5-22-2014

Energy-Efficient STDP-Based Learning Circuits with Memristor Synapses

Xinyu Wu
Boise State University

Vishal Saxena
Boise State University

Kristy A. Campbell
Boise State University

 doi: 10.1117/12.2053359

Energy-efficient STDP-based Learning Circuits with Memristor Synapses

Xinyu Wu*, Vishal Saxena, Kristy A. Campbell
Dept. of Electrical and Computer Engineering, Boise State University
1910 University Dr., Boise, ID USA 83725

ABSTRACT

It is now accepted that the traditional von Neumann architecture, with processor and memory separation, is ill suited to process parallel data streams which a mammalian brain can efficiently handle. Moreover, researchers now envision computing architectures which enable cognitive processing of massive amounts of data by identifying spatio-temporal relationships in real-time and solving complex pattern recognition problems. Memristor cross-point arrays, integrated with standard CMOS technology, are expected to result in massively parallel and low-power Neuromorphic computing architectures. Recently, significant progress has been made in spiking neural networks (SNN) which emulate data processing in the cortical brain. These architectures comprise of a dense network of neurons and the synapses formed between the axons and dendrites. Further, unsupervised or supervised competitive learning schemes are being investigated for global training of the network. In contrast to a software implementation, hardware realization of these networks requires massive circuit overhead for addressing and individually updating network weights. Instead, we employ bio-inspired learning rules such as the spike-timing-dependent plasticity (STDP) to efficiently update the network weights locally. To realize SNNs on a chip, we propose to use densely integrating mixed-signal integrate-and-fire neurons (IFNs) and cross-point arrays of memristors in back-end-of-the-line (BEOL) of CMOS chips. Novel IFN circuits have been designed to drive memristive synapses in parallel while maintaining overall power efficiency (<1 pJ/spike/synapse), even at spike rate greater than 10 MHz. We present circuit design details and simulation results of the IFN with memristor synapses, its response to incoming spike trains and STDP learning characterization.

Keywords: Machine Learning, Memristors, Spiking Neural Networks, Spike-Timing-Dependent Plasticity (STDP).

1. INTRODUCTION

Society has realized enormous processing gains leveraging the von Neumann computing architecture. However, maintaining the growth trend now faces significant challenges at both the fundamental and practical levels. The von Neumann computer architecture is built upon the separation of memory from the processing units, which makes it a multi-purpose and generic design that has successfully leveraged the rapid exponential growth of modern electronics in the past decades; but at the same time limits the performance of the computing system architecture. This so called von Neumann bottleneck renders a processor unable to execute a program faster than it can fetch instructions and data from the memory. Many methods have been invented and used to alleviate the von Neumann bottleneck, including adding cache hierarchies between the processor and main memory, providing separate cache and separate access paths for data and instructions, using branch predictor algorithms and logic, providing a limited CPU stack or other on-chip scratchpad memory to reduce memory access, and so on. For many years, simply increasing clock speeds allowed chips to improve performance without addressing the von Neumann bottleneck, though frequency scaling has likely reached its limit following the transistor scaling in semiconductor technology is approaching atomic scale. Multiprocessor designs were used to improved processor performance in the absence of further frequency scaling as well, however, simple multicore scaling of this type will be ultimately limited by power constraints, as well as the parallelizability of the applications that will run on them. Finally, with several decades of development, trough the von Neumann computers have been very powerful, they are still incredibly inefficient at several tasks that are easy for even the simplest brains, such as recognizing images and navigating in unfamiliar spaces. Machines found in research labs or vast data centers can perform such tasks, but they are huge and energy-hungry, and they need specialized programming.

* xinyuwu@u.boisestate.edu; phone 1 208 426-3824; fax 1 208 426-2470; coen.boisestate.edu/ams

Machine Intelligence and Bio-inspired Computation: Theory and Applications VIII,
edited by Misty Blowers, Jonathan Williams, Proc. of SPIE Vol. 9119, 911906

Recently, neural inspired computing architecture and the hardware implementation of neural networks, called Neuromorphic system, has become a frontrunner for inspiring non- von Neumann computing systems. The re-emergence of Neuromorphic systems is fueled by two factors. First, more understanding has been obtained on both biological neural networks and artificial networks through experimental and modeling based studies. Second, the emergence of new classes of nano-devices, particularly two-terminal resistive switching devices (memristive devices), makes it possible to build functional Neuromorphic hardware that will not only serve to test the various neural network models but also can directly lead to new, effective, high-performance computing hardware [1,2].

Radically different from the von Neumann computer, the brain, and more specifically the mammalian cerebral cortex, is capable of harnessing a large number of inherently faulty components, is highly parallel, energy efficient, and fault tolerant. Complex pattern recognition tasks such as recognizing images, speech, text, and finding the connections among massive amount of information are almost trivial to humans. For these reasons, computing models inspired by the cortex have become a promising candidate model for future computing devices. Instead of separating the memory and processing elements, this biological system stores memory and performs computation in the same elements. Neurons perform computation by propagating spikes, and their synapses (connections between neurons) store memories through particular connectivity and their relative weights (or strengths). Specially, recent advances in computational neuroscience discovered and demonstrated unsupervised and supervised learning with spiking neural networks [3,4]. These spiking neural networks are not only useful for large-scale cortical network simulation, but also demonstrate powerful computational capabilities for engineering applications. This interest in spiking neurons has also inspired the development of many Neuromorphic hardware implementations, specifically designed for simulating cortical networks or executing neural-inspired applications. For a hardware implementation of the large-scale cortical network, it is very crucial to have very compact designs of silicon synapse and neuron.

The key to the high efficiency of biological systems is the large connectivity between neurons that offers highly parallel processing power. The synaptic weight between two neurons can be precisely adjusted by the ionic flow through them and it is widely believed that the adaptation of synaptic weights enables the biological systems to learn and function. A synapse is essentially a two-terminal device and its high-level functionality bears striking resemblance to an electrical device termed memristor. Similar to a biological synapse, the conductance of a memristor can be incrementally modified by controlling charge or flux through it. Researchers recently reported several silicon nano-scale memristive devices demonstrating such variable resistance, and then can be used in implementation of synaptic functions. In particular, compatible with this structure is the weight setting paradigm spike-timing-dependent plasticity (STDP) [5].

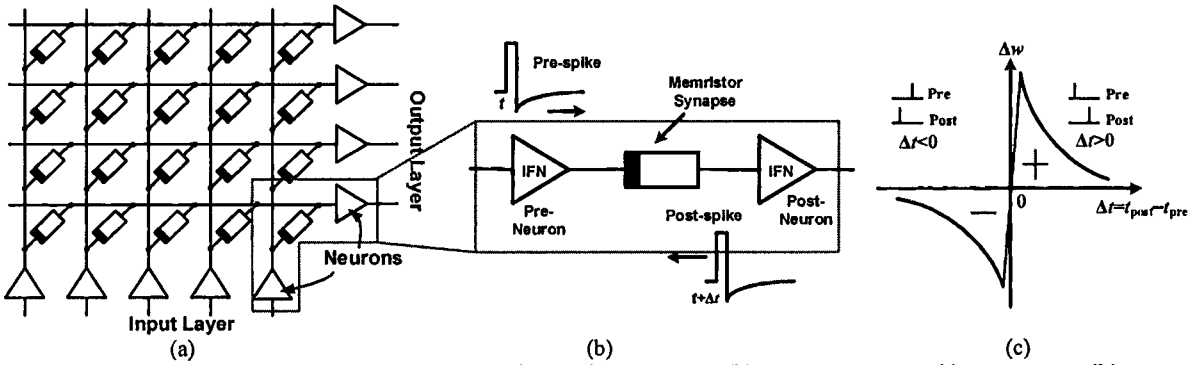

Figure 1 (a) A cross-point spiking neural network with memristor synapses, (b) a synapse connected between two spiking neurons showing pre- and post-synaptic spikes, (c) graphical depiction of the STDP rule.

STDP is an important synaptic modification rule for competitive Hebbian learning can be achieved in a hybrid synapse/neuron circuit composed of complementary metal-oxide semiconductor (CMOS) neurons and nano-scale memristor synapses [6]. Under the STDP learning rule, the node preceding the synapse fires as does the node following (post) the synapse, both influencing the weight value. The difference in timing of the pre- and post- synaptic pulses determines the weight change Δw. The difference of the pre- and post- synaptic pulses is equivalent to a traditional voltage pulse, which researchers have shown to be able to set a memristor's resistance. Memristor amenable with STDP has led to fast, low-energy, high-endurance and powerful plastic devices that can be scaled down to nanometer scale and has potential to stack in three dimensions, then supporting for the implementation large-scale memristor-based Neuromorphic systems. A promising hardware structure of such Neuromorphic systems is cross-point network: a two-

terminal memristor synapse is formed at each cross-point and connects CMOS-based pre- and postsynaptic neurons. Figure 1 shows how STDP works with the memristor in a cross-point network. The hybrid memristor/CMOS circuits discussed here can be fabricated using similar techniques developed recently for memristor-based memory and logic. The cross-point synapse network can potentially offer connectivity and function density comparable to those of biological systems and operate in a way analogous to biological systems rather than digital computers. In this case, every CMOS neuron in the pre-neuron layer of the cross-point configuration is directly connected to every neuron in post-neuron layer with unique synaptic weights. A high synaptic density of $10^{10}/cm^2$ can also be potentially obtained for cross-point networks with 100 nm pitch, a feature size readily achievable with advanced lithography approaches [6].

Since the emergence of nano-scale memristors, there has been a growing interest in integrating these devices with CMOS circuits to realize novel Neuromorphic functionality. Since the brain uses analog computation principles based on spiking neural networks (SNNs) that significantly vary from the digital von Neumann paradigm, researchers have made several investigations into VLSI (Very Large Scale Integration) implementation of neural networks [7]. These investigations of silicon neuron (SiN) circuits initiated the field of Neuromorphic Engineering intending to emulate the electrophysiological behavior of mammalian neurons and synapses. The silicon neurons are comprised of several synapse blocks and a soma block. The synapses receive the synaptic spikes from the other connected neurons and convert them into currents, and the soma blocks perform spatio-temporal integration of the spiking pulses and generate output spikes (or action potentials). Further, the dendrites and axon blocks are implemented using interconnect circuits which model the spiking signal propagation through the passive neuronal fibers and help realize larger signal processing networks, comprised of hierarchical neural blocks. To provide a background on the Neuromorphic circuits, several silicon neuron design styles have appeared in the literature, which model certain aspects of biological neuron, such as sub-threshold biophysically realistic models, compact IFN circuits, switched-capacitor neuron and digital VLSI implementations [7]. The sub-threshold biophysically realistic SiNs work on the equivalence between the transport of ions in biological channels and charge carriers in transistor channels when biased near a threshold. A prominent silicon neuron implementation is the thalamic relay neuron which models the dynamics of calcium and potassium ion relay channels in the biological neurons. Another is the sub-threshold Hodgkin-Huxley neuron which implements a non-linear channel kinetics model [7]. Faithful modeling of biological spiking neurons typically consumes a large amount of silicon layout area and are useful for biomimetic emulation of neural dynamics. However, for building large spiking Neuromorphic systems that can perform tasks faster than with a von Neumann architecture, only an empirical spiking neuron model is needed which can sufficiently abstract the underlying biophysiological processes while capturing the computational functionality of the biological neurons. Researchers have used this bio-inspired approach as an alternative to biomimetic neuron design to implement large arrays of interconnected spiking neurons. In these systems, researchers have used a compact integrate-and-fire neuron (IFN) circuit as an abstraction for the biological neuron. Further, these IFN circuits are designed to capture the transient spiking behavior of the neurons with reasonable accuracy to be useful for neural learning and require a far lower number of transistors to implement. The IFN employs modifications of the basic spike-event generation concept, called the Axon-Hillock circuit.

2. INTEGRATE AND FIRE NEURONS WITH MEMRISTIVE SYNAPSES

As discussed earlier, several silicon integrate-and-fire neurons (IFN) which mimic the electrophysiological functionality of the biological neurons have emerged in literature. However, barely any of them has leveraged the STDP-learning characteristics of the memristors.

The memristor used here was designed using carefully chosen layers of silver (Ag), metal-chalcogenide (M-Se), and chalcogenide glass, stacked between tungsten electrodes. We identify the electrode nearest the Ag layer as the "oxidizable electrode." The device operates by changing resistance according to the quantity of Ag forced into the chalcogenide glass layer[8]. First, to program the device, we force Ag into the chalcogenide glass layer by applying a potential across the device. When we apply a positive voltage above a threshold voltage ($V_{t,p}$) to the oxidizable electrode, Ag in the Ag layer is oxidized to Ag^+, which then migrates with the applied electric field, into the glass layer. When these Ag^+ ions make electrical contact with the electrode at the lower potential, Ag^+ is reduced to Ag and eventually forms a conductive filament which spans the two electrodes. Conversely, when we erase the device, we apply a negative potential of larger magnitude than a threshold voltage ($V_{t,n}$) to the oxidizable electrode, Ag^+ is generated at the more positive potential electrode and migrates with the applied electric field towards the top electrode. This process reduces the metallic contact between the two electrodes, and eventually severs the contact; both outcomes increase device

resistance. A typical range of device resistances (or memristance) is from 1 kΩ to 10 MΩ with quasi-static (DC) threshold voltage magnitude ranging from 150 mV to 250 mV [8]. The threshold voltage magnitude is, however, a function of the duration of the applied pulse. The shorter the applied pulse, the higher the threshold voltage.

To work with the above memristor, the IFNs are expected to generate spikes with the desired action potential waveform and drive the memristive synapses with pre- and post-synaptic potentials. In order to integrate currents across several memristors (with 1kΩ to10MΩ resistance range) and drive thousands of these in parallel, the conventional current-input IFN architecture cannot be directly employed; current summing and the large current drive required would be prohibitive. Instead, an opamp-based IFN is desirable as it provides the required current summing node and a large current drive capability.

Figure 2 shows a schematic diagram for the proposed leaky IFN block. Here, a low-power opamp operates in two asynchronous phases: integration and firing phases, as shown in Figure 3. In the integration phase, the opamp realizes a leaky integrator with leak rate controlled by the transistor R_{leaky}, and charges the capacitor C_{mem} resulting in "membrane potential" V_{mem}. The potential V_{mem} is compared with a threshold V_{thr} crossing which causes the spike-generation circuit to create the required action potential waveform V_{spk} and forces the opamp into the "firing phase." During the firing-phase, the opamp is reconfigured as a buffer, propagating the pre-synaptic spikes in the forward direction and post-synaptic spikes in the direction of the input synapses. In the absence of an input spike, the membrane potential rests at the refractory potential (V_{refr}). We employed folded-cascade opamp with class-AB output stage for realizing low power operation at 1MHz spiking rate. Circuit structures to further reduce standby power by shutting down sections of IFN circuit when not in integration or firing phases were also employed. Using these techniques, the IFN is realized an energy consumption less than 1 pJ/spike/synapse and less than 30μm^2 layout area in 0.18μm CMOS process.

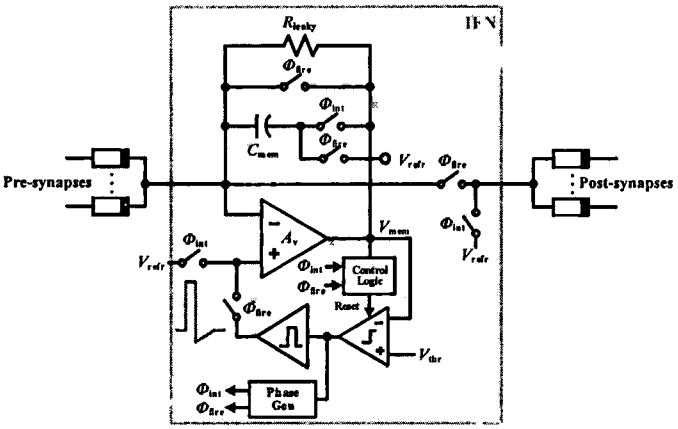

Figure 2. A leaky integrate-and-fire neuron (IFN) circuit, compatible with STDP-learning memristive synapses.

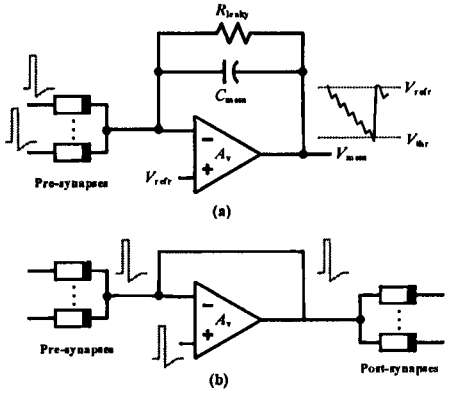

Figure 3. (a) IFN in integration phase where the opamp integrates the input spikes. (b) IFN in firing phase where the opamp drives all the memristive synapses with the spiking action potential.

The shape of the action potential function V_{spk} strongly influences the shape of the resulting STDP function. A biological-like STDP pulse with exponential rising edges is not friendly for circuit implementation. However, a biological-like STDP learning function can be achieved with a simpler action potential shape by implementing narrow short positive pulse of large amplitude and a longer slowly decreasing negative tail [9]. Moreover, with analog/digital configurable capacitor and resistor banks, an external tunability of IFNs was designed to optimize their response to the memristor characteristics (e.g., threshold voltage and the STDP program/erase pulse shape required by the fabricated Ag memristors with varied stack compositions) and the action potential shape required by learning algorithms.

To realize unprecedented on-chip integration density, we can employ dense cross-point arrays of memristive devices in the BEOL of a CMOS process, physically on top of the mixed-signal neuron arrays. The concurrent mixed-signal neural architecture, where currents through all the memristors are simultaneously integrated as opposed to a random-access memory architecture, alleviates the sneak-path issues typically associated with sensing in such dense cross-point device arrays. The Neuromorphic chips should consist of regular structures of cross-point arrays (synapses), with tiled integrate-and-fire neurons. Further, peripheral circuits to implement the SNN learning algorithms such as rank order coding of spikes will be needed to encode and decode neuron firing patterns.

3. SIMULATION RESULTS

The proposed structure shown in Figure 2 was simulated using Cadence Spectre. Figure 4 illustrates the neuron's member voltage V_{mem} and the STDP spike V_{spk} generated by the neuron. This spike is sent to pre- and post-synapses and then creates STDP of memristors' resistance. The shape of spike were adjustable to fit the characters of memristors and effective learning algorithm. Because of the opamp reconfiguration, the membrane voltage demonstrated the integration on the input spikes during the integration phase, and switched to drive a spike during firing phase later. As an example, the tunable spike amplitudes were selected to be able to produce net potentials across a memristor with write and erase threshold voltages around 0.16V and -0.15V respectively.

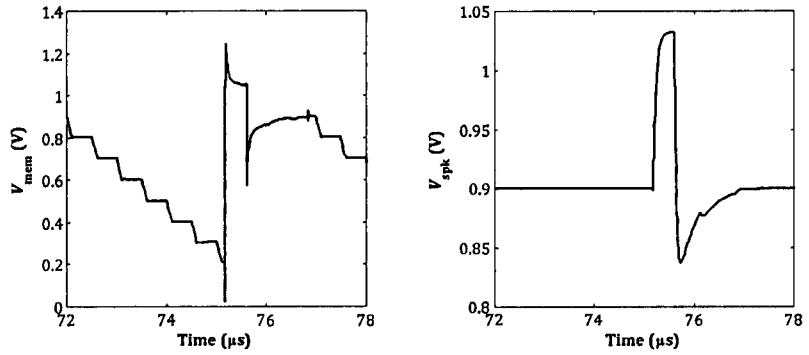

Figure 4. (a) Response of 'membrane voltage' V_{mem} and (b) output firing spike V_{spk}.

A published device model [10] has been fitted to our device and implemented using Verilog-A for circuit simulation in Cadence Spectre. To prepare the memristive device model for STDP simulation, the memristor was warming up by given a 10 kHz 6V sinusoid potential across the device for three to five cycles. This warming up procedure a transient simulation to move the memristive device model from an "initial state" to a "steady state". Then the parameter of the memristive device model to set the initial device resistance in a simulation in the range of 1 kΩ to 1 MΩ, and the device was connected into the target neural network with two neurons: a pre-synapse neuron firing pre-spikes, and a post-synapse neuron to integrate and fire post-spikes.

According to STDP learning rule, if the spike (action potential) from the pre-synaptic neuron arrives at the synaptic cleft before that of the post-synaptic neuron, potentiation will be induced. Otherwise depression will be induced. How effectively the potentiation and depression take place in turn depends on the relative timing of the pre- and post-synaptic spikes. The synaptic weight in such a circuit is effectively the conductance of the memristor; the lower the resistance of a memristor is (or the higher its conductance is) the stronger the synaptic efficiency will be, as it will let more current

through and thus affect more strongly the effective synaptic potential (ESP) of the post neuron. In the simulation presented here, post-synaptic spikes are always after the pre- synaptic spikes, simply due to post-synaptic spikes are generated once the integration of pre- synaptic spikes crossed threshold. Also with the same reason, pre-synaptic spikes and post-synaptic spikes here are correlated.

Figure 5 (a) shows the plasticity of memristor synapse with the STDP spike pairs from pre and post-neuron run across it. Each time when there is a pair of correlated pre-synaptic and post-synaptic spikes run across memristor, its resistance is reduced. In other words, the correlated pre-synaptic and post-synaptic spike pair strengthens the synaptic connection between the pre and post neurons. Figure 6 zooms in a cycle of correlated pre-synaptic and post-synaptic spike pair. The figure shows a post-synaptic spike was generated just after the last pre-synaptic came in and integrated by the neuron. As results, the action voltage of the post-synapse neuron crossed the firing threshold, and then one post-synapse spike was generated with the neuron's firing. The shape of STDP spike generated was carefully designed to be small enough to avoid to disturb memristor, at the same time, be large enough to be able create a net potential across memristor with potential above the programing threshold of the memristor. In our simulation, the STDP spike was designed with positive amplitude $A_{mp}^{+} = 140\,\text{mV}$, negative amplitude $A_{mp}^{-} = 100\,\text{mV}$, positive tail $\tau^{+} = 0.5\,\mu s$ and negative tail $\tau^{-} = 1.5\,\mu s$. With these parameters, a pre/post-synapse spike pairs with 0.5 μs arriving time difference created a 0.25 μs synapse updating time with the overlap of their negative and positive tails respectively, and resulting a decrease of resistance around 1.5 kΩ.

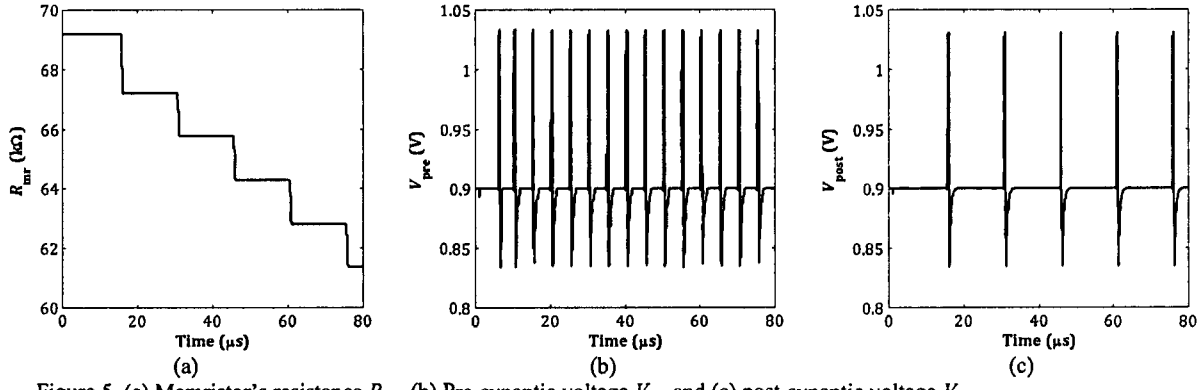

Figure 5. (a) Memristor's resistance R_{mr}, (b) Pre-synaptic voltage V_{pre} and (c) post-synaptic voltage V_{post}

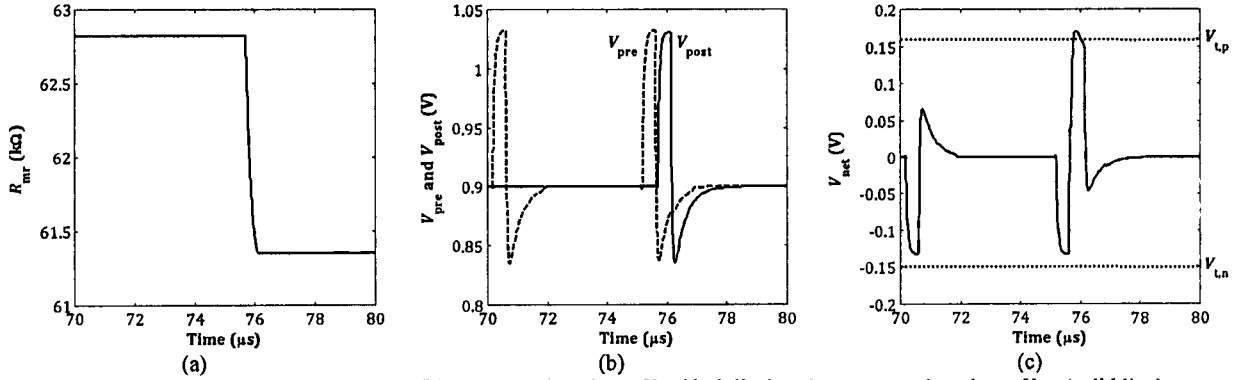

Figure 6. (a) Memristor's resistance R_{mr}, (b) pre-synaptic voltage V_{pre} (dash line) and post-synaptic voltage V_{post} (solid line), and (c) net potential across memristor $V_{net} = V_{post} - V_{pre}$. Noting the memristor programing thresholds are $V_{t,p} = 0.16$ V and $V_{t,n} = -0.15$ V in the model.

4. CONCLUSION

Significant progress has been made in spiking neural networks (SNNs) which expect to emulate data processing in the cortical brain recently. To realize SNNs on a chip, we are densely integrating mixed-signal integrate-and-fire neurons (IFNs) and cross-point arrays of memristors in back-end-of-the-line (BEOL) of CMOS chips. Further, we employ bio-inspired learning rules such as the spike-timing-dependent plasticity (STDP) to efficiently update the network weights. Novel IFN circuits have been designed to drive >1000 memristive synapses in parallel while maintaining overall power efficiency (<1 pJ/spike/synapse), even at spike rate greater than 10 MHz.

ACKNOWLEDGMENT

This work is partially supported by the National Science Foundation grant CCF-1320987.

REFERENCES

[1] Chang, T., Yang, Y., Lu, W., "Building Neuromorphic Circuits with Memristive Devices," IEEE Circuits Syst. Mag. 13(2), 56–73 (2013).

[2] Serrano-Gotarredona, T., "A proposal for hybrid memristor-CMOS spiking neuromorphic learning systems," IEEE Circuits Syst. Mag., 74–88 (2013).

[3] Masquelier, T., Guyonneau, R.., Thorpe, S. J., "Competitive STDP-based spike pattern learning," Neural Comput. 21(5), 1259–1276 (2009).

[4] Yu, Q., Tang, H., Tan, K.., Li, H., "Rapid feedforward computation by temporal encoding and learning with spiking neurons," IEEE Trans. Neural Networks Learn. Syst. 24(10), 1539–1552 (2013).

[5] Linares-barranco, B.., Serrano-gotarredona, T., "Memristance can explain spike-time-dependent-plasticity in neural synapses," Nature precedings, 1–4 (2009).

[6] Jo, S. H., Chang, T., Ebong, I., Bhadviya, B. B., Mazumder, P.., Lu, W., "Nanoscale memristor device as synapse in neuromorphic systems.," Nano Lett. 10(4), 1297–1301 (2010).

[7] Indiveri, G., Linares-Barranco, B., Hamilton, T. J., van Schaik, A., Etienne-Cummings, R., Delbruck, T., Liu, S.-C., Dudek, P., Häfliger, P., et al., "Neuromorphic silicon neuron circuits.," Front. Neurosci. 5(May), 73 (2011).

[8] Oblea, A., Timilsina, A., Moore, D., Campbell, K., "Silver chalcogenide based memristor devices," 2010 Int. Jt. Conf. Neural Networks 3, 4–6 (2010).

[9] Serrano-Gotarredona, T., Masquelier, T., Prodromakis, T., Indiveri, G., Linares-Barranco, B., "STDP and STDP variations with memristors for spiking neuromorphic learning systems," Front. Neurosci. 7(February), 2 (2013).

[10] Yakopcic, C., Taha, T., "A memristor device model," IEEE electron device, 32(10), 1436–1438 (2011).