Optimization of Conductance Change in $Pr_{1-x}Ca_xMnO_3$-Based Synaptic Devices for Neuromorphic Systems

Jun-Woo Jang, *Student Member, IEEE*, Sangsu Park, Geoffrey W. Burr, *Senior Member, IEEE*, Hyunsang Hwang, *Senior Member, IEEE*, and Yoon-Ha Jeong, *Fellow, IEEE*

***Abstract*— The optimization of conductance change behavior in synaptic devices based on analog resistive memory is studied for the use in neuromorphic systems. Resistive memory based on $Pr_{1-x}Ca_xMnO_3$ (PCMO) is applied to a neural network application (classification of Modified National Institute of Standards and Technology handwritten digits using a multilayer perceptron trained with backpropagation) under a wide variety of simulated conductance change behaviors. Linear and symmetric conductance changes (e.g., self-similar response during both increasing and decreasing device conductance) are shown to offer the highest classification accuracies. Further improvements can be obtained using nonidentical training pulses, at the cost of requiring measurement of individual conductance during training. Such a system can be expected to achieve, with our existing PCMO-based synaptic devices, a generalization accuracy on a previously-unseen test set of 90.55%. These results are promising for hardware demonstration of high neuromorphic accuracies using existing synaptic devices.**

***Index Terms*— Resistive random-access memory (ReRAM), memristor, long-term potentiation (LTP), long-term depression (LTD), hardware neural network (HNN), bio-inspired system.**

I. Introduction

ARTIFICIAL neural networks (ANNs) are computing architectures that consist of an extremely large number of simple processors with many interconnections [1]. Because of their massively parallel structure, such ANNs can perform complex computations such as perception and cognition [1]. While ANNs can be implemented in software or hardware, software implementations are energy-inefficient for large systems because the underlying hardware is still based on sequential Von Neumann-based processors. Hardware implementations of ANNs, known as neuromorphic systems, could potentially realize massive-parallelism approaching those of biological nervous systems, but the large number of neurons and synapses mandate high device density and extremely low power consumption. To date, appropriate synaptic devices for the neuromorphic systems have not yet been conclusively identified [2].

Resistive memory technology is a promising synaptic device due to its analog memory characteristics offering many "intermediate" conductance states, the high-density of the cross-point array structure, and low power consumption [3]. In some resistive memory devices, the application of successive pulses can smoothly change analog conductance in one direction (increasing), but conductance change in the other direction (decreasing) is regrettably abrupt, returning to the conductance extrema after a single pulse. A neuromorphic solution that minimizes the effects of this abrupt switching has been reported [4], but requires additional neuron or external-control circuits. Resistive memory devices that offer bidirectional analog conductance change should lead to more power- and area-efficient systems. However, such devices can exhibit various conductive-change behaviors, and the resulting impact on neuromorphic systems has not previously been examined.

In this letter, we study the optimum potentiation and depression characteristics of bidirectional synaptic devices for neuromorphic systems. To investigate the effects of the potentiation and depression characteristics on the system, we simulate a multilayer perceptron [5] for the application of handwritten digit classification, using the measured bidirectional switching characteristics of $Pr_{1-x}Ca_xMnO_3$ (PCMO)-based synaptic devices.

II. Methods

We fabricated 1k-bit PCMO-based resistive memory arrays for evaluation as synaptic devices (Fig. 1), extending upon earlier work [6]. For device fabrication, a 50-nm-thick Pt layer for a bottom electrode and a 30-nm-thick polycrystalline PCMO film were deposited and patterned using conventional lithography and reactive ion etching. Next, an 80-nm-thick SiN_x layer was deposited by chemical vapor deposition, and via-holes (ranging in size from 0.15 to 1.0 μm) were

Manuscript received March 17, 2015; accepted March 28, 2015. Date of publication April 2, 2015; date of current version April 22, 2015. This work was supported in part by the Ministry of Science, ICT and Future Planning (MSIP), Korea, under the "IT Consilience Creative Program" (NIPA-2014-H0201-14-1001) supervised by the National IT Industry Promotion Agency (NIPA), and in part by the Pioneer Research Center Program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (2012-0009460). The review of this letter was arranged by Editor T. Wang.

J.-W. Jang is with the Department of Creative IT Engineering, Pohang University of Science and Technology, Pohang 790-784, Korea (e-mail: junwoo410@postech.ac.kr).

S. Park is with the Department of Nanobio Material and Electronic, Gwangju Institute of Science and Technology, Gwangju 500-712, Korea.

G. W. Burr is with IBM Research-Almaden, San Jose, CA 95120 USA.

H. Hwang is with the Department of Material Science and Engineering, Pohang University of Science and Technology, Pohang 790-784, Korea (e-mail: hwanghs@postech.ac.kr).

Y.-H. Jeong is with the Department of Electrical Engineering, Pohang University of Science and Technology, Pohang 790-784, Korea (e-mail: yhjeong@postech.ac.kr).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LED.2015.2418342

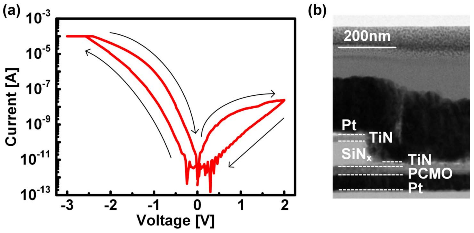

Fig. 1. (a) Current-voltage characteristics of TiN/PCMO based resistive memory and (b) a TEM image of the device.

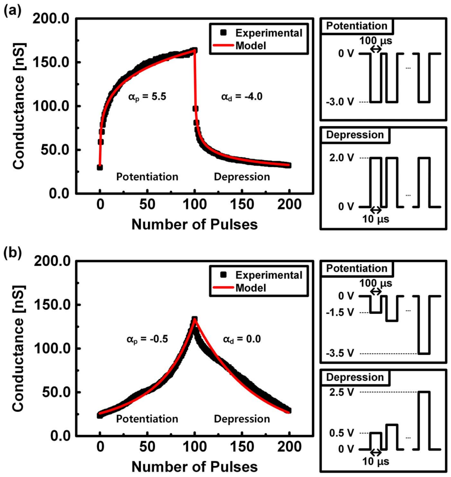

Fig. 2. Potentiation and depression characteristics of the experimental TiN/PCMO resistive memory and the proposed resistive-memory-based synaptic device model when (a) identical pulses are applied and (b) non-identical (increasing amplitude) pulses are applied.

formed by conventional lithography and reactive ion etching. A 10-nm-thick TiN layer and an 80-nm-thick Pt layer for a top electrode were deposited and patterned by conventional lithography. Electrical characteristics of the resistive memory devices based on PCMO were measured using an Agilent B1500A (Fig. 2). Reads were performed at −1.0 V; write currents ranged from ∼0.1 nA to ∼1.0 mA.

Based on the measured device characteristics, we performed simulations of a neural network with three layers of synapses (Fig. 3ab) using the backpropagation supervised learning algorithm [5]. During the forward propagation step (Fig. 3a), the activations (x) of each neuron are propagated through synaptic weights (W) to the next neurons and passed through a nonlinear function $f()$. In this letter, we do not specify any particular encoding of x (using some combination of analog voltage and/or read-pulse duration/count), nor the associated CMOS circuitry. Instead, our focus remains solely on whether the conductance change of the PCMO devices could potentially deliver adequate neural network performance even with ideal encoding of x values.

In the output layer, the neural network computes error terms (δ) between the desired outputs (d) and the actual outputs (y), which are then back-propagated through synapses to all previous layers. Synaptic weights are updated by programming pulses that depend on a scalar learning rate (η), the activation of the pre-synaptic neuron (x_i), and the back-propagated error term (δ_j) at the post-synaptic neuron [7].

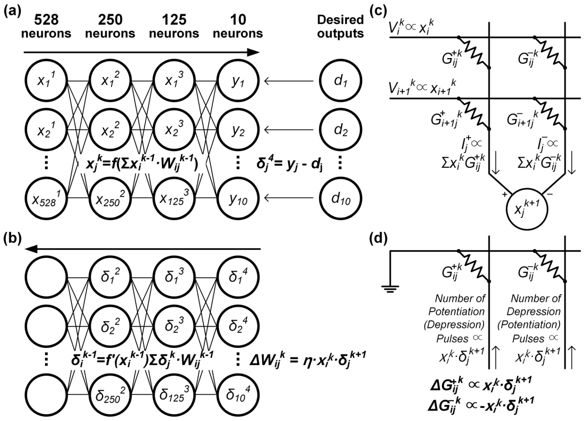
(a)
528 neurons
250 neurons
125 neurons
10 neurons
Desired outputs
x_1^1
x_2^1
x_{528}^1
x_1^2
x_2^2
x_{250}^2
x_1^3
x_2^3
x_{125}^3
y_1
y_2
y_{10}
d_1
d_2
d_{10}
$x_j^k=f(\Sigma x_i^{k-1}\cdot W_{ij}^{k-1})$
$\delta_j^4= y_j - d_j$
(b)
δ_1^2
δ_2^2
δ_{250}^2
δ_1^3
δ_2^3
δ_{125}^3
δ_1^4
δ_2^4
δ_{10}^4
$\delta_i^{k-1}=f'(x_i^{k-1})\Sigma\delta_j^k\cdot W_{ij}^{k-1}$
$\Delta W_{ij}^k = \eta\cdot x_i^k\cdot\delta_j^{k+1}$
(c)
$V_i^k \propto x_i^k$
G_{ij}^{+k}
G_{ij}^{-k}
$V_{i+1}^k \propto x_{i+1}^k$
G_{i+1j}^{+k}
G_{i+1j}^{-k}
$I_j^+ \propto \Sigma x_i^k G_{ij}^{+k}$
$I_j^- \propto \Sigma x_i^k G_{ij}^{-k}$
+
−
x_j^{k+1}
(d)
G_{ij}^{+k}
G_{ij}^{-k}
Number of Potentiation (Depression) Pulses $\propto x_i^k\cdot\delta_j^{k+1}$
Number of Depression (Potentiation) Pulses $\propto x_i^k\cdot\delta_j^{k+1}$
$\Delta G_{ij}^{+k} \propto x_i^k\cdot\delta_j^{k+1}$
$\Delta G_{ij}^{-k} \propto -x_i^k\cdot\delta_j^{k+1}$

Fig. 3. (a) Forward propagation and (b) backward propagation of a multilayer neural network. In circuitry for (c) the propagations and (d) conductance (weight) updates, they can be implemented by encoding the variables or their combinations to voltage or current signals.

Although conductance cannot be negative in real devices, the signed weights required by this learning algorithm can be realized on a crossbar array of PCMO-based resistive memories by using two devices, G^+ and G^-, for each synapse [4]. The conductance difference, $W = G^+ - G^-$, serves as the synaptic weight (Fig. 3c). When the algorithm calls for a weight increase, for instance, a number of potentiating pulses (negative pulse voltage, Fig. 2) proportional to the desired synaptic weight update are applied to G^+ while depressing pulses (positive pulse voltage, Fig. 2) are applied to G^- (Fig. 3d) [7]. Weight decreases are handled analogously, except that G^+ depresses and G^- potentiates.

Neural network simulations were performed using a simulator designed for evaluating resistive memory devices as synapses, which was previously shown to accurately predict the performance of a network trained on actual memory hardware [7]. We used the Modified National Institute of Standards and Technology handwritten dataset [8], which has a training set of 60,000 digits and a test set of 10,000 digits. Each 28×28 pixel image was cropped to the central 22×24 pixels, and the neural network was repeatedly trained on first 5,000 digits in the training set (for 50 "epochs"). Test accuracies were calculated using the full test set. For comparison purposes, note that the ideal version of this network achieves a "test" accuracy of 94% with this number of training examples [7].

III. Synaptic Device Modeling

Previously, we proposed a resistive-memory-based synaptic device model (1) for various potentiation and depression characteristics to find conductance change behavior which can optimize the performance of a neuromorphic system [9].

$$G = \begin{cases} \left(\left(G_{LRS}^{\alpha} - G_{HRS}^{\alpha}\right) \times w + G_{HRS}^{\alpha}\right)^{1/\alpha} & \text{if } \alpha \neq 0, \\ G_{HRS} \times \left(G_{LRS}/G_{HRS}\right)^{w} & \text{if } \alpha = 0. \end{cases} \quad (1)$$

where G_{LRS} and G_{HRS} are low resistance state (LRS) and high resistance state (HRS) conductance respectively, α is a parameter that controls potentiation (α_p) or depression (α_d) characteristics, and w is an internal variable which ranges from 0 to 1. During learning, w increases or decreases as

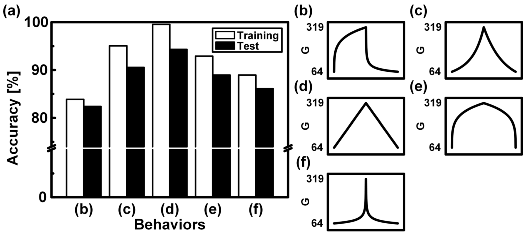

Fig. 4. (a) Calculated classification accuracies when (b) identical pulses are used ($a_p = 5.5, a_d = -4.0$), (c) non-identical pulses are used ($a_p = -0.5, a_d = 0.0$), (d) the conductance-change behavior is perfectly linear and symmetric ($a_p = a_d = 1.0$), (e) and (f) the conductance-change behaviors are non-linear but are mirror-symmetric ($a_p = a_d = 5.5$, and $a_p = a_d = -4.0$, respectively).

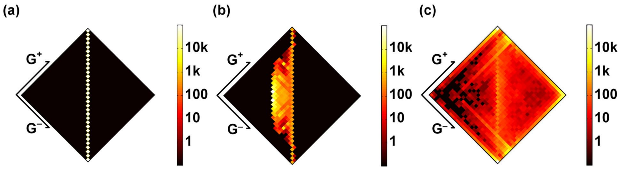

Fig. 5. Diamond-shaped plots of G^+ vs. G^- (weight is vertical position [7]) (a) before training, and (b) after training when identical pulses are used ($a_p = 5.5$, $a_d = -4.0$) or (c) after training when non-identical pulses are used ($a_p = -0.5$, $a_d = 0.0$).

potentiating (depressing) pulses are applied to the resistive-memory-based synaptic device. The potentiation and depression characteristics of the resistive-memory-based synaptic device model are concave-down if $\alpha > 1$, and concave-up if $\alpha < 1$.

IV. Results and Discussions

To investigate the effect of α on the neuromorphic system, we evaluated neural network accuracies as both α_p and α_d were varied. The fixed, unit-less G_{LRS} and G_{HRS} values (64 and 319), and the size of the smallest change in w (0.004), were based on the measurement data (Fig. 2), resulting in an on-off ratio of 5 and 256 effective multiple-conductance-levels. From these simulations, we expect the conductance response in these PCMO devices to lead to a classification "test" accuracy of 82.38% when identical pulses are used (Fig. 4b), and 90.55% when non-identical pulses are used (Fig. 4c). The highest possible accuracy, which occurs when the switching behavior is perfectly linear and symmetric (Fig. 4d), is 94.31%. (At this point, the only remaining non-ideality is the constraint on weight magnitude imposed by the finite G_{LRS} [7].) However, similarly high accuracies, 88.93% (Fig. 4e), and 86.12% (Fig. 4f), can be obtained for non-linear conductance responses, so long as the increasing and decreasing conductance responses are mirror-symmetric.

G-diamond plots (Fig. 5), first introduced in [7], can be used to represent distributions of G^+ and G^- in the neural network. Such plots represent both conductance values together with the resulting synaptic weight, $G = G^+ - G^-$ (as vertical position within the diamond) [7]. The weights, initially distributed uniformly along the center axis of the G-diamond (Fig. 5a), spread out during neural network training. When identical pulses are used, the resulting G^+ and G^- values tend to concentrate around low weight values (Fig. 5b), preventing the neural network from utilizing the full-range of possible weights. On the other hand, when non-identical pulses are used, G^+ and G^- are spread out more (Fig. 5c), allowing the neural network to utilize the full-range of weights.

However, the use of non-identical training pulses in such a neuromorphic system requires additional external circuits, because the system has to read the conductance of a synaptic device before programming it, in order to identify which non-identical training pulse to apply. Thus there is a trade-off between the higher accuracy, and the resulting lower chip-area efficiency, higher power, and longer training time associated with the need to repeatedly measure individual conductances.

Total power is difficult to estimate without specifying the CMOS circuitry. With our current devices, training power would certainly be dominated by the large PCMO write energy (currently, 300 nJ/pulse = (100 μs)×(3 V)×(1 mA)). However, further increases in scaling (from 1 μm diameter or $\sim 8\times10^5$ nm^2 down to 20 nm diameter $\sim$300 nm^2) can be expected to reduce this by at least three orders of magnitude [10].

V. Conclusion

The effects of conductance-change behavior in synaptic devices on the performance of neuromorphic systems are demonstrated. The switching behavior of a resistive memory device based on PCMO is measured, followed by neural network simulations using this conductance-change behavior. Handwritten digit classification accuracies are high for resistive-memory-based synaptic device with symmetric switching behavior, and are maximized when that response is linear. We can achieve nearly symmetric and linear behavior by using non-identical training pulses, albeit at some cost in area efficiency, training time and training power. Because synaptic switching behavior can strongly affect the resulting system performance, these results can be used to guide the development of PCMO-based resistive memories as synaptic devices towards high performance neuromorphic systems.

References

[1] A. K. Jain, J. Mao, and K. M. Mohiuddin, "Artificial neural networks: A tutorial," *Computer*, vol. 29, no. 3, pp. 31–44, Mar. 1996.

[2] B. Rajendran *et al.*, "Specifications of nanoscale devices and circuits for neuromorphic computational systems," *IEEE Trans. Electron Devices*, vol. 60, no. 1, pp. 246–253, Jan. 2013.

[3] S. H. Jo *et al.*, "Nanoscale memristor device as synapse in neuromorphic systems," *Nano Lett.*, vol. 10, no. 4, pp. 1297–1301, Apr. 2010.

[4] O. Bichler *et al.*, "Visual pattern extraction using energy-efficient '2-PCM synapse' neuromorphic architecture," *IEEE Trans. Electron Devices*, vol. 59, no. 8, pp. 2206–2214, Aug. 2012.

[5] S. O. Haykin, "Multilayer perceptrons," in *Neural Networks and Learning Machines*, 3rd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2008.

[6] S. Park *et al.*, "RRAM-based synapse for neuromorphic system with pattern recognition function," in *IEEE IEDM Tech. Dig.*, Dec. 2012, pp. 10.2.1–10.2.4.

[7] G. W. Burr *et al.*, "Experimental demonstration and tolerancing of a large-scale neural network (165,000 synapses), using phase-change memory as the synaptic weight element," in *IEEE IEDM Tech. Dig.*, Dec. 2014, pp. 29.5.1–29.5.4.

[8] Y. LeCun *et al.*, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[9] J.-W. Jang *et al.*, "ReRAM-based synaptic device for neuromorphic computing," in *Proc. IEEE ISCAS*, Jun. 2014, pp. 1054–1057.

[10] J. Lee *et al.*, "Materials and process aspect of cross-point RRAM (invited)," *Microelectron. Eng.*, vol. 88, no. 7, pp. 1113–1118, Jul. 2011.