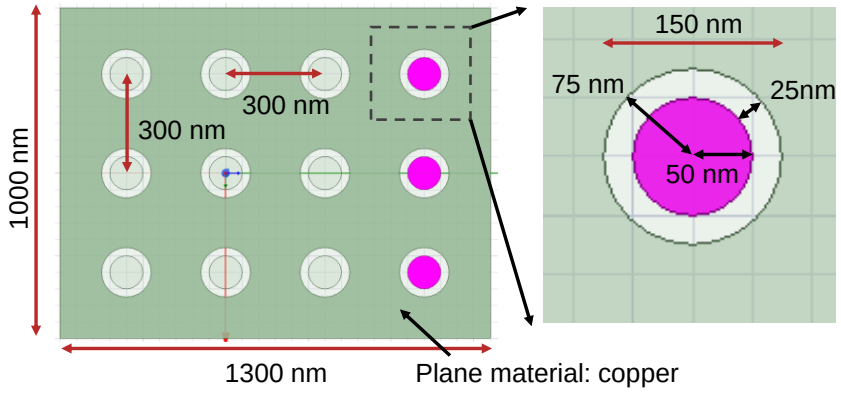


Figure 5-12: Top view of the 3D vertical memristive synapse structure

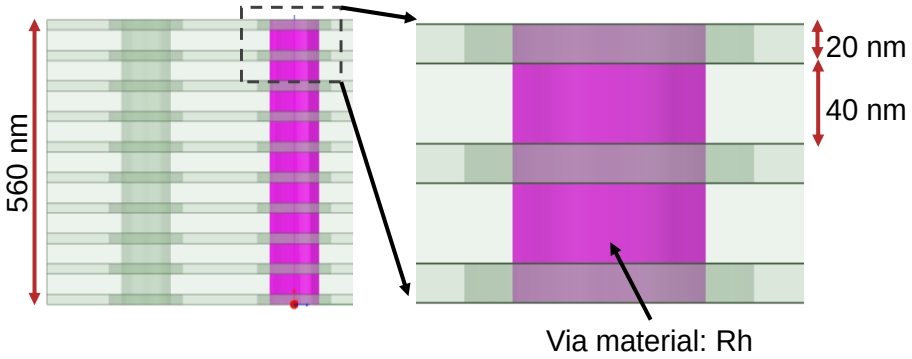


Figure 5-13: Side view of the 3D vertical memristive synapse structure

Table 5-4: Parameters of the vertical 3D memristive synapse model

Parameters	Descriptions	Values
R_{plane}	The resistance of the plane	1.179 Ω
R_{via}	The resistance of inter-layer via	0.406 Ω
C_{v_v}	The parasitic capacitance between the vias	1.19 E-8 pF
C_{p_v}	The parasitic capacitance between the plane and the via	7.43 E-6 pF
C_{p_p}	The parasitic capacitance between the planes	7.6 E-5 pF

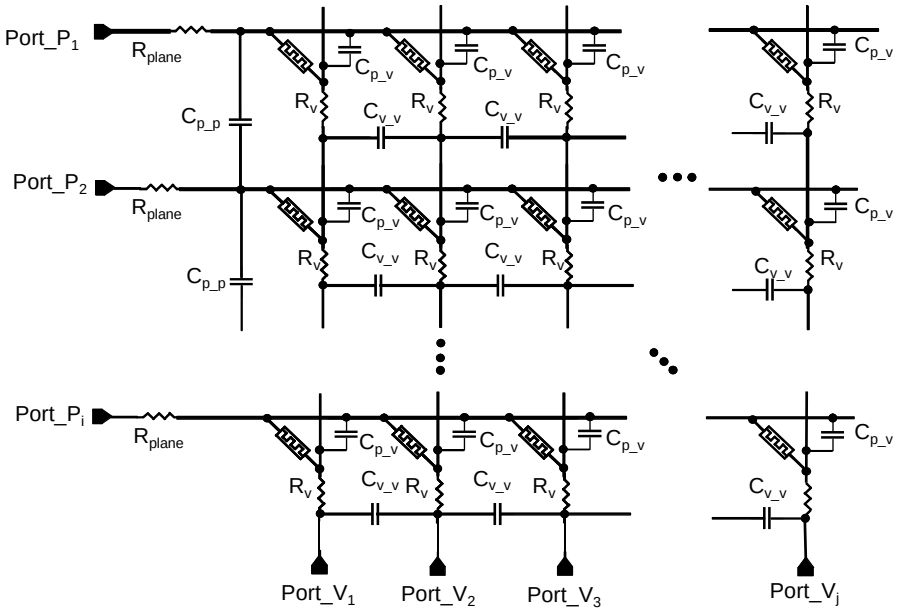


Figure 5-14: Model of the vertical memristive synapse array

Table 5-5: The geometry and materials of the vertical 3D memristive synapse

Parameters/Descriptions	Values
The distance between the MIVs	300 nm
The radius of the MIVs	50 nm
The distance between the MIVs and the anti-pads	25 nm
The size of the plane	1000 nm × 1300 nm
The distance between the planes	40 nm
The thickness of the plane	20 nm
The material of the plane	Copper
The material of the via	Rhodium
The insulator between the planes	SiO_2

5.6 Cellular Level Small-scale Associative Memory Learning

The cellular level small-scale associative memory model with memristor (see Figure 5-4) requires additional nanowires and adders for the signal coupling, which increases the circuit design area [207]. To address this issue, a novel memristor weight (resistance) updating scheme (see Figure 5-15) is proposed without adding modules. Furthermore, the memristor resistance updating behavior of the proposed scheme is controlled by the applied voltage at its two terminals rather than through a selector device. Thus, the proposed memristor updating scheme makes a nanoscale synaptic array practicable, since the design area of the memristor array is mainly limited by the large selector device, e.g., transistors or diodes [208].

As depicted in Figure 5-15, the memristor in this scheme receives two opposite polarity signals at its terminals whose voltage potential difference is the stimulus signal for triggering resistance updating of the memristor. The spiking signals from neuron B1 and A1 can be considered as the waveforms propagating in the wires. With the impedance matched terminals, no reflection signals would cause a distortion of the spiking signals. The resistance of the memristor would be modified when the voltage potential at the terminals exceeds its set voltage.

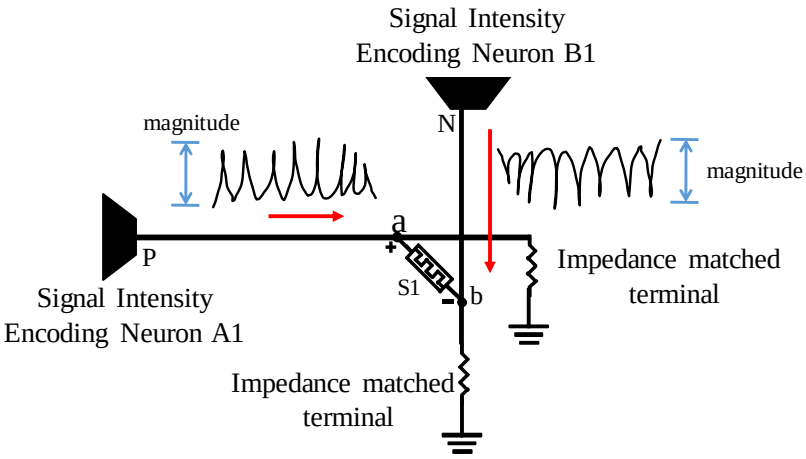


Figure 5-15: Novel memristor weight updating scheme

Figure 5-16 and Figure 5-17 illustrate the simulation results of the proposed memristor weight updating scheme. The output spiking signal of SIEN B1 is negative. In Figure 5-16, two square inputs of SIENs are not perfectly synchronized and only partially overlapped. At the non-

overlapping part, both signals are small, and cannot trigger the memristor switching alone (see Figure 5-17).

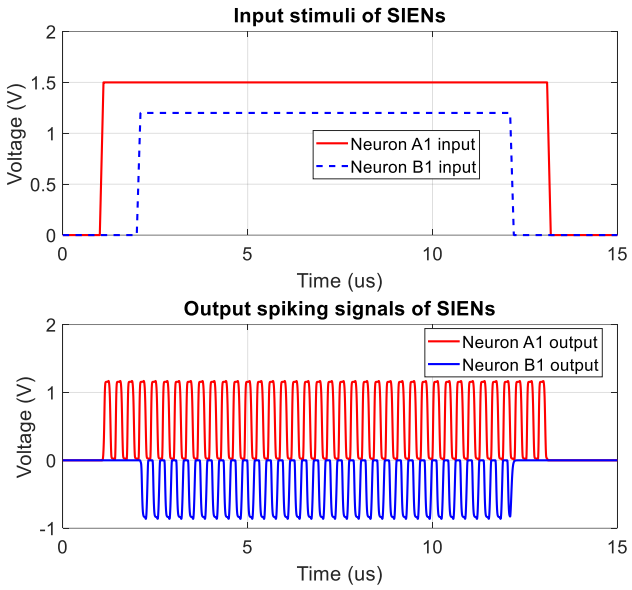


Figure 5-16：Input analog signals and output spiking signals of Neuron A1 and Neuron B1

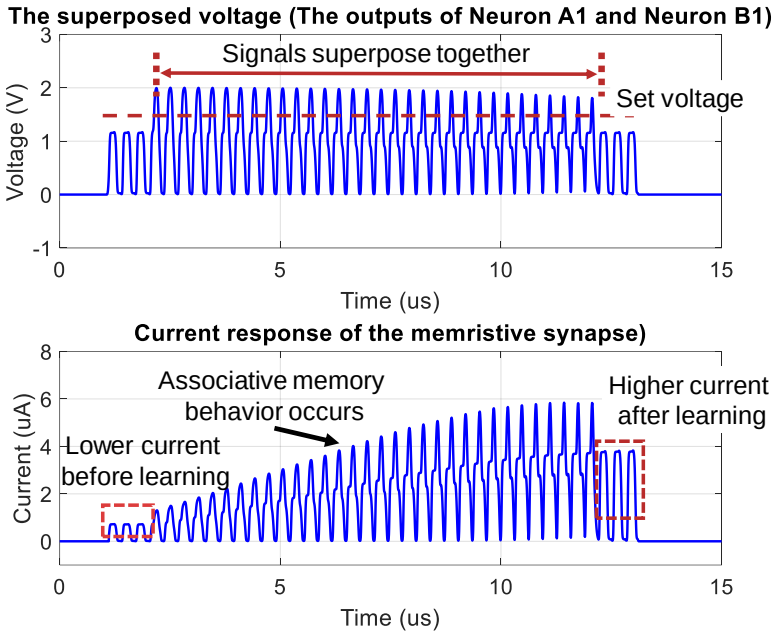


Figure 5-17: Voltage potential at terminals of the memristor, which is the superposed voltage of Neuron A and B outputs, and the corresponding current.

At the overlapping part, two signals are superposing their peak values with each other. Consequently, the magnitude of the superposed spiking signal will be larger than the set voltage of the memristive synapse, resulting in a resistance modification behavior. As illustrated in Figure 5-17, the current after learning is larger than the current before learning indicating a successful associative memory learning behavior at the cellular level in Aplysia (see Figure 5-3).

5.7 Behavior Level Large-scale Associative Memory Learning

The next approach of realizing associative memory learning is to extend the scale of the system from small to large for enhancing the learning capability. In the model of behavior level large-scale associative memory learning (see Figure 5-5), the original data is first processed by the ANNs. The information carried by the original data is abstracted into the output scores of ANNs. Then the scores are further imported into the SIENs, Next, SIENs encode the scores into a series of spiking signals whose magnitudes and frequencies corresponding to the values of the scores. The highest scores would be transferred into a spiking signal with the highest peak magnitudes and the shortest interval between spikes, accordingly. At last, the spiking signal outputs of SIENs are delivered to a synaptic array for a large-scale association. The size of the synaptic array is $i \times j$ which are the index of SIENs at two stimulus pathways as illustrated in Figure 5-5. For instance, in the experiments of Pavlov's dog, the input original data of these two stimulus pathways could be visual and auditory signals, corresponding to the presence of food and sound of bells. In this paper, we associate the visual (image) and auditory data (pronunciation) of digits together.

In the synaptic array, the spiking signals couple and superpose with each other at the synaptic cells described with the equation:

$$V_{synapse} = V_{coupled} = V_{A_i} + V_{B_j}, \tag{5-5}$$

where V_{A_i} and V_{B_j} are the output spiking signals from neuron A_i and B_j, respectively. $V_{synapse}$ is the voltage potential between the terminals of the synapse, which is the sum of the V_{A_i} and V_{B_j}. Since the scores from ANNs are different (within the interval [0 1]), the magnitudes of the V_{A_i} (V_{B_j}) are various accordingly. Apparently, the largest spiking signals $V_{synpase_{max}}$ would be generated from the largest signals of SIEN A_i ($V_{A_{max}}$) and B_j ($V_{B_{max}}$). An associative memory learning behavior

would occur under the condition of $V_{synpase_{max}} > V_{set}$, where V_{set} is the set voltage of the synaptic.

By employing the SIENs and memristive synapse array, the behavior level large-scale associative memory learning is reproduced and mimicked in this section. As illustrated in Figure 5-5, unlike the cellular level associative memory with two simple nanowires, the US and CS signal pathways in the proposed behavior level large-scale associative memory learning system are constructed by two ANNs that can preprocess and inference the visual and auditory signals respectively. These two ANNs are both serving as operating phases. In Figure 5-18, the auditory signal and the visual signal of digit number "3" are separately imported into the ANNs for preprocessing. The output is ten scores indicating the probability of the input original data belongs to a specific category. The scores for auditory and visual information of digit 3 are listed in Figure 5-18 (a). In this paper, we use MNIST [204] and SDCD for the visual and auditory input data, respectively. SDCD is a subset of the Speech Commands Dataset from Google containing spoken digits from 0 to 9 [205]. We can observe that the scores for "3", marked in red, are highest among other scores. The values of these scores would be further mapped into corresponding spiking signals by SIENs.

In Figure 5-18, the SIENs from visual data is notated as A_i within the unconditional signal pathways. Meanwhile, the sensory neurons (B_j) at conditional signal pathways are connected to the response neurons through a memristive synapse array. Through the SIENs, the largest scores would generate a spiking signal with the largest magnitudes and highest frequencies and vice versa. The memristive synapses connecting the sensory neuron A_i and B_j are notated as $M\_A_i\_B_j$. The memristive synapse array for the unconditional pathways (red-dash lines) is modeled by the 3D vertical memristor structure. As illustrated in Figure 5-18 (a), the memristive associative memory network contains 20 neurons and 100 memristive synapse.

Figure 5-18 (b) and (c) depict the simulation results. With different analog input signals corresponding the scores, the superposed voltage difference at the memristive synapses is different accordingly. The synapse of $M\_A_4\_B_4$ has the largest input stimulus due to the corresponding highest scores.

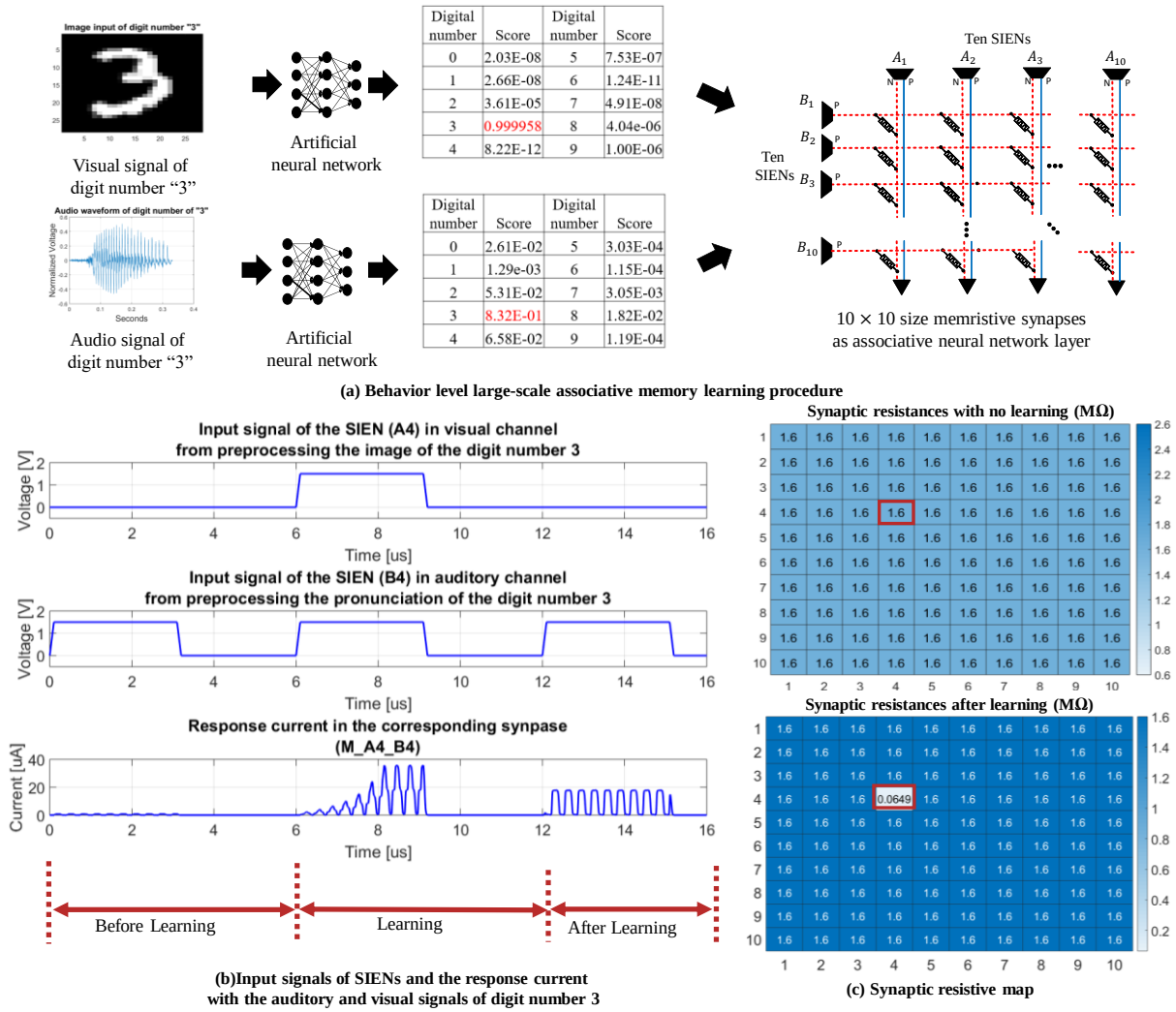


Figure 5-18: (a) Behavior level large-scale associative memory learning procedure. (b) the detailed associative memory learning signals at the memristive synapse of M_A4_B4. (c) the resistance values of the memristive synapses (HRS and LRS) before and after associative memory learning. The associative memory learning only occurs at M_A4_B4 marked in the red square.

Figure 5-18 (b) illustrates the detailed current response in memristive synapse $M\_A_4\_B_4$. When only the auditory signal is provided (no firing behavior in A_i neurons), the current in $M\_A_4\_B_4$ is very small (<1uA). During the learning process, the visual and auditory input are presented simultaneously (firing behavior occur in A_i and B_j neurons), the current in $M\_A_4\_B_4$ is gradually increasing, which indicates the resistance reduction of the memristor and the associative memory learning behavior is accomplished. In Figure 5-18 (c), the memristive synapse of $M\_A_4\_B_4$ switches from its HRS (1.6 MΩ) to its LRS (64KΩ). On the contrary, other memristive

synapses, connecting the sensory neurons receiving lower input analog stimulus signals, do not switch since the voltage potentials of the spiking signals at their terminals are lower than the set voltage of the memristors

5.8 Discussion

The novel behavior-level associative memory learning methodology has been proposed and analyzed with the corresponding neuromorphic circuitry designs including SIENs, 3D memristive synapse array, and a novel memristive synapse updating scheme. The approach successfully associates two large-scale ANNs together, realized by associating the outputs of ANNs with an extra layer of neural network referred to an associative memory network. The outputs of the ANNs, representing the probabilities of the input belonging to a particular category or prediction, would be encoded into the magnitudes and frequencies of spiking signals and associated together for updating the corresponding memristive synapse weights. The coupling signal from the two highest values of the outputs of ANNs would decrease the resistance of the memristive synapse from HRS to LRS. The decrease of the resistance of the memristors demonstrates that the connection between presynaptic and postsynaptic neurons is becoming strong, which further indicates an accomplishment of successful associative memory behavior.

Compared with other state-of-the-art memristor-based associative memory models (<10 synapses) listed in TABLE 5-6, the proposed large-scale memristive synapse model successfully relates the signals from 20 neurons together with 100 memristive synapses, realizing a behavior level large-scale associative memory learning of associating the auditory and visual information of digits together like our brain.

Table 5-6: Comparisons of scales and Association Capability with other related works

	Scale		**Synapse**		**Neuron**	**Association Methodology**	**Association capability**
	Neurons	**Synapse**	**Device**	**Structure**			
[175]	6	3	RRAM	2D memristor	Binary neuron model	Hopfield network	Associate signals

							(cellular level)
[173]	3	1	RRAM	2D	leaky integrate-and-fire	Spike-rate-dependent plasticity	Associate signals (cellular level)
[174]	5	6	RRAM	2D/1R	N/A	N/A	Associate signals (cellular level)
[172]	3	1	RRAM	2D/1R	N/A	N/A	Associate signals (cellular level)
[170]	3	1	RRAM	2D/1R	N/A	Adding	Associate signals (cellular level)
[167]	3	2	RRAM + ADC + digital controller	N/A	Electronic neuron (ADC + microcontroller)	Hebbian rule	Associate signals (cellular level)
[171]	N/A	N/A	PCM	N/A	Integrate-and-fire neurons	Spike timing dependent plasticity	Associate signals (cellular level)
This work	10 + 10	10 × 10	RRAM	3D RRAM structure	SIEN (Ver. 2)	Associate the output of multiple neural networks	Associate two ANNs together

Chapter 6. Conclusions and Future Work

6.1 Conclusions

People have been exploring the surrounding world for thousands of years. The knowledge learned in this exploration has promoted our civilization to unprecedented prosperity. Now, it is time for us to switch the focus from surroundings into an interior world of neural science and brains and cogitation. the research direction of mimicking neural networks of brains through algorithms and circuitry, which is referred to as neuromorphic computing, has a great potential to be the foundation of next-generation artificial intelligence platform.

In this dissertation, three emerging architectures of next-generation neuromorphic systems are proposed showing a promising path to realizing next-generation platform of artificial intelligence with self-learning capability and high energy efficiency. Additionally, memristors and three-dimensional integration technology are applied for designing a high-performance neuromorphic computing system. In this work, a Deep-DFR model is used for evaluating the VT low variation memristors as the weight storing devices. The datasets CIFAR-10 and CIFAR-100 are used for training the Deep-DFR model. The design area, power consumption, and latency of the Deep-DFR system with VT low-variation memristor are reduced by ~48%, ~42%, and ~67% compared to conventional SRAM memory technique. At last, these hardware parameters are also improved at various degrees (~13%-73%) compared to other state-of-the-art memristors [134, 135].

Furthermore, a new learning method with solid biological rationale *Associative Memory Learning* is exhibited which has the capability of remembering and correlating two concurrent events together. The novel design of the behavior-level associative memory learning system includes SIENs, 3D memristive synapse array, and a novel memristive synapse updating scheme. The main contribution of this work successfully associates the signals from 20 neurons together with 100 memristive synapses, realizing a behavior level large-scale associative memory learning of associating the auditory and visual information of digits together like the brains.

6.2 Future Work

Artificial Intelligence is always one of the most challenging and exciting scientific missions in mankind history. Neuromorphic computing is an emerging approach to realizing an artificial intelligence system through mimicking neural systems mathematically and physically. In my future research, I will devote myself to build a self-learning neuromorphic system with brain-comparable energy efficiency. The details of my research plan are illustrated in Figure 6-1. I focused on designing a Spiking Associative Memory Learning System through Three-dimensional Memristors in my Ph.D. period. Moving forward, I will consummate these works and investigate their board impact on artificial intelligence and seek more board applications.

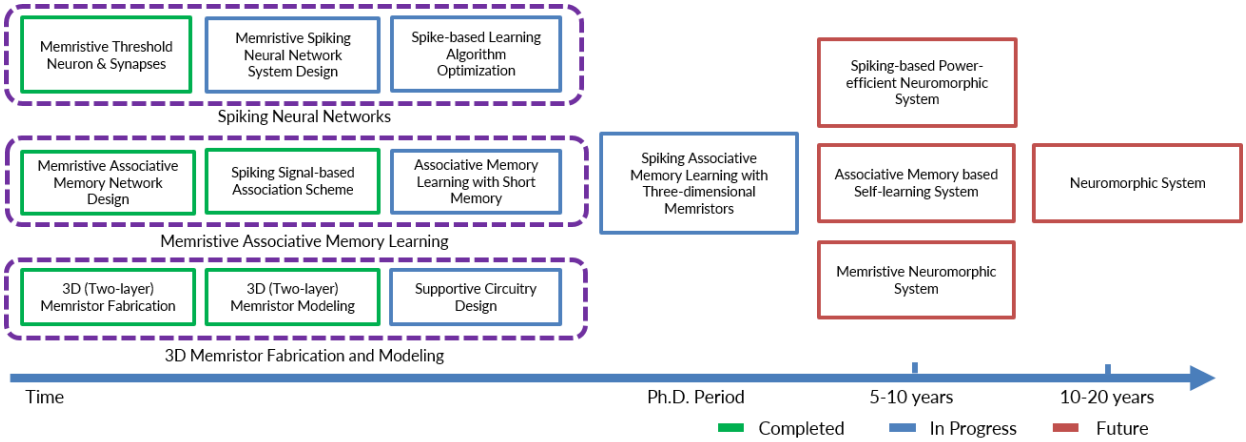


Figure 6-1: Overview of ongoing and future research

The energy efficiency of human brains mainly stems from their low firing rate of neurons and the low magnitude (several tens of millivolts) of membrane potentials (spikes). Inspired by these unique features of human brains, I want to design a spike-based neuromorphic system with a low firing rate (both on forwarding inference and backward learning) to achieve the brain-comparable energy efficiency. For now, I am developing a method to convert the traditional ANNs into a spiking neural network with more biological threshold neurons during the training process. This method can achieve a binary communication scheme in a well-trained neural network with satisfying accuracy. The binary communication between neurons can be represented by one single spike. As a result, the energy spent on communication is significantly reduced. My future work will address this limitation of training algorithms for BNNs and investigate the spike-based

backpropagation algorithm. In this way, the communication signals among neurons, including forward and backward, are entirely in the form of spiking signals rather than high computational digital signals.

Building a neuromorphic computing system with a self-learning capability like the brain has been investigated for a long time [19]. The animals learn skills on their own through experience based on a unique learning method so-called associative memory learning [3]. Through this learning method, dogs can learn the sound of bells as a sign of food; people can remember a word representing an object [3, 165]. In my work [209], I have designed a memristor-based associative memory system that successfully correlates the visual and auditory signals together forming a behavior level associative memory learning.

Based on associative memory learning, the neuromorphic system should have the capability of constantly collecting data from surroundings and learning through their own experiences. The learning process is no longer dependent on the manually processed datasets, but the signals from real-world. In the future, I want to further investigate associative memory learning on theory, implementations, and applications. In theory, I plan to exploit the relationship between the associative memory learning and the short-term/long-term memory. The self-learning neuromorphic system requires an organ-like sensory system to constantly capture and process signals from surroundings, which is not available now. In my work [210], I have preliminarily investigated the functionality of synapse on the attention mechanism of the visual system. In the future, I will further design a memristor-based visual sensor system capturing the objection motion with low power consumption and real-time information processing capability. In applications, I would investigate the self-learning machine in the autonomous system, like self-driving cars.

Furthermore, the research on Spiking Associative Memory Leaning provides a solution to avoid large datasets by enabling neuromorphic systems to learn from their surroundings and experience. Since the self-learning neuromorphic systems with associative memory learning can interact with the surroundings, they will have some level of self-adaptivity and the capability of independent working. The neuromorphic system with self-adaptivity is competent to work in some environments where are not suitable for humans, like volcanos, deep oceans, high radiation environments, and outer space.

Additionally, as a next-generation platform of artificial intelligence, neuromorphic chips potentially power the society into the next level of the industrial revolution (Industry 4.0 [211]). Three Industrial Revolutions occurred in human history. The First Industrial Revolution occurred in the 18th century accompanying the invention of a steam engine. The extensive utilization of steam engines successfully transited the production activities of human society from hand production to machine manufacturing. Next, the Second Industrial Revolution at the beginning of the 20th century further improved productivity through the massive employment of electrification and the production line. Lastly, the Third Industrial Revolution started in the late 1950s has propelled our society into the information age, which is built upon integrated circuits, digital computers, and the internet. Since the Third Industrial Revolution, the computer-powered machines in modern factories have enhanced our manufacturing capability to an unprecedented level. However, the machines still lack one of the essential features to further free people from tedious works, which is the capability of independent, smart, and autonomous manufacturing without any human intervention. The thrust of realizing this level of autonomous manufacturing is called the Fourth Industrial Revolution, also referred to as Industry 4.0 [211]. Industry 4.0 requires a smart and autonomous manufacturing system that is capable of performing tasks on their own and making necessary decisions independently [11, 79, 209, 212, 213]. This novel neuromorphic-powered Industry 4.0 system replaces conventional von Neumann-based chips invented in the Third Industrial Revolution with neuromorphic chips. Then each machine at the so-called light-out factory will deploy self-learning and adaptive neuromorphic chips that can seamlessly connect each other through the Internet of Things. Moreover, the control system built upon the neural networks can make rapid and adaptive responses to the changes in the environment. For example, birds can constantly adjust the flying height and direction to avoid obstacles. The capabilities of real-time response and adaptivity of the neural network-based motion system are highly suitable for complex manufacturing tasks in Industry 4.0. The neural network-based robotics would help us to design next-generation advanced robotics for autonomous manufacturing in Industry 4.0 [213-217].

At last, I always believe that the significances of the neuromorphic system are not only limited in engineering but also to neuroscience, such as potential explanations to an optical illusion, visual agnosia, and Alzheimer's disease [3, 218-221]. For example, recent research proves that synapse

loss highly correlates to the cognitive deficits of human brains (Alzheimer's disease) [222-224]. Modeling the neural system of human brains through neuromorphic systems can provide a simulation platform to comprehensively understand the mechanism of these diseases.

Bibliography

[1] K. Roy, A. Jaiswal, and P. Panda, "Towards spike-based machine intelligence with neuromorphic computing," *Nature*, vol. 575, no. 7784, pp. 607-617, 2019.

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[3] E. R. Kandel, J. H. Schwartz, T. M. Jessell, S. A. Siegelbaum, and A. Hudspeth, *Principles of neural science*. McGraw-hill New York, 2000.

[4] A. Ghani, "Neuro-inspired speech recognition based on reservoir computing," in *Advances in Speech Recognition*: InTech, 2010.

[5] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*, 2015, pp. 2048-2057.

[6] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[8] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016.

[9] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295-2329, 2017.

[10] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-Octob, pp. 843-852, 2017, doi: 10.1109/ICCV.2017.97.

[11] H. An, K. Bai, and Y. Yi, "The Roadmap to Realize Memristive Three-Dimensional Neuromorphic Computing System," in *Advances in Memristor Neural Networks-Modeling and Applications*: IntechOpen, 2018.

[12] B. Soediono, "The Handbook of Brain Theory and Neural Networks," *Journal of Chemical Information and Modeling*, vol. 53, pp. 719-725, 1989, doi: 10.1017/CBO9781107415324.004.

[13] S. R. y Cajal, *Comparative study of the sensory areas of the human cortex*. Clark University, 1899.

[14] I. B. Levitan and L. K. Kaczmarek, *The neuron: cell and molecular biology*. Oxford University Press, USA, 2015.

[15] F. A. Khan, D. Almohazey, M. Alomari, and S. A. Almofty, "Impact of nanoparticles on neuron biology: current research trends," *International journal of nanomedicine,* vol. 13, p. 2767, 2018.

[16] T. D. Albright, T. M. Jessell, E. R. Kandel, and M. I. Posner, "Neural science: a century of progress and the mysteries that remain," *Neuron,* vol. 25, no. 1, pp. S1-S55, 2000.

[17] T. D. Albright, T. M. Jessell, E. R. Kandel, and M. I. Posner, "Progress in the neural sciences in the century after Cajal (and the mysteries that remain)," *Annals of the New York Academy of Sciences,* vol. 929, no. 1, pp. 11-40, 2001.

[18] M. F. Bear, B. W. Connors, and M. A. Paradiso, *Neuroscience*. Lippincott Williams & Wilkins, 2007.

[19] C. Mead, "Neuromorphic electronic systems," *Proceedings of the IEEE,* vol. 78, no. 10, pp. 1629-1636, 1990.

[20] M. Davies *et al.*, "Loihi : A Neuromorphic Manycore Processor with On-Chip Learning," 2018.

[21] K. Bai and Y. Y. Bradley, "A path to energy-efficient spiking delayed feedback reservoir computing system for brain-inspired neuromorphic processors," in *2018 19th International Symposium on Quality Electronic Design (ISQED)*, 2018: IEEE, pp. 322-328.

[22] C. Zhao, Y. Yi, J. Li, X. Fu, and L. Liu, "Interspike-Interval-Based Analog Spike-Time-Dependent Encoder for Neuromorphic Processors," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, pp. 2193-2205, 2017, doi: 10.1109/TVLSI.2017.2683260.

[23] C. D. Schuman *et al.*, "A Survey of Neuromorphic Computing and Neural Networks in Hardware," pp. 1-88, 2017, doi: 10.1016/j.neucom.2010.03.021.

[24] M. Osswald, S. H. Ieng, R. Benosman, and G. Indiveri, "A spiking neural network model of 3D perception for event-based neuromorphic stereo vision systems," *Sci Rep,* vol. 7, p. 40703, Jan 12 2017, doi: 10.1038/srep40703.

[25] M. A. Lastras-Montaño, B. Chakrabarti, D. B. Strukov, and K.-T. Cheng, "3D-DPE: A 3D high-bandwidth dot-product engine for high-performance neuromorphic computing," in *2017 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2017: IEEE, pp. 1257-1260.

[26] T. Ferreira de Lima, B. J. Shastri, A. N. Tait, M. A. Nahmias, and P. R. Prucnal, "Progress in neuromorphic photonics," *Nanophotonics*, vol. 6, no. 3, 2017, doi: 10.1515/nanoph-2016-0139.

[27] M. A. Ehsan, Z. Zhou, and Y. Yi, "Neuromorphic 3D Integrated Circuit," pp. 221-226, 2017, doi: 10.1145/3060403.3060470.

[28] C. D. Schuman, "Roadmap for Neuromorphic Computing : A Computer Science Perspective," 2016.

[29] T. Hylton, "Perspectives on Neuromorphic Computing," in *Neuromorphic Computing Symposium on Architectures, Models, and Applications*, 2016.

[30] F. Walter, F. Röhrbein, and A. Knoll, "Neuromorphic implementations of neurobiological learning algorithms for spiking neural networks," *Neural Networks,* vol. 72, pp. 152-167, 2015.

[31] C. D. Schuman, O. Ridge, and A. Disney, "Dynamic Adaptive Neural Network Arrays : A Neuromorphic Architecture," *Proceedings of the Workshop on Machine Learning in High-Performance Computing Environments - MLHPC '15*, pp. 1-4, 2015, doi: 10.1145/2834892.2834895.

[32] R. S. Williams, "Neuromorphic circuits : nonlinearity & neuristors Director , Foundational Technologies Background HP Labs Foundational Tech Group Systems Integration Group Systems Software Group photonics & nanoelectronics," 2014.

[33] A. Joubert, B. Belhadj, O. Temam, H. R, x00E, and liot, "Hardware spiking neurons design: Analog or digital?," in *The 2012 International Joint Conference on Neural Networks (IJCNN)*, 10-15 June 2012 2012, pp. 1-5, doi: 10.1109/IJCNN.2012.6252600.

[34] L. S. Smith, "Neuromorphic Systems: past, present and future," in *Brain Inspired Cognitive Systems 2008*: Springer, 2010, pp. 167-182.

[35] B. Benjamin *et al.*, "Neurogrid: A Mixed-Analog-Digital Multichip System for Large-Scale Neural Simulations," (in English), *Proceedings of the Ieee,* vol. 102, no. 5, pp. 699-716, May 2014, doi: 10.1109/Jproc.2014.2313565.

[36] K. Ramanaiah and S. Sridhar, "Hardware Implementation of Artificial Neural Networks," *i-Manager's Journal on Embedded Systems,* vol. 3, no. 4, p. 31, 2014.

[37] P. Merolla, J. Arthur, F. Akopyan, N. Imam, R. Manohar, and D. S. Modha, "A digital neurosynaptic core using embedded crossbar memory with 45pJ per spike in 45nm," in *Custom Integrated Circuits Conference (CICC), 2011 IEEE*, 2011: IEEE, pp. 1-4.

[38] B. Rajendran *et al.*, "Specifications of Nanoscale Devices and Circuits for Neuromorphic Computational Systems," *Ieee T Electron Dev,* vol. 60, no. 1, pp. 246-253, 2013, doi: 10.1109/TED.2012.2227969.

[39] N. Qiao *et al.*, "A Re-configurable On-line Learning Spiking Neuromorphic Processor comprising 256 neurons and 128K synapses," (in English), *Frontiers in Neuroscience*, Original Research vol. 9, 2015-April-29 2015, doi: 10.3389/fnins.2015.00141.

[40] K. Ovtcharov, O. Ruwase, J.-Y. Kim, J. Fowers, K. Strauss, and E. S. Chung, "Accelerating deep convolutional neural networks using specialized hardware," *Microsoft Research Whitepaper,* vol. 2, 2015.

[41] G. Indiveri, F. Corradi, and N. Qiao, "Neuromorphic Architectures for Spiking Deep Neural Networks," (in English), *2015 Ieee International Electron Devices Meeting (Iedm)*, 2015. [Online]. Available: <Go to ISI>://WOS:000380472500017.

[42] A. Putnam *et al.*, "A reconfigurable fabric for accelerating large-scale datacenter services," in *Computer Architecture (ISCA), 2014 ACM/IEEE 41st International Symposium on*, 2014: IEEE, pp. 13-24.

[43] P. A. Merolla *et al.*, "A million spiking-neuron integrated circuit with a scalable communication network and interface," (in English), *Science*, Article vol. 345, no. 6197, pp. 668-673, Aug 2014, doi: 10.1126/science.1254642.

[44] M. Hock, "Modern Semiconductor Technologies for Neuromorphic Hardware," 2014.

[45] Y. Chen *et al.*, "Dadiannao: A machine-learning supercomputer," in *Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture*, 2014: IEEE Computer Society, pp. 609-622.

[46] E. Stromatias, F. Galluppi, C. Patterson, and S. Furber, "Power analysis of large-scale, real-time neural networks on SpiNNaker," in *Neural Networks (IJCNN), The 2013 International Joint Conference on*, 2013: IEEE, pp. 1-8.

[47] E. Painkras *et al.*, "SpiNNaker: A 1-W 18-core system-on-chip for massively-parallel neural network simulation," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 8, pp. 1943-1953, 2013.

[48] S. B. Furber *et al.*, "Overview of the SpiNNaker system architecture," *Computers, IEEE Transactions on*, vol. 62, no. 12, pp. 2454-2467, 2013.

[49] S. K. Esser *et al.*, "Cognitive computing systems: Algorithms and applications for networks of neurosynaptic cores," in *Neural Networks (IJCNN), The 2013 International Joint Conference on*, 2013: IEEE, pp. 1-10.

[50] T. Pfeil *et al.*, "Six networks on a universal neuromorphic computing substrate," *arXiv preprint arXiv:1210.7083*, 2012.

[51] J.-s. Seo *et al.*, "A 45nm CMOS neuromorphic chip with a scalable architecture for learning in networks of spiking neurons," in *Custom Integrated Circuits Conference (CICC), 2011 IEEE*, 2011: IEEE, pp. 1-4.

[52] D. Brüderle *et al.*, "A comprehensive workflow for general-purpose neural modeling with highly configurable neuromorphic hardware systems," *Biological cybernetics*, vol. 104, no. 4-5, pp. 263-296, 2011.

[53] H. De Garis, C. Shuo, B. Goertzel, and L. Ruiting, "A world survey of artificial brain projects, Part I: Large-scale brain simulations," *Neurocomputing*, vol. 74, no. 1, pp. 3-29, 2010.

[54] W. Gerstner and R. Naud, "How good are neuron models?," *Science*, vol. 326, no. LCN-ARTICLE-2009-014, pp. 379-380, 2009.

[55] J. Schemmel, J. Fieres, and K. Meier, "Wafer-scale integration of analog neural networks," in *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, 2008: IEEE, pp. 431-438.

[56] M. Versace and B. Chandler, "The brain of a new machine," *IEEE spectrum*, vol. 47, no. 12, pp. 30-37, 2010.

[57] F. A. Azevedo *et al.*, "Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain," *Journal of Comparative Neurology*, vol. 513, no. 5, pp. 532-541, 2009.

[58] J. Hasler and B. Marr, "Finding a roadmap to achieve large neuromorphic hardware systems," *Front. Neurosci*, vol. 7, no. 118, p. 10.3389, 2013.

[59] B. Goertzel, R. Lian, I. Arel, H. de Garis, and S. Chen, "A world survey of artificial brain projects, Part II: Biologically inspired cognitive architectures," *Neurocomputing*, vol. 74, no. 1–3, pp. 30-49, 12// 2010, doi: http://dx.doi.org/10.1016/j.neucom.2010.08.012.

[60] J. von Neumann, "First Draft of a Report on the EDAVAC," *American Mathematical Society*, vol. 15, pp. 1-10, 1945.

[61] J. Von Neumann, *The computer and the brain*. Yale University Press, 2012.

[62] R. Preissl *et al.*, "Compass: A scalable simulator for an architecture for cognitive computing," *International Conference for High Performance Computing, Networking, Storage and Analysis, SC*, 2012, doi: 10.1109/SC.2012.34.

[63] D. Harris and S. Harris, *Digital design and computer architecture*. Elsevier, 2012.

[64] J. L. Hennessy and D. A. Patterson, *Computer architecture: a quantitative approach*. Elsevier, 2011.

[65] J. Hasler and H. B. Marr, "Finding a Roadmap to achieve Large Neuromorphic Hardware Systems," (in English), *Frontiers in Neuroscience*, Hypothesis & Theory vol. 7, 2013-September-10 2013, doi: 10.3389/fnins.2013.00118.

[66] F. Akopyan *et al.*, "True North: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip," (in English), *IEEE Trans. Comput-Aided Des. Integr. Circuits Syst.*, Article vol. 34, no. 10, pp. 1537-1557, Oct 2015, doi: 10.1109/tcad.2015.2474396.

[67] J. Schemmel, D. Bruderle, A. Grubl, M. Hock, K. Meier, and S. Millner, "A wafer-scale neuromorphic hardware system for large-scale neural modeling," in *Circuits and systems (ISCAS), proceedings of 2010 IEEE international symposium on*, 2010: IEEE, pp. 1947-1950.

[68] D. S. Modha, R. Ananthanarayanan, S. K. Esser, A. Ndirango, A. J. Sherbondy, and R. Singh, "Cognitive Computing," (in English), *Communications of the Acm*, vol. 54, no. 8, pp. 62-71, Aug 2011, doi: 10.1145/1978542.1978559.

[69] J. Simmons and R. Verderber, "New conduction and reversible memory phenomena in thin insulating films," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 301, no. 1464, pp. 77-102, 1967.

[70] F. Argall, "Switching phenomena in titanium oxide thin films," *Solid State Electronics*, vol. 11, pp. 535-541, 1968, doi: 10.1016/0038-1101(68)90092-0.

[71] C. A. Balanis, *Advanced engineering electromagnetics*. John Wiley & Sons, 2012.

[72] L. J. I. T. o. c. t. Chua, "Memristor-the missing circuit element," vol. 18, no. 5, pp. 507-519, 1971.

[73] B. Swaroop, W. West, G. Martinez, M. Kozicki, and L. Akers, "Programmable current mode Hebbian learning neural network using programmable metallization cell," in *ISCAS'98. Proceedings of the 1998 IEEE International Symposium on Circuits and Systems (Cat. No. 98CH36187)*, 1998, vol. 3: IEEE, pp. 33-36.

[74] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, "The missing memristor found," *nature*, vol. 453, no. 7191, pp. 80-83, 2008.

[75] S. R. Williams, "How we found the missing memristor," *Spectrum, IEEE*, vol. 45, no. 12, pp. 28-35, 2008.

[76] V. Keshmiri, "A Study of the Memristor Models and Applications," 2014.

[77] B. Govoreanu *et al.*, "10× 10nm 2 Hf/HfO x crossbar resistive RAM with excellent performance, reliability and low-energy operation," in *Electron Devices Meeting (IEDM), 2011 IEEE International*, 2011: IEEE, pp. 31.6. 1-31.6. 4.

[78] H. S. P. Wong *et al.*, "Metal-oxide RRAM," *Proceedings of the IEEE*, vol. 100, pp. 1951-1970, 2012, doi: 10.1109/JPROC.2012.2190369.

[79] H. An, M. S. Al-Mamun, M. K. Orlowski, and Y. Yi, "Learning Accuracy Analysis of Memristor-based Nonlinear Computing Module on Long Short-term Memory," in *Proceedings of the International Conference on Neuromorphic Systems*, 2018: ACM, p. 5.

[80] Y. Long, T. Na, and S. Mukhopadhyay, "ReRAM-based processing-in-memory architecture for recurrent neural network acceleration," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 26, no. 12, pp. 2781-2794, 2018.

[81] P.-Y. Chen, X. Peng, and S. Yu, "NeuroSim: A circuit-level macro model for benchmarking neuro-inspired architectures in online learning," *IEEE Trans. Comput-Aided Des. Integr. Circuits Syst.*, vol. 37, no. 12, pp. 3067-3080, 2018.

[82] X. Guan, S. Yu, and H.-S. P. Wong, "A SPICE Compact Model of Metal Oxide Resistive Switching Memory With Variations," *IEEE Electron Device Letters*, vol. 33, pp. 1405-1407, 2012, doi: 10.1109/LED.2012.2210856.

[83] A. Chen and M.-R. Lin, "Variability of resistive switching memories and its impact on crossbar array performance," in *2011 International Reliability Physics Symposium*, 2011: IEEE, pp. MY. 7.1-MY. 7.4.

[84] B. Liu, H. Li, Y. Chen, X. Li, Q. Wu, and T. Huang, "Vortex: variation-aware training for memristor x-bar," in *Proceedings of the 52nd Annual Design Automation Conference*, 2015: ACM, p. 15.

[85] L. Chen *et al.*, "Accelerator-friendly neural-network training: Learning variations and defects in RRAM crossbar," in *Proceedings of the Conference on Design, Automation & Test in Europe*, 2017: European Design and Automation Association, pp. 19-24.

[86] Y. Long, X. She, and S. Mukhopadhyay, "Design of reliable DNN accelerator with un-reliable ReRAM," in *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2019: IEEE, pp. 1769-1774.

[87] D. Ielmini, "Modeling the universal set/reset characteristics of bipolar RRAM by field-and temperature-driven filament growth," *Ieee T Electron Dev*, vol. 58, no. 12, pp. 4309-4317, 2011.

[88] G. Ghosh and M. K. Orlowski, "Write and erase threshold voltage interdependence in resistive switching memory cells," *IEEE transactions on Electron Devices*, vol. 62, no. 9, pp. 2850-2856, 2015.

[89] W. Wu, H. Wu, B. Gao, N. Deng, S. Yu, and H. Qian, "Improving analog switching in HfO x-based resistive memory with a thermal enhanced layer," *IEEE Electron Device Letters*, vol. 38, no. 8, pp. 1019-1022, 2017.

[90] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu, "Nanoscale memristor device as synapse in neuromorphic systems," *Nano letters*, vol. 10, no. 4, pp. 1297-1301, 2010.

[91] T. Serrano-Gotarredona, T. Masquelier, T. Prodromakis, G. Indiveri, and B. Linares-Barranco, "STDP and STDP variations with memristors for spiking neuromorphic learning systems," *Front Neurosci*, vol. 7, p. 2, 2013, doi: 10.3389/fnins.2013.00002.

[92] A. Adamatzky and L. Chua, *Memristor Networks*. Springer Science & Business Media, 2013.

[93] M. M. S. Aly *et al.*, "The N3XT approach to energy-efficient abundant-data computing," *Proceedings of the IEEE*, vol. 107, no. 1, pp. 19-48, 2019.

[94] D. Ielmini and H.-S. P. Wong, "In-memory computing with resistive switching devices," *Nature Electronics*, vol. 1, pp. 333-343, 2018, doi: 10.1038/s41928-018-0092-2.

[95] M. M. Shulaker, T. F. Wu, M. M. Sabry, H. Wei, H.-S. P. Wong, and S. Mitra, "Monolithic 3D integration: a path from concept to reality," in *Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition*, 2015: EDA Consortium, pp. 1197-1202.

[96] H. Li *et al.*, "3-D resistive memory arrays: From intrinsic switching behaviors to optimization guidelines," *Ieee T Electron Dev*, vol. 62, pp. 3160-3167, 2015, doi: 10.1109/TED.2015.2468602.

[97] M. M. Shulaker *et al.*, "Monolithic 3D integration of logic and memory: Carbon nanotube FETs, resistive RAM, and silicon FETs," in *Electron Devices Meeting (IEDM), 2014 IEEE International*, 2014: IEEE, pp. 27.4. 1-27.4. 4, doi: 10.1109/IEDM.2014.7047120.

[98] J. Sohn, S. Lee, Z. Jiang, H. Y. Chen, and H. S. P. Wong, "Atomically thin graphene plane electrode for 3D RRAM," in *2014 IEEE International Electron Devices Meeting*, 15-17 Dec. 2014 2014, pp. 5.3.1-5.3.4, doi: 10.1109/IEDM.2014.7046988.

[99] D. Kuzum, S. Yu, and H. P. Wong, "Synaptic electronics: materials, devices and applications," *Nanotechnology*, vol. 24, no. 38, p. 382001, 2013.

[100] A. Amir *et al.*, "Cognitive computing programming paradigm: A Corelet Language for composing networks of neurosynaptic cores," *Proceedings of the International Joint Conference on Neural Networks*, 2013, doi: 10.1109/IJCNN.2013.6707078.

[101] H. H.-y. Chen and H. P. Wong, "3D Vertical RRAM," *Flash Memory Summit 2013*, pp. 1-27, 2013.

[102] M. S. Al-Mamun, "Impact of Inert-electrode on the Performance and Electro-thermal Reliability of ReRAM Memory Array," Virginia Tech, 2019.

[103] G. Stefanovich, A. Pergament, and D. Stefanovich, "Electrical switching and Mott transition in VO2," *Journal of Physics: Condensed Matter*, vol. 12, no. 41, p. 8837, 2000.

[104] J. Honig and T. Reed, "Electrical properties of Ti 2 O 3 single crystals," *Physical Review*, vol. 174, no. 3, p. 1020, 1968.

[105] M. Janousch, G. I. Meijer, U. Staub, B. Delley, S. F. Karg, and B. P. Andreasson, "Role of oxygen vacancies in Cr-doped SrTiO3 for resistance-change memory," *Advanced materials*, vol. 19, no. 17, pp. 2232-2235, 2007.

[106] G.-S. Park, X.-S. Li, D.-C. Kim, R.-J. Jung, M.-J. Lee, and S. Seo, "Observation of electric-field induced Ni filament channels in polycrystalline Ni O x film," vol. 91, no. 22, p. 222103, 2007.

[107] M. Al-Mamun, S. W. King, S. Meda, and M. K. Orlowski, "Impact of the Heat Conductivity of the Inert Electrode on ReRAM Performance and Endurance," *ECS Transactions*, vol. 85, no. 8, pp. 207-212, 2018.

[108] M. Al-Mamun, S. W. King, and M. K. Orlowski, "Impact of the Heat Conductivity of the Inert Electrode on Reram Memory Cell Performance and Endurance," in *Meeting Abstracts*, 2018, no. 24: The Electrochemical Society, pp. 1476-1476.

[109] Y. Fan, M. Al-Mamun, B. Conlon, S. W. King, and M. K. Orlowski, "Resistive Switching Comparison between Cu/TaOx/Ru and Cu/TaOx/Pt Memory Cells," *ECS Transactions*, vol. 75, no. 32, p. 13, 2017.

[110] C. Walczyk *et al.*, "Impact of Temperature on the Resistive Switching Behavior of Embedded HfO_2-Based RRAM Devices," *IEEE transactions on electron devices*, vol. 58, no. 9, pp. 3124-3131, 2011.

[111] C. D. Landon *et al.*, "Thermal transport in tantalum oxide films for memristive applications," *Applied Physics Letters*, vol. 107, no. 2, p. 023108, 2015.

[112] P. Sun *et al.*, "Physical model of dynamic Joule heating effect for reset process in conductive-bridge random access memory," *Journal of Computational Electronics*, vol. 13, no. 2, pp. 432-438, 2014.

[113] S. Kaeriyama *et al.*, "A nonvolatile programmable solid-electrolyte nanometer switch," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 1, pp. 168-176, 2005.

[114] H. An, M. A. Ehsan, Z. Zhou, and Y. Yi, "Electrical modeling and analysis of 3D synaptic array using vertical RRAM structure," in *Quality Electronic Design (ISQED), 2017 18th International Symposium on*, 2017: IEEE, pp. 1-6.

[115] Z. Jiang, S. Yu, Y. Wu, J. H. Engel, X. Guan, and H. S. P. Wong, "Verilog-A Compact Model for Oxide-based Resistive Random Access Memory (RRAM)," pp. 41-44, 2014.

[116] H. Li, P. Huang, B. Gao, B. Chen, X. Liu, and J. Kang, "A SPICE model of resistive random access memory for large-scale memory array simulation," *IEEE Electron Device Letters*, vol. 35, pp. 211-213, 2014, doi: 10.1109/LED.2013.2293354.

[117] H. Li, P. Huang, B. Gao, B. Chen, X. Liu, and J. Kang, "A SPICE model of resistive random access memory for large-scale memory array simulation," *IEEE Electron Device Letters*, vol. 35, no. 2, pp. 211-213, 2014.

[118] P. Huang *et al.*, "A physical based analytic model of RRAM operation for circuit simulation," *Technical Digest - International Electron Devices Meeting, IEDM*, vol. 5, pp. 605-608, 2012, doi: 10.1109/IEDM.2012.6479110.

[119] K. Bai, Q. An, and Y. Yi, "Deep-DFR: A Memristive Deep Delayed Feedback Reservoir Computing System with Hybrid Neural Network Topology," *56th ACM/ESDA/IEEE Design Automation Conference (DAC) IEEE*, 2019.

[120] C. Gallicchio, A. Micheli, and L. Pedrelli, "Deep reservoir computing: A critical experimental analysis," *Neurocomputing*, vol. 268, pp. 87-99, 2017, doi: 10.1016/j.neucom.2016.12.089.

[121] J. Li, K. Bai, L. Liu, and Y. Yi. A Deep Learning Based Approach for Analog Hardware Implementation of Delayed Feedback Reservoir Computing System.

[122] K. Bai, J. Li, K. Hamedani, and Y. Yi. Enabling An New Era of Brain-inspired Computing: Energy-efficient Spiking Neural Network with Ring Topology.

[123] K. Bai and Y. Yi, "DFR: An Energy-efficient Analog Delay Feedback Reservoir Computing System for Brain-inspired Computing," *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 14, no. 4, p. 45, 2018.

[124] K. Bai, Q. An, and Y. Yi, "Deep-DFR: A Memristive Deep Delayed Feedback Reservoir Computing System with Hybrid Neural Network Topology," in *Proceedings of the 56th Annual Design Automation Conference 2019*, 2019: ACM, p. 54.

[125] K. Bai and Y. Yi, "Opening the “Black Box” of Silicon Chip Design in Neuromorphic Computing," in *Bio-Inspired Technology*: IntechOpen, 2019.

[126] L. Appeltant *et al.*, "Information processing using a single dynamical node as complex system," *Nat Commun*, vol. 2, p. 468, 2011.

[127] L. Appeltant, G. Van der Sande, J. Danckaert, and I. Fischer, "Constructing optimized binary masks for reservoir computing with delay systems," *Scientific reports*, vol. 4, p. 3629, 2014.

[128] K. Hamedani, L. Liu, S. Hu, J. Ashdown, J. Wu, and Y. Yi, "Detecting Dynamic Attacks in Smart Grids Using Reservoir Computing: A Spiking Delayed Feedback Reservoir Based Approach," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2019.

[129] S. Ortín and L. Pesquera, "Reservoir Computing with an Ensemble of Time-Delay Reservoirs," *Cognitive Computation,* vol. 9, no. 3, pp. 327-336, 2017, doi: 10.1007/s12559-017-9463-7.

[130] S. Mosleh, L. Liu, C. Sahin, Y. R. Zheng, and Y. Yi, "Brain-Inspired Wireless Communications: Where Reservoir Computing Meets MIMO-OFDM," in *IEEE Transactions on Neural Networks and Learning Systems,* 2017.

[131] L. Keuninckx, J. Danckaert, and G. Van der Sande, "Real-time Audio Processing with a Cascade of Discrete-Time Delay Line-Based Reservoir Computers," *Cognitive Computation,* vol. 9, pp. 315-326, 2017, doi: 10.1007/s12559-017-9457-5.

[132] K. Hamedani, L. Liu, A. Rachad, J. Wu, and Y. Yi, "Reservoir Computing Meets Smart Grids: Attack Detection using Delayed Feedback Networks," in *IEEE Transactions on Industrial Informatics,* 2017.

[133] N. D. Lane, P. Georgiev, and L. Qendro, "DeepEar: robust smartphone audio sensing in unconstrained acoustic environments using deep learning," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2015, pp. 283-294.

[134] S. Choi *et al.*, "SiGe epitaxial memory for neuromorphic computing with reproducible high performance based on engineered dislocations," *Nature materials,* vol. 17, no. 4, pp. 335-340, 2018.

[135] J. Woo *et al.*, "Improved synaptic behavior under identical pulses using AlOx/HfO2 bilayer RRAM array for neuromorphic systems," *IEEE Electron Device Letters,* vol. 37, no. 8, pp. 994-997, 2016.

[136] P.-Y. Chen and S. Yu, "Technological Benchmark of Analog Synaptic Devices for Neuroinspired Architectures," *IEEE Design & Test,* vol. 36, no. 3, pp. 31-38, 2018.

[137] W. Severa, C. M. Vineyard, R. Dellana, S. J. Verzi, and J. B. Aimone, "Training deep neural networks for binary communication with the Whetstone method," *Nature Machine Intelligence,* vol. 1, no. 2, p. 86, 2019.

[138] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature,* vol. 521, pp. 436-444, 2015, doi: 10.1038/nature14539.

[139] I. Goodfellow, B. Yoshua, and C. Aaron, "Deep Learning," *Deep Learning,* p. 785, 2016, doi: 10.1016/B978-0-12-391420-0.09987-X.

[140] F. Akopyan *et al.*, "Truenorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* vol. 34, no. 10, pp. 1537-1557, 2015.

[141] J. Woo *et al.*, "Improved synaptic behavior under identical pulses using AlO x/HfO 2 bilayer RRAM array for neuromorphic systems," *IEEE Electron Device Letters,* vol. 37, no. 8, pp. 994-997, 2016.

[142] S. Panth, S. Samal, Y. S. Yu, and S. K. Lim, "Design challenges and solutions for ultra-high-density monolithic 3D ICs," in *2014 SOI-3D-Subthreshold Microelectronics*

Technology Unified Conference (S3S), 6-9 Oct. 2014 2014, pp. 1-2, doi: 10.1109/S3S.2014.7028195.

[143] M. A. Ehsan, H. An, Z. Zhou, and Y. Yi, "A Novel Approach for using TSVs as Membrane Capacitance in Neuromorphic 3D IC," *IEEE Trans. Comput-Aided Des. Integr. Circuits Syst.*, vol. 0070, 2017, doi: 10.1109/TCAD.2017.2760506.

[144] M. A. Ehsan, H. An, Z. Zhou, and Y. Yi. Adaptation of enhanced TSV capacitance as membrane property in 3D brain-inspired computing system.

[145] M. A. Ehsan, Z. Zhou, and Y. Yi, "Modeling and optimization of TSV for crosstalk mitigation in 3D neuromorphic system," in *Electromagnetic Compatibility (EMC), 2016 IEEE International Symposium on*, 2016: IEEE, pp. 621-626.

[146] M. A. Ehsan, Z. Zhou, and Y. Yi, "Analytical Modeling and Analysis of through Silicon Vias (Tsvs) in High Speed Three-Dimensional System Integration," *Progress In Electromagnetics Research M*, vol. 42, pp. 49-59, 2015, doi: 10.2528/pierm15021404.

[147] M. A. Ehsan, Z. Zhou, L. Liu, and Y. Yi, "An analytical through silicon via (TSV) surface roughness model applied to a millimeter wave 3-D IC," *IEEE Transactions on Electromagnetic Compatibility*, vol. 57, no. 4, pp. 815-826, 2015.

[148] A. Ehsan, Z. Zhou, and Y. Yi, "Analytical Modeling and Analysis of Through Silicon Vias (TSVs) in High Speed Three-Dimensional System Integration," *Progress in Electromagnetics Research Journal (PIER)*, vol. 42, pp. 49-59, 2015.

[149] M. Amimul Ehsan, Z. Zhou, L. Liu, and Y. Yi, "An Analytical Through Silicon Via (TSV) Surface Roughness Model Applied to a Millimeter Wave 3-D IC," *IEEE Transactions on Electromagnetic Compatibility*, vol. 57, pp. 815-826, 2015, doi: 10.1109/TEMC.2015.2408262.

[150] T. Song *et al.*, "Full-chip multiple TSV-to-TSV coupling extraction and optimization in 3D ICs," *Proceedings of the 50th Annual Design Automation Conference on - DAC '13*, pp. 1-7, 2013, doi: 10.1145/2463209.2488956.

[151] I. Savidis, "Characterization and Modeling of TSV Based 3-D Integrated Circuits," p. 443, 2013.

[152] Y. Peng, T. Song, D. Petranovic, and S. K. Lim, "On accurate full-chip extraction and optimization of TSV-to-TSV coupling elements in 3D ICs," *IEEE/ACM International Conference on Computer-Aided Design, Digest of Technical Papers, ICCAD*, pp. 281-288, 2013, doi: 10.1109/ICCAD.2013.6691133.

[153] R. Wang, G. Charles, and P. Franzon, "Modeling and compare of through-silicon-via (TSV) in high frequency," *2011 IEEE International 3D Systems Integration Conference, 3DIC 2011*, pp. 3-8, 2011, doi: 10.1109/3DIC.2012.6263048.

[154] B. Rajendran, A. K. Henning, B. Cronquist, and Z. Or-Bach, "Pulsed laser annealing: A scalable and practical technology for monolithic 3D IC," in *3DIC*, 2013, pp. 1-5.

[155] S. K. Lim and Y. Shi, "Design Challenges and Solutions for Monolithic 3D ICs," in *Physical Design for 3D Integrated Circuits*: CRC Press, 2015, pp. 321-335.

[156] M. Swaminathan, "Electrical design and modeling challenges for 3D system integration," in *Design Conference 2012*, 2012.

[157] C. Su, H. Ke, and T. Hubing, "Overview of electromagnetic modeling software," in *Proc. of the 25th International Review of Progress in Applied Computational Electromagnetics*, 2009, pp. 736-741.

[158] R. J. Baker, *CMOS: circuit design, layout, and simulation*. John Wiley & Sons, 2019.

[159] P. Lin *et al.*, "Three-dimensional memristor circuits as complex neural networks," *Nature Electronics*, vol. 3, no. 4, pp. 225-232, 2020.

[160] Y. C. Chen, H. Li, W. Zhang, and R. E. Pino, "3D-HIM: A 3D High-density interleaved memory for bipolar RRAM design," *Proceedings of the 2011 IEEE/ACM International Symposium on Nanoscale Architectures, NANOARCH 2011*, vol. 1, pp. 59-64, 2011, doi: 10.1109/NANOARCH.2011.5941484.

[161] M. A. Ehsan, Z. Zhou, and Y. Yi, "Three Dimensional Integration Technology Applied to Neuromorphic Hardware Implementation," *Proceedings - 2015 IEEE International Symposium on Nanoelectronic and Information Systems, iNIS 2015*, pp. 203-206, 2016, doi: 10.1109/iNIS.2015.72.

[162] M. M. Shulaker *et al.*, "Three-dimensional integration of nanotechnologies for computing and data storage on a single chip," *Nature*, vol. 547, no. 7661, pp. 74-78, Jul 5 2017, doi: 10.1038/nature22994.

[163] M. M. S. Aly *et al.*, "Energy-efficient abundant-data computing: The N3XT 1,000," *Computer*, vol. 48, pp. 24-33, 2015, doi: 10.1063/1.4913279.

[164] M. M. Shulaker *et al.*, "Monolithic 3D integration of logic and memory: Carbon nanotube FETs, resistive RAM, and silicon FETs," *Technical Digest - International Electron Devices Meeting, IEDM*, vol. 2015-Febru, pp. 27.4.1-27.4.4, 2015, doi: 10.1109/IEDM.2014.7047120.

[165] P. I. Pavlov, "Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex," *Annals of neurosciences*, vol. 17, no. 3, p. 136, Jul 2010, doi: 10.5214/ans.0972-7531.1017309.

[166] C. Eliasmith *et al.*, "A large-scale model of the functioning brain," *Science*, vol. 338, no. 6111, pp. 1202-1205, 2012.

[167] Y. V. Pershin and M. Di Ventra, "Experimental demonstration of associative memory with memristive neural networks," *Neural Networks*, vol. 23, no. 7, pp. 881-886, 2010.

[168] A. W. Roe, S. L. Pallas, Y. H. Kwon, and M. Sur, "Visual projections routed to the auditory pathway in ferrets: receptive fields of visual neurons in primary auditory cortex," *Journal of Neuroscience*, vol. 12, no. 9, pp. 3651-3664, 1992.

[169] D. Kuzum, R. G. Jeyasingh, and H.-S. P. Wong, "Energy efficient programming of nanoelectronic synaptic devices for large-scale implementation of associative and temporal sequence learning," in *Electron Devices Meeting (IEDM), 2011 IEEE International*, 2011: IEEE, pp. 30.3. 1-30.3. 4.

[170] M. Ziegler *et al.*, "An electronic version of Pavlov's dog," *Adv Funct Mater,* vol. 22, no. 13, pp. 2744-2749, 2012.

[171] S. B. Eryilmaz *et al.*, "Brain-like associative learning using a nanoscale non-volatile phase change synaptic device array," *Frontiers in neuroscience,* vol. 8, 2014.

[172] K. Moon *et al.*, "Hardware implementation of associative memory characteristics with analogue-type resistive-switching device," *Nanotechnology,* vol. 25, no. 49, p. 495204, 2014.

[173] X. Liu, Z. Zeng, and S. Wen, "Implementation of memristive neural network with full-function pavlov associative memory," *IEEE Transactions on Circuits and Systems I: Regular Papers,* vol. 63, no. 9, pp. 1454-1463, 2016.

[174] X. Hu, S. Duan, G. Chen, and L. Chen, "Modeling affections with memristor-based associative memory neural networks," *Neurocomputing,* vol. 223, pp. 129-137, 2017, doi: 10.1016/j.neucom.2016.10.028.

[175] J. Yang, L. Wang, Y. Wang, and T. Guo, "A novel memristive Hopfield neural network with application in associative memory," *Neurocomputing,* vol. 227, pp. 142-148, 2017, doi: 10.1016/j.neucom.2016.07.065.

[176] H. An, Z. Zhou, and Y. Yi, "Opportunities and challenges on nanoscale 3D neuromorphic computing system," in *Electromagnetic Compatibility & Signal/Power Integrity (EMCSI), 2017 IEEE International Symposium on,* 2017: IEEE, pp. 416-421.

[177] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature,* vol. 521, no. 7553, pp. 436-44, May 28 2015, doi: 10.1038/nature14539.

[178] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks,* vol. 3361, no. 10, p. 1995, 1995.

[179] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on,* 2013: IEEE, pp. 6645-6649.

[180] L. F. Abbott, "Lapicque's introduction of the integrate-and-fire model neuron (1907)," *Brain research bulletin,* vol. 50, no. 5, pp. 303-304, 1999.

[181] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *The Journal of physiology,* vol. 117, no. 4, p. 500, 1952.

[182] R. B. Stein, "A theoretical analysis of neuronal variability," *Biophysical Journal,* vol. 5, no. 2, p. 173, 1965.

[183] E. M. Izhikevich, "Simple model of spiking neurons," *IEEE Transactions on neural networks,* vol. 14, no. 6, pp. 1569-1572, 2003.

[184] I. Darian-Smith, K. Johnson, and R. Dykes, "" Cold" fiber population innervating palmar and digital skin of the monkey: responses to cooling pulses," *Journal of Neurophysiology,* vol. 36, no. 2, pp. 325-346, 1973.

[185] H. An, M. A. Ehsan, Z. Zhou, F. Shen, and Y. Yi, "Monolithic 3D neuromorphic computing system with hybrid CMOS and memristor-based synapses and neurons," *Integration, the VLSI Journal*, 2017.

[186] H. Lim *et al.*, "Relaxation oscillator-realized artificial electronic neurons, their responses, and noise," *Nanoscale*, vol. 8, no. 18, pp. 9629-9640, 2016.

[187] S. Dutta *et al.*, "Dynamics, Design, and Application of a Silicon-on-Insulator Technology Based Neuron," *MRS Advances*, vol. 3, no. 57-58, pp. 3347-3357, 2018, doi: 10.1557/adv.2018.490.

[188] C. D. Schuman *et al.*, "A survey of neuromorphic computing and neural networks in hardware," 2017.

[189] M. A. Ehsan, H. An, Z. Zhou, and Y. Yi, "Design challenges and methodologies in 3D integration for neuromorphic computing systems," in *Quality Electronic Design (ISQED), 2016 17th International Symposium on*, 2016: IEEE, pp. 24-28.

[190] H. An, M. A. Ehsan, Z. Zhou, and Y. Yi, "Electrical modeling and analysis of 3D Neuromorphic IC with Monolithic Inter-tier Vias," in *Electrical Performance Of Electronic Packaging And Systems (EPEPS), 2016 IEEE 25th Conference on*, 2016: IEEE, pp. 87-90.

[191] C. Zhao, J. li, H. An, and Y. Yi, "When Energy Efficient Spike-Based Temporal Encoding Meets Resistive Crossbar: From Circuit Design to Application," *Proceedings of Neuromorphic Computing Symposium*, pp. 1-6, July 17-19, 2017 2017.

[192] C. Zhao, J. Li, H. An, and Y. Yi, "Energy efficient analog spiking temporal encoder with verification and recovery scheme for neuromorphic computing systems," in *Quality Electronic Design (ISQED), 2017 18th International Symposium on*, 2017: IEEE, pp. 138-143.

[193] M. A. Ehsan, H. An, Z. Zhou, and Y. Yi, "Adaptation of Enhanced TSV Capacitance as Membrane Property in 3D Brain-inspired Computing System," in *Proceedings of the 54th Annual Design Automation Conference 2017*, 2017: ACM, p. 86.

[194] C. Zhao, B. T. Wysocki, Y. Liu, C. D. Thiem, N. R. McDonald, and Y. Yi, "Spike-time-dependent encoding for neuromorphic processors," *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 12, no. 3, p. 23, 2015.

[195] C. Zhao, J. Li, and Y. Yi, "Making neural encoding robust and energy efficient: an advanced analog temporal encoder for brain-inspired computing systems," in *Proceedings of the 35th International Conference on Computer-Aided Design*, 2016, pp. 1-6.

[196] C. Zhao, J. Liu, F. Shen, and Y. Yi, "Low power CMOS power amplifier design for RFID and the Internet of Things," *Computers & Electrical Engineering*, vol. 52, pp. 157-170, 2016.

[197] C. Zhao *et al.*, "Energy efficient spiking temporal encoder design for neuromorphic computing systems," *IEEE Transactions on Multi-Scale Computing Systems*, vol. 2, no. 4, pp. 265-276, 2016.

[198] C. Zhao, K. Hamedani, J. Li, and Y. Yi, "Analog Spike-timing-dependent Resistive Crossbar Design for Brain Inspired Computing," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems,* pp. 1-1, 2017, doi: 10.1109/jetcas.2017.2765892.

[199] C. Zhao, Y. Yi, J. Li, X. Fu, and L. Liu, "Interspike-interval-based analog spike-time-dependent encoder for neuromorphic processors," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* vol. 25, no. 8, pp. 2193-2205, 2017.

[200] B. Prince, *Vertical 3D memory technologies*. John Wiley & Sons, 2014.

[201] M. D. Pickett and R. S. Williams, "Phase transitions enable computational universality in neuristor-based cellular automata," *Nanotechnology,* vol. 24, no. 38, p. 384002, 2013.

[202] M. D. Pickett and R. S. Williams, "Sub-100 fJ and sub-nanosecond thermally driven threshold switching in niobium oxide crosspoint nanodevices," *Nanotechnology,* vol. 23, no. 21, p. 215202, 2012.

[203] R. J. Baker, *CMOS: circuit design, layout, and simulation*. John Wiley & Sons, 2008.

[204] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," (in English), *Proceedings of the Ieee,* vol. 86, no. 11, pp. 2278-2324, Nov 1998, doi: Doi 10.1109/5.726791.

[205] P. J. Warden, "Speech Commands: A public dataset for single-word speech recognition," vol. 1, 2017.

[206] H.-S. P. Wong *et al.*, "Metal–oxide RRAM," *Proceedings of the IEEE,* vol. 100, no. 6, pp. 1951-1970, 2012.

[207] H. An, Z. Zhou, and Y. Yi, "Memristor-based 3D neuromorphic computing system and its application to associative memory learning," in *2017 IEEE 17th International Conference on Nanotechnology (IEEE-NANO),* 25-28 July 2017 2017, pp. 555-560, doi: 10.1109/NANO.2017.8117459.

[208] B. Hudec *et al.*, "3D resistive RAM cell design for high-density storage class memory—a review," *Science China Information Sciences,* vol. 59, no. 6, 2016, doi: 10.1007/s11432-016-5566-0.

[209] H. An, Q. An, and Y. Yi, "Realizing Behavior Level Associative Memory Learning Through Three-Dimensional Memristor-Based Neuromorphic Circuits," *IEEE Transactions on Emerging Topics in Computational Intelligence,* 2019.

[210] H. An, Z. Zhou, and Y. Yi, "3D memristor-based adjustable deep recurrent neural network with programmable attention mechanism," in *Proceedings of the Neuromorphic Computing Symposium*, 2017: ACM, p. 11.

[211] T. Stock and G. Seliger, "Opportunities of sustainable manufacturing in industry 4.0," *Procedia Cirp,* vol. 40, pp. 536-541, 2016.

[212] N. Gronau, M. Grum, and B. Bender, "Determining the optimal level of autonomy in cyber-physical production systems," in *2016 IEEE 14th International Conference on Industrial Informatics (INDIN)*, 2016: IEEE, pp. 1293-1299.

[213] H. An, J. Li, Y. Li, X. Fu, and Y. Yi, "Three dimensional memristor-based neuromorphic computing system and its application to cloud robotics," *Computers & Electrical Engineering*, vol. 63, pp. 99-113, 2017, doi: 10.1016/j.compeleceng.2017.06.023.

[214] J. L. Krichmar and H. Wagatsuma, *Neuromorphic and brain-based robots*. Cambridge University Press, 2011.

[215] S. Menon, S. Fok, A. Neckar, O. Khatib, and K. Boahen, "Controlling articulated robots in task-space with spiking silicon neurons," in *5th IEEE RAS/EMBS International Conference on Biomedical Robotics and Biomechatronics*, 12-15 Aug. 2014 2014, pp. 181-186, doi: 10.1109/BIOROB.2014.6913773.

[216] S. Davies, C. Patterson, F. Galluppi, A. Rast, D. Lester, and S. B. Furber, "Interfacing real-time spiking I/O with the SpiNNaker neuromimetic architecture," in *Proceedings of the 17th International Conference on Neural Information Processing: Australian Journal of Intelligent Information Processing Systems*, 2010: Citeseer, pp. 7-11.

[217] C. Denk, F. Llobet-Blandino, F. Galluppi, L. A. Plana, S. Furber, and J. Conradt, "Real-time interface board for closed-loop robotic tasks on the spinnaker neural computing system," in *International Conference on Artificial Neural Networks*, 2013: Springer, pp. 467-474.

[218] H. J. Chiel and R. D. Beer, "The brain has a body: adaptive behavior emerges from interactions of nervous system, body and environment," *Trends in neurosciences*, vol. 20, no. 12, pp. 553-557, 1997.

[219] E. J. Izquierdo and R. D. Beer, "Connecting a connectome to behavior: an ensemble of neuroanatomical models of C. elegans klinotaxis," *PLoS Comput Biol*, vol. 9, no. 2, p. e1002890, 2013.

[220] M. Suzuki, D. Floreano, and E. A. Di Paolo, "The contribution of active body movement to visual development in evolutionary robots," *Neural Networks*, vol. 18, no. 5, pp. 656-665, 2005.

[221] N. Franceschini, J.-M. Pichon, C. Blanes, and J. Brady, "From insect vision to robot vision [and discussion]," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 337, no. 1281, pp. 283-294, 1992.

[222] M. Sheng, B. L. Sabatini, and T. C. Südhof, "Synapses and Alzheimer's disease," *Cold Spring Harbor perspectives in biology*, vol. 4, no. 5, p. a005777, 2012.

[223] E. M. Izhikevich, "Dynamical Systems in Neuroscience Computational Neuroscience," *Dynamical Systems*, vol. 25, pp. 227-256, 2007, doi: 10.1017/S0143385704000173.

[224] W. Gerstner, H. Sprekeler, and G. Deco, "Review Theory and Simulation in Neuroscience," pp. 60-66, 2012, doi: 10.1126/science.1227356.

Publications

- **Refereed Book Chapters**
  1. **H. An**, K. Bai, and Y. Yi, "The Roadmap to Realize Memristive Three-Dimensional Neuromorphic Computing System," in *Advances in Memristor Neural Networks-Modeling and Applications,* IntechOpen, 2018.
- **Refereed Journal Articles**
  1. **H. An**, Mohammad Shah Al-Mamun, Marius K. Orlowski, Lingjia Liu, Yang Yi, "Three-dimensional Neuromorphic Computing System with Two-layer and Low-variation Memristive Synapses," IEEE Transactions on Computing Aided Design of Integrated Circuits and Systems (TCAD), submitted
  2. **H. An**, Mohammad Shah Al-Mamun, Marius K. Orlowski, Lingjia Liu, Yang Yi, "Robust Deep Reservoir Computing through Reliable Memristor with Improved Heat Dissipation Capability," IEEE Transactions on Computing Aided Design of Integrated Circuits and Systems (TCAD), 2020.
  3. **H. An**, Q. An, Yang Yi, "Realizing Behavior Level Associative Memory Learning Through Three-Dimensional Memristor-Based Neuromorphic Circuits," IEEE Transactions on Emerging Topics in Computational Intelligence (TETCI). 2019 Jul 2.
  4. **H. An**, M. A. Ehsan, Z. Zhou, F. Shen, and Y. Yi, "Monolithic 3D neuromorphic computing system with hybrid CMOS and memristor-based synapses and neurons," *Integration, the VLSI Journal,* 2019.
  5. M. A. Ehsan, **H. An**, Z. Zhou, and Y. Yi, "A Novel Approach for using TSVs as Membrane Capacitance in Neuromorphic 3D IC," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), 2017.
  6. **H. An**, J. Li, Y. Li, X. Fu, and Y. Yi, "Three-dimensional memristor-based neuromorphic computing system and its application to cloud robotics," *Computers & Electrical Engineering,* vol. 63, pp. 99-113, 2017.
- **Refereed Conference Proceedings**
  1. **H. An**, Dong Ha, Yang Yi, "Powering Next-generation Industry 4.0 by a Self-learning and Low-power Neuromorphic System", ACM International Conference on Nanoscale Computing and Communication, 2020, in publishing
  2. **H. An**, M. S. Al-Mamun, M. K. Orlowski, and Y. Yi, "Learning Accuracy Analysis of Memristor-based Nonlinear Computing Module on Long Short-term Memory," in *Proceedings of the International Conference on Neuromorphic Systems*, 2018, p. 5.
  3. M. A. Ehsan, **H. An**, Z. Zhou, and Y. Yi, "Adaptation of Enhanced TSV Capacitance as Membrane Property in 3D Brain-inspired Computing System," in Proceedings of the 54th Annual Design Automation Conference (DAC) 2017, 2017, p. 86.

4. **H. An**, M. A. Ehsan, Z. Zhou, and Y. Yi, "Electrical modeling and analysis of 3D synaptic array using vertical RRAM structure," in Quality Electronic Design (ISQED), 2017 18th International Symposium on, 2017, pp. 1-6. (**Best Paper Nomination**)
5. C. Zhao, J. Li, **H. An**, and Y. Yi, "Energy-efficient analog spiking temporal encoder with verification and recovery scheme for neuromorphic computing systems," in Quality Electronic Design (ISQED), 2017 18th International Symposium on, 2017, pp. 138-143.
6. **H. An**, Z. Zhou, and Y. Yi, "Memristor-based 3D neuromorphic computing system and its application to associative memory learning," in 2017 IEEE 17th International Conference on Nanotechnology (IEEE-NANO), 2017, pp. 555-560. (**NSF Student Travel Fellowship Reward**)
7. **H. An**, Z. Zhou, and Y. Yi, "Opportunities and challenges on nanoscale 3D neuromorphic computing system," in Electromagnetic Compatibility & Signal/Power Integrity (EMCSI), 2017 IEEE International Symposium on, 2017, pp. 416-421.
8. **H. An**, Z. Zhou, and Y. Yi, "3D Memristor-based Adjustable Deep Recurrent Neural Network with Programmable Attention Mechanism," in Proceedings of Neuromorphic Computing Symposium (NCS), 2017.
9. C. Zhao, J. Li, **H. An**, and Y. Yi, "When Energy Efficient Spike-Based Temporal Encoding Meets Resistive Crossbar: From Circuit Design to Application," *in Proceedings of Neuromorphic Computing Symposium, 2017.*
10. **H. An**, M. A. Ehsan, Z. Zhou, and Y. Yi, "Electrical modeling and analysis of 3D Neuromorphic IC with Monolithic Inter-tier Vias," in *Electrical Performance of Electronic Packaging and Systems (EPEPS), 2016 IEEE 25th Conference on, 2016, pp. 87-90.*
11. M. A. Ehsan, **H. An**, Z. Zhou, and Y. Yi, "Design challenges and methodologies in 3D integration for neuromorphic computing systems," in *Quality Electronic Design (ISQED), 2016 17th International Symposium on, 2016, pp. 24-28.*