CMOS and Memristor Technologies for Neuromorphic Computing Applications

Jeff Sun

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2015-219
http://www.eecs.berkeley.edu/Pubs/TechRpts/2015/EECS-2015-219.html

December 1, 2015

Copyright © 2015, by the author(s).
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

CMOS and Memristor Technologies for Neuromorphic Computing Applications

by Jeff K.Y. Sun

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:

Professor Tsu-Jae King Liu
Research Advisor

(Date)

* * * * * * *

Professor Vladimir Stojanovic
Second Reader

(Date)

Abstract

In this work, I present a CMOS implementation of a neuromorphic system that aims to mimic the behavior of biological neurons and synapses in the human brain. The synapse is modeled with a memristor-resistor voltage divider, while the neuron-emulating circuit ("CMOS Neuron") comprises transistors and capacitors. The input aggregation and output firing characteristics of a CMOS Neuron are based on observations from studies in neuroscience, and achieved using both analog and digital circuit design principles. The important Spike Timing Dependent Plasticity (STDP) learning scheme is explored in detail, and a simple adaptive learning experiment is performed to demonstrate the CMOS Neuron's potential for future artificial intelligence applications.

Contents

Acknowledgements

I would like to first express my sincere gratitude to my research advisor Professor Tsu-Jae King Liu for her guidance and support throughout my undergraduate and graduate years at UC Berkeley. Her EE130 (Integrated-Circuit Devices) course was the first upper division EE course I took as an undergrad, and taught me about CMOS technologies that allowed me to venture into my area of research. She is my favorite Professor who's taught me at Cal, and has always provided me with encouragement and feedback that allowed me to succeed in my research. Her insight and ideas allowed me to make progress in weeks that would normally take months, giving me the chance to learn as much about Neuromorphic Computing as I possibly could over the past two years.

I am also grateful to Dr. Nuo Xu for advising me on my research throughout my time at Berkeley, serving as my mentor, recognizing my work, and being a knowledgeable friend I could always rely on. Thank you for giving me the opportunity to help you with your research as an undergrad, and thank you for believing in my abilities.

I would like to thank all the amazing Professors I've had at Berkeley from the MSE, CS, and EE departments for helping me become a multi-disciplined student. Thank you Professor Ronald Gronsky for teaching me three different courses, Prof. Daryl Chrzan, Prof. A.M. Glaeser, Dr. Chris Kumai, Prof. Dan Garcia, Prof. Jonathan Shewchuk, Prof. John Wawrzynek, Prof. Michel Maharbiz, Prof. Sayeef Salahuddin, Prof. Laura Waller, Prof. Clark Nguyen, Prof. Connie Chang-Hasnain, and Prof. Jaijeet Roychowdhury. I would also like to especially thank Professor Vladimir Stojanovic for reading this project report, as well as teaching me EE241A (Intro to Digital Integrated Circuits) to give me insight on circuit design.

It has been a pleasure to meet and work with everyone in King Group, and special thanks to Dr. I-Ru "Tim" Chen for helping me with computer and office setup, Yenhao "Phillip" Chen for lending me his account when I didn't have one to perform simulations, Peng Zheng for teaching me in EE105, and Chuang Qian for teaching me how to use probe stations for device characterization. I would like to thank all my old friends I've continued to keep in touch with from Vancouver, new ones I've met since coming over to the bay area, and Jason Lau for being the best friend anyone can ask for over the past ten years.

Finally, I would like to thank my parents, Echo and Peter for their dedication and frequent visits to Berkeley over the past five years. Everything would not have been possible without their support for my education, and trust in giving me the freedom to pursue my passions.

Chapter 1. Introduction

Improvements in integrated circuit (IC) performance, cost and energy efficiency have driven semiconductor manufacturing technology advancement to enable ever more functional electronic devices. Due to fundamental limitations in the energy efficiency of conventional complementary metal-oxide-semiconductor (CMOS) digital logic circuits [1], however, alternative computer architectures (*i.e.* other than the von Neumann architecture) eventually will be needed. Among these, a promising candidate is the artificial neural network and associated neuromorphic computing principles. The motivation for developing neuromorphic systems is that the human brain is capable of processing information and performing a wide variety of functions while consuming only ~20 W of power [2]. Particularly for noise reduction, pattern recognition and image detection applications, neuromorphic computing can be superior in performance [3]-[5]. Therefore, there is tremendous interest to develop information processing systems that embody analogues of the physical processes of neural computation [6].

A human brain comprises ~10^{11} neurons interconnected with ~10^{15} synapses [7]; the neurons function as signal-processing units, while the synapses act as signal transmission units and also store information. Each neuron communicates with ~10^{4} other neurons via action potentials, allowing the brain to excel at computations that are parallel in nature. Previous attempts at developing neuromorphic systems exhibiting such characteristics have relied on analog CMOS circuits to implement neuron behavior, and SRAM cells to implement the synapses [8]-[10]. With the recent development of the memristor [11], which is also known as resistive random-access memory (RRAM), more compact implementations of a synaptic connection have been proposed for reductions in power consumption and area, as compared against CMOS transistor implementations [12].

This report outlines the essential characteristics of biological neurons and proposes a novel circuit design comprising CMOS transistors and memristors for electronic implementation of a neuromorphic system. Each neuron-emulating circuit ("CMOS neuron") aggregates input signals from sensory receptors or other neurons, and generates action potentials based on "leaky integrate-and-fire" (LIF) principles. A two-step neuron-to-neuron interaction scheme is proposed: firstly, the action potential output is converted to a post-synaptic input signal for the next layer of neurons; secondly, during the refractory period where a second action potential cannot be fired, the value of the synaptic memristor element is modified according to Spike Timing Dependent Plasticity (STDP) learning rules, *i.e.* a decrease in resistance corresponds to long-term synaptic potentiation (LTP) whereas an increase in resistance corresponds to long-term synaptic depression (LTD). Finally, this report demonstrates the capability of the proposed neuromorphic circuit for adaptive learning.

Chapter 2. Neuromorphic Computing Principles

The concept of neuromorphic computing was developed in the late 1980s by Carver Mead, who described the use of very-large-scale integrated (VLSI) systems to process analog signals in order to mimic neuro-biological processes [6]. This chapter outlines neuromorphic computing principles based on observations from studies in neuroscience. It then briefly mentions previous work by other research groups to develop and implement computational models of the human brain at various levels of abstraction.

2.1 Characteristics of Biological Neurons

The majority of neurons in the human brain are multipolar neurons, which have a single axon that acts as an output terminal and multiple dendrites that behave as input terminals. Axons can be connected to multiple dendrites, and each axon-dendrite interface comprises a synapse that facilitates electrochemical interactions between neurons. Dendrites also can be connected to receptor cells which respond to external stimulus. Almost all neural network implementations (hardware and software) are modeled after a system of multipolar neurons (Figure 1).

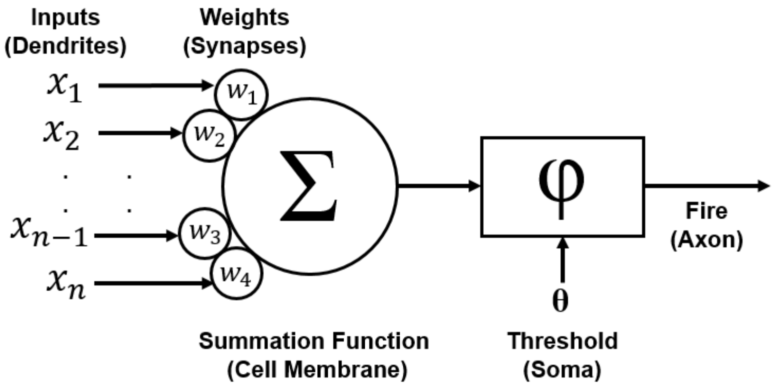


Figure 1: Computational model of a biological multipolar neuron [13].

The cell body of a neuron is known as the soma, and it produces the necessary proteins for a neuron to function. From a computational standpoint, the soma integrates information received from the dendrites, and outputs an action potential (*i.e.* it “fires”) if the soma’s membrane potential exceeds a threshold. Dendritic inputs from other neurons are categorized as either excitatory postsynaptic potentials (EPSPs) or inhibitory postsynaptic potentials (IPSPs). EPSPs increase a neuron’s likelihood of firing by increasing the soma’s membrane potential, while IPSPs decrease a neuron’s likelihood of firing. The postsynaptic input potentials are aggregated by the soma via spatial summation and temporal summation (Figure 2).

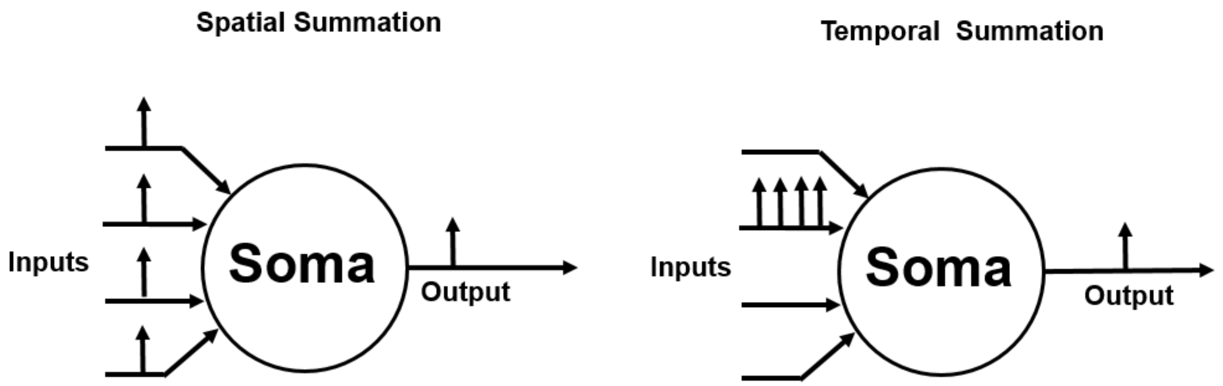


Figure 2: Biological neurons generate action potentials (spikes in electric potential *vs.* time, denoted by the vertical arrows) based on spatial and temporal summation [14].

Spatial summation refers to the addition of input signals from different dendrites at a specific point in time. Most software artificial neural network (ANN) models implement spatial summation to determine whether an action potential should be fired and transmitted to the next layer of neurons. Traditionally, synaptic inputs have been summed linearly:

$$A_{out} = step\left(\sum_{i=1}^{N} w_i A_i - \theta\right)$$

where A_i is an action potential from a previous layer of neurons, w is the synaptic weight, and θ is the threshold potential. However, recent studies in neuroscience suggest that the integration of EPSPs and IPSPs may contain nonlinear terms proportional to the product of EPSPs and IPSPs [15]. It remains unclear whether simple arithmetic rules are applicable for the spatial summation of inputs from a dendritic tree [15].

Temporal summation refers to the addition of successive input signals over time from a single dendrite. The capacitive nature of a neuron's cell membrane prevents abrupt changes to the membrane potential. Successive weak EPSPs (below the threshold potential) from a single dendritic input may still generate an action potential if they arrive within a short period of time.

2.2 Action Potentials and Signal Propagation

An action potential is a spiking of the membrane potential when a threshold is reached. It is characterized as having steep depolarization and repolarization phases, followed by a refractory period where the membrane potential undershoots before reaching back to its resting potential (Figure 3).

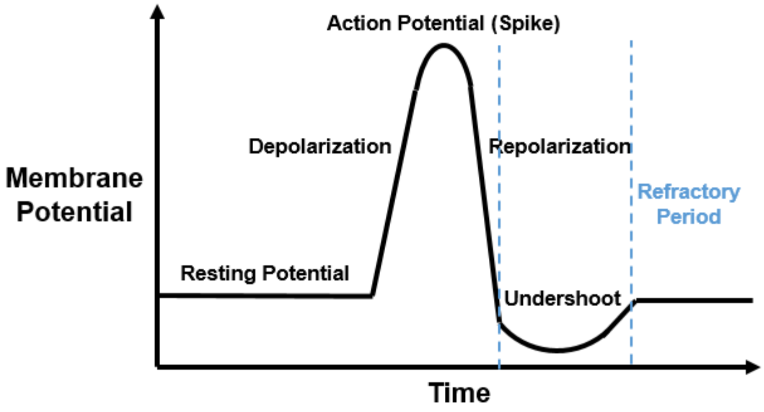


Figure 3: Generic features of an action potential in biological neurons [16].

Action potentials are sometimes referred to as a train of pulses. They follow two important principles that should be considered in neuromorphic systems [17]:

1) All-or-none principle: All action potentials fired from the same neuron are identical, in the sense that stronger inputs do not generate larger action potentials. As long as the sum of input signals causes the membrane potential to exceed the threshold, the exact same action potential will be fired.
2) Absolute refractory period: After an action potential is fired, there is a period of time when a second action potential is prohibited from firing, regardless of the strength of the sum of input signals.

The firing frequencies of biological neurons can range from ~1 Hz to ~100 Hz, depending on the type of neuron and the intensity of the input stimulus [18]. Stronger inputs cause action potentials to be fired at faster rates. Due to the refractory period, there exists an upper limit for the firing frequency. Neuromorphic systems can be designed to operate at ~GHz frequencies [12], since CMOS device capacitances (~fF) and memristor switching delays (~ns) naturally allow CMOS neurons to operate faster than biological neurons.

2.3 Spike Timing Dependent Plasticity (STDP)

The ultimate goal of neuromorphic computing is to achieve a system that is capable of unsupervised learning. Over time, neuroscientists have discovered that human cognition and memory are attributable to a network of synapses in the brain with tunable strengths. Hebb's postulate is a well-known learning rule, which states that when "Cell A" repeatedly contributes to the firing of "Cell B," a metabolic change occurs such that "Cell A" becomes more efficient in contributing to the firing of "Cell B" [19]. The strengthening and weakening of synaptic connections are known respectively as long-term potentiation (LTP) and long-term depression (LTD). A stronger synaptic connection between two neurons increases the likelihood of the pre-neuron inducing the post-neuron to fire, consistent with Hebb's postulate of learning.

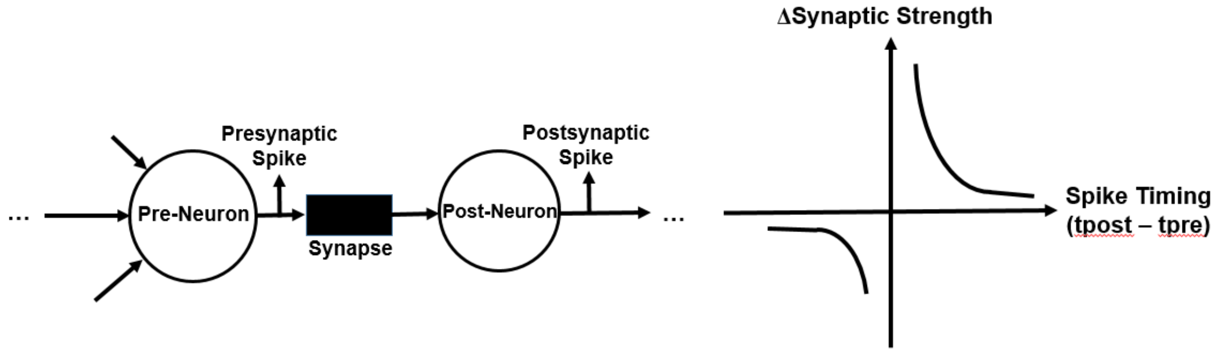


Figure 4: Spike timing dependent plasticity (STDP) [20].

In 1998, Bi and Poo discovered that LTP occurs when presynaptic spikes lead postsynaptic firing, while LTD occurs when postsynaptic firing leads presynaptic spikes [20]. Moreover, pairs of action potentials fired at close points in time affected the synaptic strength much more than those fired far apart (Figure 4). This biological process is now known as spike timing dependent plasticity (STDP). While STDP is only one of several factors that contribute to synaptic weight changes, it has now almost become a universal kernel for associative learning due to its simplicity and occurrence among >20 different types of mammals and insects [21]. Designing a neuromorphic system capable of exhibiting STDP may be the key to future advancements in artificial intelligence.

2.4 Previous Work on Electronic Neuromorphic Systems

Neuromorphic computing was originally postulated as a computational alternative to digital computers. In 1995, Mead noticed that when CMOS devices operated in their subthreshold regime, transistor gate voltages controlled the energy barriers over which charge carriers flowed across the channel in similar ways as neurons regulated charge movements across their membranes [6]. He argued that human brains exceled at localizing computations and being accurate in the presence of noise. Digital systems require precision from individual bits, whereas neural systems rely on feedback so that multiple signals combine to collectively achieve precision [6]. Early neuromorphic circuits were mostly analog, and included attempts of developing artificial vision [5] as well as studies on neuron interactions in the spinal cord [37].

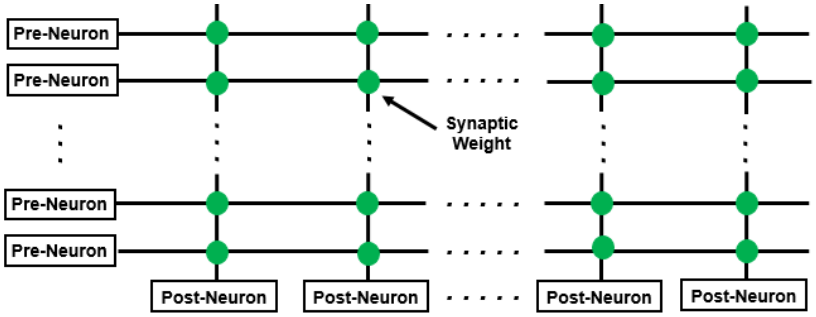


Figure 5: Common neuromorphic crossbar array for high-density synaptic storage [8].

The emergence of digital computers and the electronic industry's efforts on related research decelerated neuromorphic computing advancements until recent years. Focus was shifted to developing efficient pattern recognition algorithms, including software implementations like artificial neural networks [38]. Due to the increasing need for energy efficient computing systems and the discovery of scalable memristive devices, neuromorphic computing has resurfaced, with an emphasis on the development of hardware that mimics the human brain. Modern neuromorphic systems can be application-based or biology-based; designs driven by applications tend to be digital and compatible with gate-level logic, while designs focused on modeling the electrochemical interactions between neurons incorporated analog principles [12]. As device engineers develop reliable devices compatible with neuromorphic circuits, circuit designers have begun exploring novel architectures to implement neuromorphic systems.

In 2011, Seo *et al.* at IBM Research designed a digital neuromorphic chip for pattern classification using 45nm CMOS technology and transposable crossbar 8T SRAM arrays as binary synaptic weights [8]. Recently, this crossbar configuration (Figure 5) has been extended by other research groups to incorporate memristors as the synaptic storage element for better scaling and efficiency [27], [35]-[36]. Kim *et al.* at the University of Michigan designed and fabricated a fully operational high-density hybrid crossbar/CMOS storage system with multilevel memristors and CMOS decoders [36]. Research has also been conducted on the design of CMOS circuits that can produce action potential firings using capacitors and transistors [31]. Studies suggest that while digital implementations at the 10nm node can consume less power, analog circuits and scaled memristive devices provide an advantage in overall system area [12].

This work incorporates both analog and digital principles from previous research to design a neural system that interacts via action potential firings. Dendritic input aggregation and neural spiking are achieved using CMOS amplifier circuits, while synaptic plasticity is implemented with memristive technologies.

3. The Memristor as an Artificial Synapse

The memristor is a passive two-terminal electrical component postulated by Chua in 1971 [22], and realized recently by Williams *et al.* in 2008 [11]. Due to the analogous relationships between voltage and current, charge and flux, the memristor has been regarded as the fourth fundamental circuit element along with the resistor, capacitor and inductor (Figure 6).

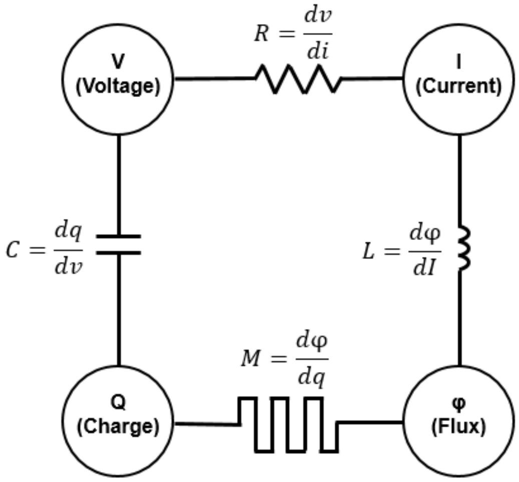


Figure 6: Circuit elements relating fundamental electrical quantities [11].

The memristance *M* describes a flux and charge relation:

$$M(q) = \frac{d\varphi}{dq}$$

If charge changes as a function of time, this relation can be rewritten as [11], [22]:

$$M(q(t)) = \frac{d\varphi/dt}{dq/dt} = \frac{V(t)}{I(t)}$$

Hence, the memristor can be described as a device whose resistance varies based on changes in its charge profile over time. It has been proposed that this property can effectively model synaptic behavior for neuromorphic computing applications [23].

3.1 RRAM as Memristor: Modeling and Simulations

Researchers around the world have been investigating Resistive random-access memory (RRAM) devices for high-density non-volatile memory applications due to their scalability (down to ~10 nm cell size), reliability (~10^{12} SET/RESET cycles), short programming time (~ns) and low energy consumption (~0.1 pJ/bit) [24]-[26].

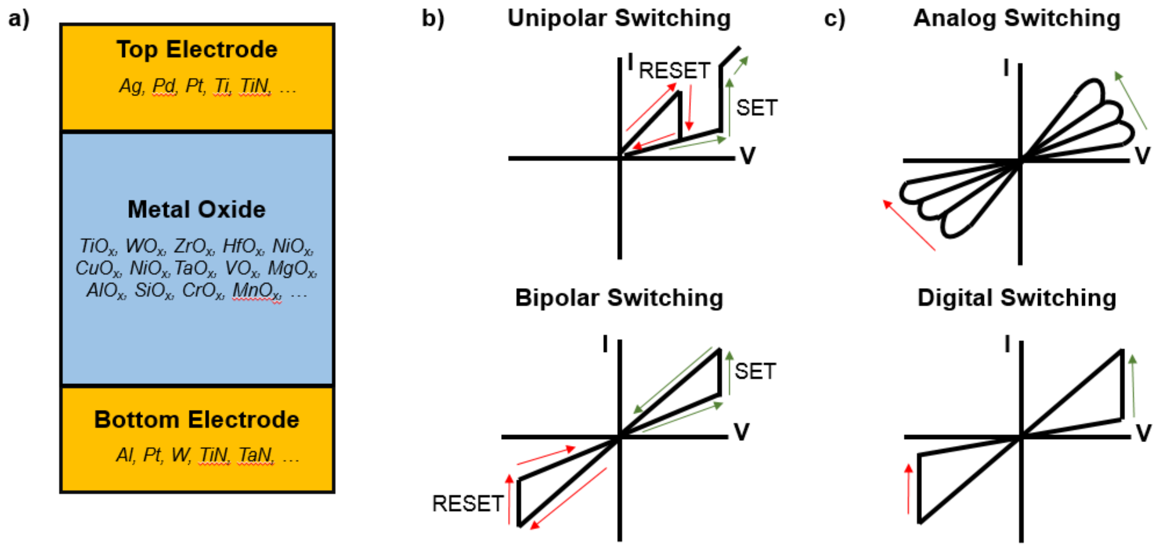


Figure 7: a) schematic illustration of an RRAM cell showing the various material choices [24] b) representative current *vs*. voltage (*I*-*V*) characteristics for unipolar and bipolar switching mechanisms. c) *I*-*V* curve for analog and digital programming of an RRAM cell.

An RRAM cell typically consists of a transition metal oxide layer sandwiched between two electrodes (Figure 7a). The resistive switching behavior is attributed to the formation and rupture of conductive filaments that facilitate current flow through the oxide layer. If a threshold electric field or/and current is achieved, the RRAM cell can transition from a high-resistance state (HRS) to a low-resistance state (LRS) during the SET process, or from the LRS to HRS during the RESET process. For a unipolar RRAM cell, both SET and RESET operations can be achieved with the same polarity of applied voltage; for a bipolar RRAM cell, SET and RESET operations require opposite polarities of applied voltage (Figure 7b). The resistance of an RRAM cell can be switched abruptly between two states, or gradually between multiple resistance states. (Some RRAM devices exhibit mixed analog-digital switching behavior, such as the HfO_x-based device fabricated by Yu *et al.* which undergoes abrupt SET, but gradual RESET operations [27].) In the latter case, the RRAM cell behaves as a memristor. Therefore, in this chapter, a bipolar RRAM device with analog programming behavior is investigated as a synaptic device for neuromorphic circuits.

Using a compact model for RRAM devices developed by the University of Michigan [28], SPICE simulations were performed to study the analog behavior of a bipolar RRAM device with ~ns programming time and ~1.2 V threshold voltage (Figure 8). A 10 ns potentiation pulse was applied across the RRAM device to SET it from HRS to LRS, followed by a 10 ns depression pulse to RESET it back to its original HRS. For voltages below 1.0 V, no change in resistance state occurs, as there are insufficient charge carriers created to increase current flow; however, for a peak voltage (V_p) of 1.2 V, a 10 ns pulse causes the RRAM device to change its resistance by an order of magnitude (Figure 7c). Devices designed to mimic synaptic behavior should output voltage signals of similar magnitude to take advantage of the analog behavior of the RRAM device. For V_p = 1.4 V, the RRAM device undergoes binary switching.

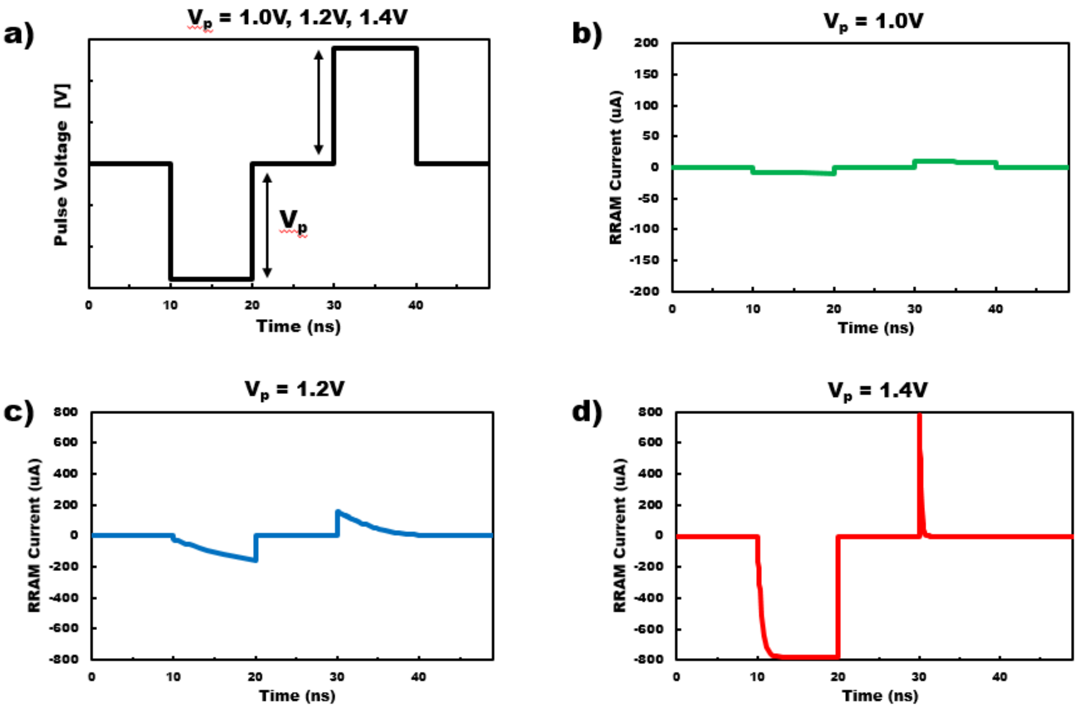


Figure 8: SPICE simulations of an RRAM cell, using the compact model from [28]: a) Pulse voltage signal applied across a RRAM cell. b) Simulated current flowing through the RRAM cell for V_p = 1.0 V. c) Simulated current flowing through the RRAM cell for V_p = 1.2 V. d) Simulated current flowing through the RRAM cell for V_p = 1.4 V.

Similarly, DC voltage sweeps were simulated using LTSpice to illustrate the *I-V* characteristics of the RRAM device (Figure 9). For sweeps below the threshold voltage, there is no change in the resistance of the RRAM. The device can be SET or RESET (depending on the polarity) by applying a voltage of 1.2 V; the application of higher voltages results in abrupt changes in resistance. To further demonstrate its suitability as a synaptic device, a series of 1 ns pulses was applied across the RRAM every 10 ns to mimic action potential firings (Figure 10). Depending on the polarity of the bias voltage, the RRAM can be gradually switched between LRS and HRS states with a series of short pulses. Most RRAM devices can reliably change their resistances continuously across a range of two orders of magnitude [25]-[27]; this characteristic is verified by the simulation results herein.

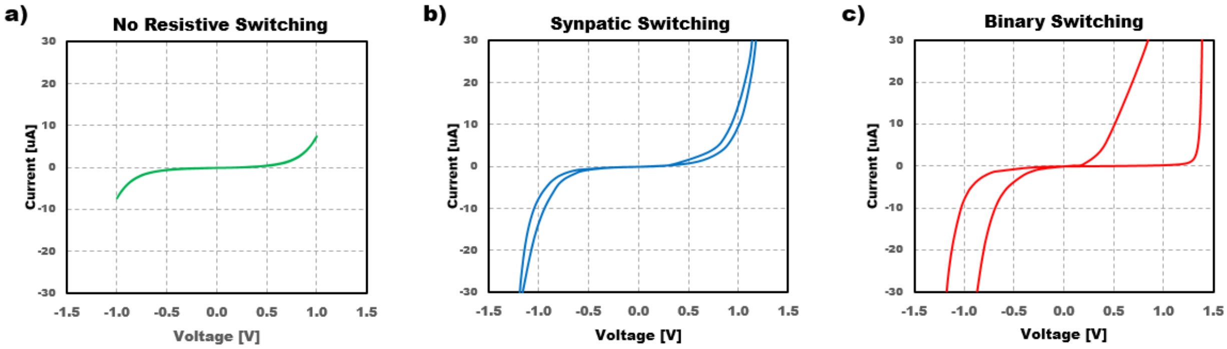


Figure 9: Simulated *I-V* characteristics of a RRAM device operating in different regimes: a) Sub-threshold voltage operation (no resistive switching); b) Analog switching for intermediate applied voltage. c) Digital switching for high applied voltage.

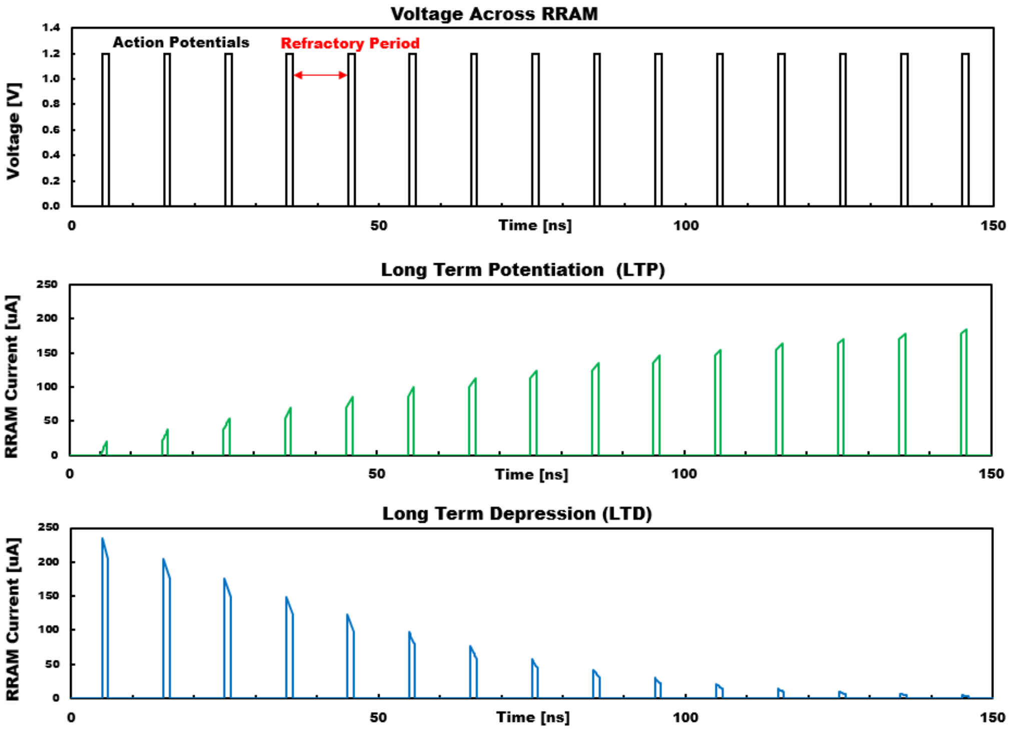


Figure 10: Simulations showing Long Term Potentiation (LTP) and Long Term Depression (LTD) behavior of an RRAM device. The RRAM resistance can be decreased or increased by a series of 1 ns pulses, depending on the polarity of the applied voltage.

3.2 Synapse Configuration

A voltage divider comprising a memristor (analog bipolar RRAM device) connected in series with a resistor can emulate the behavior of a synapse (Figure 11). In the biological process, an action potential may cause neurotransmitters to electrochemically transmit strong or weak postsynaptic input potentials to the next layer of neurons, depending on the synaptic strength. With the artificial synapse configuration shown, the strength of the postsynaptic potential varies in a similar fashion depending on the resistance of the memristor. For an analog RRAM device with two orders of magnitude difference in resistance between HRS and LRS, the resistance *R* should be selected such that the resistance of the memristor can vary between 0.1R to 10R. Under such conditions, the strength of the postsynaptic potential can vary from ~10% to ~90% of the action potential (Figure 12). This potential acts as the input signal to the next layer of neurons.

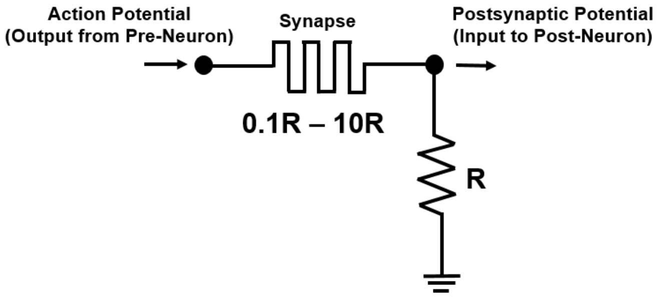


Figure 11: Synapse implementation using memristor-resistor voltage divider configuration.

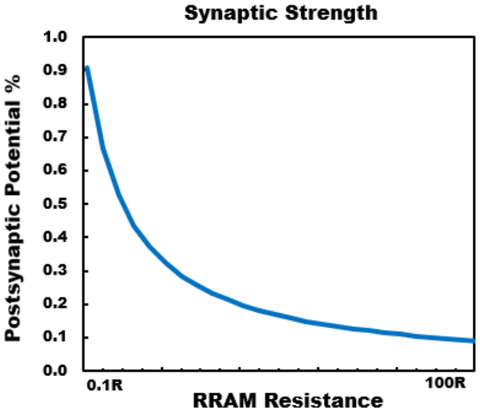


Figure 12: Synaptic strength based on RRAM resistance value.

Chapter 4. CMOS Neuron Design

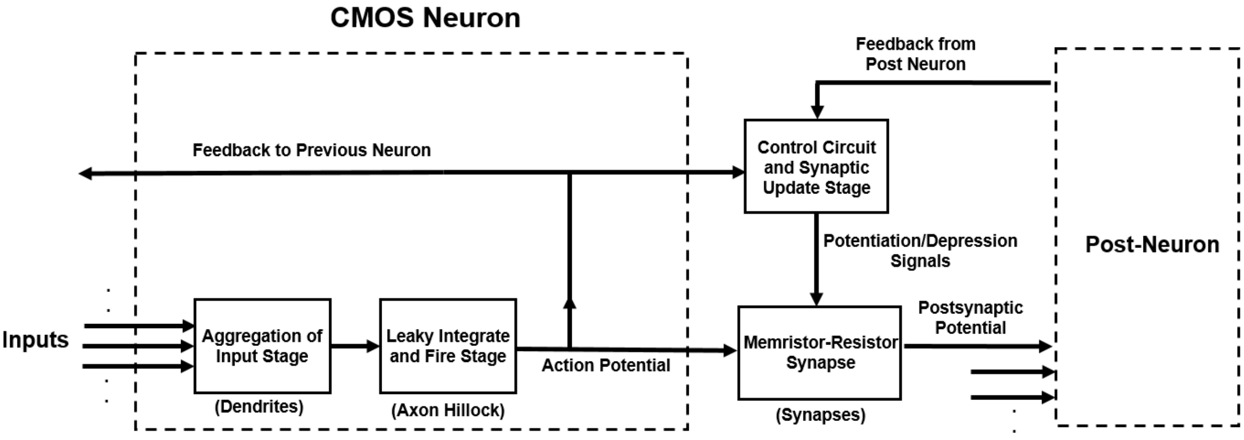

CMOS Neuron
Feedback from Post Neuron
Feedback to Previous Neuron
Control Circuit and Synaptic Update Stage
Potentiation/Depression Signals
Post-Neuron
Inputs
Aggregation of Input Stage
(Dendrites)
Leaky Integrate and Fire Stage
(Axon Hillock)
Action Potential
Memristor-Resistor Synapse
(Synapses)
Postsynaptic Potential

Figure 13: Block diagram of the CMOS Neuron.

The CMOS neuron design presented in this chapter accounts for the anatomy of biological neurons, the way neurons communicate with each other, and synaptic behavior that supports memory and learning.

To satisfy the neuromorphic computing principles described in Chapter 2, a CMOS Neuron should consist of multiple stages, each implementing a function corresponding to a behavioral component of a biological neuron. As shown in the block diagram of Figure 13, input signals are spatially and temporally summed via a source follower aggregation circuit [29] that outputs a signal to a leaky integrate and fire (LIF) stage [12] which generates action potentials with absolute refractory periods. The postsynaptic action potential is weighted by the memristor-resistor synapse circuit as it is transmitted to a neuron in the next layer. A control circuit takes action potentials from both the pre-neuron and the post-neuron to modify the synaptic strength between the pre-neuron and post-neuron, according to STDP rules. The strength of the synapse is updated during the refractory period, after the action potential output of the pre-neuron passes through to the dendrite of the post-neuron.

4.1 Aggregation of Input Stage

A series of source followers connected in parallel can be used to implement the dendrites for the CMOS neuron design (Figure 14). The source follower has high input impedance with an output voltage given by

$$V_{out} \approx \frac{\sum_{i=1}^{n} G_i V_i}{\sum_{i=1}^{n} G_i}$$

where G_i is the transconductance of the operational transconductance amplifier (OTA) and V_i is the input signal. A transistor level schematic of each OTA is shown in Figure 15. For input voltages that deviate significantly from the average, the OTA current saturates, limiting the contributions of extreme input voltage signals [29]. The individual source followers can be tuned to have greater transconductances so that different types of inputs contribute differently to the likelihood of the neuron's firing. For example, larger transconductances can model dendrites connected to receptor cells that offer strong stimulus, while smaller transconductances can model minor excitations from pre-neuron firings.

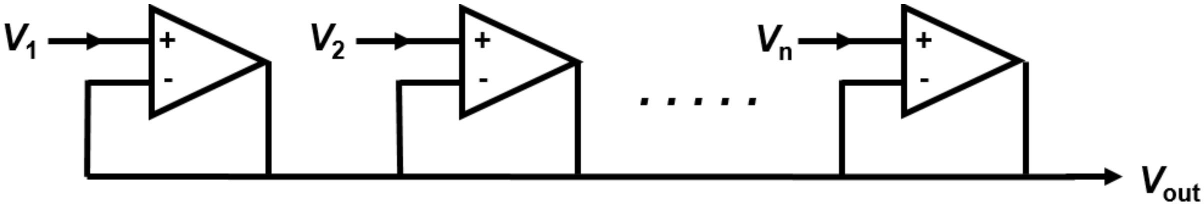


Figure 14: Source Follower Aggregation Circuit [29].

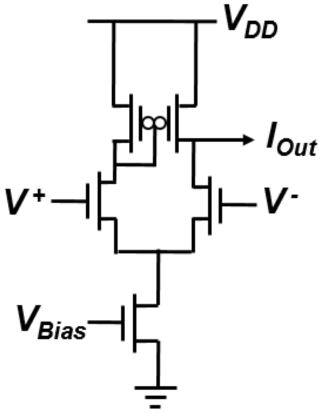


Figure 15: Operational Transconductance Amplifier (OTA) [29].

Using the Predictive Technology Model (PTM) for 90 nm-generation CMOS technology [30], SPICE simulations were performed on a follower aggregation circuit with three inputs (Figure 16). In this experiment, a neuron first receives an input signal from a single dendrite 10 ns before receiving subsequent stimulus from the other two dendrites. As the number of inputs increases, the follower aggregation output increases accordingly to account for the spatial summation capabilities of biological dendrites. This weighted average is converted into a current via a transconductance amplifier, and the output current is transmitted to a LIF circuit to mimic action potential firings from the axon hillock.

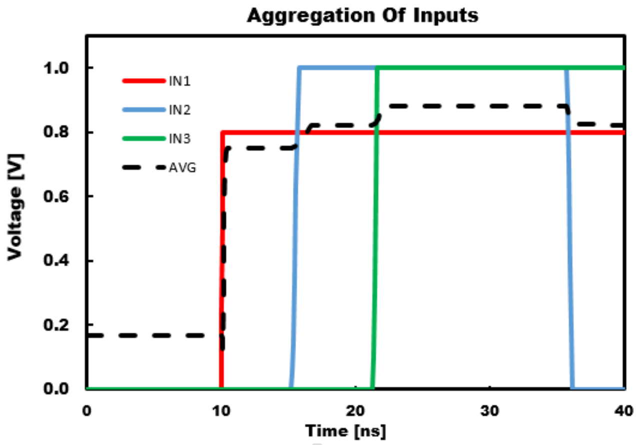


Figure 16: Aggregation of Input Simulation.

4.2 Leaky Integrate and Fire Stage

The LIF model is commonly used to emulate the behavior of the neuron cell body. The membrane potential of the CMOS neuron can be described with differential equation [12],

$$I_{In}(t) = C_{Mem}\frac{dV_{Mem}(t)}{dt} + \frac{V_{Mem}(t)}{R_{Mem}} + C_{Fb}\frac{d(V_{Mem}(t) - V_{Spk}(t))}{dt}$$

where C_{Mem} is the membrane capacitance, V_{Mem} is the membrane potential, $\mathrm{R}_{\mathrm{Mem}}$ is the membrane resistance, I_{In} is the input current, and V_{Spk} is the output spike. Based on LIF principles described by Indiveri *et al.* [31], the axon hillock circuit can be implemented using a combination of capacitors, inverters, and differential amplifiers (Figure 17). First, the membrane capacitance (C_{Mem}) is charged by the incoming current (I_{In}) from the dendritic input aggregation stage. If the membrane potential (V_{Mem}) exceeds the threshold potential (V_{TH}), V_{Spk} rapidly changes from 0 V to V_{DD}. When V_{Spk} goes high, a reset transistor (N_{Rst}) is switched on, activating positive feedback through capacitor C_{Fb}. Once the membrane capacitor is discharged, V_{Spk} swings back to 0 V, switching the reset transistor off before repeating the integration cycle. The rate of firing can be tuned by adjusting the ratio between the capacitances of C_{Fb} and C_{Mem}, while the pulse width of the output spike can be controlled by the bias voltage V_{C}.

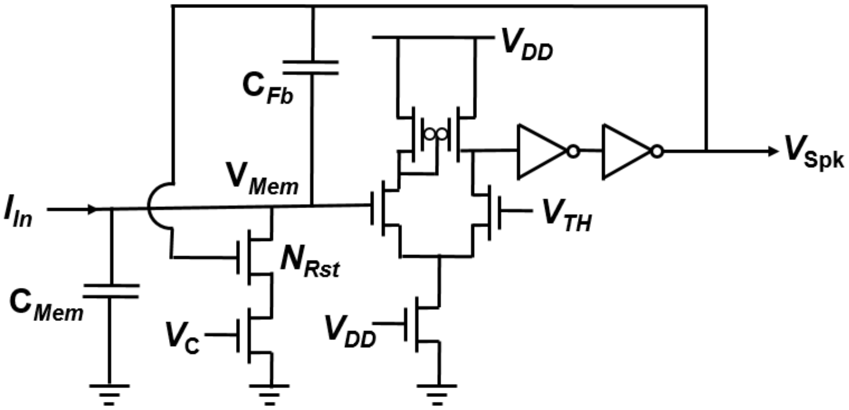


Figure 17: Axon Hillock LIF circuit design.

A three-input neuron can be implemented by connecting the axon hillock circuit in series with the aggregation of input stage to demonstrate the CMOS neuron's capabilities of performing spatial and temporal summation (Figure 18). When the CMOS neuron receives a signal from one dendrite, the stimulus is insufficient to cause action potentials to be fired. When all three dendrites receive signals simultaneously, the membrane potential exceeds the threshold potential and firing occurs. Similarly, an action potential is generated when a single dendrite receives three successive input signals, upon the third input.

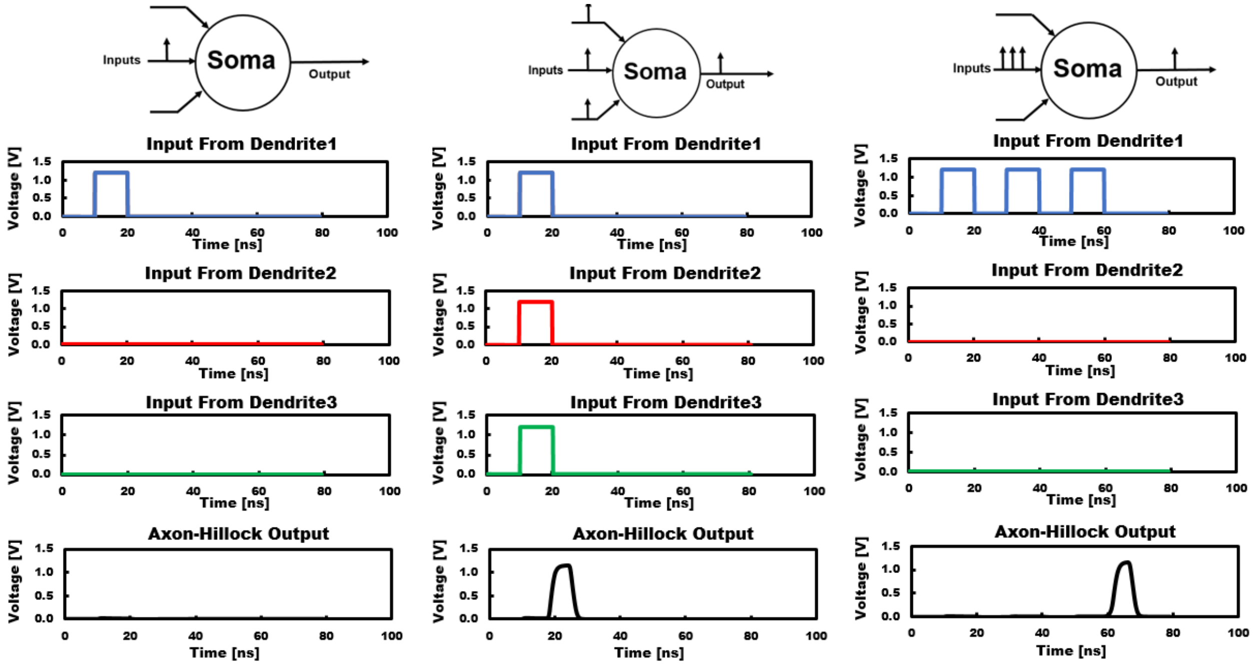


Figure 18: SPICE simulation showing temporal and spatial summation for axon hillock circuit design.

In addition, the rate of firing of the axon hillock circuit is dependent on the strength of the input stimulus such that strong inputs cause action potentials to be fired at faster rates than weaker inputs (Figure 19). Simulation results also demonstrate that regardless of the strength of the input stimulus, the magnitude of the action potential is the same, and the strength of transmission to the next layer of neurons is entirely dependent on the synaptic weight of the memristor. Due to the discharging of the membrane capacitor, there is always a refractory period between action potential firings. These results are consistent with the important features observed in biological neurons as described in Chapter 2.

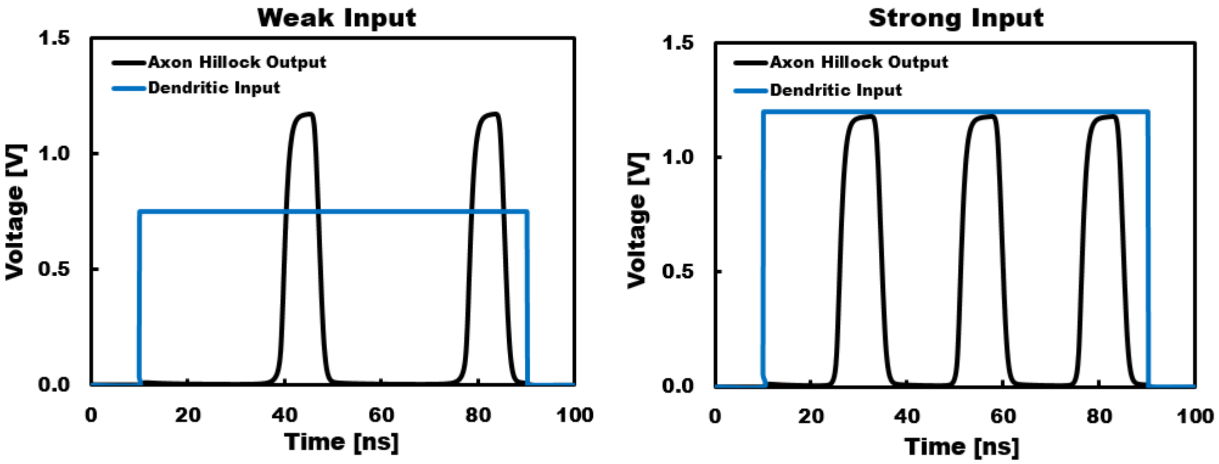


Figure 19: SPICE simulations of CMOS neuron with absolute refractory period and stimulus dependent firing rates.

4.3 Synaptic Strength Update Stage

In order to update the resistance of the memristor during the refractory period, potentiating and depressing pulses must be applied across the memristor according to STDP learning rules. One possible way of achieving STDP is to design circuitry that control the currents flowing across the memristor as seen in Figure 20. Here, the memristor synapse configuration as described in Chapter 3.2 is connected to three sets of complementary pass gates. When an action potential is fired from the pre-neuron (V_{pre}), V_{Fire} goes high and the action potential is converted to a postsynaptic input potential (V_{out}). In this case, V_{pre} is below the threshold voltage and resistive switching does not occur.

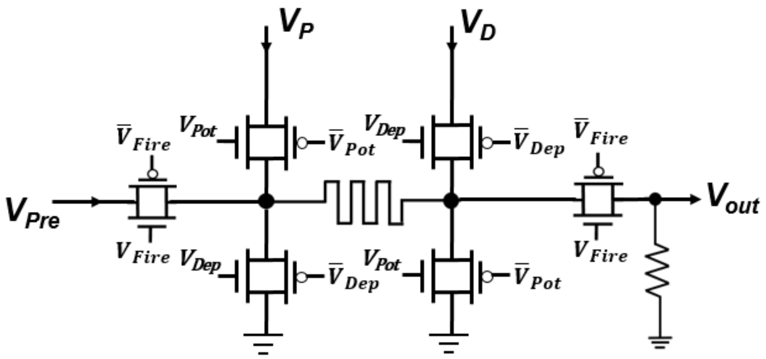


Figure 20: Synaptic weight update circuit for memristor.

If post-neuron firing precedes pre-neuron firing, V_{Dep} is switched high during the refractory period to apply a strong depressing pulse (V_{D}) across the memristor to update its resistance. Here, V_{Fire} is switched low so that the depressing pulse is not transmitted to the next layer of neurons. When post-neuron firing follows pre-neuron firing, a similar scheme occurs with V_{Pot} switching high, applying V_{P} across the memristor.

The control signals can be generated using logic and memory circuits similar to the ones shown in Figure 21. It is assumed that the CMOS neuron is configured such that all action potentials have comparable pulse widths and that the pulse widths are much shorter than the refractory period of the neuron. According to STDP, the memristor synapse should only update its resistance when action potential firings from pre- and post-neurons occur at close points in time. A memory circuit should be used to determine whether potentiation or depression pulses should be generated. In this work, a simple capacitor is used to store information regarding the order of the pre- and post-neuron firings. For the potentiation circuit, the capacitor is charged whenever a pre-neuron fires in the absence of a post-neuron. The capacitor always operates in its “write” mode, and only changes to its “read” mode when the pre- and post-neurons both fire at the same time. This configuration constantly resets the capacitor so that the AND gate (Figure 21) goes high only when there is a pre-before-post firing followed immediately by a pre-AND-post firing. When firing conditions are met, a chain of inverters can be used to delay the overlap of the two signals in order to generate the appropriate control signal (V_{Pot} or V_{Dep}) during the refractory period. Another chain of inverters with high gate capacitances is used to generate the actual potentiating and depressing signals (V_{P} and V_{D}). The gate capacitances serve to convert the pulse width of the overlap between pre- and post-neurons into a difference in strength of voltage applied across the memristor. SPICE simulation results for the synaptic strength update stage with are shown in Figure 22. Principles described in this chapter serve as general guidelines, and can be further optimized for power, area and reliability.

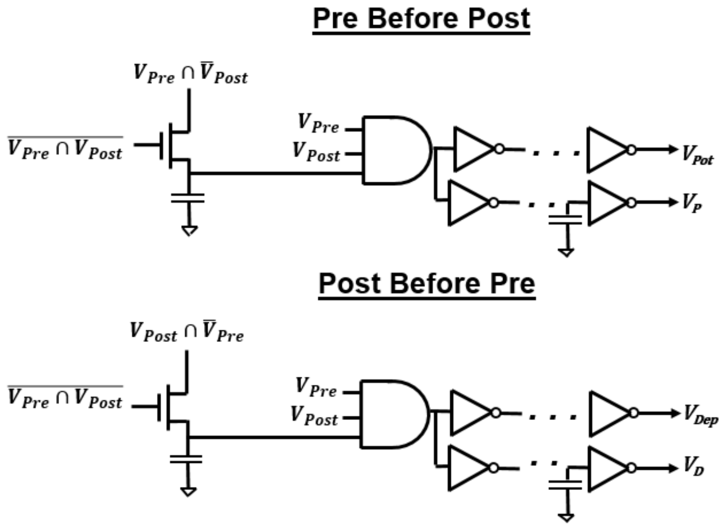

Pre Before Post
$V_{Pre} \cap \bar{V}_{Post}$
$\overline{V_{Pre} \cap V_{Post}}$
V_{Pre}
V_{Post}
V_{Pot}
V_P
Post Before Pre
$V_{Post} \cap \bar{V}_{Pre}$
$\overline{V_{Pre} \cap V_{Post}}$
V_{Pre}
V_{Post}
V_{Dep}
V_D

Figure 21: Control circuit that generates potentiating and depressing pulses for STDP.

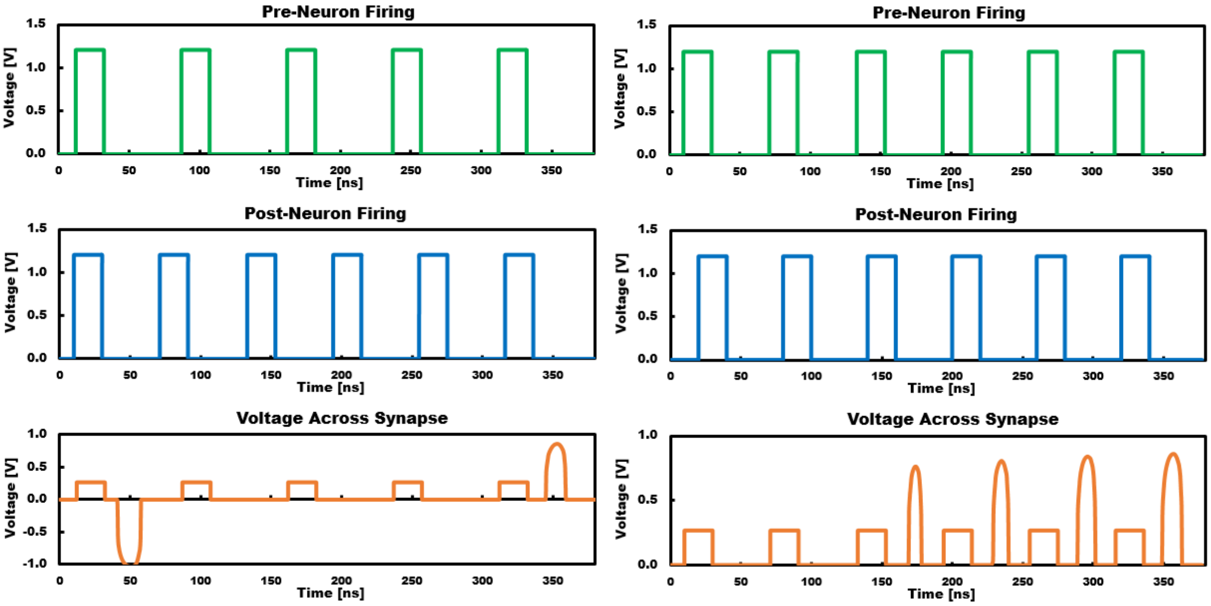


Figure 22: SPICE simulation results for synaptic strength update circuit. A resistor is used in place of a memristor as synapse to avoid convergence issues in SPICE. Simulation only aims to show how voltages with different duration, polarity, and magnitude can be applied across a synapse during the refractory period depending on spike-timing between pre and post-neurons.

Chapter 5. CMOS Neuron Applications

At present, neuromorphic systems provide performance advantages over traditional computers in certain noise reduction [3], pattern recognition [4], and image detection applications [5]. There have also been efforts among the semiconductor community to explore systems that are capable of mimicking the human brain's ability to learn and adapt to complex environments [32], [33]. The CMOS neuron designed in this report integrates computation and memory with high parallelism, and also possesses capacity for adaptive learning.

5.1 Associative Learning

During the 1890s, Russian physiologist Ivan Pavlov noticed that after feeding his dogs for a long period of time, his dogs would salivate upon seeing him without food [34]. This led to his initial experiments, where Pavlov repeatedly rang a bell before giving his dog food and discovered that his dog eventually salivated in response to the bell even in absence of food. This is known as classical conditioning, and has helped shape modern-day understanding of associative learning.

In classical conditioning and in context of Pavlov's dog, food is an unconditioned stimulus that always triggers an unconditioned response (salivation), while the bell is normally a neutral stimulus that doesn't cause any physical reactions [34]. After a few repetitions of ringing the bell and feeding the dog simultaneously, the bell became a conditioned stimulus that is able to initiate a conditioned response by itself. Pavlov also observed that the shorter the time interval between the ringing of the bell and the appearance of food, the quicker his dog became conditioned [34].

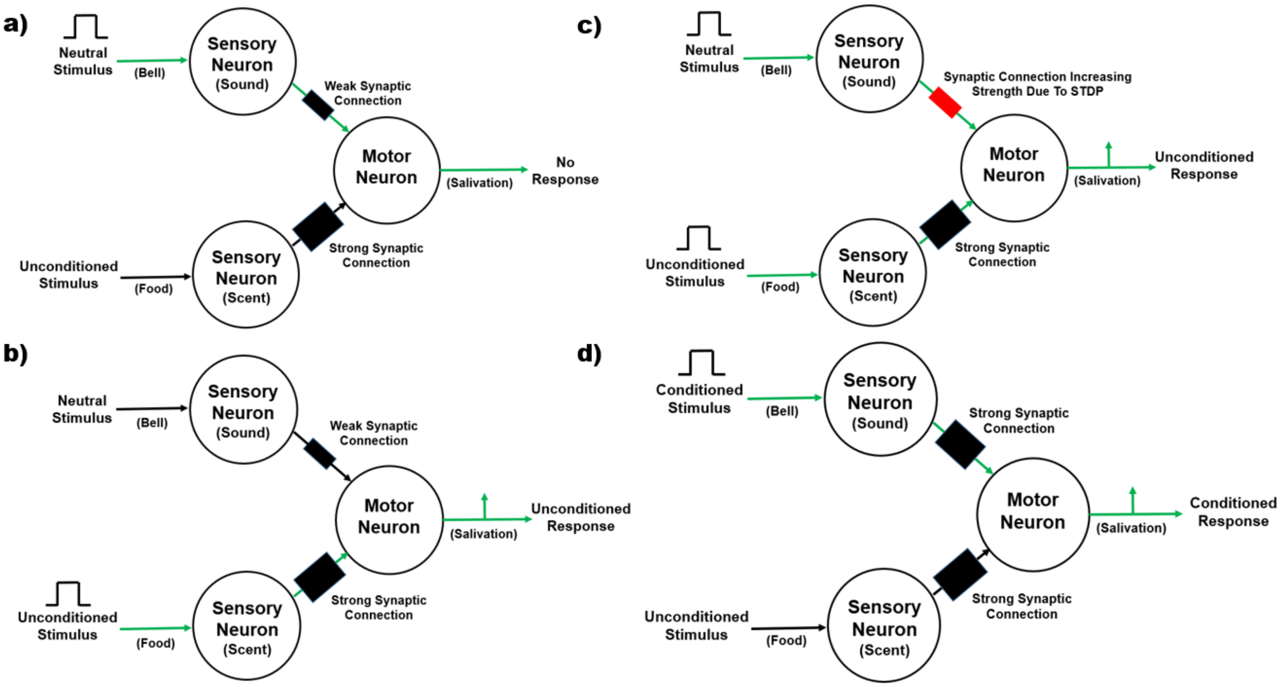


Figure 23: CMOS neuron simulation based on Pavlov's experiments [34]: a) Neutral stimulus triggers no response. b) Unconditioned stimulus triggers unconditioned response. c) Synaptic strengthening due to STDP. d) Conditioned stimulus triggers conditioned response.

To demonstrate the adaptive learning capabilities of the CMOS neuron, Pavlov's experiment was simulated in SPICE with three CMOS neurons (Figure 23) and Stanford's compact model for HfO_x-based memristors [39]. In this simplified model, it is assumed that the sound of the bell and the scent of the food act as inputs for two independent single-input sensory neurons, and that the two sensory neurons are synaptically connected to a single motor neuron that outputs the salivation response. Initially, the synaptic connection between the sound sensory neuron and the motor neuron is weak so that input to the sound sensory neuron is unable to trigger a response from the motor neuron (Figure 23a). Due to the strong synaptic connection between the scent sensory neuron and the motor neuron, food acts as an unconditioned stimulus and is able to trigger an unconditioned response (Figure 23b). When both sensory neurons receive stimulus simultaneously, the scent sensory neuron causes the motor neuron to fire, and STDP from the sound sensory neuron firing before the motor neuron strengthens the synaptic connection between the two neurons (Figure 23c). When the strength of the synaptic connection between the sound sensory neuron and the motor neuron exceeds a certain threshold, the sound of the bell becomes a conditioned stimulus that triggers a conditioned response (Figure 23d). The SPICE simulation shown in Figure 24 demonstrates the associative learning ability of the CMOS neuron.

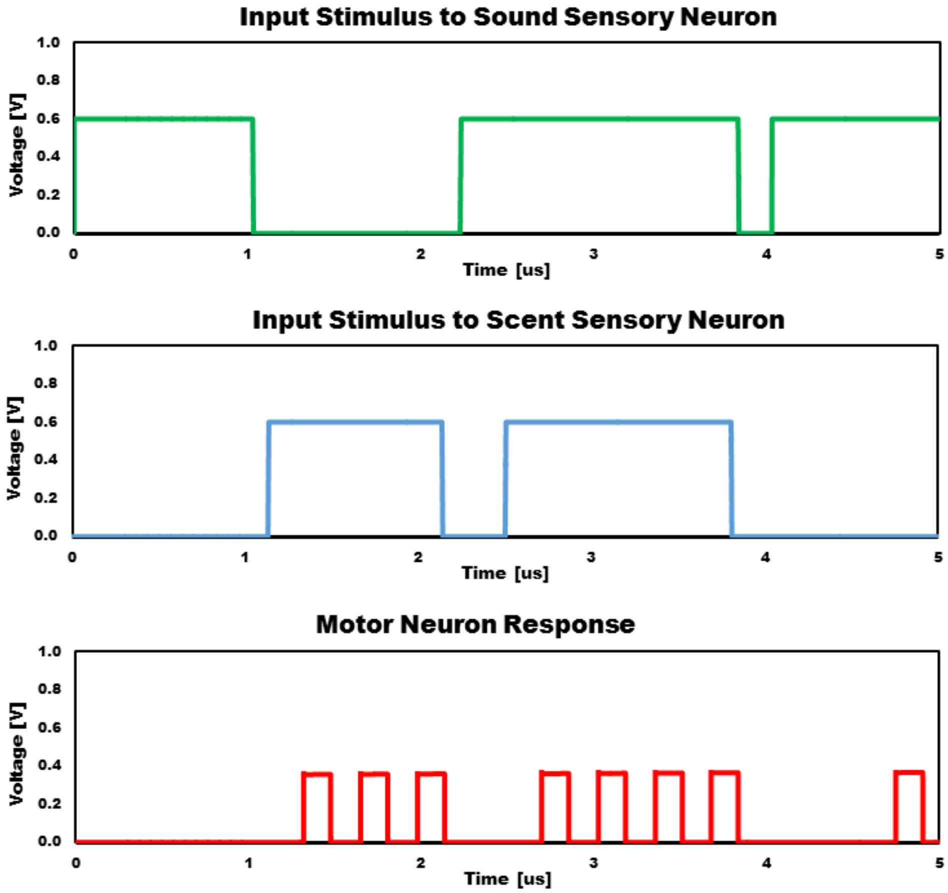


Figure 24: SPICE simulation for associative learning experiment using Stanford's compact model for HfO_x memristors [39].

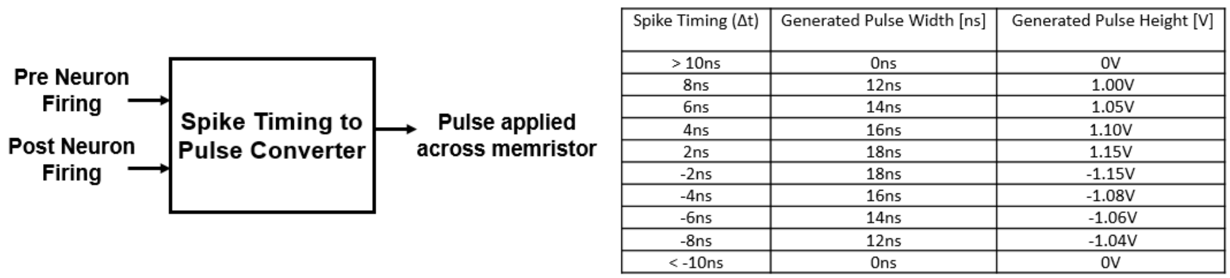
5.2 Future Prospects

This report provides design principles for a CMOS neuron using 90nm technology that closely mimics key characteristics observed in biological neurons. The present design requires ~25 transistors for each CMOS neuron, 5 transistors for each dendritic input, and ~70 transistors + 1 memristor for each synapse. While CMOS neurons (~GHz) can operate faster than the human brain (~Hz), it remains important to decrease the number of transistors per synapse to increase the feasibility of developing more complex neuromorphic systems (*i.e.* human brain with 10^{11} neurons interconnected with ~10^{15} synapses). As the switching voltage of memristors continue to decrease, it is also worth revisiting Mead's original idea of utilizing transistors operating in the sub-threshold regime in designing CMOS neurons to facilitate further reductions in power consumption [6]. Emerging technologies with zero off-state current (*i.e.* nanoelectromechanical relays) may address some of the scaling challenges faced by analog neuromorphic systems, as the number of synapses an axon can drive is presently limited by sub-threshold leakage.

It is also particularly important for circuit designers to develop simple and reliable means of achieving STDP in artificial neurons, as STDP is the distinguishing feature of neuromorphic systems that are capable of unsupervised adaptive learning. At present, a popular way of achieving STDP is to design a neuron circuit that converts the spike timing between two action potentials (Δt) into a pulse voltage via time-division multiplexing (TDM), such that longer pulses are applied across the memristor at smaller Δt [40]. The synaptic strength update stage described in Chapter 4.3 of this report implements a variation by converting smaller Δt to longer pulses with higher voltages applied across the memristor. Figure 25 shows a mapping scheme that can theoretically be achieved by refining the synaptic strength update circuit described in this report. SPICE simulations using University of Michigan's memristive compact model [28] were performed by applying pulses with different magnitudes and durations across the memristor and observing the resulting resistance change. To parallel LTD and LTP in biological synapses, the change in the synaptic weight of the Memristor-resistor synapse is defined as:

$$\Delta G = \frac{G_{Memristor\ Final} - G_{Memristor\ Initial}}{G_{Memristor\ Initial}}$$

Figure 26 shows STDP achieved with University of Michigan's compact model and the Δt to pulse mapping scheme described in Figure 25.


Pre Neuron Firing
Post Neuron Firing
Spike Timing to Pulse Converter
Pulse applied across memristor

Spike Timing (Δt)	Generated Pulse Width [ns]	Generated Pulse Height [V]
> 10ns	0ns	0V
8ns	12ns	1.00V
6ns	14ns	1.05V
4ns	16ns	1.10V
2ns	18ns	1.15V
-2ns	18ns	-1.15V
-4ns	16ns	-1.08V
-6ns	14ns	-1.06V
-8ns	12ns	-1.04V
< -10ns	0ns	0V

Figure 25: STDP spike mapping scheme. Different spike timings generate pulses with different magnitudes and durations.

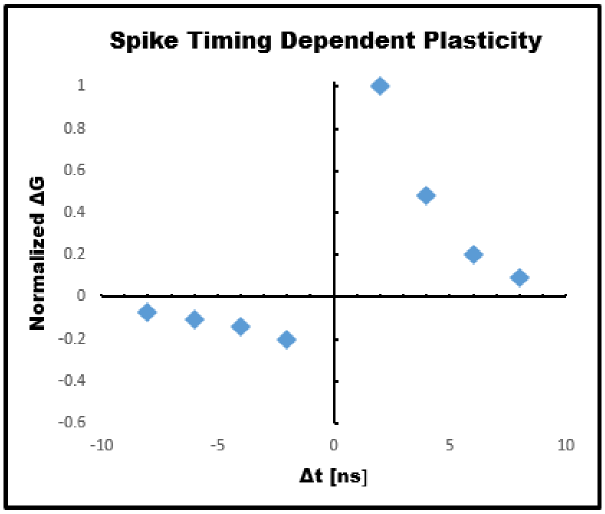


Figure 26: ΔG vs. Δt curve derived from SPICE simulations of the pulse mapping scheme shown in Figure 25.

The use of antisymmetric exponential spikes as artificial action potentials has been proposed by circuit designers and device engineers [12], [41]. Under such configuration (Figure 27), the effective voltage across the memristor varies according to the timing of the presynaptic and postsynaptic potentials. It remains a challenge to design reliable spike generation circuits with low output impedance to drive antisymmetric action potentials across the memristor. A neuromorphic system that interacts with spikes mimicking biological action potentials can eliminate the need for a separate synaptic update circuit, as the action potentials themselves are sufficient to cause STDP. This type of design can reduce the number of transistors per neuron significantly, paving way for future developments in large scale neural network systems.

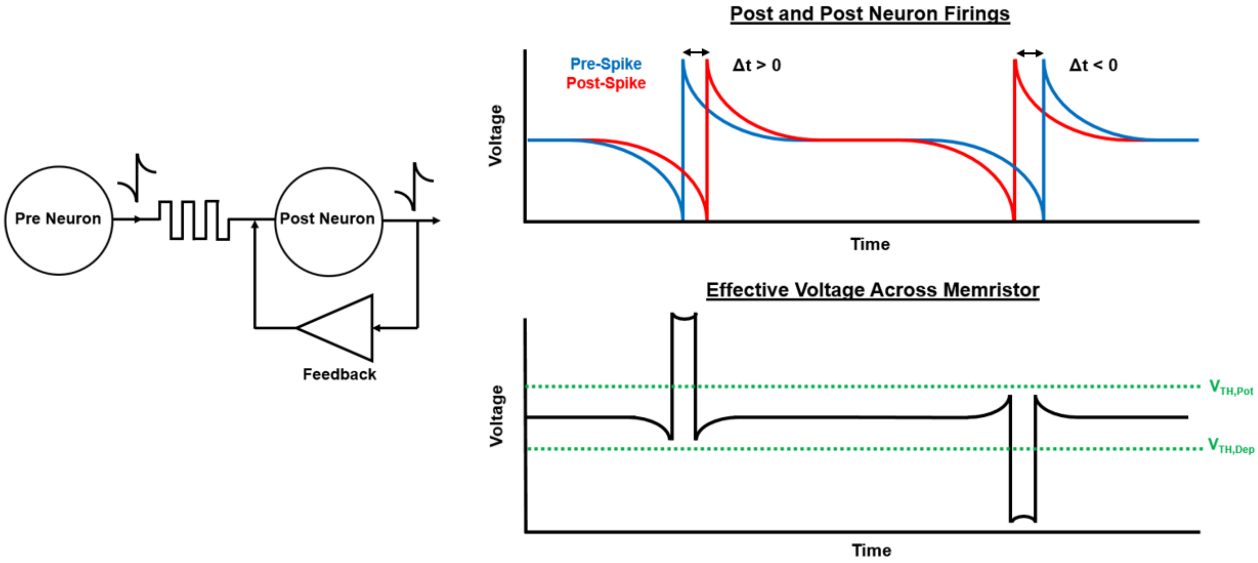


Figure 27: Future neuromorphic systems incorporating action potential-like waveforms [41]. The effective voltage across the memristor will have similar characteristics as mapping scheme described in Figure 25. This type of implementation can reduce the number of transistors per neuron/synapse, and mimics biological neural behavior more closely.

References:

[1] B.H. Calhoun et al., “Modeling and Sizing for Minimum Energy Operation in Subthreshold Circuits,” *IEEE Journal of Solid-State Circuits*, vol.40, no.9, pp.1778-1786, 2005.

[2] G. J. Siegel, B. W. Agranoff, and R. W. Albers, Eds., Basic Neurochemistry: Molecular, Cellular and Medical Aspects”. Philadelphia, PA: Lippincott-Raven, 1999.

[3] Beiye Liu; Miao Hu; Hai Li; Zhi-Hong Mao; Yiran Chen; Tingwen Huang; Wei Zhang, "Digital-assisted noise-eliminating training for memristor crossbar-based analog neuromorphic computing engine," *Design Automation Conference (DAC), 2013 50th ACM / EDAC / IEEE*, vol., no., pp.1,6, May 29 2013-June 7 2013.

[4] P. Sheridan, M. Wen, W. Lu, "Pattern recognition with memristor networks," *Circuits and Systems (ISCAS), 2014 IEEE International Symposium*, vol., no., pp.1078,1081, 1-5 June 2014.

[5] Koyanagi, M.; Nakagawa, Y.; Kang-Wook Lee; Nakamura, T.; Yamada, Y.; Inamura, K.; Ki-Tae Park; Kurino, H., "Neuromorphic vision chip fabricated using three-dimensional integration technology," *Solid-State Circuits Conference, 2001. Digest of Technical Papers. ISSCC. 2001 IEEE International*, vol., no., pp.270,271, 7-7 Feb. 2001.

[6] R. Douglas, M. Mahowald, and C. Mead, “Neuromorphic analogue vlsi,” Annu Rev Neurosci, vol. 18, pp. 255–281, 1995.

[7] P. Lennie, “The cost of cortical computation,” Curr. Biol., vol. 13, no. 6, pp. 493–497, Mar. 2003.

[8] J. Seo, B. Brezzo, Y. Liu, B. Parker, S. Esser, R. Montoye, B. Rajendran, J. Tierno, L. Chang, and D. Modha, “A 45 nm CMOS neuromorphic chip with a scalable architecture for learning in networks of spiking neurons,” in Proc. IEEE CICC, 2011, pp. 1–4.

[9] Andrew S. Cassidy, Julius Georgiou, Andreas G. Andreou, Design of silicon brains in the nano-CMOS era: Spiking neurons, learning synapses and neural architecture optimization, Neural Networks, Volume 45, September 2013, Pages 4-26, ISSN 0893-6080.

[10] Merolla, P.; Arthur, J.; Akopyan, F.; Imam, N.; Manohar, R.; Modha, D.S., "A digital neurosynaptic core using embedded crossbar memory with 45pJ per spike in 45nm," *Custom Integrated Circuits Conference (CICC), 2011 IEEE* , vol., no., pp.1,4, 19-21 Sept. 2011.

[11] D. Strukov, G. Snider, G. Stewart, and R. Williams, “The Missing Memristor Found,” Nature, vol. 453, pp. 80–83, 2008.

[12] Rajendran, *et al.* "Specifications of Nanoscale Devices and Circuits for Neuromorphic Computational Systems," *Electron Devices, IEEE Transactions on*, vol.60, no.1, pp.246, 253, 2013

[13] McCulloch, W. S., and Pitts, W., “A Logical Calculus of Ideas Imminent in Nervous Activity,” Bulletin ofMathematica2 Biophysics, 5, 115-133, 1943.

[14] Sherrington C S. “The Integrative Action of the Nervous System,” New York: Charles Scribner’s Sons, 1906.

[15] Hao J, Wang XD, Dan Y, Poo MM, Zhang XH, “An arithmetic rule for spatial summation of excitatory and inhibitory inputs in pyramidal neurons”. Proc Natl Acad Sci USA 2009, 106:21906-21911.

[16] Hodgkin A.L., and A.F. Huxley, “Action potentials recorded from inside a nerve fibre,” Nature 144, 710–711, 1939.

[17] John A. White, “Action Potential”, In *Encyclopedia of the Human Brain*, edited by V.S. Ramachandran, Academic Press, New York, 2002.

[18] Hodgkin, A. L., and A. F. Huxley, “A Quantitative Description of Membrane Current and Its Application to Conduction and Excitation in Nerve,” *The Journal of Physiology* 117.4 (1952): 500–544.

[19] Hebb, D.O, “The Organization of Behavior” (New York: Wiley), 1949.

[20] Bi, G.Q., and Poo, M.M., “Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type,” J. Neurosci. 18, 10464–10472, 1998.

[21] Feldman, D.E. “The spike-timing dependence of plasticity,” Neuron 75, 556–571, 2012.

[22] L. Chua, “Memristor—The missing circuit element,” IEEE Trans. Circuit Theory, vol. 18, no. 5, pp. 507–519, Sep. 1971.

[23] Chua, L.O.; Sung Mo Kang, "Memristive devices and systems," *Proceedings of the IEEE* , vol.64, no.2, pp.209,223, Feb. 1976

[24] Pan, F., Gao, S., Chen, C., Song, C. & Zeng, F. “Recent progress in resistive random access memories: Materials, switching mechanisms, and performance”. Mat. Sci. Eng. R 83, 1-59 10

[25] Govoreanu, B *et al*., "10×10nm^2 Hf/HfOx crossbar resistive RAM with excellent performance, reliability and low-energy operation," *Electron Devices Meeting (IEDM), 2011 IEEE International* , vol., no., pp.31.6.1,31.6.4, 5-7 Dec. 2011

[26] Chung-Wei Hsu; I-Ting Wang; Chun-Li Lo; Ming-Chung Chiang; Wen-Yueh Jang; Chen-Hsi Lin; Tuo-Hung Hou, "Self-rectifying bipolar TaOx/TiO2 RRAM with superior endurance over 10^{12} cycles for 3D high-density storage-class memory," *VLSI Technology (VLSIT), 2013 Symposium on* , vol., no., pp.T166,T167, 11-13 June 2013

[27] Yu, S.; Guan, X.; Wong, H.-S. P. “On the Switching Parameter Variation of Metal Oxide RRAM - Part II: Model Corroboration and Device Design Strategy”. IEEE Trans. Electron Devices 2012, 59, 1183–1189.

[28] Chang T, Jo S H, Kim K-H, Sheridan P, Gaba S and Lu W 2011 “Synaptic behaviors and modeling of a metal oxide memristive device” Appl. Phys. A 102 857–63.

[29] M. A. C. Maher, S. P. Deweerth, M. A. Mahowald, and C. A. Mead, “Implementing neural architectures using analog VLSI circuits,” IEEE Trans. Circuits Syst., vol. 36, no. 5, pp. 643–652, May 1989.

[30] Predictive Technology Model: 90nm_bulk. http://ptm.asu.edu/

[31] Indiveri G, *et al*. “Neuromorphic silicon neuron circuits”. Front Neurosci 5:73, 2011.

[32] Bichler O, *et al.* “Pavlov’s dog associative learning demonstrated on synaptic-like organic transistors Neural Comput”. 24 1–18, 2012.

[33] Ziegler, M. et al. “An electronic version of Pavlov’s dog”. Adv. Funct. Mater. 22, 2744–2749, 2012.

[34] Pavlov IP: “Conditioned Reflexes”. New York, Oxford University Press, 1927

[35] Yu, Shimeng, et al. "HfOx-based vertical resistive switching random access memory suitable for bit-cost-effective three-dimensional cross-point architecture." *ACS nano* 7.3 (2013): 2320-2325

[36] Kim, Kuk-Hwan, et al. "A functional hybrid memristor crossbar-array/CMOS system for data storage and neuromorphic applications." *Nano letters* 12.1 (2011): 389-395.

[37] Jung, R.; Brauer, E.J.; Abbas, J.J., "Real-time interaction between a neuromorphic electronic circuit and the spinal cord," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on* , vol.9, no.3, pp.319,326, Sept. 2001

[38] Wang, Sun-Chong. "Artificial neural network." *Interdisciplinary Computing in Java Programming*. Springer US, 2003. 81-100.

[39] Jiang, Zizhen, et al. "Verilog-A Compact Model for Oxide-based Resistive Random Access Memory (RRAM)." *Simulation of Semiconductor Processes and Devices (SISPAD), 2014 International Conference on*. IEEE, 2014.

[40] Jo, Sung Hyun, et. al, "Nanoscale Memristor Device as Synapse in Neuromorphic Systems". *NanoLetters* 2010 *10* (4), 1297-1301

[41] Wang I-Ting, *et al.*, "3D synaptic architecture with ultralow sub-10 fJ energy per spike for neuromorphic computation," *Electron Devices Meeting (IEDM), 2014 IEEE International* , vol., no., pp.28.5.1,28.5.4, 15-17 Dec. 2014