STDP Implementation Using

CBRAM Devices in CMOS

by

Mahraj Sivaraj

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved June 2015 by the
Graduate Supervisory Committee:

Hugh Barnaby, Chair
Michael Kozicki
Jenifer Blain Christen

ARIZONA STATE UNIVERSITY

August 2015

ABSTRACT

Alternative computation based on neural systems on a nanoscale device are of increasing interest because of the massive parallelism and scalability they provide. Neural based computation systems also offer defect finding and self healing capabilities. Traditional Von Neumann based architectures (which separate the memory and computation units) inherently suffer from the Von Neumann bottleneck whereby the processor is limited by the number of instructions it fetches. The clock driven based Von Neumann computer survived because of technology scaling. However as transistor scaling is slowly coming to an end with channel lengths becoming a few nanometers in length, processor speeds are beginning to saturate. This lead to the development of multi-core systems which process data in parallel, with each core being based on the Von Neumann architecture.

The human brain has always been a mystery to scientists. Modern day super computers are outperformed by the human brain in certain computations. The brain occupies far less space and consumes a fraction of the power a super computer does with certain processes such as pattern recognition. Neuromorphic computing aims to mimic biological neural systems on silicon to exploit the massive parallelism that neural systems offer. Neuromorphic systems are event driven systems rather than being clock driven. One of the issues faced by neuromorphic computing was the area occupied by these circuits. With recent developments in the field of nanotechnology, memristive devices on a nanoscale have been developed and show a promising solution. Memristor based synapses can be up to three times smaller than Complementary Metal Oxide Semiconductor (CMOS) based synapses.

In this thesis, the Programmable Metallization Cell (a memristive device) is used to prove a learning algorithm known as Spike Time Dependant Plasticity (STDP). This learning algorithm is an extension to Hebb's learning rule in which the synapses weight can be altered by the relative timing of spikes across it. The synaptic weight with the memristor will be its conductance, and CMOS oscillator based circuits will be used to produce spikes that can modulate the memristor conductance by firing with different phases differences.

ACKNOWLEDGEMENTS

I would like to first and foremost thank my advisor, Dr. Hugh Barnaby, for his invaluable guidance and technical discussions provided throughout the completion of work described in this thesis. I have learnt a great deal from him while being his student, teaching assistant and research student. I am deeply grateful to Dr. Jennifer Blain Christen who helped me tape out my circuits with ON semiconductor's educational program. I would also like to thank Debayan Mahalanabis, Sahil Shah and Jennifer Taggart, PhD students of my research group, for their technical contributions and assistance in the lab which helped present the results in this thesis. A special thanks goes for Wenhao "Vincent" Chen for fabricating the memristive devices on my chip.

Finally I would like to thank my father, mother and brother, for their continuous support. I will always be forever grateful for your support. Additionally I would like to thank all my close friends who have always been there to support me and intellectually challenge me.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION TO NEUROMORPHIC COMPUTING

1.1 *Historical Perspective*

Most modern computers and electronic systems used in almost in every aspect of our daily lives are based on the traditional Von Neumann architecture. This architecture proposed in 1945 (for the EDVAC), separates the memory and computation unit. A high speed clock is required for its operation making it a clock driven architecture. The reason for the Von Neumann architecture to be so popular and stand the test of time is the inexorable rise in the density of transistors in CMOS and reduction in cost over the years. This increasing number of transistors due to the aggressive down scaling of the feature size, allowed the Von Neumann Architecture to handle and process large amounts of data. However transistor scaling is reaching it limits. According to the 2011 ITRS (International Technology Roadmap for Semiconductors) [1] there is a need for new devices, and "beyond CMOS information processing technology" in which a new type of data representation may also be required. The reason for scaling to reach its limits is because the channel lengths are becoming just a few atoms wide, where quantum physics comes into play. This inhibits the transistor functioning as a switch. This limitation at the lowest abstraction level causes performance limitations of the architecture. The Von Neumann architecture itself suffers from a bottleneck called the Von Neumann bottleneck [2]. The processor cannot execute programs faster than the required data and instructions are fetched. Engineers got around this problem by using a memory hierarchy, such as cache and registers and more recently multicores on a single die. It had been speculated and even proved that an architecture which can process and store simultaneously at the same physical location consumes less power and is significantly

faster. These advantages formed the basis for searching newer devices, materials and system architectures to provide a whole new alternative to the conventional computing paradigms.

The human brain is an extraordinary computing machine. It constantly monitors and regulates vital biological systems and performs the computations required for day to day tasks such reading, writing or driving. These tasks and functions handled by a brain prove to be daunting even for modern supercomputers. A significant difference between any computer and a brain is the amount of power consumed by each core computational unit. The neuron serves as the computational unit in a brain and central processing unit (CPU) for a computer. The brain weighing approximately 1.5 Kg consumes about 20W of power while driving in rush hour in peak traffic. On the other hand, simulating a few seconds of rush hour driving with IBM's state of the art supercomputer requires 3 MW. It occupies 5500ft square of floor space and weighs in at 227 metric tons [3]. The brain is power efficient because neural spikes charge only a small fraction of a neuron as they travel. By contrast digital computers have their transmission lines at a certain voltage at all times. The brain's neurons operate in milliseconds, and are capable of image and pattern recognition tasks in the ten to hundred millisecond range. Powerful computers requires hours, if not days to process the same work load. This low power consuming and fast performing neural architecture is very much worth exploring in an effort to overcome the bottlenecks being faced by digital computers.

Neuromorphic computing aims to use VLSI and analog circuits to mimic biological neural systems such as the brain which is a colossal parallel network of biological processing elements: neurons. Each of these neurons is connected to thousands of other

neurons through conducting channels called synapses. These neurons fire after reaching a certain threshold. This makes the system event driven instead of a clock driven architecture used in digital computers. The human brain does not have a distinct separation between memory and a computation element as in the Von Neumann architecture. Synaptic connections in the brain retain their state in the absence of inputs. Such a behavior explains why the brain performs extremely well with complex tasks such as pattern recognition, maintaining and regulating vital human systems, and performing day to day activities while consuming a small amount of power. On the other hand, CMOS based computers take several hours, if not days to process the same tasks and can occupy thousands of square feet of land. CMOS technology suffers from its sensitivity to defects and failures. While aggressive scaling over 20 years has reduced the feature size of the MOS, newer issues relating to the robustness and reliability of CMOS technology has surfaced. It has been shown that neuromorphic circuits can adapt to such defects because of the inherent parallel nature of their architecture [4]. One of the major design challenges with neuromorphic circuits was the design of the synapse in CMOS. CMOS based synapses consumed more power and area in comparison to biological systems, making the Boolean based digital computer far more efficient.

With advancements in nanotechnology, newer devices have re-ignited interest in bridging the gap between a neural and digital architecture. The memristor device, a contraction of "memory resistor" was developed by a team led by R. Stanley Williams at the Hewlett-Packard (HP) Laboratories, Palo Alto, California [5] in 2008 although its existence had been postulated as the fourth basic element in 1971 by Leon Chua [6]. The memristor is a passive circuit element with a unique I-V relationship resembling a

pinched hysteresis loop. The memristor operates analogously to the neural synapse. This helps alleviate the density issue, mentioned previously, with traditional CMOS circuits. Memristors show much promise to reach the densities found in neural systems. Hardware architectures shown by Turel [7] and Zhao [8], the "Crossnets" approach, provides a solution to the design problem and method of integrating resistive nanoscale devices in a crossbar topology with CMOS circuitry to design neuromorphic circuitry. Crossnets have been shown to dramatically increase device integration [9]. It can be shown that these neuromorphic networks can maintain relationships based on learning rules [10]. Due to the plethora of neurons, their behavior and synaptic weight combinations can produce a state which is unique at any given point in time. The memristor based synapse can be as small as three orders compared to a CMOS-based design as well as being both a non-volatile and a variable resistive element [11].

1.2 *The Neuron*

The neuron is a biological cell which is electrically excitable. It is the primary mode of transmitting information through electrical and chemical processes. Neurons are connected to each other via a synapse. Each neuron connects to other neurons forming the neural network. Neurons are the core elements of the neural system. The nervous system consists of the brain, spinal cord and the peripheral nervous system. Motor neurons or neurons from sensory organs which respond to touch, sound or light depending in which organ they reside, send signals to the brain via the spinal cord. Neurons can also receive signals from the spinal cord or brain which can affect muscle contractions, glandular outputs and other neurons within the same neural network.

Neurons can be very diverse in nature but there is a general description to identify a neuron. The neuron is typically divided into three parts, the cell body or soma, dendrites and axon. The cell body is usually the smallest in size with the axon and dendrites protruding out from the soma. The soma contains the nucleus and is where protein synthesis occurs. The axon and dendrites are filament-like structures. Dendrites branch profusely, thinning out with each branch. They reach just a few hundred micrometers from the soma, giving rise to a “dendritic tree”. The axon leaves the cell at the axon hillock and extends out over very large distances (as much as 1 meter in humans), branching out in the hundreds. The axon hillock has the greatest density of voltage dependent sodium channels. Electrical signals from other neurons in the neural network are received by the cell body and dendrites. A typical synapse is formed between the axon of one neuron and the dendrite, or soma, of another neuron in the neural network. Fig. 1.1 illustrates the components of the neuron system.

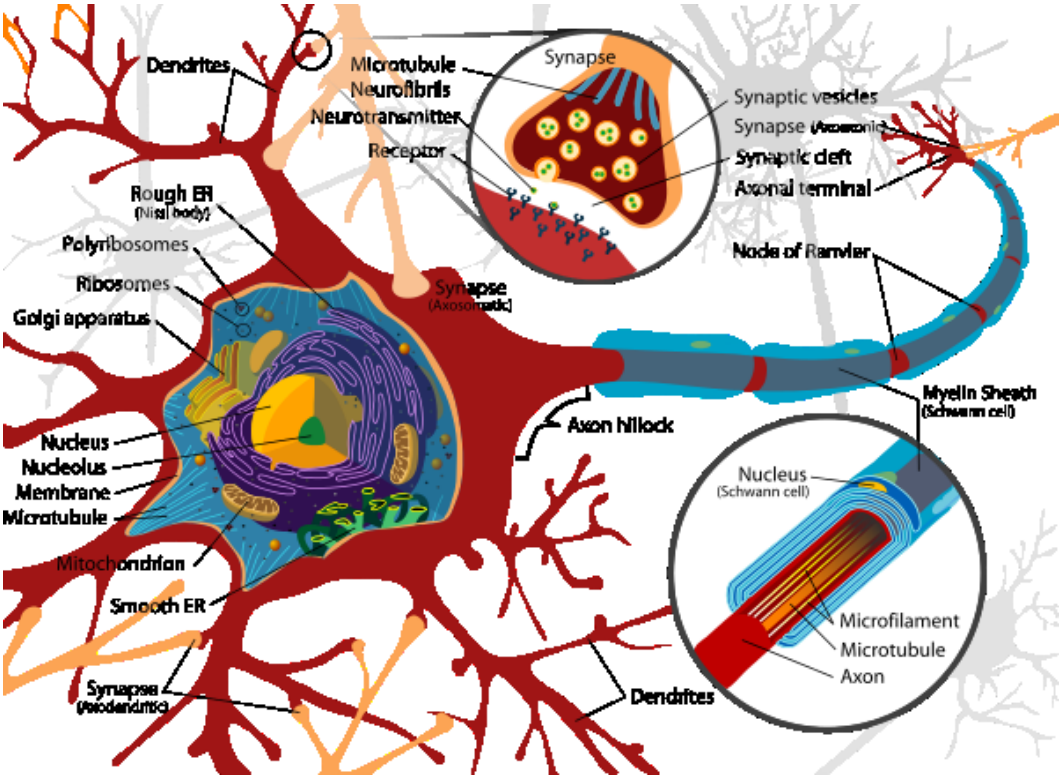

Fig. 1.1. Diagram of typical neuron system [51]

Synaptic signals fall into two types: excitatory and inhibitory. If the overall excitation over a short period of time is large enough, a short spike or pulse called an action potential originates at the cell body and quickly moves across the axon, activating synapses of other neurons. This is known as salutatory conduction. The cell body and the axon contain voltage-gated ion channels in their cell membrane which allows the neuron to generate and carry electrical signals. All neurons are electrically excitable and maintain a voltage gradient across their membrane. The potential across the membrane is maintained by ion pumps which are metabolically driven. Electrical signals are produced and propagated with the help of the ion channels embedded in the membrane to produce an intracellular-extracellular concentration of charge carrying ions, namely sodium (Na^{+}), potassium (K^{+}), chloride (Cl^{-}), and calcium (Ca^{2+}). To activate a neuron, electrical

stimuli may be applied to it. These include pressure, stretching of the neuron, chemical transmitters and changes in electrical potential across the cell membrane [12]. The membrane potential changes when the ion channels within the cell membrane open allowing a flow of ions across it. Changes in voltage across this membrane modify the function of the voltage dependent ion channels. If the voltage changes are significant, then an all-or-none pulse, the action potential, is created.

For a neuron to respond, it must *completely* respond. In other words the strength by which the nerve responds (the magnitude of the action potential) is not a simple analog function of the strength of the stimulus. The stimulus must surpass a threshold potential for the neuron to respond. If the threshold potential is not met by the stimuli, a response is not produced. This property is known as the all-or-none relationship. Stimuli which are too weak to produce any spike create a local electrotonus. The magnitude of the electric potential of this electronus progressively increases with the strength of the stimuli. This means that a greater intensity of stimulation does not produce a stronger response but rather produces a faster firing rate. If a nerve containing several thousands of neurons is excited with a progressively increasing stimulus, the neurons which have their threshold potentials satisfied by the stimuli only respond. Increasing the magnitude of the stimuli further increases the response of the entire nerve. Fig. 1.2 shows the equivalent circuital model of a neuron [33].

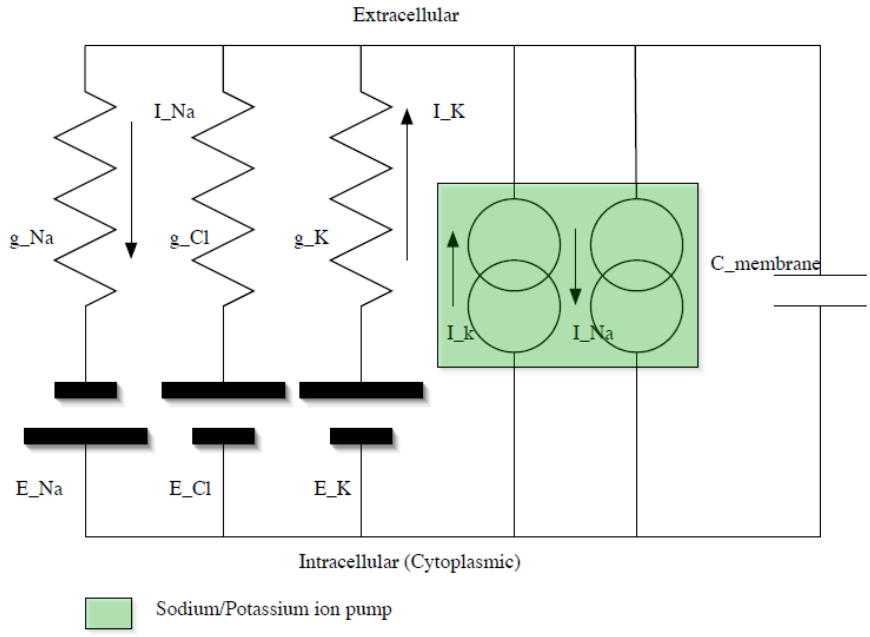

Fig. 1.2. Equivalent circuital model of a neuron model. The arrows indicate the direction of movement of ions [33].

In the absence of inputs, a potential inside the neuron relative to the outside will be maintained by the membrane. This resting potential is about -65mV, which can vary across different neurons. An increase in the potential is known as depolarization and a decrease is referred to as hyperpolarisation.

The human brain contains an enormous number of synapses. There are about one hundred billion neurons each having about 7000 synaptic connection with other neurons. Synaptic plasticity is the ability of the synapse to strengthen or weaken depending on their activity. The plasticity of the synapse can also change with the number of neurotransmitters in the synapse. The plasticity of both excitatory and inhibitory synapses depends on the post synaptic calcium release [14]. Synaptic plasticity forms the basic mechanism for learning and memory. Actual synaptic operation is very complex. Many operate by releasing microscopic bubbles called vesicles of a chemical that aids in

conductance, known as a neurotransmitter into the cleft (the space between the presynaptic axon and postsynaptic dendrite). A synapse becomes excitatory when the potential alteration is depolarizing and the synapse is inhibitory when hyperpolarization occurs. When alterations are made they are linearly summed on the dendrites. This is true for a certain range of potentials above which the linearity ceases to exist. This nonlinearity may occur at some part of the dendrite. Studies [15] have shown that nonlinear processing occurs on the neuron providing other neurons with considerable processing power.

The potential at the axon hillock is of great importance. This acts as a trigger zone, where there is a large concentration of sodium channels. As a result, when the potential increases beyond a threshold (usually about -50mV), voltage sensitive ion channels open, allowing an influx of Na^+ ions increasing the depolarization. This results in more channels opening up, which causes in an exponential rise in the membrane potential. The increased depolarization causes the sodium ion channels to close and another set of channels to open, causing efflux of K^+ ions causing the potential to drop as fast as it increased. This phenomenon produces a spike like wave that rapidly propagates across the axon. The sodium channels can be immediately reopened creating a maximum limit at which the spike can be generated. This delay is caused by the sodium channels and gives rise to a refractory period, during which the neuron cannot fire again even if inputs are being constantly applied. Fig.1.3 illustrates the depolarization on an axon [33].

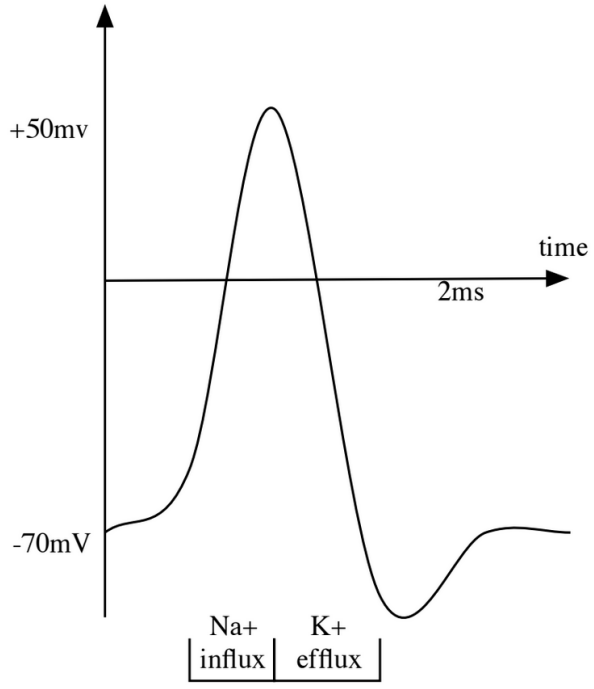

Fig. 1.3. Depolarization on an axon during an action potential model [33].

Neural systems alter their response to their inputs, and this adaptation occurs so that identical inputs can be differentiated if they occur at different times. This adaptation can occur over different time scales. In mammals synapses are formed but usually do not last forever. Hormones and diffusible neurotransmitters can cause the growth or decay of synapses.

Neural models focus mainly on the synaptic changes. Synaptic changes can be short term (dynamic synapse behavior) or long term. Long term changes are termed long-term potentiation (long term strengthening of the synapse) and long-term depression (long term weakening of the synapse). These mechanisms form the basis for neural learning models.

This thesis aims to demonstrate Spike-Timing-Dependent Plasticity (STDP) with CMOS and memristors made from silver chalcogenide films integrated on the same die. The Leaky Integrate and Fire (LIF) circuit will be used create neural spikes required to prove this learning rule. Chapter 1 will provide an overview of neuromorphic engineering. Chapter 2 will discuss STDP and the memristor. Chapter 3 will discuss the CMOS aspects of this work, with an in depth analysis of spiking neuron circuit design. Chapter 4 will cover the analysis and present results of this study. Chapter 5 will present conclusions as a result of this study.

CHAPTER 2 SPIKE TIME DEPENDENT PLASTICITY

2.1 Spike Time Dependent Plasticity

Abstract models of synaptic plasticity aim to demonstrate the learning phenomenon by basing these models on the relative timing between presynaptic and postsynaptic neurons. When choosing a synaptic plasticity model, it is important to consider the target's application requirements. Examples of such models are spike-timing-dependent plasticity (STDP) and the triplet-based STDP (TSTDP). Other models aim to explain neural system behavior in more detail by attempting to explain neuroscience experimental data by adopting neuron and synaptic state variables. Table 2.1 shows the various synaptic plastic rules that exist, ranging to abstract to physically implementable models. Comparisons are made in terms of synaptic variables used.

Table 2.1 Comparison of STDP models [16]

Model	Spike Time	Membrane Potential	Calcium
Pair Based STDP	Yes	No	No
Triplet Based STDP	Yes	No	No
Spike Driven Synaptic Plasticity	Yes	Yes	Yes
Modified ion channel based plasticity	No	Yes	Yes
Iono-neuromorphic intracellular calcium mediated plasticity	No	Yes	Yes

Donald Hebb wrote in his book "*The Organization of Behavior"* about a learning rule called the Hebbian Theory. This theory describes a mechanism for synaptic plasticity. An increase in the synaptic efficacy is caused by the presynaptic neuron's constant stimulation of the postsynaptic neuron. When an axon of cell *A* is near enough to excite a cell *B* and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that *A*'s efficiency, as one of the cells firing *B*, is increased." [17]. He emphasized that causality should exist such that the cell *A* needs to fire before cell *B*. This particular emphasis of his theory created a refined theory known as the Spike-Timing-Dependent Plasticity (STDP) which requires temporal precedence [18].

STDP imposes tighter temporal correlations between the pre-and postsynaptic neuron's spike. This biological process works as follows. Recurring presynaptic spikes which arrive a few milliseconds before postsynaptic action potential causes long-term potentiation (LTP) of the synapse. Conversely, recurring spike arrival after postsynaptic spikes leads to long-term depression (LTD). The STDP function relates the change in the weight of the synapse as a function of the relative timing of pre- and postsynaptic action potentials. The change of the STDP function with relative timing of pre- and postsynaptic spikes enables a temporal coding scheme in the Kilohertz range.

The experimental form of the STDP protocol is as follows. A synapse is activated by stimulating a presynaptic neuron either before or after making the postsynaptic neuron fire. After this, the weight of the synapse is measured as the amplitude or initial slope of the postsynaptic potential. The learning window can then be plotted as the change of the

synaptic weight as a function of the relative timing between the presynaptic spike arrival and postsynaptic firing.

The basic model of the STDP learning rule is explained below. The change in weight of the synapse Δw_x of a presynatpic neuron x is determined by the relative timings between the presynaptic spike arrival and postsynaptic spike. Let the presynaptic spike arrival times at synapse x be t_x^f where is f is a real number. Similarly, t_y^n where n is an integer are the spike times of the postsynaptic neuron. The total change in weight Δw_x caused by this protocol [48] is

$$\Delta w_x = \sum_{f=1}^{N} \sum_{n=1}^{N} W\left(t_y^n - t_x^f\right). \tag{2.1}$$

The function *W(x)* is called the learning window and can be expressed as

$$W(x) = A + exp\left(-\frac{x}{\tau} +\right) \qquad for \quad x > 0 \tag{2.2}$$

$$W(x) = -A - exp\left(\frac{x}{\tau} -\right) \qquad for \quad x < 0 \tag{2.3}$$

[48], where $A+$ and $A-$ can depend on the present values of the synaptic weight w_x. The time constants are on the order 10 ms.

It becomes clear that rate-based Hebbian learning is inherently unstable because synaptic inputs that drive a neuron become stronger. Although this instability is required to make the neuron sensitive to correlations in the input, there is a rapid increase in the firing rate of neuron. For practical models renormalization of the weights as well as upper and lower bounds are introduced to control the rapid growth of firing rates. STDP is sensitive to spike arrivals and their timing in the millisecond range and hence it can be used for temporal coding schemes. Synaptic weights are a continuous variable in most

models although it can be discrete. There are several advantages with being discrete such as bistability, which would ensure long-term stability (weeks to years).

2.2 The Memristor

In 1971, Leo Chua had postulated the fourth circuital element called the memristor. The three circuital elements, resistor, capacitor and resistor have well defined relationships between the fundamental electricity and magnetism variables, namely: current i, voltage v, the charge q and the flux linkage φ. The resistor has a linear relationship between current i and voltage v. The capacitor has a linear relationship formed between voltage v and charge q. The inductor forms its linear relationship with flux linkage φ and current i. For the sake of completeness Leo Chua proposed a two terminal device which relates flux linkage φ and charge q. He called this circuit element the memristor because it behaves like a nonlinear resistor with memory [6]. The equation that establishes the relation between flux linkage φ and charge q is given by [6]

$$d\varphi = \mathrm{M}\, \mathrm{dq} \tag{2.4}$$

Substituting the flux with the time integral of the voltage, and charge with the time integral of current [6],

$$M\big(q(t)\big) = \frac{d\varphi\ /\mathrm{dt}}{dq/dt} = \frac{v(t)}{i(t)} \tag{2.5}$$

The voltage across the memristor is given by [6],

$$v(t) = M(q(t))i(t) \tag{2.6}$$

where

$$M\big(q(t)\big) = d\varphi\,(\mathrm{q})/\mathrm{dq}. \tag{2.7}$$

The current through the memristor is given by [6],

$$i(t) = W\big(\varphi(\mathrm{t})\big)\mathrm{v}(\mathrm{t}) \tag{2.8}$$

where

$$W\big(\varphi(\mathrm{t})\big) = dq(\varphi)/\mathrm{d}\varphi \tag{2.9}$$

M(q) is known as the incremental memristance and W(φ) known as the incremental memductance. From these set of equations it can be observed that incremental memristance (memductance) at any instant of time is equal to the integral of the memristor current (voltage) from t = 0 to t_o. What this means is that the memristor behaves like an ordinary resistor at t_o, but its resistance (conductance) depends on the history of the current (voltage) through (across) it. This means that device has a memory aspect to it, and can remember its most recent resistance when switched on [6]. The memristor is physically a two terminal device having a low conductive material in between an active electrode (anode) and a passive electrode (cathode).

The memristor has an interesting property of a pinched hysteresis effect, shown in Fig. 2.1. For a current-controlled memristive element, let the input be x(t) (being the current) and the output be y(t) (being the voltage). The slope of this curve gives the resistance. The change in slope of the pinched hysteresis shows different resistance states of the memristive device.

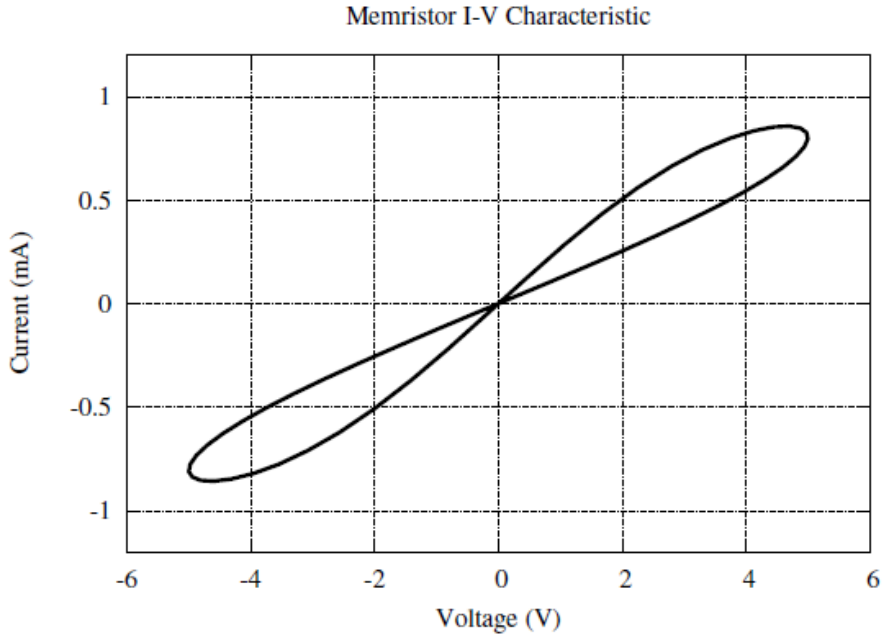

Fig. 2.1. I-V plot of a memristor showing hystersis [48].

2.3 Programmable Metallization Cell (PMC)

Only a few groups have managed to characterize their memristor for SPICE modeling. The memristor used in this thesis is the Programmable Metallization Cell (PMC) or Conductive Bridging RAM. Developed by Dr. Michael Kozicki of Arizona State University at the university's Center for Applied Nanoionics, the PMC can be used as a memory or synapse element in neuromorphic circuits. The PMC is a resistive switching memory which works via the transport of metal ions moving through a solid electrolyte and forming a conductive filament between two electrodes. PMCs have a low programming current and display multilevel resistance state capabilities [23]. Fig. 2.2 [23] shows the cross-sectional representation of the PMC.

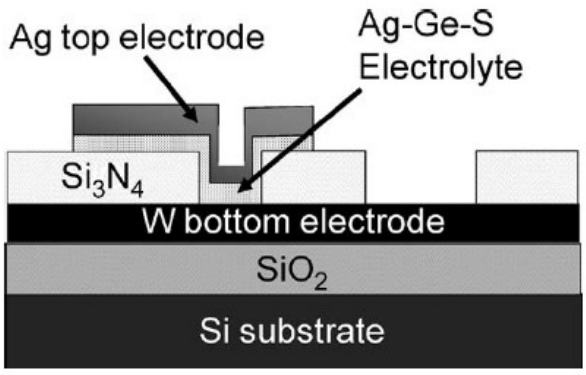

Fig. 2.2. Schematic view of a PMC [23].

The PMC works on the physical re-location of ions in a solid electrolyte. A PMC memory cell consists of two solid electrodes, an inert metal such as tungsten and an active electrode such as silver. The two electrodes are placed with a thin film of chalcogenide glass (ChG) electrolyte between them, as shown in Fig. 2.2. The ChG film in the PMC device used in this work is made of $AgGe_{0.3}Se_{0.7}$. The cell has metal-insulator-metal structure, since the ChG layer is inherently insulating, i.e., highly resistive.

When a few hundred millivolts is applied from anode to cathode, oxidation of the anode occurs, which releases metal ions (Ag+) into the ChG film. Ions that transport to the inert cathode are reduced and electro-deposited there. Charge neutrality is maintained due to the balance between oxidation and reduction. Over a short period of time, the ions flowing across the ChG solid electrolyte form a nanowire between the electrodes. This nanowire reduces the overall resistance of the PMC. The formation of the nanowire or conductive filament is shown in Fig. 2.3. The nanowire need not be a continuous; it can be a chain of electrodeposits islands or nanocrystals.

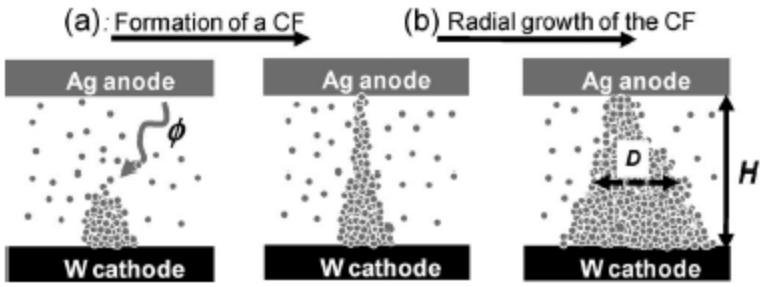

Fig. 2.3. Schematic diagram showing (a) formation of conductive filament and (b) radial growth of the conductive filament [23].

The resistivity of the electrodeposit is about eight orders of magnitude lower than that of the surrounding electrolyte. Once the electrodeposit has formed, the cell resistance falls considerably. The final resistance value is reached once the current flowing through the device reaches the limit of the source. When this happens the voltage drops below the threshold for electro deposition. It should be noted that the threshold for electro deposition falls once the deposition has started to form [49].

A device with an electrolyte thickness of 50 nm will switch to a low resistance state within tens of nanoseconds. When a reverse bias is applied, the inert cathode becomes the new anode and the electrodeposit will dissolve [49]. Reading the state of the PMC requires a voltage reading across it. This can be achieved by means of a transistor acting as a switch. The PMC's basic operation is illustrated in Fig 2.4 [23]. The current-voltage characteristics shown were obtained are measurements of a PMC with a via width of 1 μm. An external compliance current of 100 μA is set to prevent any damage to the device, and set the low resistance level. The reset state or high resistance state can be achieved by applying a negative voltage of about -0.2 V [23].

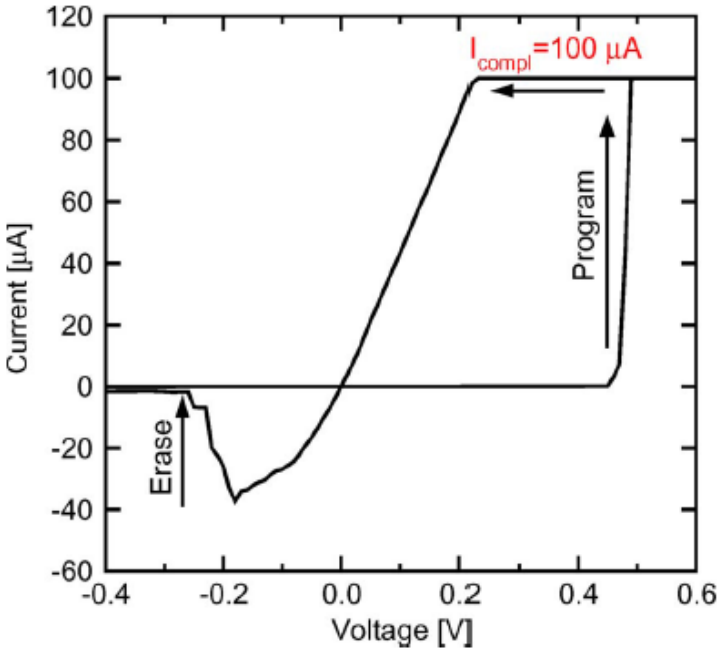

Fig. 2.4. I-V characteristics of a PMC [23].

A 1T1R (one transistor- one resistor device) can be used to program the device, shown on the inset in Fig. 2.5. The cell resistance R_C is initially very large, most of the applied voltage V_A drops across the device, forcing the transistor in the triode region. As the programming occurs and R_C drops, the transistor is driven into saturation. This forces a constant current through the cell, acting like a compliance current [23]. Fig. 2.5 shows the effect of the pulse width and applied voltage on the programming window. It can be seen that a larger voltage (V_A) allows a faster programming time. If V_A is 1.5V, a pulse width in the order of milliseconds is required for the conductive filament formation whereas if V_A is 2V, a pulse width in the order of nanoseconds suffices to initiate a drop in the resistance of the device. A larger compliance current allows for a lower resistance programming. The initial rapid fall of resistance is caused by the formation of the

filament and the gradual decrease of the resistance that follows it is caused by the radial growth of the conductive filament [23].

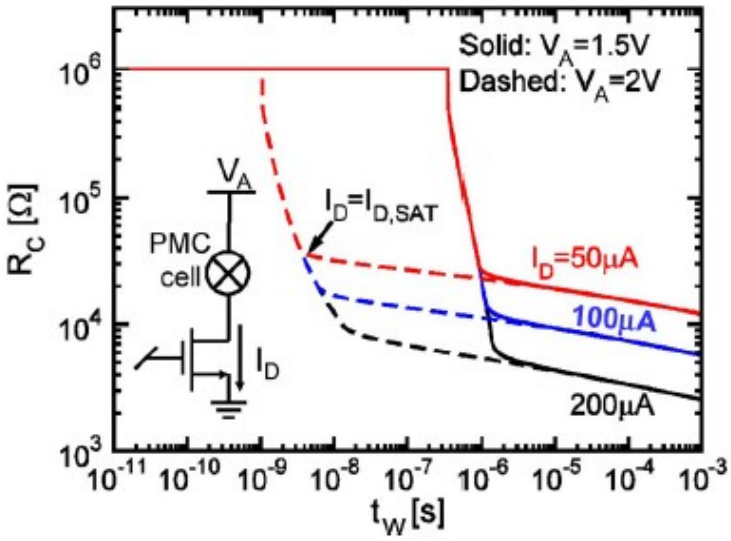

Fig. 2.5. Resistance of the PMC as a function of the applied pulse width for different I_D values [23].

Fig. 2.6 shows a current-voltage plot of a 240 nm diameter device with a 60 nm thick Ag-Ge-S electrolyte which has been annealed at 300 °C. Fig. 2.7 shows the resistance versus voltage plot of the same device. The voltage is swept in both figures from -1.0 V to +1.0 V and back to -1.0 V with a current limit of 10 μA. It can be seen that around 450 mV the device switches from an off-state resistance (Roff) which is in the order of 10^{11} Ω to an on-state resistance (Ron) in the order of 10^3 Ω. As mentioned previously, once electro deposition occurs, the threshold for further electro deposition drops [47].

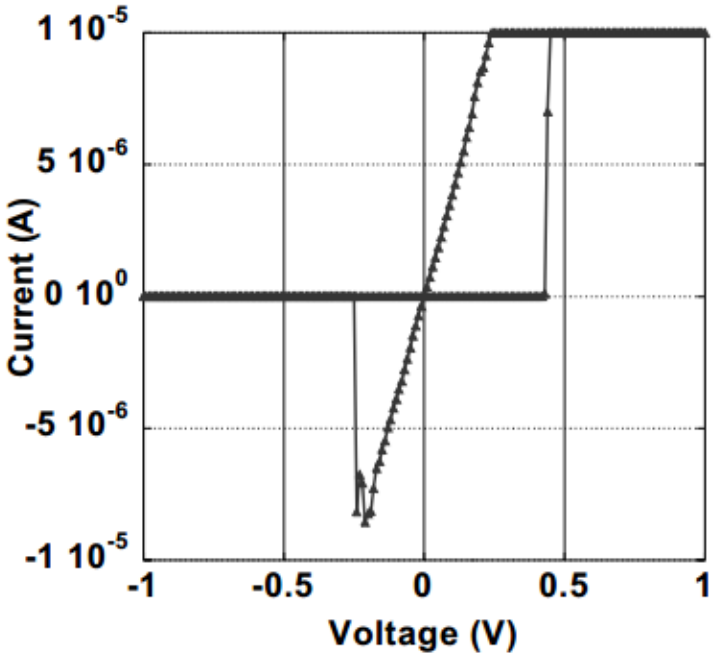

Fig. 2.6. I-V plot of 240nm diameter device with a 60nm thick Ag-Ge-Si electrolyte annealed at 300 °C. Voltage swept between -1 to 1 V with a 10 µA compliance current [49].

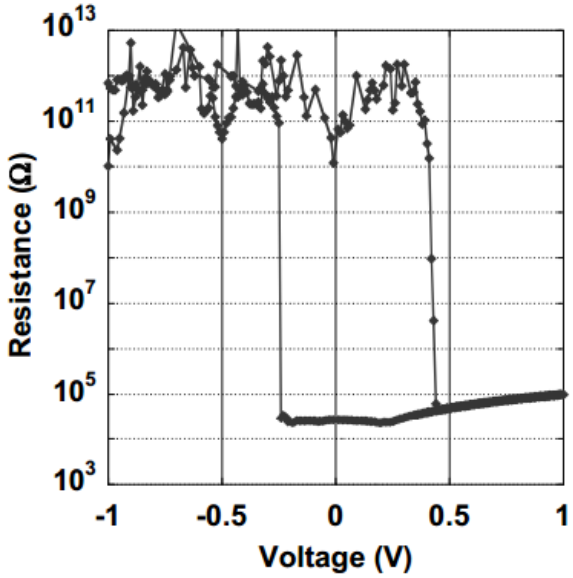

Fig. 2.7. Resistance vs Voltage plot of the device of Figure 2.6 [49].

2.4 SNNs and STDP

Spiking Neural Networks (SNNs) are closely associated with biological neural networks. As discussed before, the action potential of neurons in the brain are used to process information in biological neural networks. SNNs emulate this neural network. Theoretical results show that temporal coding of these spikes allows for a great amount of information processing [25]. The widely used neuron model in CMOS is the Leaky-Integrate-and-Fire Neuron (I&F). The silicon based I&F neuron integrates incoming currents with the help of a capacitor. Once this voltage (membrane potential) reaches a spiking threshold potential, a spike is fired, after which no more firing takes place. This models the refractory period of the spike of the neuron, which cannot fire successively. Here, the leaky integrate and fire circuit will be used to model the neuron.

Neural networks in the brain are capable of learning by "programming" the synapses. The synapses state or weight changes in response to the neurons behavior. The equation relating the weight changes of the synapse (state change) is given by the Hebbian learning rule as mentioned before. The Hebbian rule as function of pre- and post-synaptic spikes can be written as [25]

$$\frac{d\omega}{dt} = \mu X_{pre} X_{post}. \qquad (2.10)$$

In equation (2.10), μ is the learning rate. Xpre and Xpost are pre- and post-synaptic spikes, respectively. The equation relates the rate of change of the synaptic weight of the neuron. Although this is valid, there is no order of the spikes specified. The learning rate is used to scale the individual magnitude of the weight changes, $0 \leq \mu \leq 1$ [25].

PMCs find their use in making synapses for neuromorphic circuits. Neuromorphic circuits mimic biological synapses by using a crossbar network of nanodevices. This integration of CMOS and semiconductor, nanowire and or molecular circuits is termed as "CMOL" integrated circuits. CMOL circuits are suitable for implementing defect-tolerant architectures including embedded, stand alone memories and artificial neural networks (ANNs) [20]. Distributed Crossbar Networks ("CrossNets") are found to be suitable for CMOL implementation. Fig. 2.8 shows the I-V curve of an ideal synaptic switch, equivalent circuit model of a synaptic switch and array of synaptic switches. Here the synaptic connections are used as switches, having only two states, ON and OFF. The CMOS neuron circuits lies at the bottom layer and the two-terminal synaptic switches are formed at each crosspoint.

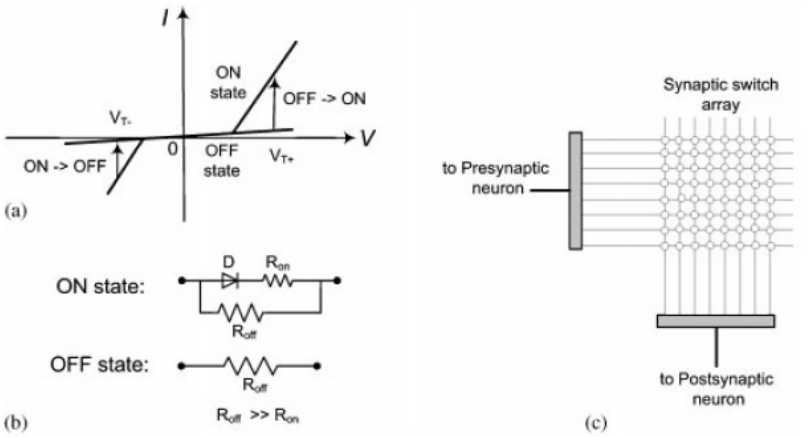

Fig. 2.8. (a) I-V plot of a synaptic switch, (b) equivalent circuit diagrams of the ON and OFF state, (c) square array of synaptic switches [25].

The switches at the crosspoints behave as a programmable memory cell, having diode like characteristics at low voltages and switch like characteristics are higher applied voltages. Turning the synaptic switches ON and OFF is done quickly and is a statistical

process. When the voltage across the synaptic switch exceeds a threshold voltage ($\Delta V > V_T$), its conductance increases rapidly. The synaptic switch can be turned OFF when a large negative potential is applied such that ($\Delta V < -V_T$). It now becomes evident that the device is polarized. ΔV is defined as Vaxon − Vdendrite.

For implementing STDP using the crossbar network, each neuron input is a virtual input. Currents arriving from other neurons via the nanowire network are summed on a capacitor. When the voltage buildup exceeds a threshold a voltage spike is produced and travels to other connected neurons. The synapse is modelled using the memristor device. A weaker synapse is modelled with a greater the resistance of the connection. The resistance of the synapse is set by the voltages arriving at its terminals. As the voltage spike propagates through the synaptic connections the resistance of the connection changes. To integrate CMOS with the crossbar architecture requires a hybrid CMOS stack in which the network of the memristor devices will be fabricated on top of the CMOS circuits as shown in Fig. 2.9.

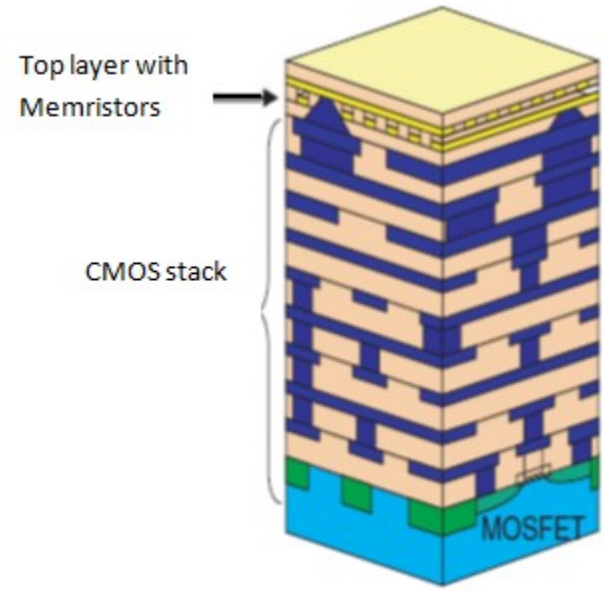

Fig 2.9. A conceptual diagram of a 3-dimensional CMOS-memristor hybrid structure [24].

2.4.1 Spike shapes

Unlike conventional neural spikes, the spike waveform required for STDP require a negative triangular wave followed by an exponential or linear decaying positive pulse. The spikes are asymmetric. The positive part dictates the width of the learning window. This is shown in Fig. 2.10.

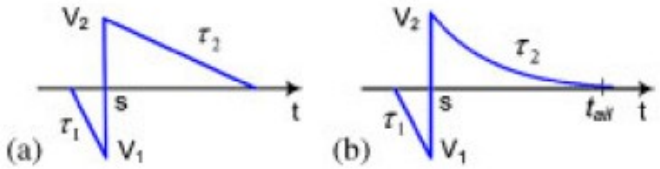

Fig. 2.10. A negative triangular pulse followed by a (a) linear or a (b) exponential decaying positive pulse [25].

Fig.2.11 illustrates how presynaptic and post synaptic spikes can be sent both forward and backward to program synaptic switches. The direction of the spike can be selected using a switch.

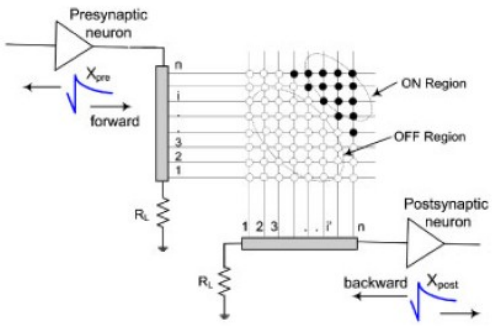

Fig. 2.11. Pre- and post-synaptic neurons generate spikes that are sent both forward and backward to program the synaptic switches [25].

Fig. 2.12 from [25], shows that there are three possible cases that can occur when a pre-synaptic and post-synaptic fire across a synapse. Assume that X_{pre} occurs at time t_{pre}= 0 and X_{post} occurs at time t_{post}=s, giving $\Delta T = t_{post} - t_{pre} = t_{post} = s$. Here "s" is the "spacing" between fire signals.

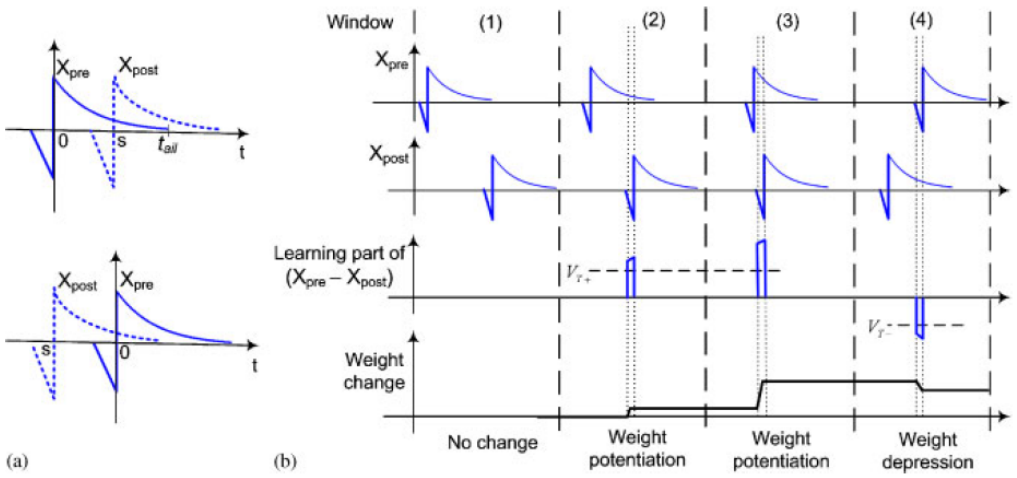

Fig. 2.12.(a) Shows the relative timing of the pre and post neuron spikes, (b) shows the different weight changes that occur due to different overlaps of the pre and post neuron spikes [25].

Case 1: If $s > t$ail or $s < -t$ail, the voltage drop across the memristor is not sufficient to change the conductance of the device and the weight has no change.

Case 2: If $0 < s < t$ail, or when the pre-synaptic spike precedes post-synaptic spike, the synaptic weight is potentiated. When this occurs, the applied exceeds the positive threshold of synaptic switches, $(X\text{pre} - X\text{post}) > V_{T+}$ for a fraction of the learning window.

Case 3: If $-t\text{ail} < s < 0$, or when the pre-synaptic spike follows post-synaptic spike, the synaptic weight is depressed. This means that the applied voltage exceeds the negative threshold of the device $(X\text{pre} - X\text{post}) < V_{T-}$ [25].

It now becomes evident that a learning window exists where the spikes are close together and an overlap of sufficient voltage is required to cause a weight potentiation or depression. This is similar to the synaptic weight changing in a biological neural network. The idea for using memristors as synaptic connections is that their resistance changes negligibly for voltage below the set threshold, beyond which there is a significant change.

Fig.2.13 shows the influence of the spike shape on the resulting STDP memristor weight update function $\xi(T)$. . For this work, a spike shape similar to the action potential shown in C1 of Fig.2.13 is used ,thus a weight update function similar to the update function shown in C2 of the same figure is expected.

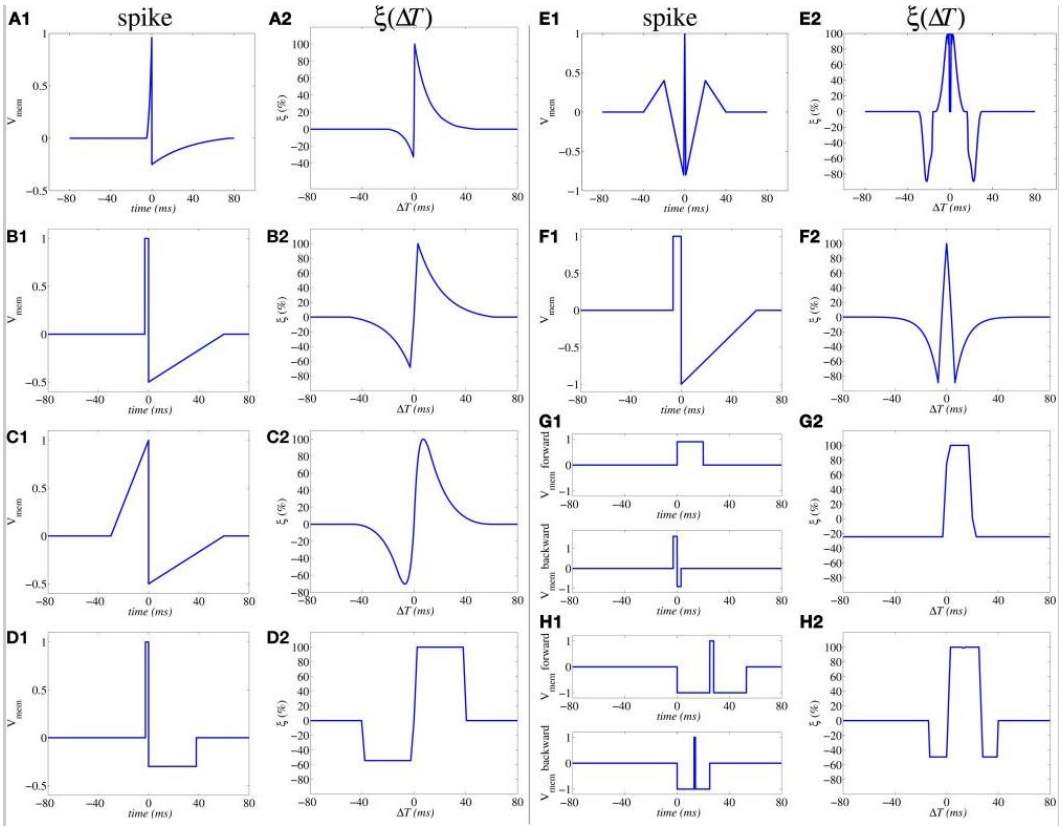

Fig.2.13 Illustration of influence of action potential shapes on the resulting STDP memristor weight update function ξ(T). [52].

Synaptic weight changes need not be discrete. Reference [48] shows a continuous, asynchronous and a deterministic implementation of STDP using TiO_2 device. Fig. 2.14 shows the STDP simulations with a TiO_2 device. This thesis work aims to produce a similar implementation using silver chalcogenide devices with actual silicon results.

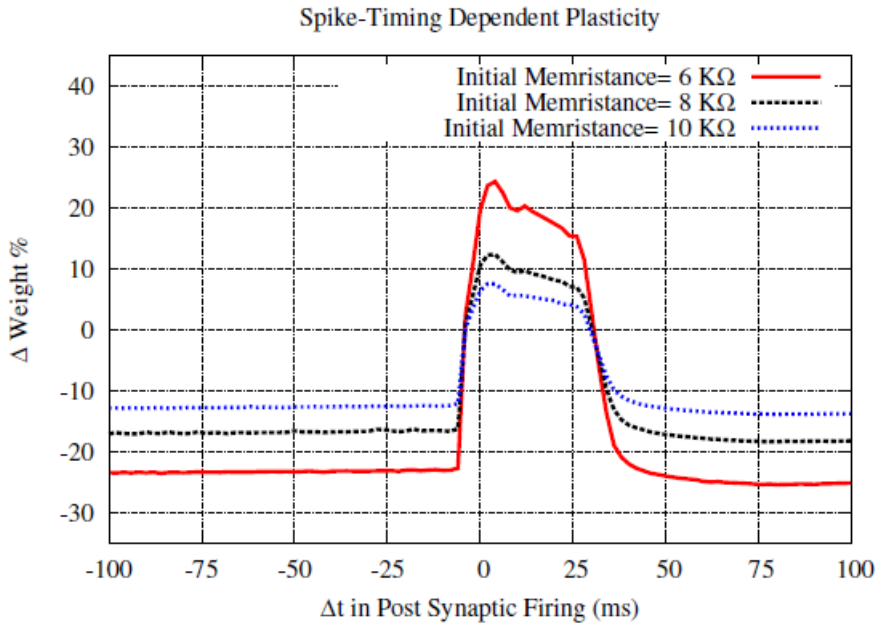

Fig. 2.14. STDP simulations using a TiO_2 device [48].

CHAPTER 3 IMPLEMENTING NEURAL MODELS IN SILICON

3.1 Case Study

Neural models are found in computational neuroscience and pattern recognition. While computational neuroscience aims to provide a better understanding of neural systems, patter recognition strives to outperform systems working sequentially. Hardware based neural models (analog or digital) are faster than a sequential machine which uses complex and slow software because hardware is inherently faster than software and neural systems are massively parallel.

A neuron can be viewed functionally as a circuit having one or more synapse, which acts as input ports responsible for receiving spikes from other neurons. These spikes are integrated over time with help of the soma block. The soma block produces an analog action potential. The soma block and synapse block can be interfaced with circuits modeling the neuron's spatial structure. This allows the signal processing that takes place in dendritic trees and axons to be mimicked on silicon.

The synapse in silicon can carry out either linear or non-linear integration of the spikes it receives from other neurons. Different temporal dynamics can be modeled along with short or long term plasticity mechanisms. This work focuses on implementing the synapse by using its plasticity mechanism. The plasticity mechanism described in STDP is achieved with the inherent property of the memristor: the memristor's resistance change is a function of its previous resistance.

The soma block in silicon can be further divided into sub-blocks each serving its own function. Typically they are divided into the following stages. A (linear or non-linear) temporal integration block, a spike generation block, a spike-timing block and a

spike-inhibiting block (or refractory period block). Each of these blocks can be designed using various methods from linear threshold circuits to complex systems. In this work, the neuron will be implemented using a leaky I&F circuit which implements the memory of the neuron as well as the refractory period.

There are different design styles to realize the blocks described above in silicon. There are sub-threshold (weak inversion) design approaches as well as strong inversion designs. The weak inversion design aims for a low power operation. There are voltage and current mode design styles which generally depend on the type of input or output signals. Switch capacitor designs implement discrete time signal processing and require a clock. Real and accelerated-time implementations refer to designs that run at biologically plausible rates and rates of 10 times faster, respectively. In this work the circuits are designed so that the MOSFETs operate in strong inversion and are real-time implementations of neurons.

3.1.1 Time Independent Neuron Models

The time free neuron model is the simplest neuron model. The inputs are vectors and the output is computed from these input vectors without considering past inputs or outputs, making the model independent of time. Even if the network contains loops, the state information is sensitive to the order of the inputs and not on the actual timing of the inputs. This type of model can be easily implemented using digital logic and serves as the basis for most of the current work in pattern recognition. An early model of a time independent neuron was the McCulloch-Pitts neuron [26]. The model can be written as [26]

$$A = \sum_{i=1}^{n} w_i X_i \, . \qquad (3.1)$$

According to this model, if the sum of the vector input exceeds a threshold a binary output of 1 is produced, otherwise a 0 is produced. Here A is the activity factor of the neuron, w_i is the weight of the synapse and X_i is the I'th input. A binary output Y = 1 is produced if the activity factor A is greater than a threshold Θ. If the threshold is not reached, the neuron outputs logic 0. The activity A represents the overall effect of all the presynaptic neurons. The effect of the presynaptic neuron on the postsynaptic neuron is computed by simple multiplication of the synaptic weight and the input. An excitatory synapse has $w_i > 0$, and an inhibitory synapse has $w_i < 0$. This makes computing the activity factor A a linear operation. The model can designed using digital gates (NOT, AND and OR gates). A digital computer can then be built if a clock is added [26].

3.1.2 Time dependent neuron models

Time dependent neuron models are sensitive to the timing of the inputs as opposed to the order, in other words they have a time varying state. According to [27] a simple first order differential equation can be used to represent the neurons conductance dynamics and synaptic transmission mechanisms. The equation takes the form

$$\tau \frac{dy}{dt} = -y + x, \qquad (3.2)$$

where y represents the output voltage or current and x represents the input current or voltage. This equation can be used to describe all the behavior of passive ionic channels. As mentioned before, there are different approaches to implement this differential equation in circuits. Reference [28] proposes a circuit that models neurons as passive leak conductance and follower integrator circuits. The follower integrator basically consists of a transconductance amplifier configured in negative feedback and connected to a

capacitor. It behaves as a low-pass filter with a tunable conductance. Log domain circuits help realize the first order differential equation. The Bernoulli cell described in [29] is an example of a log domain circuit. Current mode circuits are an efficient way of implementing neurons on silicon. The “tau cell” was first proposed in [30] as a BiCMOS log-domain filter, later characterized as a subthreshold log domain circuits. This circuit is also used in the tau-cell neuron.

Another way to model channel conductance is to solve the channel dynamics and conductance described by differential equations using analog circuits. Systematic synthesis methods are used to map non-linear differential equations in analog circuits. These circuits use voltage mode, strong inversion biased MOSFETS with integrators usually implemented in switch capacitor techniques.

3.2 Techniques For Hardware Implementation And Challenges

Implementing neural models in hardware allows real time operation and parallelism. Neural systems can be implemented in a plethora of ways, each having its own advantages and disadvantages. The different implementation methods can be compared as summarized in the table [32] below,

Table 3.1 Comparison of different hardware implementations of neurons.

Implementation Method	Degree of implementation	Speed	Real Time	Power consumed
Subthreshold VLSI	High	High	Yes	Very Low
Strong Inversion based VLSI	High	Very High	Yes	Low
Digital VLSI	Low	High	Possible	Medium to high
FPGA	Low to Medium	Medium	Possible	Medium to high
DSP based Software	Low	Medium to High	Possible	High

The implementations can be broadly classified into digital and analog. The digital implementation can be further divided into fields ranging programmable gate array (FPGA) based to software based implementations. Analog based implementations largely fall between subthreshold designs and strong inversion based designs. The only disadvantage of hardware based designs (except for FPGA – which has the best of both worlds) is that there exists a lengthy time between design and testing. Any change of design in hardware requires a re-fabrication (taking several months), whereas in software it can be done in minutes if not seconds.

Hardware implementation of neurons have been also already tried using discrete components [33], however modern neural hardware engineers are generally more interested in chip-based implementations. They are smaller compared to discrete circuits and can easily be incorporated into other systems. Testing of chip-based designs is also easier since the designer will receive a greater number of systems to test. The only limitation with chip based designs is there are no process technology kits dedicated for

neural implementations. This means a neural hardware designer must use technologies dedicated for high speed digital processors.

A good starting point for an on chip design is to choose between an analog or digital implementation. A number of factors can determine which implementation to choose. The primary differences between analog and digital systems are the way the signals are represented and coded. Analog signals are continuous in both time and magnitude where are digital signals are discrete in time and magnitude. Analog systems work with a different set of parameters such as bandwidth, slew rate and noise. These parameters are of paramount interest when the accuracy of the system is considered. For a digital system, the record length (sample length) of the analog value sets the accuracy. A longer digital word requires more transistors, meaning more area and power.

In order to implement a simple neuron model described above, adders and multipliers are required. A digital implementation would typically have adders and multipliers whereas an analog implementation would use circuit elements to add voltages (currents). Digital adders can be extremely fast and small, but multipliers tend to be quiet large and power hungry. The issue becomes a practical one and using a separate multiplier for each synapse is not feasible, especially when the word lengths are longer (longer words make multipliers slower and larger). Analog multipliers on the other hand can be made with ease provided they are to multiply within the first quadrant where both values are positive. Analog implementations also allow easier ways of creating exponential, integration and differentiation functions.

Neural systems require memory to hold constants such as thresholds or variables such as synaptic weights. Digital memories include dynamic and static Random Access

Memories (RAM) for variables and Electrically Erasable Programmable Read Only Memory (EEPROM) for constants. Dynamic RAM and static RAM are volatile and dynamic RAM requires constant refresh cycles. Analog memory implementations require resistors or capacitors to retain values. On chip resistors are usually not fabricated with very high accuracy and capacitors require a lot of area. A simple MOSFET can be used to store a charge on its capacitor but this tends to leak away so a refreshing system may be required. Thus these values may need to be refreshed by using Digital to Analog Converters (DACs) or external digital values. These techniques can be used to store constants. However the issue of storing variables in an analog implementation is not so straightforward. A digital system can easily change a variable value by changing the binary value by say a counter, the value of which comes from some computational element such as an adder and/or multiplier. Many novel solutions have been proposed such as motor driven potentiometers and memristors. Meador in [35] suggests a mechanism to transfer charges using floating gate transistors, however checking the value of the weight may be required since process variations exist.

This thesis aims to use memristors as a solution for the leakage problem discussed above. Memristors possess non volatility, low power and high density characteristics; a promising solution for synaptic weight change. Memristors do pose their own problems. There are large device to device variations making circuit design a bit more challenging.

3.3 Building Blocks for Implementing Synaptic Plasticity Rules In VLSI

To implement neurons in CMOS different building blocks are required. Synaptic weights and other weights need to be stored in storage devices. Capacitors serve as a storage element for analog based implementations. In CMOS different capacitor

implementations exist. Almost all capacitors in CMOS are parallel plate capacitor composed of two conductive parallel plates having equal dimensions. VLSI dielectrics which are used to increase the capacitance significantly are made of an insulator. Integrated capacitors can be MOS capacitors, poly-poly capacitors, metal-metal or metal – insulator – metal (MIM) capacitors. The choice of the capacitor type depends on linearity and accuracy. Thin film capacitors can have less parasitic capacitors compared to MOS capacitors but require additional processing steps. Capacitors can vary up to +/- 30 % and are not easy to trim because trimming forms parasitic capacitors. Poly-Poly capacitors are formed by two poly layers separated by a silicon dioxide layer. The temperature coefficient of poly-poly capacitor is typically less than 250ppm/°C which depends on the doping levels of the poly. MIM capacitors are made with two metal layers, usually the top most layer, keeping parasitic capacitances formed with the substrate to a minimum. MIM capacitors offer a high quality-factor (Q) of about 100 to 1000. MOS gate capacitances are made of a thin layer of oxide on the diffusion. The capacitance value depends on the bias values on the MOSFET. Junction capacitors are formed across the depletion region of reverse biased diode and have the high capacitance per unit area. The major disadvantage of junction capacitors is that they suffer from nonlinearities caused by variations of the depletion region.

A very commonly used neuromorphic circuit block is the differential pair or long tail differential pair [41]. A differential pair circuit is typically made of a bias transistor on which two input transistors are stacked upon. The large signal analysis of such a circuit gives a sigmodial relationship. Differential circuits, shown in Fig. 3.1 can also function as comparators, multipliers (Gilbert cell) and form the basis for operational amplifiers.

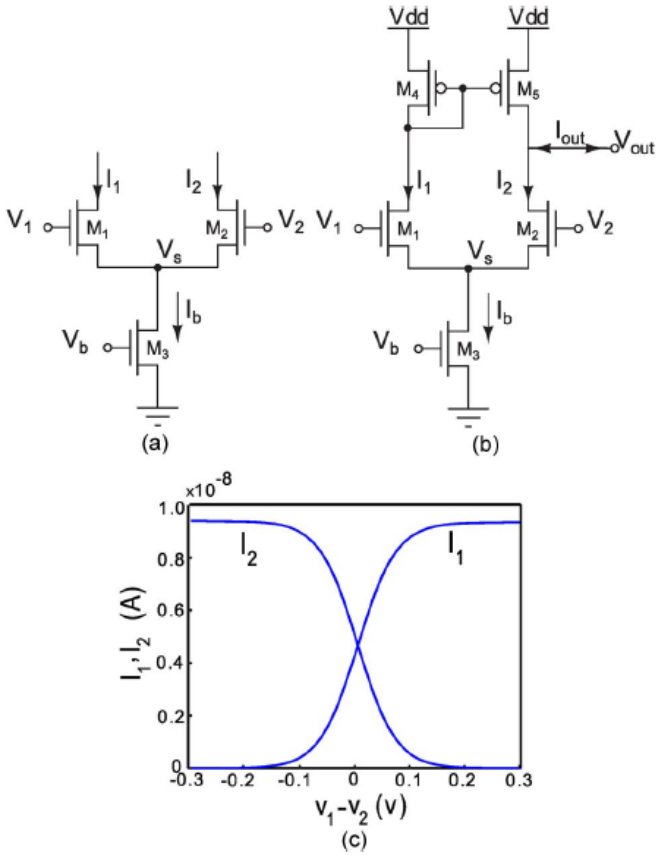

Fig 3.1 (a) A differential pair circuit consists of three transistors. (b) The OTA acts a transconductance circuit that converts the difference between its input voltages to a corresponding current at its output, (c) The sigmodial relationship between differential input voltages and the currents flowing in each transistor of the differential pair. This is a useful behavior for implementing similar sigmodial behavior observed in neural systems [45, 48].

Operational amplifiers or operational transconductance amplifiers (OTA) are another important building block. They are basically a differential amplifier with a very high gain, typically used with voltage inputs and current outputs. OTA's can be used in a number of ways in neuromorphic circuits [42]. Typically they are used as comparators, but find themselves in innovative circuits such as active resistors in leaky integrator circuits [43]. OTAs generally are stable from noise and process variations proving to be a very useful building block.

3.4 Spike Generating Circuits

To implement the potentiation and depression characteristics of plasticity rules, circuits which emulate these dynamics are required. Spiking circuits can cause potentiation or depression of a synaptic weight. Neurons produce analog waveforms, continuous in time however in many neuron models the action potential is described to a discrete even generated when a threshold is exceeded.

One of the original circuits proposed for generating discrete events in VLSI implementations of silicon neurons is the *Axon-Hillock* circuit [44]. Fig. 3.2 (a) shows a schematic diagram of this circuit. The amplifier block A is typically implemented using two inverters in series. Input currents I_{in} are integrated on the membrane input capacitance C_{mem}, and the analog voltage V_{mem} increases linearly until it reaches the amplifier switching threshold as shown in Fig.3.2 (b). At this point V_{out} quickly changes from 0 to V_{dd}, switching on the reset transistor and activating a positive feedback through the capacitor divider implemented by C_{mem} and the feedback capacitor C_{fb}. If the reset current set by V_{pw} is larger than the input current, the membrane capacitor is discharged, until it reaches the amplifier's switching threshold again. At this point V_{out} swings back to 0 and the cycle repeats. The inter-spike interval t_L is inversely proportional to the input current, while the pulse duration period t_H depends on both the input and reset currents. A comprehensive description of the circuit operation is presented in Mead [44].

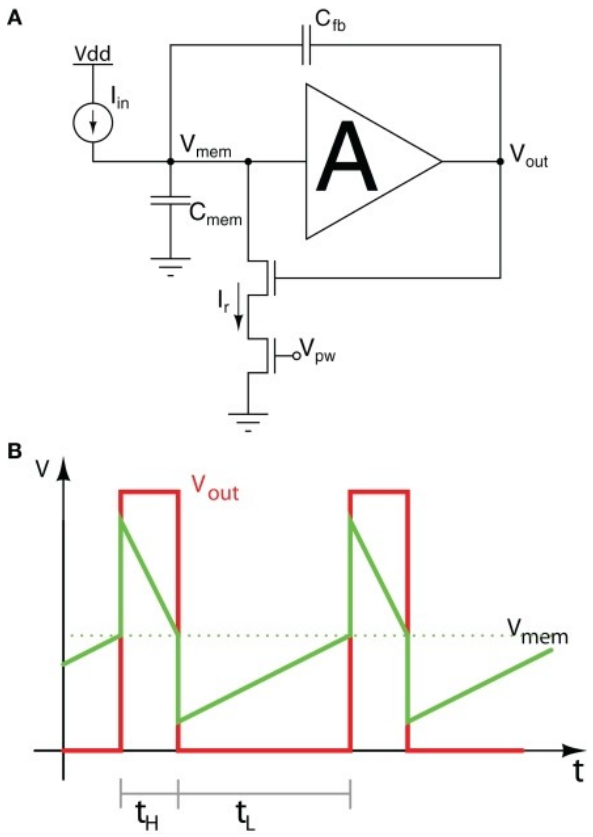

Fig. 3.2 Axon-hillock circuit. (a) Schematic diagram of the Axon-hillock circuit and (b) membrane voltage and output voltage traces over time [50].

The main advantage of such a circuit is that it has very good matching properties. The mismatches do not depend on the transistors, but rather on the sizes of the capacitors.

A spiking event in the Axon-Hillock circuit is produced when the membrane voltage reaches a threshold, which has a strong dependence on the widths and lengths of the transistors and the process characteristics. To improve on the neuron model, a refractory period can also be incorporated. Such a circuit is present in Fig. 3.3 (a) which uses a simple five transistor OTA as a low cost comparator. Fig. 3.3 (b) [50] shows the various events that occur during the spike generation.

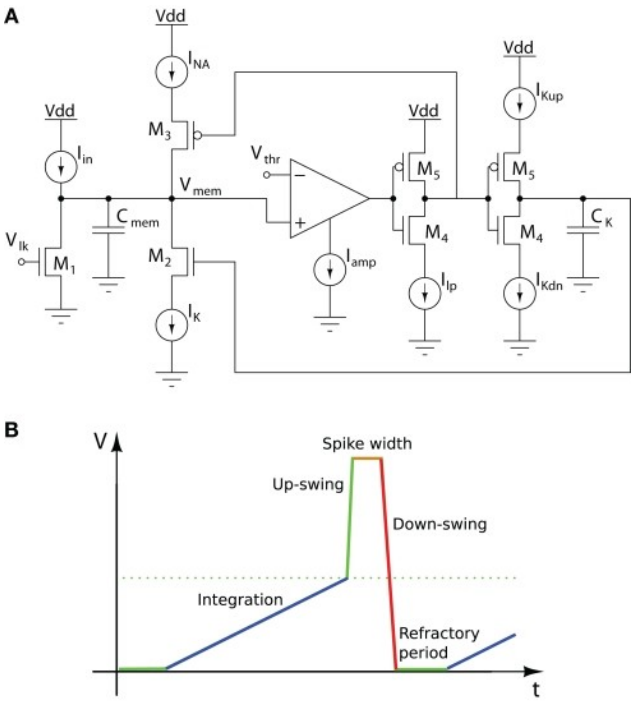

Fig. 3.3. (a)Schematic diagram of the I&F circuit, (b) membrane voltage trace over time [50].

The membrane of the biological circuits is represented by the capacitor C_{mem}. When no input is given to this circuit a membrane leakage current controlled by V_{leak} will draw all the charge on the membrane capacitor bringing to a resting potential, ground in this case. Excitatory inputs in the form of a current I_{in} will build charge on the membrane capacitance as long as it is greater than the leakage current. The membrane voltage is compared to an external threshold (V_{th}). This comparison between the two potentials is made by a comparator. If V_{mem} exceeds the externally controlled V_{th} a spiking event occurs. This is similar to a biological neuron in which the increase in sodium conductance causes the waveform to rise and a slightly delayed potassium conductance increase creates a downswing.

As V_{mem} rises above V_{th} the comparator outputs a positive voltage equal to the supply. The inverter that follows will go low allowing the "sodium" current I_{NA} to pull up the membrane potential. While this occurs the second inverter charges capacitor C_K, controlled by I_{kup}. Once the voltage on C_K reaches a large enough, M2 turns on allowing the potassium current I_K to discharge the membrane capacitance. These two potassium channel currents govern the opening and closing of the potassium channels. I_{KUP} controls the spike width, and the delay between the opening of the sodium channels and potassium channels is inversely proportional to $I_{KUP.}$ Once the voltage of the membrane capacitor V_{mem} drops below the threshold voltage V_{th} the sodium current becomes 0. The second inverter causes C_k to discharge to ground through I_{KDN}. This current controls the refractory period of the circuit.

The advantage of this circuit over the Axon-Hillock circuit is that it consumes less power. The amplifier used in this circuit, typically two inverters in series dissipates lots of power for slow varying input signals, which occurs when the NMOS and PMOS are both conducting (short circuit power).

CHAPTER 4 ANALYSIS AND SIMULATION RESULTS

4.1 Design and Evaluation approach.

The design begins with a block level simulation of the circuit. This is done to evaluate critical parameters of the circuit such as frequency and power consumption. It also serves as starting point for the design. From there on, transistor level design is done followed by layout and post-layout verifications. The designs are also tested across Process, Voltage and Temperature (PVT). Since the CBRAM device does not have a specified threshold voltage for programming, the I&F circuit should be able to produce a wide range of spikes with different pulse widths. This was the main requirement for the I&F block. Characterization of the memristor is done to get a general idea of the spike amplitude and pulse width that would be able to program the device. The programming window is chosen to be in microseconds. From [23] a microsecond programming time frame requires spike amplitude of about 1.5 V to 2 V.

4.1.1 Integrate and Fire Circuit

As mentioned before different circuit blocks can be used to model the biological neuron in CMOS. The spiking neuron in CMOS will be implemented using the LIF circuit. In a biological system, a neuron is surrounded with water having ions. These ions maintain a concentration across the cell membrane of the neuron. This is modeled as a capacitance C_{mem} . As mentioned before, the firing of a neuron takes place when these ions move across the cell membrane. This can be described as a current $i(t)$. The electrical potential across the cell membrane which manifests as a spike (action potential) is defined as $V_{mem}(t)$. This can be described using

$$I(t) = C_{mem} \frac{dV_{mem}}{dt} \tag{4.1}$$

When the voltage across C_{mem} reaches a threshold voltage V_{th}, a spike occurs and the potential is reset to the resting potential. The frequence of firing here is clearly unbounded. The neuron has a refactory period t_{rp} ,during which no firing can take place. The linear time invarient frequency bounded model for frequency of firing for the neuron is given by

$$f(I) = \frac{I}{C_{mem}V_{th} + t_{ref}I}. \tag{4.2}$$

The shortcoming of equation (4.2) is that it fails to incorporate the memory aspect of the neuron. The memory of the nueron is time dependent. The leaky integrate and fire model accounts for this memory phenomenon by adding a "leaking" current. This accounts for the diffusion of ions across the membrane when the equilibiruim is not reached. The equation can be expressed as

$$i(t) - \frac{V_{mem}(t)}{R_{mem}} = C_{mem} \frac{dV_{mem}}{dt} \tag{4.3}$$

Here R_{mem} is the membrane resistance. The cell membrane does not act as perfect insulator. If the threshold is not exceeded, there will be a leak of charge. The frequency of firing now becomes

$$f(I) = \begin{cases} 0, & I \leq I_{th} \\ \dfrac{1}{t_{ref} - R_{mem}C_{mem}\log\left(1 - \frac{V_{th}}{IR_{mem}}\right)} & I \leq I_{th} \end{cases} . \tag{4.4}$$

The I&F circuit architecture used in this thesis is shown in Fig. 3.5.The I&F circuit is first modeled in Verilog AMS to start with. The block level simulation of the I&F gives an idea of the frequency and power it consumes.

4.1.2 Operational Amplifier Design

The operational amplifier (op-amp) here is used as a simple comparator. Input offsets are not of a major concern-shifting the threshold by a few tens of mill volts has no effect on the spiking width or frequency.

The op-amp is designed to have a gain of approximately 40 dB. Its architecture is kept simple since a high gain is not required. A 5 transistor design is shown in Fig. 4.1. It functions as the comparator.

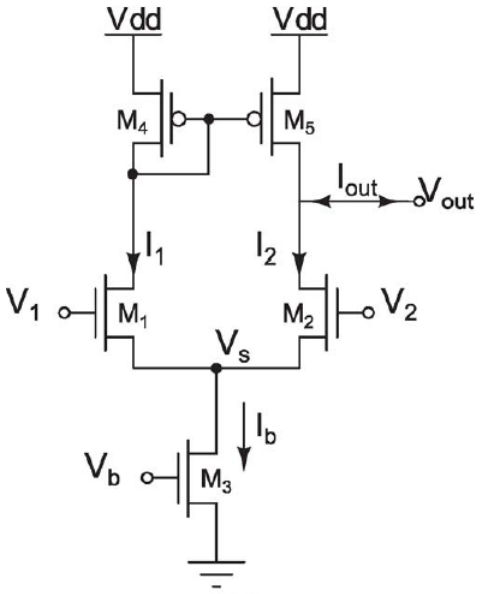

Fig. 4.1. OTA architecture used to function as a comparator [42].

The gain of the op-amp is given by the transconductance (g_m) of the input pair multiplied by the output impedance. To achieve the gain of 40 dB, (which is sufficient to resolve the input voltage difference), the bias current to the op-amp and the widths of the

input pair can be sized accordingly. The unity gain bandwidth of the op-amp is designed to be above 100 MHz. The gain of the op-amp is given as

$$A(s) = \sqrt{2I_{bias}\frac{W_n}{L_n}\mu_n C_{ox}} * \left(\frac{L_n}{I_{bias}} || \frac{L_p}{I_{bias}}\right), \qquad (4.5)$$

where, I_{bias} is the bias current, W_n, L_n are the width and the length of the input pair's transistor respectively, μ_n is the mobility of the NMOS device and C_{ox} is the capacitance of the gate oxide of the input transistor pair. The bias voltages for the op-amp are generated using a constant g_m biasing circuit with a start up circuit shown in Fig. 4.2.

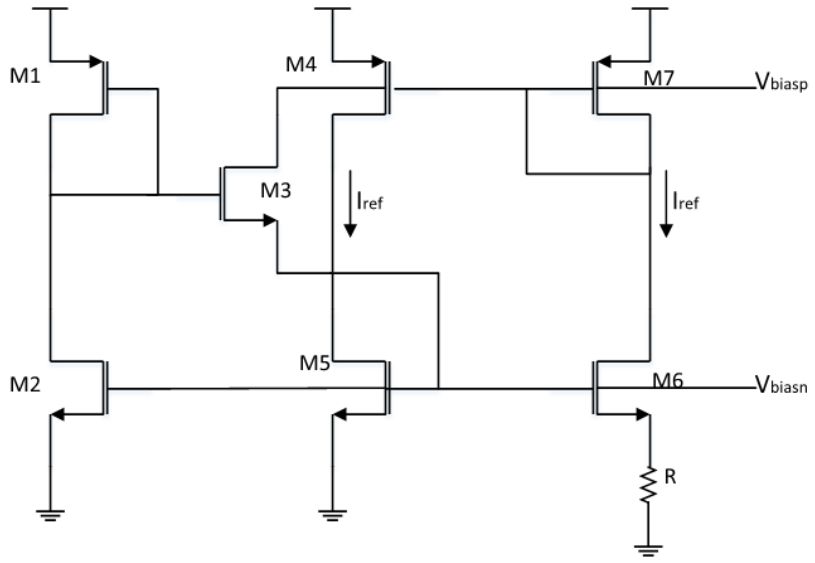

Fig. 4.2. Constant g_m bias.

The constant g_m bias circuit offers several advantages. The current reference is derived from V_{GS6} divided by the on chip resistance R. The resistance R is shielded from the supply hence the reference current is supply insensitive to a first order. By sizing the width of transistor M_6 to be four times as the width of transistor M_5, the transconductance of M_6 becomes equal to the inverse of R. On chip resistors have a very good temperature co-efficient .Thus a N-channel MOSFET referenced from M_6 will have a g_m with a

temperature coefficient of R. This circuit has a positive feedback loop which causes a stability concern. To prevent oscillations, R must be selected such that the source degenerated common source amplifier formed by R,M_6 and M_5 have a gain of less than unity. Another issue with this circuit is that it has a second stable where all the currents are zero. To guarantee this condition does not happen, a start up circuit (M_1 ,M_2 and M_3) is implemented to ensure non zero current flow during start up.

The capacitors are sized according to equations (4.2) and (4.3) to achieve the required frequency. The capacitor is chosen to be 9 pF so that the frequency of operation is lies in the MHz range. Fig. 4.3 shows the circuit.

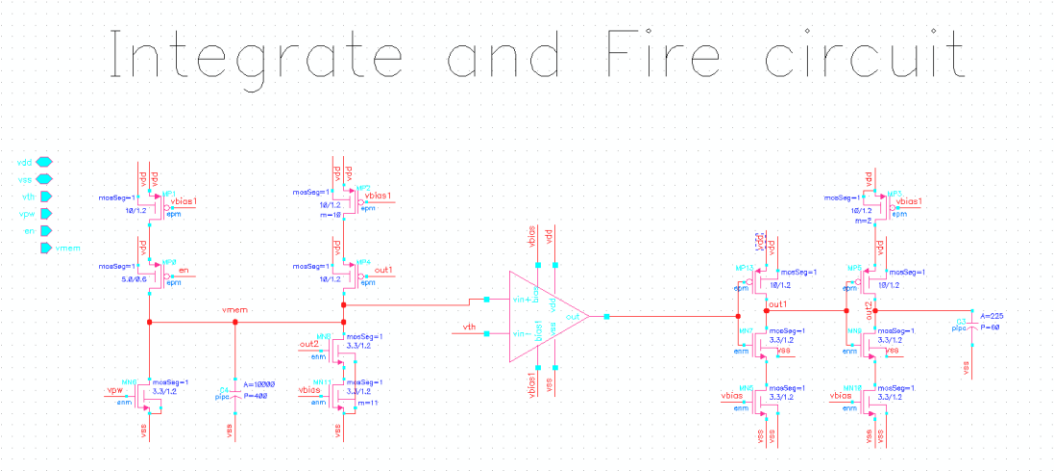

Fig. 4.3. I&F circuit schematic.

The threshold can be adjusted externally. The current required for the pulse width adjustment can also be adjusted externally. All other bias voltages required for modeling the sodium and potassium currents are taken from the constant g_m bias. A spike swing of about 1.2 V is required to modulate the CBRAM device. This spike swing can be set by adjusting capacitor C_{mem} and the operating V_{dd}. The V_{dd} is chosen to be 3.6V so that the required swing can be achieved. Table 4.1 shows the pin list for the I&F circuit.

Table 4.1 Pin list of I&F circuit.

Name	I/O	Function
VDD	I/O	Supply
GND	I/O	Ground
ENABLE	I	Enables the circuit
V_{PW}	I	Adjust spike pulse width
V_{TH}	I	Set threshold for spike to occur
V_{MEM}	O	Output

4.1.3 Layout and Floor Planning

Fig. 4.4 shows the layout of the I&F circuit. The input pairs are laid out using a common centroid to improve matching (although it is not a strict requirement). The rail devices used for current mirroring are laid out using interdigitization to reduce mismatches up to a first order.

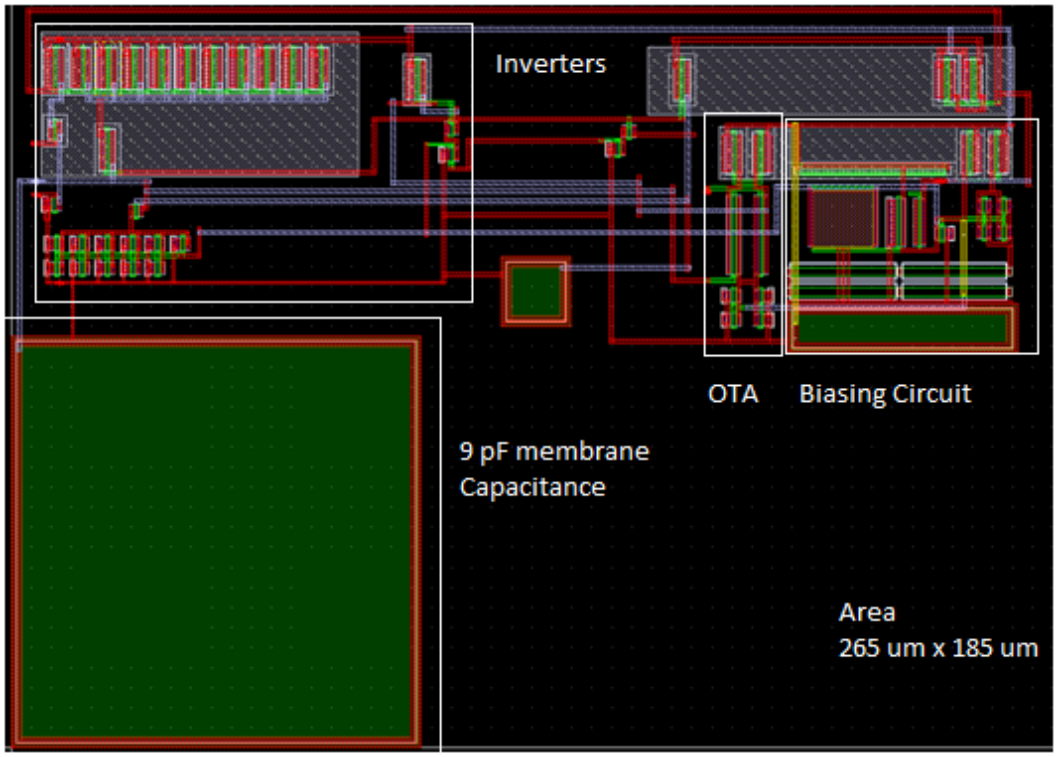

Fig. 4.4. I&F circuit layout.

The die size is 1.5 mm x 3 mm with pads of 30um x 30um. This allows 21 pads on the sides and 9 pads on the top and bottom. The CBRAM devices are deposited by hand with masks made out of tape. To allow sufficient spacing for the CBRAM deposition, two pad spacing is used. More pads were placed in the center on the die as required. Fig.4.5 shows the layout of the chip.Fig.4.6 shows the two pad spacing to for the CBRAM device deposition.

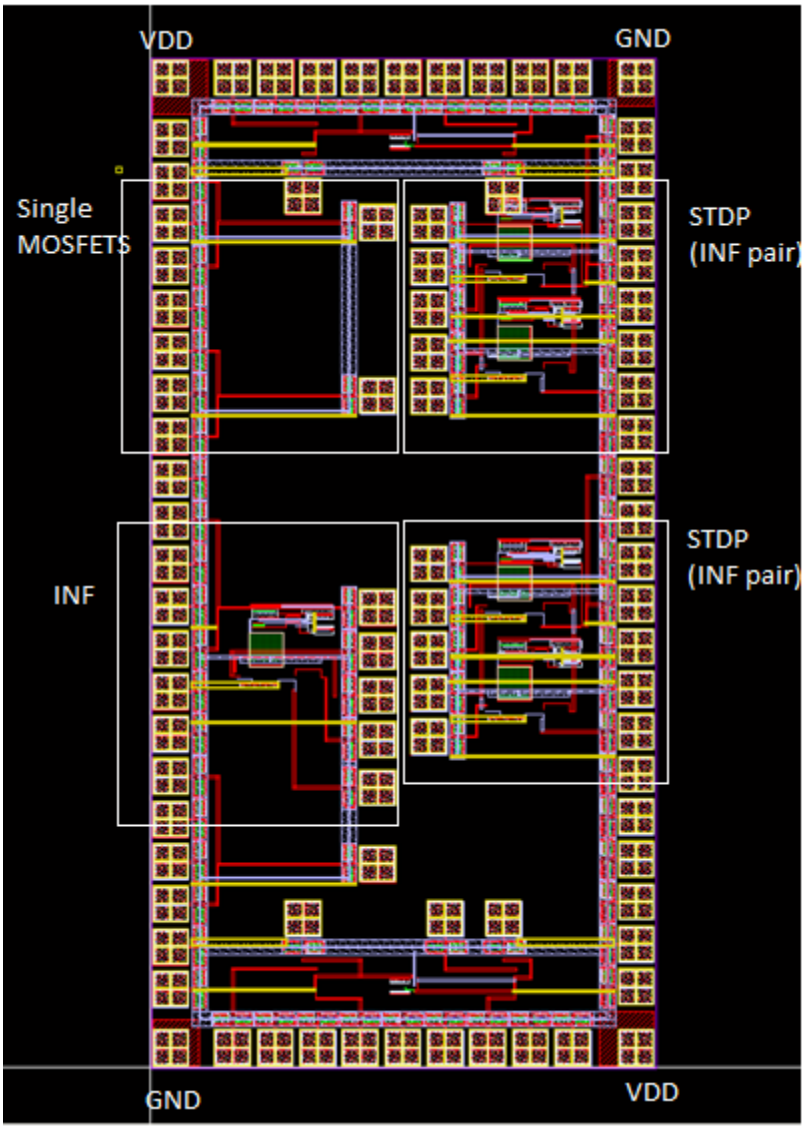

Fig. 4.5. Chip layout.

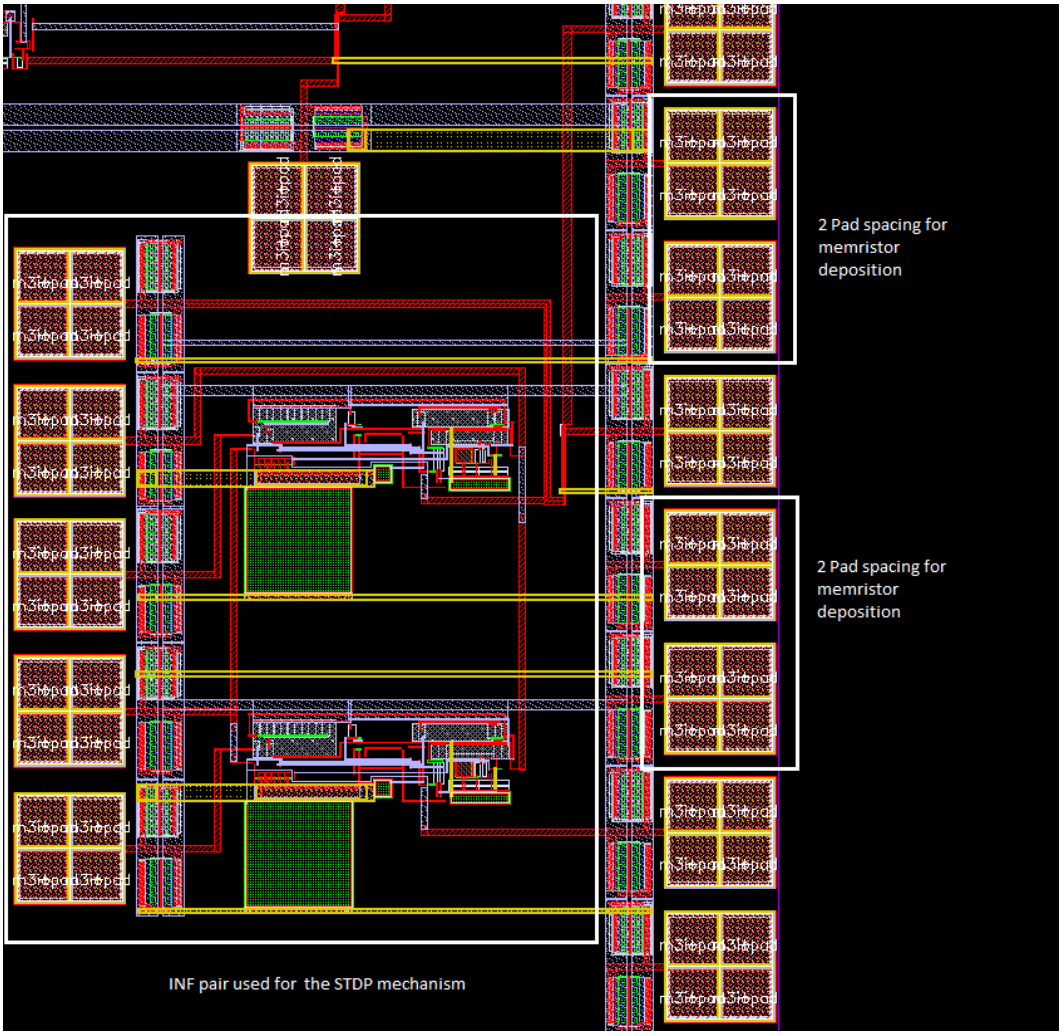

Fig. 4.6. Layout of the I&F pair used for programming the CBRAM device. The figure also shows the 2 pad spacing used for memristor deposition done by hand.

4.1.3 Die Post Processing

The CBRAM device is deposited on the required pads on the die. A 100nm layer of chalcogenide glass is deposited by thermal evaporation over the chip, followed by a 400 nm Ag deposition by thermal evaporation over the ChG. Photo doping of silver is done by exposing the depositions to ultraviolet light for 20 minutes. Another 100 nm layer of Ag is deposited by thermal evaporation. Finally, a 1 micron layer of aluminum and 1

micron layer of copper are deposited over the silver by e-beam evaporation respectively. Fig. 4.7 below summarizes the post processing to deposit the CBRAM device on the CMOS die.

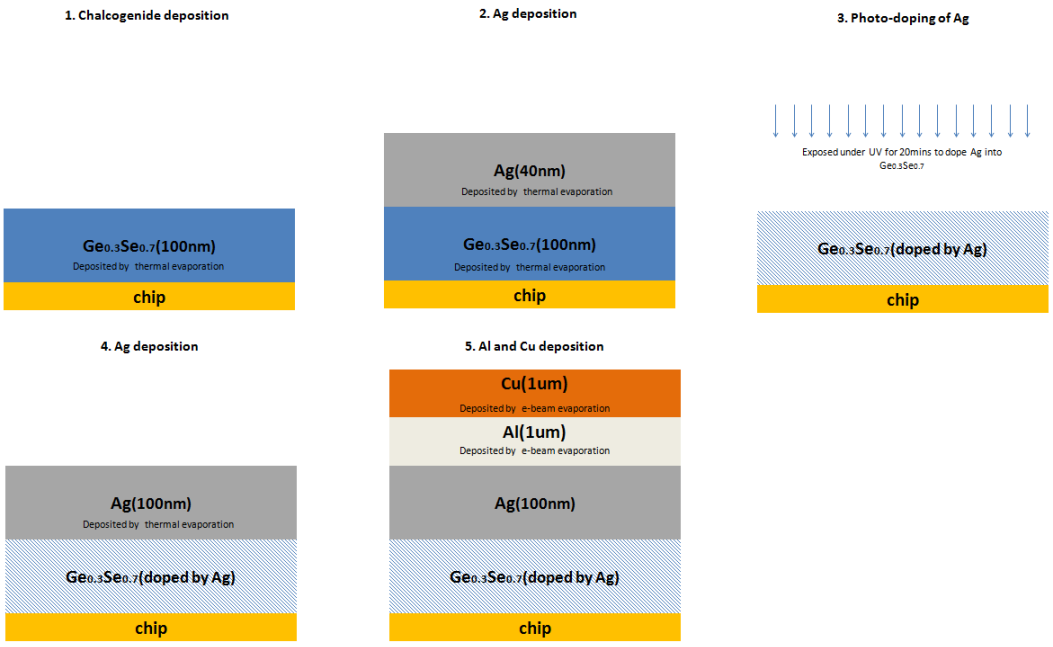

Fig.4.7. Steps involved in the CBRAM device deposition on the chip pads.

4.2 Evaluation

The memristor after fabrication has unknown resistance and I-V characteristic. Since there is batch to batch variation of these devices, characterization is required. Characterization of the device shows the positive threshold voltage at which the conductance change occurs and the negative threshold at which the erase occurs. These voltages are required to measure the memristor's resistance by applying a subthreshold pulse, and measuring the current flowing through the device. After the characterization of the device, the device is programmed to different low resistance states of 1 kΩ, 5 kΩ and 10 kΩ. The device can be programmed using the 1T-1R circuit. After the device has been programmed, the pre and post spikes can be applied across the device. The spike can be

produced by enabling the I&F circuit using a pulse. The post-spike is then delayed by 0.5 µs. For each delay, the resistance (conductance) of the device is measured using a semiconductor parameter analyzer and the 1T-1R circuit. The flow chart shown in Fig. 4.8 shows the steps that are carried out to acquire the readings. Fig.4.9 illustrates the test setup.

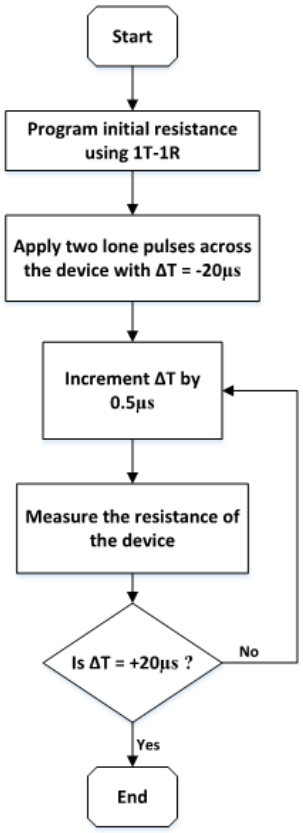

Fig.4.8. Flow chart showing steps followed to prove STDP in CBRAM devices.

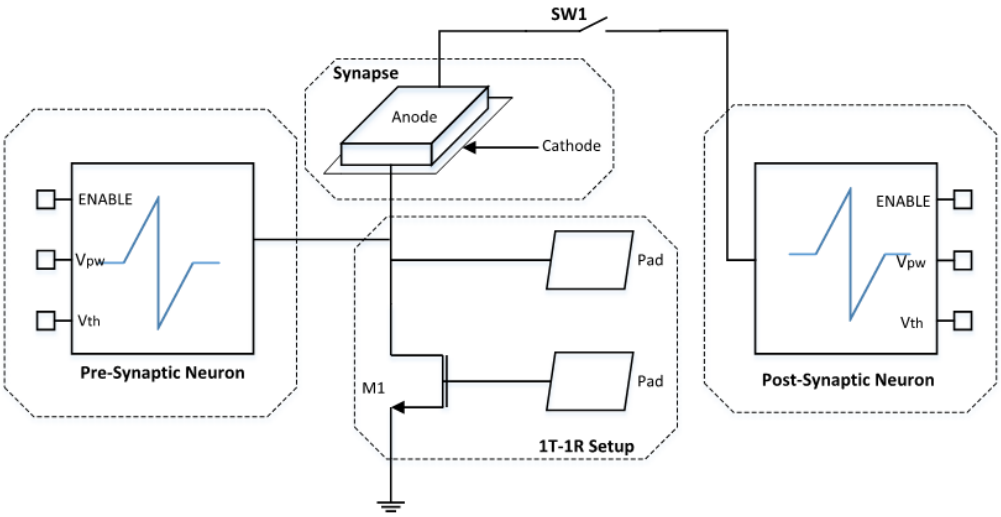

Fig.4.9. Test setup used to modulate the resistance(conductance) of the CBRAM device.

Fig. 4.10 shows the 1T-1R circuit that is used to program the memristor to its initial resistance.

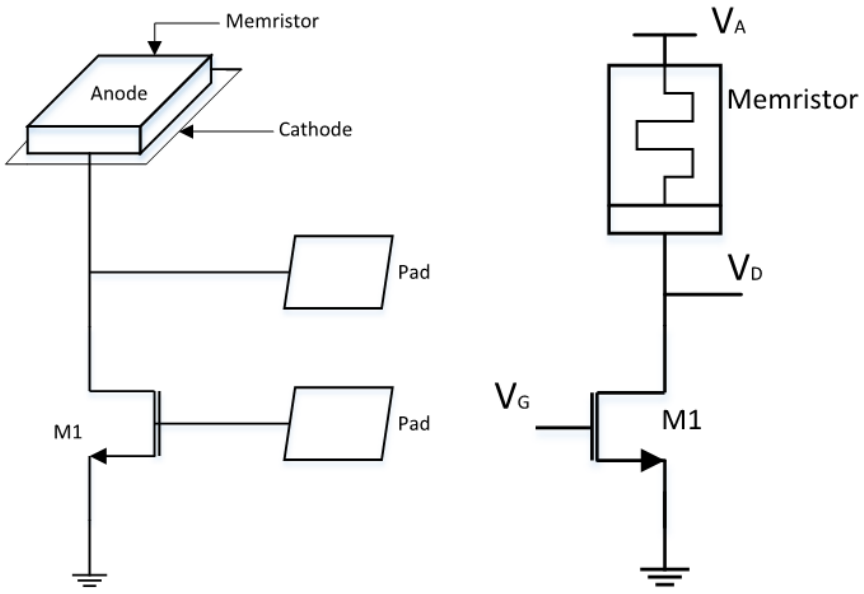

Fig.4.10. 1T-1R used to set the initial resistance of the memristor device.

The I&F outputs are applied across the memristor once the switch SW1 is closed (shown in Fig. 4.9). The pad at the drain of the transistor is also connected to the cathode of the device. The 1T-1R circuit is used to program the device to a low resistance state

and also measure the resistance after the spikes are applied. By enabling the I&F circuits with a phase difference, relative spike timings can be achieved. A dual channel signal generator is used for this purpose, which offers better synchronous signals than using two individual signal generators externally synchronized. All other inputs (V_{TH} and V_{PW}) are common to both I&F circuits so that similar spikes shapes are produced. DC power supply units connected to a common ground are used to provide input voltages.

4.3 Results

The Spice and experimental results for each block and the STDP rule are presented in this section.

4.3.1 I&F

The simulations were run at optimum conditions (room temperature and typical models). The results presented are post-layout simulations with extracted parasitic capacitances. Fig. 4.11 shows the SPICE waveform required to prove the STDP mechanism using CBRAM devices. This waveform conforms to the required specifications to modulate the conductance of the memristor. The waveform shape is similar to those in Fig. 2.12 and Fig. 3.5 (b). Fig. 4.11 shows a spike that is ideal to modulate the resistance of the memristor. Fig. 4.12 shows spike of different widths and amplitudes that can be produced with the I&F circuit as required.

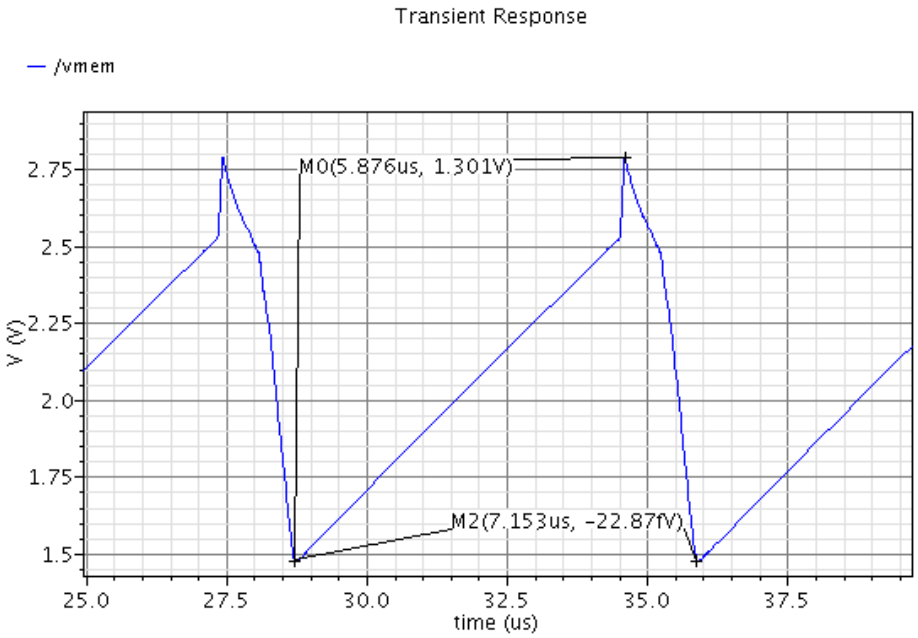

Fig. 4.11. Ideal spike to modulate the conductance of the CBRAM device.

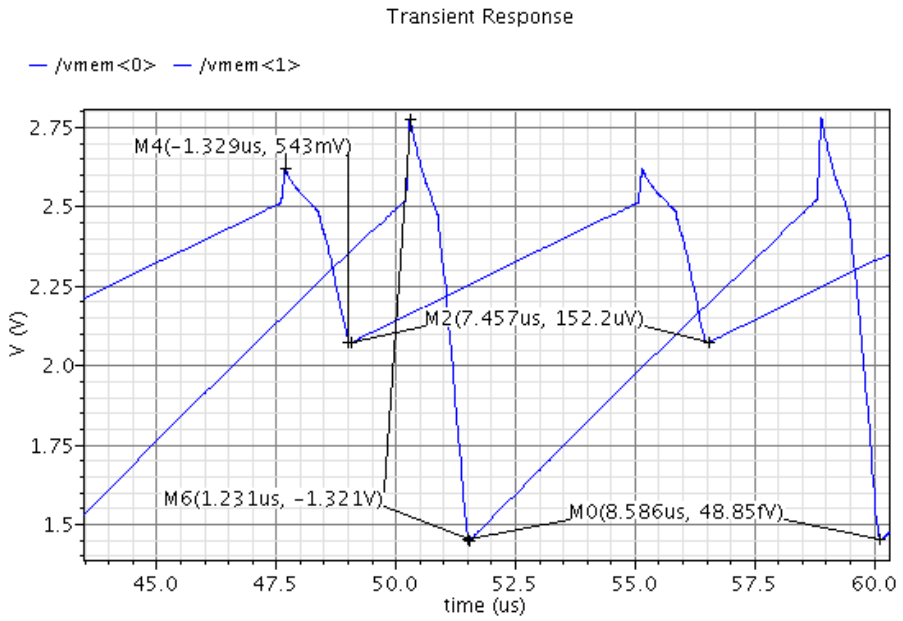

Fig. 4.12. Example of spikes that can be produced by the I&F circuit.

Fig. 4.13 shows the oscilloscope trace of the real hardware I&F circuit's output.

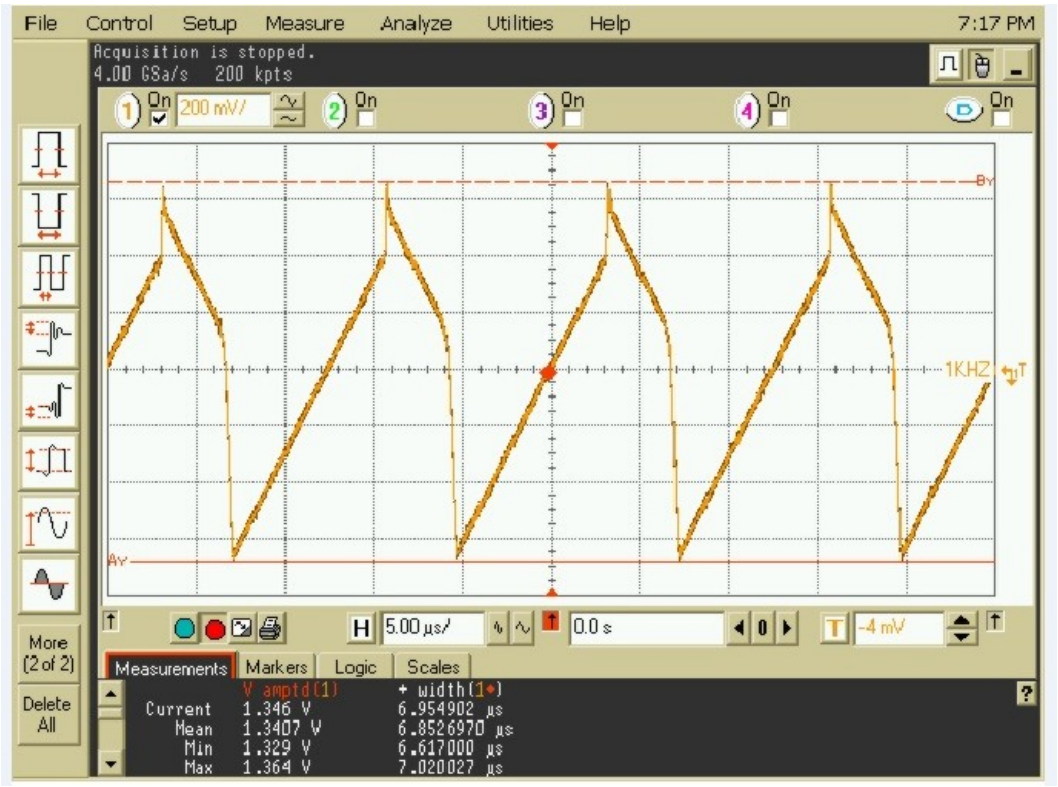

Fig. 4.13. Hardware test result showing the I&F output.

Fig. 4.14 shows the hardware test results of individual spikes from pre-synaptic neuron and post-synaptic neuron that are used to modulate the conductance of the device. As mentioned before, pulse width modulated signals are applied to the enable pin to produce the lone pulses.

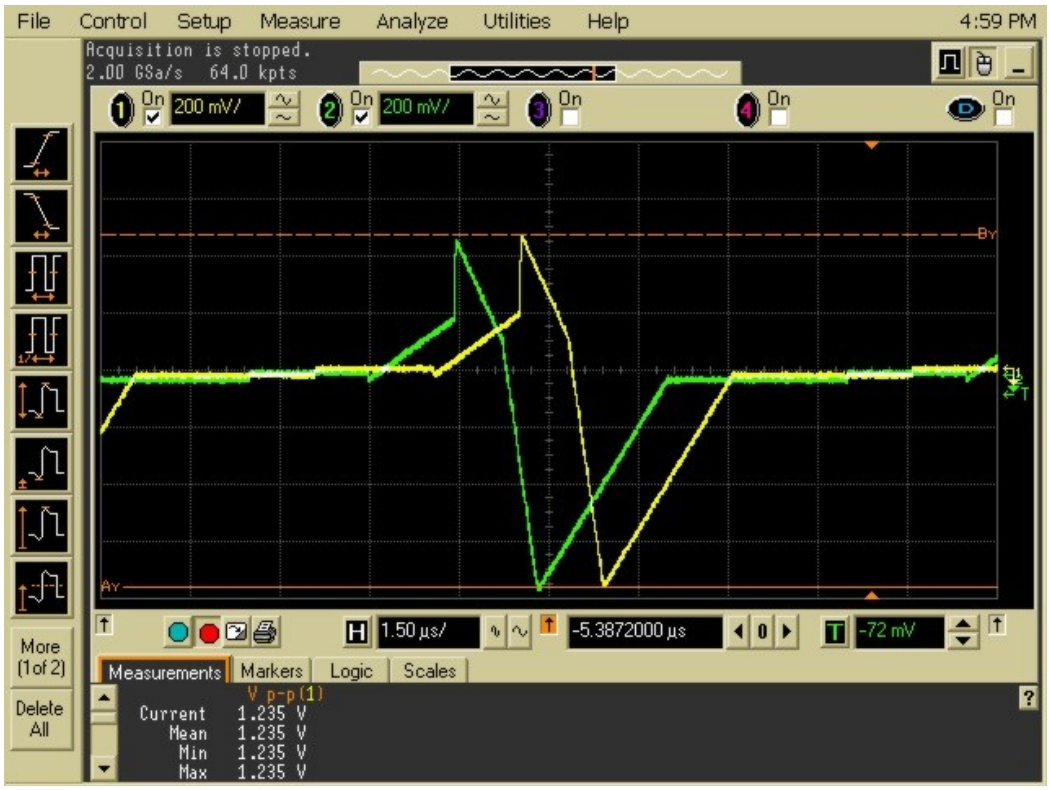

Fig 4.14 Hardware test results showing lone spikes used to modulate the device resistance (conductance).

4.3.2 Memristor

The memristor's I-V characteristic is shown below in Fig.4.15. The characterization of the memristor was done using a 1T-1R circuit. The parametric semiconductor analyzer is used to for this purpose. The high resistance state of the device was measure to be 14.74 MΩ. A compliance current of 100µA was used when obtaining the I-V curve.

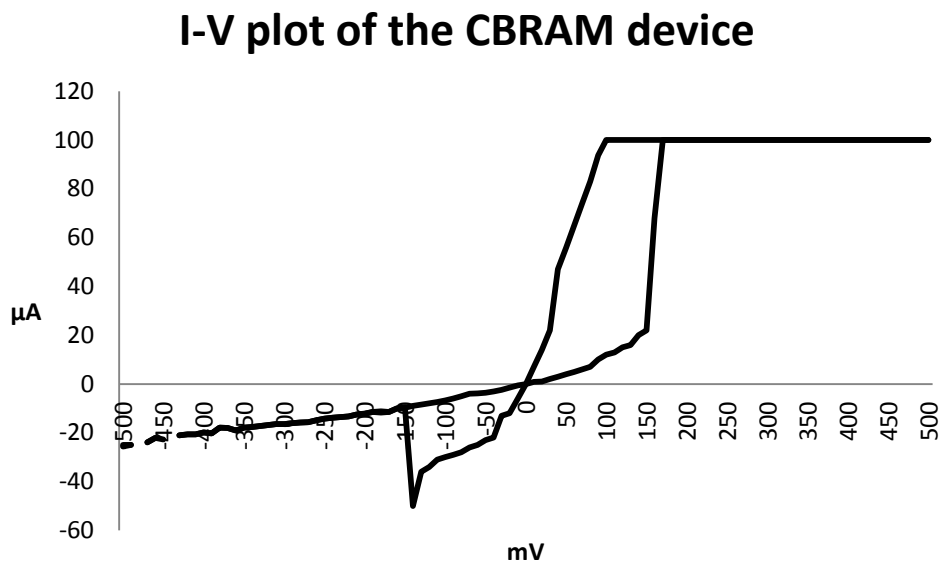

Fig.4.15 I-V characteristic of the CBRAM device deposited on hardware.

4.3.3 STDP

The die was packaged using a 28 pin ceramic package. The input pads of the I&F circuit were wirebonded leaving the anode of the memristor exposed for probing. The memristor was first programmed to a low resistance value of 10kΩ using the 1T – 1R circuit. When the external switch SW1(shown in Fig. 4.9) is closed , the pre-synaptic and post synaptic spikes are applied across the memristor. Switch SW1 is closed by probing the anode of the memristor and connecting it to the post-synaptic I&F circuit's output. The resistance of the device was measured using a semiconductor parametric analyzer. The resistances were recorded for different phase differences between the pre-synaptic spike and post-synaptic spike. The process was then repeated for an initial resistance of 5kΩ and 1kΩ. Fig. 4.15 shows the results of the weight change due to the relative timings between the pre and post synaptic spikes. The post-synaptic neuron's spike is shifted by ΔT. A positive ΔT indicates that the post-synaptic spike occurs after the pre-synaptic spike, and a negative ΔT indicates that the post-synaptic spike occurs before the pre-

synaptic spike. Fig. 4.16 (a) shows the results with a spike width of 7μs and (b) shows results with a spike width of15μs. From the results presented in Fig.4.16 a spike width of 7 μs causes a maximum weight change of about 20% whereas a spike width of 15 μs causes a maximum weight change of about 40 % for LTP and -30% for LTD. It is also observed that a lower initial resistance has a slightly greater weight change. The weight update function is similar to the update function shown in Fig. 2.13. From the results presented above, it can be concluded that a hybrid neural network using CMOS circuits and CBRAM devices can be used to implement a STDP weight update function

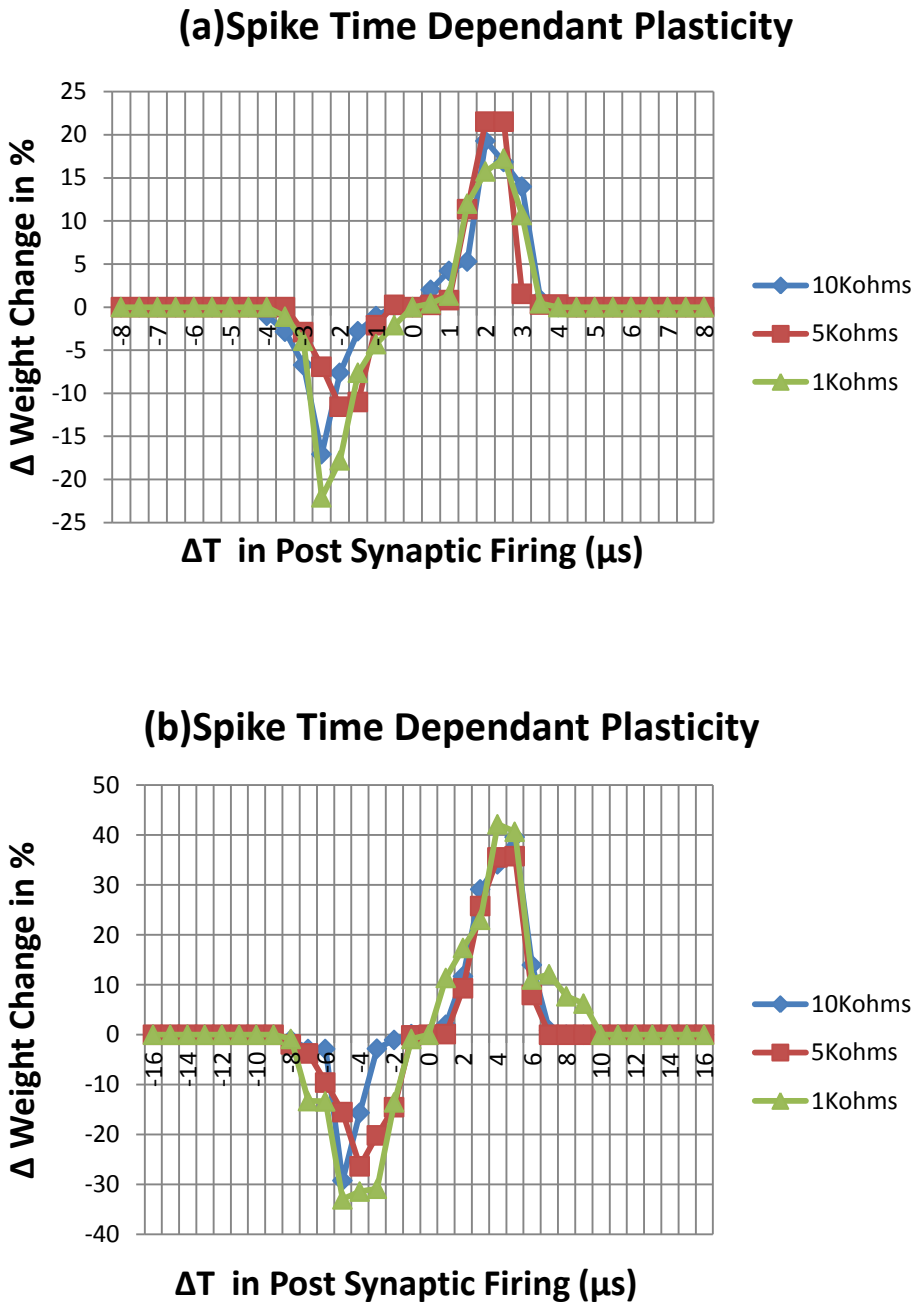

Fig.4.16. Plots showing results for STDP using CBRAM devices made of silver chalcogenide using spikes widths of (a) 7 µs and (b) 15 µs.

CHAPTER 5 CONCLUSIONS AND FUTURE WORK

5.1 Summary and Conclusions

Alternative methods of computation will continue to advance in an effort to process complex problems while consuming little energy per computation. Biological systems provide a model from which such a processor can be designed. The challenges to emulate a neural system on chip had been an engineering challenge but with recent developments such as memristors, bio-inspired processors are becoming more feasible. Most implementations demonstrate proofs of principles. A successful application of neuromorphic circuits in large-scale requires a system that goes beyond single core solutions. The major challenges include interchip communication and EDA tools that allow design, verification and testing. Synaptic weight storage circuits tended to consume a large area but memristors are new alternatives. Memristors consume less power, occupy smaller area and are compatible with CMOS technology. Although memristors show a promising alternative to circuit based synaptic weight implementation, memristors suffer from wide process variations.

This thesis presents a proof of concept in which CBRAM devices made of Ag ChG materials are used with CMOS circuits to prove a learning algorithm, i.e., Spike Time Dependant Plasticity. The learning algorithm is proved by implementing a hybrid circuit of CMOS based neurons and a memristor based synapse.

REFERENCE

[1] "International Technology Roadmap for Semiconductors, Executive Summary, Emerging Research Devices Grand Challenges, p. 30; Available: www.itrs.net/Links/2011ITRS/Home2011.htm.

[2] "Can Programming Be Liberated from the von Neumann Style? A Functional Style and Its Algebra of Programs J. Bachus", Commun. ACM 1978, pp 613.

[3] K. Koch "Roadrunner System Overview" , Los Alamos Nat. Lab., Albuquerque, NM.

[4] P.Carrasco, Z.Ramos, Gotarredona, L.Barranco, "On Neuromorphic Spiking Architectures for asynchronous memristive STDP systems", Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS).

[5] S. Williams, "How We Found the Missing Memristor," IEEE Spectrum, vol. 45, no. 12, 2008, pp. 28-35.

[6] L. Chua, "Memristor - The Missing Circuit Element," IEEE Transactions on Circuits Theory (IEEE), vol. 18, no. 5, 1971, pp. 507–519

[7] O. Turel, J. H. Lee, X. Ma, and K."Neuromorphic architectures for nanoelectronic circuits: Research articles", Int. J. Circuit Theory Appl., vol. 32, no. 5, pp. 277–302, 2004.

[8] W. S. Zhao, G. Agnus, V. Derycke, A. Filoramo, J.-P. Bourgoin, and C. Gamrat, "Nanotube devices based crossbar architecture" Toward neuromorphic computing,Nanotechnology, vol. 21, no. 17, 2010175202.

[9] M. S. Zaveri,D. Hammerstrom, "Performance/price estimates for cortex-scale Hardware A design space exploration", Neural Netw., vol. 24, no. 3, pp. 291–304, 2010.

[10] G. Cauwenberghs, Norwell, Kluwer,"Neuromorphic Learning VLSI Systems: A Survey". Norwell",1998.

[11] G.Snider,"Spike Timing Dependent Plasticity in memristive Nanodevices",IEEE International Symposium on Nanoscale Architectures, 2008, NANOARCH 2008.

[12] Patlak,Joe,Gibbons, Ray "Electrical Activity of Nerves", Aps in Nerve Cells,2000.

[13] Bernabé,Linares-Barranco, T.Serrano-Gotarredona, "Memristance can explain Spike-Time-Dependent-Plasticity in Neural Synapses", Nature Precedings, 2009.

[14] Gerrow, Kimberly; Antoine, "Synaptic stability and plasticity in a floating world",Current Opinion in Neurobiology 20 (5) pp. 631–639,2010.

[15] B. Mel, "Information processing in dendritic trees," Neural Computation, vol. 6, pp. 1031–1085, 1994.

[16] M.Rahimi Azghadi, N.Iannella, Said F. Al-Sarawi,G.Indiveri, D.Abbott "Spike-Based Synaptic Plasticity in Silicon: Design, Implementation, Application, and Challenges", Proceedings of the IEEE, Vol.102 ,pp 717 - 737,2014.

[17] Hebb, D.O. "The Organization of Behavior", New York: Wiley & Sons,1949.

[18] N.Caporale,Y.Dan,"Spike timing-dependent plasticity: a Hebbian learning rule",Annual Review of Neuroscience, 31, 25-46,2008.

[19] F. Rosenblatt, "Principles of Neurodynamics", New York,Spartan, 1962.

[20] B. Widrow,M. Hoff, "Adaptive switching circuits," in 1960 IRE WESCON Convention Record.New York: IRE,vol. 4, pp. 96–104,1960.

[21] M. Minsky,S. Papert, "Perceptrons",Cambridge: MIT Press, 1969.

[22] Stent,"A physiological mechanism for Hebb's postulate of learning", Proceedings of the National Academy of Sciences,vol.70(4),pp 997–1001, 1973.

[23] U.Russo, D.Kamalanathan, D.Ielmini, A. L. Lacaita, M. N. Kozicki "Study of Multilevel Programming in Programmable Metallization Cell (PMC) Memory",IEEE transactions on electron devices, vol. 56, no. 5, 2009.

[24] K.K. Likharev, "CrossNets: Neuromorphic Hybrid CMOS/Nanoelectronic Networks",Science of Advanced Materials,2011.

[25] "CMOL implementation of spiking neurons and spike-timing dependent plasticity", Ahmad Afifi1, Ahmad Ayatollahi1 and Farshid Raissi, International Journal Of Circuit Theory And Applications.

[26] W. McCulloch,W. Pitts, "A logical calculus of ideas immanent in nervous activity," Bulletin of Mathematical Biophysics, vol. 5, 1943.

[27] A. Destexhe, Z.F.Mainen,T.J. Sejnowski,"Kinetic models of synaptic transmission," in Methods in Neuronal Modelling, From Ions to Networks,Cambridge, MA: The MIT Press,pp 1–25, 1998.

[28] M Mahowald, R Douglas, "A silicon neuron", Nature, pp 19-26, 1991.

[29] R. T. Edwards, G Cauwenberghs," Synthesis of log-domain filters from first-order building blocks ", Int. J. Analog Integr. Circuits Signal Process,2000.

[30] A.van Schaik, C.Jin, "The Tau-Cell: a new method for the implementation of arbitrary differential equations", International Symposium on Circuits and Systems, ISCAS 2003, pp 569–572,2003.

[31] Destexhe A, Huguenard JR J, "Nonlinear thermodynamic models of voltage-dependent currents", Comput Neurosci., 9(3):259-70, Nov-Dec 2000.

[32] W. Brockman, “A simple electronic neuron model incorporating both active and passive responses,”IEEE Transactions on Biomedical Engineering, vol. BME-26, pp. 635–639, 1979.

[33] Leslie S. Smith, "Implementing neural models in silicon” May 2004.

[34] E. Vittoz, H. Oguey, M. Maher, O. Nys, E. Dijkstra, and M. Cehvroulet, “Analog storage of adjustable synaptic weights,” in VLSI design of neural networks, U. Ramacher and E. Rueckert, Eds. Kluwer Academic, 1991.

[35] J.Meador, A.Wu, C.Cole, N. Nintunze, and P.Chintrakulchai, “Programmable impulse neural circuits,”IEEE Transactions on Neural Networks, vol. 2, no. 1, pp. 101–109, 1991.

[36] E. Neftci and G. Indiveri, ‘‘A device mismatch compensation method VLSI neural networks,’’ in Proc. IEEE Biomed.Circuits Syst. Conf., 2010, pp. 262–265.

[37] S. Bamford, A. Murray, and D. Willshaw,‘‘Spike-timing dependent plasticity with weight dependence evoked from physical constraints,’’ IEEE Trans. Biomed. Circuits Syst., vol. 6, no. 4, pp. 385–398, Aug. 2012.

[38] A. Bofill-I-Petit and A. Murray, ‘‘Synchrony detection and amplification by silicon neurons with STDP synapses,’’ IEEE Trans.Neural Netw., vol. 15, no. 5, pp. 1296–1304, Sep. 2004.

[39] B. Linares-Barranco and T. Serrano-Gotarredona, ‘‘On the design and characterization of femtoampere current-mode circuits,’’ IEEE J. Solid State Circuits, vol. 38, no. 8, pp. 1353–1363, Aug. 2003.

[40] S. Moradi and G. Indiveri, ‘‘An event-based neural network architecture with an asynchronous programmable synaptic memory,’’ IEEE Trans. Biomed. Circuits Syst., vol. 8, no. 1, pp. 98–107, Feb. 2014, DOI: 10. 1109/TBCAS.2013.2255873.

[41] S.-C. Liu, T. Delbruck, J. Kramer, G. Indiveri, and R. Douglas, Analog VLSI: Circuits and Principles. Cambridge, MA, USA: MIT Press, 2002.

[42] J. M. Cruz-Albrecht, M. W. Yung, and N. Srinivasa, ‘‘Energy-efficient neuron, synapse and STDP integrated circuits,’’ IEEE Trans. Biomed. Circuits Syst., vol. 6, no. 3,pp. 246–256, Jun. 2012.

[43] T. Koickal, A. Hamilton, S. Tan, J. Covington, J. Gardner, and T. Pearce, ‘‘Analog VLSI circuit implementation of an adaptive neuromorphic olfaction chip,’’ IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 54, no. 1, pp. 60–73, Jan. 2007.

[44] Mead C. A., Analog VLSI and Neural Systems. Reading, MA: Addison-Wesley,1989.

[45] Olsson J. A., Häfliger P. (2008). "Mismatch reduction with relative reset in integrate-and-fire photo-pixel array," in Biomedical Circuits and Systems Conference, BIOCAS 2008.

[46] M.N. Kozicki, C. Gopalan, M. Balakrishnan, M. Park, and M. Mitkova, "Non-volatile memory based on solid electrolytes," Proceedings of the 2004 Non-Volatile Memory Technology Symposium,10-17, Orlando, Florida, November, 2004.

[47] S. Song, K. Miller, and L. Abbott, ''Competitive Hebbian learning through spike-timing-dependent synaptic plasticity,'' Nature Neurosci., vol. 3, pp. 919–926, 2000.

[48] W.Chan, J Lohn "Spike Timing Dependent Plasticity with Memristive Synapse in Neuromorphic Systems", WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15, 2012 - Brisbane, Australia

[49] M. N. Kozicki, M. Balakrishnan, C. Gopalan, C. Ratnakumar, and M. Mitkova, "Programmable metallization cell memory based on Ag-Ge-S and Cu-Ge-S solid electrolytes," in Proc. Non-Volatile Memory Technol.Symp., p.89,2005.

[50] Indiveri G. et al. "Neuromorphic silicon neuron circuits",Front. Neurosci.,2011,Available:http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3130465.

[51] LadyofHats (2007, July 12). Diagram of a typical myelinated vertebrate motor neuron [Online].Available: https://en.wikipedia.org/wiki/Neuron.

[52] T. Serrano-Gotarredona1, T. Masquelier , T. Prodromakis , G. Indiveri and B. Linares-Barranco1," STDP and STDP variations with memristors for spiking neuromorphic learning systems,"Front.Neurosci.,2013. *