Orientation Classification by a Winner-Take-All Network with Oxide RRAM based Synaptic Devices

Shimeng Yu
School of Computing, Informatics, and Decision Systems Engineering,
Arizona State University
Tempe, AZ 85281, USA
shimengy@asu.edu

***Abstract*— An emerging application for the oxide based resistive random access memory (RRAM) technology is to serve as the synaptic device for the hardware implementation of neuromorphic computing. The gradual resistance modulation capability in RRAM is proposed for emulating analog synapses, and the stochastic switching behavior in RRAM is proposed for emulating binary synapses. In order to evaluate the effectiveness of analog synapses and binary synapses in realizing the competitive learning algorithm, a simulation of winner-take-all network is performed based on the parameters extracted from the experiments. The simulation suggests that the orientation classification can be effectively realized using both analog synapses and binary synapses.**

***Keywords*—resistive switching, RRAM, synaptic device, winner-take-all, neural network, neuromorphic computing**

I. INTRODUCTION

The motivation of neuromorphic computing comes from the fact that the learning and intelligence have not yet been fully captured in today's computers. Today's computers are based on Boolean logic and the Von Neumann architecure with microprossers and memories seperated. The increasing performance gap between the microprosessors and the memories becomes the bottlneck of the whole system. As a result, energy efficiency is one of the major challenges of processing big data in the conventional digital computing. In such context, new computing paradigms such as neuromorphic computing are attractive since it takes advantage of massive parallelism that comes from the distributed computing and localized storage in the neural networks [1]. Neuromorphic computing is also inherently error-tolerant, thus it is especially attractive for applications such as image or speech recognition/understanding which involve input data sets in a changing and indeterministic enviroment [1]. Neuromorphic computing can be implemented in software and run by conventional digital computers. For example, the IBM team performed a cortical simulation at the complexity of the cat brain on the Blue Gene supercomputer, which required huge amount of computation resources:147,456 microprocessors and 144 TB of memories consuming a power of 1.4 MW [2]. The parallelism of a multi-core computer is still limited in comparison to the highly-distributed computing in 10^{11} neurons and 10^{15} synapses in the human brain. As an alternative approach, the hardware implementation of neuromorphic computing may physically reproduce the parallelism on chip. Previously, neuromorphic system in hardware with both neurons and synapses was implemented by CMOS circuits. The scaling-up of these systems are mainly constrained by the device density and energy consumption of the synapses since there are thousands of synapses connecting to one neuron. And each synapse is implemented with quite a few transistors, e.g. the 8-T SRAM cells [3] that occupies a huge area (>100 F^2, F is the minimum feature size of the lithography technology) and consumes substantial static power. Recently, two-terminal emerging memory devices that show electrically-triggered resistive switching phenomenon have been proposed as artificial synapse [4]. These emerging memories have the advantage of a small cell area ($4F^2$, and $4F^2/m$ if 3D stackable, m is the number of 3D stack layer). In the literature, $Ge_2Sb_2Te_5$ based phase change memory [5], Ag/a-Si [6], Ag/GeS_2 [7] and HfO_x [8] based RRAM have been reported showing synaptic behaviors. Among these candidates, oxide based RRAM is attractive for the large-scale demonstration of a neuromorphic system due to a relatively lower energy consumption (as compared to the phase change memory), the compatibility with CMOS technology and the potential for 3D integration [9]. Mb-scale to Gb-scale prototype oxide based RRAM chips have been demonstrated recently [10-11]. Therefore, a hybrid neuromorphic system with CMOS neurons and RRAM synapses can be envisioned.

The mechanism of resistive switching phenomenon in oxides has been widely attributed to the formation/rupture of the nanoscale conductive filaments which may consist of oxygen vacancies [12]. During the SET (off-to-on transition), a conductive filament is formed connecting both electrodes. During the RESET (on-to-off transition), a conductive filament is ruptured and a tunneling gap is formed between one electrode and the residual filament. The variation in the tunneling gap distance results in the multilevel resistance states. The SET transition is typically abrupt due to the positive feedback between the speed of filament growth and the increase of temperature caused by the current rise (more Joule-heating) [12]. On the other hand, the RESET transition is typically gradual due to the negative feedback between the speed of filament dissolution and the decrease of temperature

caused by the current drop (less Joule heating) [12]. Due to the asymmetry of the abrupt SET characteristics and the gradual RESET characteristics, we developed two types of synaptic behavior in oxide RRAM: 1) an analog synapse utilizing the gradual RESET for the depression learning [13]; 2) an binary synapse utilizing the abrupt SET for the stochastic learning [14]. It is believed that the analog synapse generally outperforms the binary synapse for neuromorphic computing because a limited number of synaptic states dramatically reduce the storage capacity of an artificial neural network [15]. If the synaptic strength cannot be changed by an arbitrarily small amount as in the case of the binary synapse, the newly learned patterns quickly overwrite the previously learned ones, thus the storage capacity is limited. However, this problem can be overcome by a stochastic learning rule that changes only a small fraction of synapses randomly chosen at each training cycle [15].

II. Characterization of Synaptic Behavior

Oxide RRAM synapses based on $HfO_x/TiO_x/HfO_x/TiO_x$ stack (from bottom to top) were fabricated [16]. First, we characterized the switching characteristics of the oxide synaptic device in both DC and pulse programming mode. Fig. 1 (a) shows the measured DC I-V switching curve of the fabricated device. The abrupt SET transition and gradual RESET transition is also observable in the pulse switching mode, as shown in Fig. 1 (b) and (c). When the repetitive SET pulse (+1.7 V/10 ns) was applied to the device in the off-state, the potentiation process is abrupt and only two states can be obtained. In contrast, when repetitive RESET pulse (-1.3 V/10 ns) was applied to the device in the on-state, the depression process is gradual and multilevel intermediate states can be obtained.

The RRAM device can serve as an analog synapse with the gradual RESET. Fig. 2 (a) shows the RESET transition starting from on-state (~500Ω) with different RESET pulse amplitudes: a lower amplitude leads to a more gradual RESET transition than a higher amplitude does. Therefore, -1.1 V/10 ns is chosen as an optimized programming condition for analog synapse. Then a compact model of filament dissolution was developed to capture this gradual RESET transition. The model was fitted with the experimental data, and the details of the model can be referred to [13]. The RRAM device can also serve as a binary synapse with stochastic SET because the SET transition becomes probabilistic under a weak programming condition. We measured the pulse amplitudes required for triggering the SET transition (with fixed 10 ns pulse width) during 100 cycles in one device and repeated such testing for 50 different devices. Fig. 2 (b) shows the measured statistical distribution: for a particular device, the pulse amplitude for a successful SET operation roughly follows a Gaussian distribution (a straight line in this plot indicates a Gaussian distribution) with a standard deviation about 0.3 V; across various devices, the medium pulse amplitude for a successful SET operation is centered around 1.95 V with a standard deviation about 0.15 V.

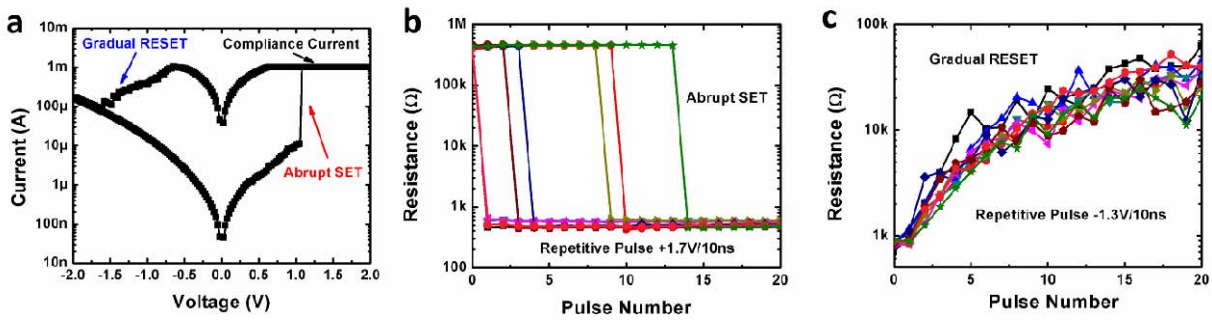

Fig. 1 (a) Measured DC I-V switching characteristics of the oxide RRAM device. (b) Measured abrupt SET transition starting from the off-state (~500 kΩ) by repetitive SET pulses (+1.7 V/10 ns), in which case the device functions as a binary synapse. (c) Measured gradual RESET transition starting from the on-state (~500 Ω) by repetitive RESET pulses (-1.3 V/10 ns), in which case the device functions as an analog synapse. Results from 10 independent testing runs are shown [14].

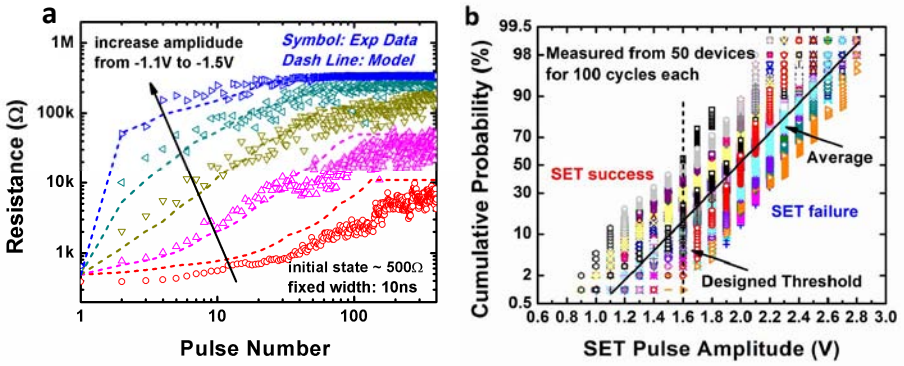

Fig. 2 (a) Measured gradual RESET transition by hundreds of RESET pulses with various amplitudes. (b) Measured statistical distribution of pulse amplitude required for triggering the SET switching. 50 different devices were measured (one type of symbol in the figure represents the data from one device), and in each device 100 continuous cycles were measured [13, 14].

III. WINNER-TAKE –ALL NETWORK SIMULATION

To validate the learning rule with both analog and binary RRAM synapses, we perform a simulation of a two-layer winner-take-all neural network [17] as a toy model. Fig. 3 shows the network architecture implemented by integrate-and-fire neurons and RRAM synapses: The input layer neurons fire according to the light intensity of the input pattern; The output layer neurons sum and integrate the input currents from all the excitatory synapses on the membrane capacitor independently, and the one with the largest input current fires first (becomes the "winner"), then it prevents all the other output layer neurons from firing ("takes all") through the inhibitory synapses. Meanwhile this winner neuron sends spikes back to all the input layer neurons and change the excitatory synapses' conductance according to the input pattern. The designed spiking scheme can be referred to [14]. In the following simulation, 32×32 neurons in the input layer are used and 2×2 neurons in the output layer are used. Thus there are 4096 oxide synaptic devices between the two layers. During the training, 200 gray-scale images of a 2D Gaussian bar with random orientation were presented to the input layer neurons. These orientations have a non-uniform distribution (centered at 0^o, 45^o, 90^o, and 135^o with a standard deviation of 7.5^o). The target of the network is to converge at these 4 dominate orientations.

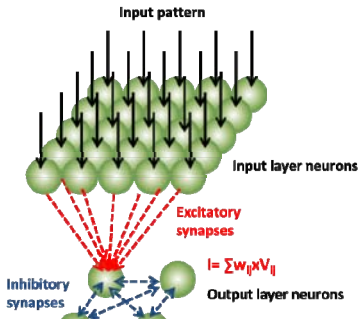

Fig. 3 Neuromorphic system based on winner-take-all neural network.

Fig. 4 shows the evolution of the normalized conductance map between the input layer neurons and the output layer neurons for the binary synapse with stochastic learning (a)-(c) and the analog synapse with depression learning (d)-(f). Initially, the resistances of all the RRAM synapses were randomized with a distribution centered at an on-state (~500 Ω). After the training, the resistances split into groups of the on-state and the off-state. With appropriate programming condition, 4 distinct orientations emerge, as shown in Fig. 4 (b) for the binary synapse using +1.6 V/10 ns SET pulse and in Fig. 4 (e) for the analog synapse using -1.1 V/10 ns RESET pulse. If the programming condition not optimized, only 3 distinct orientations emerge, as shown in Fig. 4 (c) for the binary synapse using +2 V/10 ns SET pulse and in Fig. 4 (f) for the analog synapse using -1.4 V/10 ns RESET pulse. To compare the system performance between the binary synapse and the analog synapse, three metrics were used: 1) the orientation selectivity defined as the contrast of the output layer neuron's response intensity to the 1st preferred orientation over the 2nd preferred orientation; 2) the orientation storage capacity defined as the number of distinct orientations stored in the output layer (ideally 4 distinct orientations); 3) the energy consumed on the synapses during the whole training, including the read energy for summing the current through the synapses and the write energy for programming the synapses.

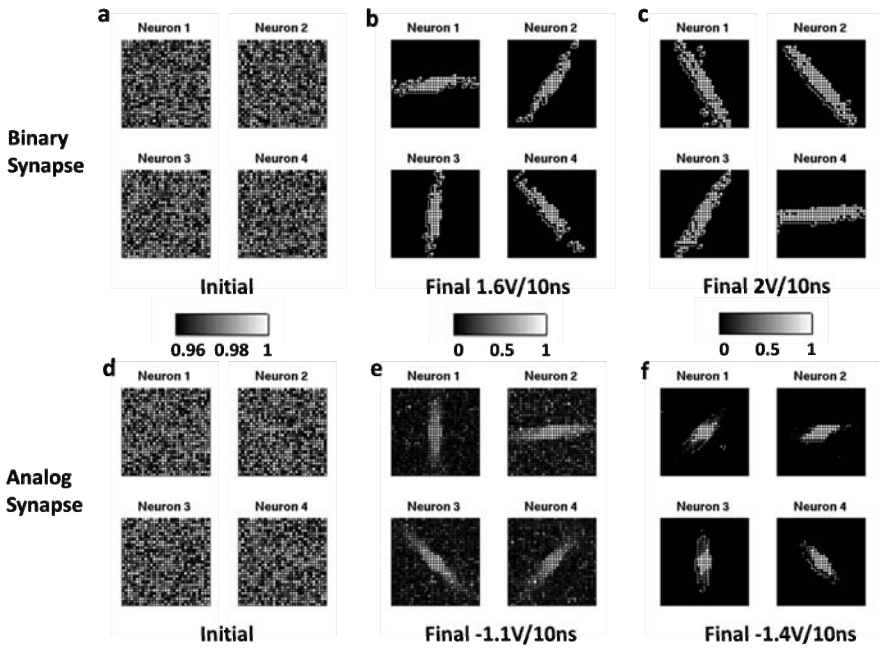

Fig. 4 Simulated normalized conductance map between the input layer neurons and the output layer neurons utilizing binary synapse with stochastic learning (a)-(c) and analog synapse with depression learning (d)-(f) [14].

Fig. 5 shows the average values of these metrics as a function of programming conditions for the binary synapse (a)-(b) and for the analog synapse (c)-(d) through 100 independent simulation runs. For the binary synapse, +1.6 V/10 ns is chosen as the optimized programming condition, which corresponds to a SET success probability ~ 12% on average. The loss of the orientation storage capacity below 1.6 V SET pulse amplitude is due to insufficient SET success probability, which limits the ability of the network to learn sufficient patterns for a limited set of training images. On the other hand, the rapid drop of the orientation storage capacity beyond 1.6 V SET pulse amplitude is due to excessive SET success probability, which hastens the network's forgetting process (overwriting the learned patterns too frequently), thus only the final patterns are remembered (see Fig. 4 (c) as an example). For the analog synapse, increasing the RESET amplitude means that the RESET transition becomes less gradual and fewer intermediate states are available (See Fig. 2 (a)). As a result, both the selectivity

and the orientation storage capacity decreases with increasing RESET pulse amplitude (see Fig. 5 (f) as an example). Here -1.1 V/10 ns is chosen as the optimized programming condition for the analog synapse. At the optimized programming condition for the binary synapse and the analog synapse respectively, the same full network storage capacity of 100% is achievable, the selectivity of the binary synapse is 14.1% and that of the analog synapse is 9.9%, and the total energy consumption of the binary synapse is 156 μJ and the that of the analog synapse is 60 μJ. The feasibility of the stochastic learning with the binary synapse is demonstrated through this system-level simulation.

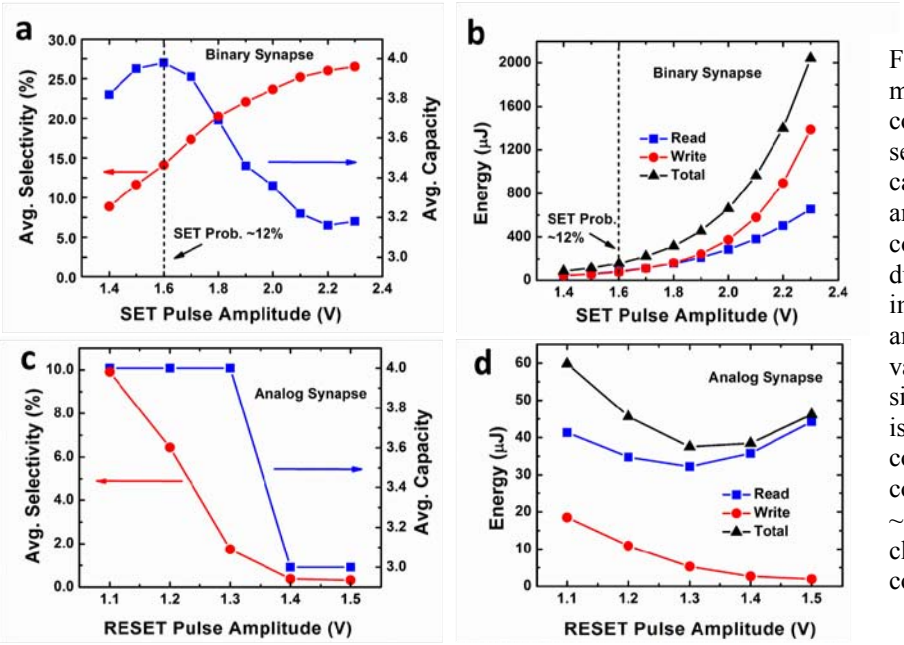

Fig. 5 Simulated system performance metrics as a function of programming conditions. Network orientation selectivity and orientation storage capacity for binary synapse in (a) and for analog synapse in (c); Energy consumption of the synaptic devices during the whole training (200 training images) for binary synapse in (b) and for analog synapse in (d). The average values through 100 independent simulation runs are shown. +1.6 V/ 10 ns is chosen as the optimized programming condition for binary synapse, which corresponds to a SET success probability ~ 12% on average. And -1.1 V/ 10 ns is chosen as the optimized programming condition for analog synapse [14].

IV. Conclusion

In this paper, we demonstrate that RRAM can function as the analog synapse utilizing the gradual RESET and the binary synapse utilizing the stochastic SET for the competitive learning. A simulation of winner-take-all network was performed for orientation classification function, showing comparable system performance between the analog synapse and the binary synapse. The significance of this demonstration is that it opens up new opportunities for a new variety of material and device choices for implementing neuromorphic computing in the hardware.

Acknowledgment

The author acknowledges the support from the Samsung GRO program.

References

[1] C.-S. Poon, K. Zhou, "Neuromorphic silicon neurons and large-scale neural networks: challenges and opportunities," *Front. Neurosci.,* vol. 5, 108, 2011.

[2] R. Preissl, et al., "Compass: A scalable simulator for an architecture for cognitive computing," *ACM/IEEE Conference High Performance Networking Computing, Storage and Analysis*, 2012.

[3] P. Merolla, et al., "A digital neurosynaptic core using embedded crossbar memory with 45pJ per spike in 45nm," *IEEE Custom Integrated Circuits Conference,* 2011.

[4] D. Kuzum, et al., "Synaptic electronics: materials, devices and applications," *Nanotechnology*, vol. 24, 382001, 2013.

[5] D. Kuzum, et al., "Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing," *Nano Lett.,* vol. 12, pp. 2179-2186, 2012.

[6] S. H. Jo, et al., "Nanoscale memristor device as synapse in neuromorphic systems," *Nano Lett.,* vol. 10, pp. 1297-1301, 2010.

[7] S. Yu, et al., "An electronic synapse device based on metal oxide resistive switching memory for neuromorphic computation," *IEEE Trans. Electron Devices,* vol. 58, pp. 2729-2737, 2011.

[8] M. Suri, et al., "CBRAM devices as binary synapses for low-power stochastic neuromorphic systems: auditory (cochlea) and visual (retina) cognitive processing applications," *IEEE IEDM*, pp. 235-238, 2012.

[9] S. Yu, et al., "A HfO_x based vertical resistive switching random access memory for bit-cost-effective three-dimensional cross-point architecture," *ACS Nano*, vol. 7, pp. 2320-2325, 2013.

[10] A. Kawahara, et al., "An 8Mb multi-layered cross-point ReRAM macro with 443MB/s write throughput," *IEEE ISSCC*, pp. 432–434, 2012.

[11] T.-Y. Liu, et al., "A 130.7mm^2 2-layer 32Gb ReRAM memory device in 24nm technology" *IEEE ISSCC*, pp. 210–212, 2013.

[12] H.-S. P. Wong, et al., "Metal oxide RRAM," *Proc. IEEE*, vol. 100, pp. 1951-1970, 2012.

[13] S. Yu, et al., "A low energy oxide-based electronic synaptic device for neuromorphic visual system with tolerance to device variation," *Adv. Mater.*, vol. 25, pp. 1774-1779, 2013.

[14] S. Yu, et al., "Stochastic learning in oxide binary synaptic device for neuromorphic computing," *Front. Neurosci.* 7, 186, 2013.

[15] W. Senn, and S. Fusi, "Convergence of stochastic learning in perceptrons with binary synapses," *Phys. Rev. E*, vol. 71, 061907, 2005.

[16] Z. Fang, et al., "$HfO_x/TiO_x/HfO_x/TiO_x$ multilayer-based forming-free RRAM devices with excellent uniformity," *IEEE Electron Device Lett.,* vol. 32, pp. 566-568, 2011.

[17] C. Zamarreno-Ramos, et al., "On spike-timing-dependent-plasticity, memristive devices, and building a self-learning visual cortex," *Front. Neurosci.,* vol. 5, 26, 2011.