Received 29 August 2017; revised 30 October 2017; accepted 10 November 2017. Date of publication 6 December 2017;
date of current version 8 January 2018. The review of this paper was arranged by Editor S. Moshkalev.

Digital Object Identifier 10.1109/JEDS.2017.2780275

Bidirectional Non-Filamentary RRAM as an Analog Neuromorphic Synapse, Part I: Al/Mo/$Pr_{0.7}Ca_{0.3}MnO_3$ Material Improvements and Device Measurements

KIBONG MOON[1] (Student Member, IEEE), ALESSANDRO FUMAROLA[2,3], SEVERIN SIDLER[2,4], JUNWOO JANG[5] (Member, IEEE), PRITISH NARAYANAN[2] (Member, IEEE), ROBERT M. SHELBY[2], GEOFFREY W. BURR[2] (Senior Member, IEEE), AND HYUNSANG HWANG[1] (Senior Member, IEEE)

1 Department of Materials Science and Engineering, Pohang University of Science and Technology, Pohang 37673, South Korea
2 IBM Research–Almaden, San Jose, CA 95120, USA
3 Max Planck Institute of Microstructure Physics, 06120 Halle, Germany
4 École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
5 Samsung, Suwon 443-803, South Korea

CORRESPONDING AUTHOR: G. W. BURR (e-mail: gwburr@us.ibm.com)

This work was supported by the PIONEER Research Center Program through the National Research Foundation of Korea funded by the Ministry of Science, ICT & Future Planning under Grant 2012-0009460.

ABSTRACT We report on material improvements to non-filamentary RRAM devices based on $Pr_{0.7}Ca_{0.3}MnO_3$ by introducing an MoOx buffer layer together with a reactive Al electrode, and on device measurements designed to help gauge the performance of these devices as bidirectional analog synapses for on-chip acceleration of the backpropagation algorithm. Previous Al/PCMO devices exhibited degraded LRS retention due to the low activation energy for oxidation of the Al electrode, and Mo/PCMO devices showed low conductance contrast. To control the redox reaction at the metal/PCMO interface, we introduce a 4-nm interfacial layer of conducting MoOx as an oxygen buffer layer. Due to the controlled redox reaction within this Al/Mo/PCMO device, we observed improvements in both retention and conductance on/off ratio. We confirm bidirectional analog synapse characteristics and measure "jump-tables" suitable for large scale neural network simulations that attempt to capture complex and stochastic device behavior [see companion paper]. Finally, switching energy measurements are shown, illustrating a path for future device research toward smaller devices, shorter pulses and lower programming voltages.

INDEX TERMS Resistive RAM, neural network hardware, nonvolatile memory.

I. INTRODUCTION

Neuromorphic systems offer strong potential for fault-tolerant, massively parallel, energy-efficient computation [2]. Tasks involving image classification, speech recognition, machine translation and other pattern recognition tasks can potentially be more efficiently implemented in such architectures than in conventional Von-Neumann hardware. For a practical VLSI implementation, both CMOS-based neurons and peripheral circuitry as well as artificial synapses for storing weight data will be required.

Numerous neuromorphic algorithms have been discussed, ranging from unproven and still immature brain-inspired Spiking Neural Network algorithms [2], [8] such as Spike-Timing-Dependent-Plasticity [9], to older and quite mature algorithms such as backpropagation [10] for Deep Neural Networks (DNNs) [11] (Fig. 1). While binary or trinary weights have been shown to be sufficient for

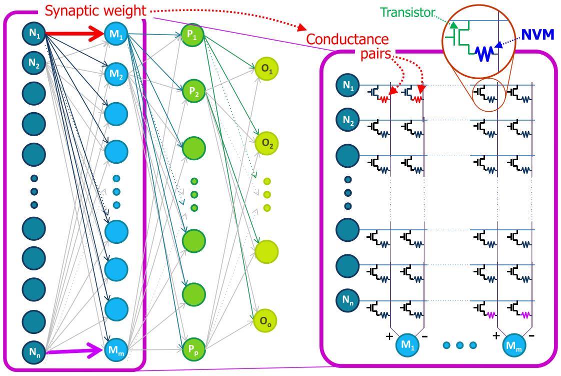


FIGURE 1. Non-Von Neumann computing [3]–[6] for implementing brain-inspired algorithms calls for multi-layer networks, in which each layer of neurons drives the next through dense networks of programmable synaptic weights. Dense crossbar arrays of nonvolatile memory (NVM) and transistor device-pairs, or potentially two-terminal selectors [7], can efficiently implement such neuromorphic networks [3].

forward-evaluation of DNNs [12], [13], weight update during training seems to require multiple bits of precision [14] or analog weights [4].

Numerous analog memory devices have been discussed for such neuromorphic applications, ranging from filamentary Resistive RAM (RRAM) [15]–[17], Phase Change Memory (PCM) [4], [18], [19], Conductive-Bridging RAM (CBRAM) [20], [21], and even Ferroelectric RAM [22]. Desirable characteristics include scalability, low power operation, high endurance, multi-level data storage, and a compact $4F^2$ cell size suitable for implementation in high-density crossbar arrays [2].

One of the weaknesses of filamentary RRAM and PCM is the asymmetry between the gradual changes of analog conductance in one direction (which depend, as is desired, on "device history"), and abrupt changes in the other direction (which are, unfortunately, nearly independent of device history). For instance, conductance increases of a PCM device by partial-SET pulses can be gradual, as successive pulses – even if identical – can crystallize more and more of an amorphous plug within the device [23]. In contrast, conductance decreases (the RESET step) are difficult to implement gradually, especially when one is constrained to a single pulse condition across a large array. For filamentary RRAM, it is the filament dissolution process (RESET) that can be gradual, while it is filament formation (SET) that is abrupt, requiring external current compliance to avoid overly–thick and conductive filaments [15]. Compliance is particularly critical when a filament is "formed" for the first time, often using significantly higher voltages than are needed subsequently. Forming is needed because such a Filamentary-RRAM device is typically a relatively thick insulator immediately after fabrication. After forming, subsequent RESET switching operations only remove a portion of the long and narrow filament that was introduced during the "forming" step.

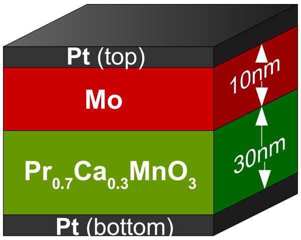


FIGURE 2. Typical stack of a Mo/PCMO device – a non-filamentary RRAM in which oxygen ions gradually drift within the device forming a thin layer of oxide on a reactive electrode.

In contrast, non-filamentary RRAM, such as devices based on $Pr_{0.7}Ca_{0.3}MnO_3$, show all the desired characteristics of analog synaptic devices including bidirectional gradual conductance change. Since there are no filaments, these non-filamentary RRAM devices do not require an initial "forming" step. PCMO-based synapse device with Mo electrodes exhibit bidirectional change but low conductance contrast [24]. Devices with Al/PCMO structure have been reported showing multi-level states of conductance and excellent uniformity in a high-density, 1 kbit cross-point array [25]. Feasibility for encoding neural network weights was shown based on fits to the median device characteristics [6]. However, Al/PCMO exhibits undesired stability issues when placed in a more conductive state, due to reactive oxidation at the metal/oxide interface [26].

In this two-part paper, we address these two issues: the need for PCMO-based devices with better stability in conductance states offering high contrast, and the need to model neural network behavior based on *measured* rather than fitted device characteristics, including their stochastic behavior. In this Part I, we introduce a thin interfacial layer of conducting MoOx as an oxygen buffer layer, and show that this helps control the redox reaction at the metal/PCMO interface by demonstrating improved retention and conductance contrast characteristics. We confirm bidirectional analog synapse characteristics and measure "jump-tables" to enable large scale neural network simulations that capture both median and stochastic device behavior. (Part II of the paper [1] describes these simulations in extensive detail.) Finally, we show switching energy measurements across a range of different size devices, and lay out a path for future device research towards smaller devices, shorter pulses and lower programming voltages.

II. $PR_{0.7}CA_{0.3}MNO_3$ MATERIAL IMPROVEMENTS

A. TYPICAL FEATURES OF $PR_{0.7}CA_{0.3}MNO_3$ MATERIALS

The mechanism of PCMO-based resistive switching is widely attributed to field-driven oxygen migration and redox reaction at a metal/PCMO interface [24]. Fig. 2 shows a simple device structure in which a thin Molybdenum (Mo) layer serves as the relevant metal, with a Platinum (Pt) cap to prevent interaction with the external environment. Under positive bias, oxygen ions in the PCMO material move toward

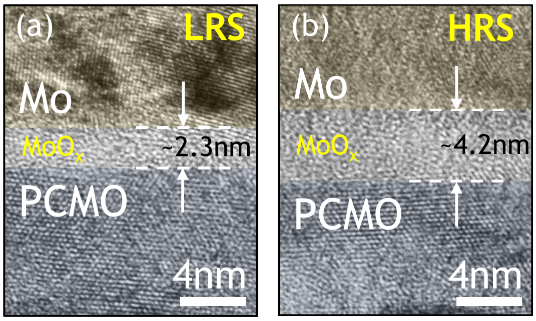


FIGURE 3. (a) Cross-sectional TEM image of a Mo/PCMO device in the low resistance state (LRS). (b) TEM image of a Mo/PCMO device in the high resistance state (HRS). The different thickness of the oxide layer is considered to be the main cause of the change in conductance [24].

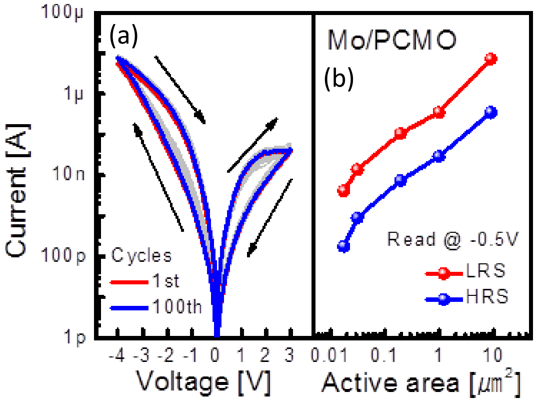


FIGURE 4. (a) Measured DC I-V characteristic for the first 100 switching cycles of a 150nm diameter Mo/PCMO RRAM device, and (b) dependence of read current in the SET and RESET states on the active device area.

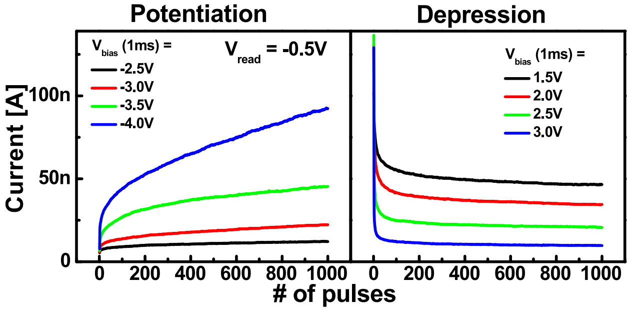


FIGURE 5. Potentiation and depression characteristics for a fresh 500nm diameter Mo/PCMO device using 1 ms pulses of various amplitudes and $V_{READ} = -0.5$ V. There was no delay introduced between the end of each 1 ms read pulse and the application of the subsequent 1 ms write pulse.

the metal electrode, which leads to the formation of an interfacial oxide layer. Conductance depression occurs due to the increased resistance of both interfacial oxide layer and the oxygen-deficient PCMO layer (Fig. 3). Conversely, by applying negative bias at the top electrode, oxygen ions in the interfacial oxide layer move back to the PCMO region, which in turn causes potentiation (conductance increase). Slow (DC) current-voltage sweeps illustrate at least two resistance states (Fig. 4(a)).

Device current, at a read voltage too low to induce switching, scales proportionally with decreasing device size (Fig. 4(b)), showing that switching occurs across the entire electrode area, not just within a local filament region. However, the double asymmetry — of the device structure and of the reduction-oxidation reaction — introduces asymmetry into the bidirectional switching characteristics. Furthermore, saturation sets in as the number of oxygen and metal ions decreases (increases) during oxidation (reduction), causing the rate of the conductance change to decrease as switching proceeds. This leads to an undesired nonlinearity in the conductance response to identical pulses (Fig. 5), the device characteristic most relevant to the training of neural networks [3]–[6]. In addition, as will be discussed shortly, the conductance response of as-fabricated devices differs from the response of extensively-exercised devices.

Note, however, that the asymmetry between the voltages needed for potentiation and depression (Fig. 5) are not a serious issue for the neuromorphic application, at least to first order. The peripheral neuron circuitry will already need to "know" whether a depression or potentiation operation is desired — in order to select the correct write polarity, for instance — so a difference in the magnitudes of these required voltages between these operations should only become a problem if the maximum voltages to be supplied by the peripheral circuitry were too large.

B. MOTIVATION FOR COMBINING AL AND MO

Recently, we demonstrated 1 kbit Al/PCMO-based synaptic arrays for use in neuromorphic systems [25]. The introduction of aluminum as the reactive metal helped improve conductance contrast when compared to older devices using molybdenum. While we confirmed the feasibility of analog memory characteristics, the SET state could unfortunately only be maintained for a short timespan. While poor retention of conductance states is not super-critical for a neuromorphic application, when the ratio between the conductance lifetime and time-for-each-training-example falls below ~10,000, accuracy falls off sharply [27]. Thus some retention improvement is needed for use in neuromorphic systems. Instability of these more conductive states was attributed to spontaneous oxidation at the $Al/Pr_{0.7}Ca_{0.3}MnO_3$ interface [25].

By introducing an ultra-thin (3-4 nm) interface layer of Molybdenum as a buffer between the reactive aluminum and PCMO, it should be possible to suppress this spontaneous oxidation. Compared to other metal oxides, MoO_x exhibits unique multivalent oxidation states with different conductances [28]. In this new structure (Fig. 6), stability of conductance states is improved by the presence of the Mo buffer layer, which increases the energy barrier to overcome when switching the state of the device. However, these Al/Mo/PCMO devices should continue to exhibit a large ON/OFF ratio, due to the large difference in resistivity between Al and its oxides.

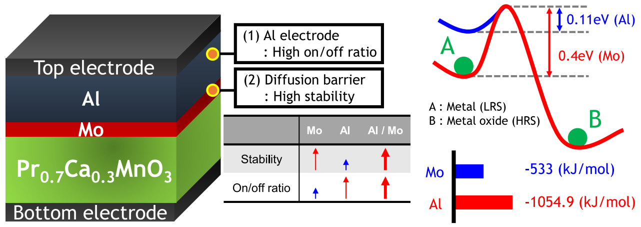


FIGURE 6. Al/Mo/PCMO-based devices are expected to show a good stability of conductance states and a high ON/OFF ratio. The former characteristic is ascribed to the presence of the Mo buffer layer, which increases the energy barrier to overcome when switching device state; the latter is due to both the high difference in resistivity between Al and its oxides.

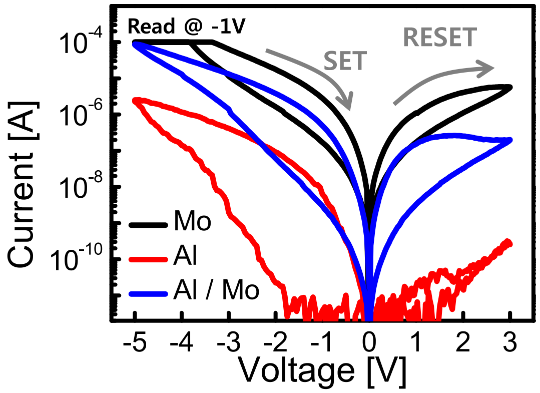


FIGURE 7. DC I-V characteristics of 500nm diameter PCMO-based synapse devices with various reactive metals such as Mo, Al, and Al/Mo electrodes.

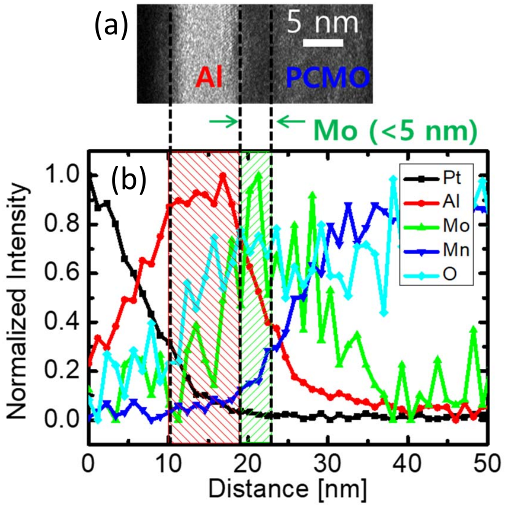


FIGURE 8. (a) TEM and (b) EDS line profile of Al/ultra-thin Mo/PCMO synapse device.

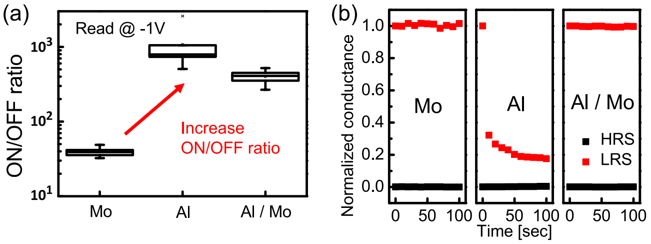


FIGURE 9. (a) ON/OFF ratio and (b) state retention for different PCMO-based RRAM variants: Mo/PCMO from [26], Al/PCMO from [25] and AlMo/PCMO from this work.

C. AL/MO/PCMO DEVICE FABRICATION AND MEASUREMENT

A polycrystalline PCMO layer was deposited by sputtering onto a Platinum bottom metal electrode. For isolation purposes, a 100nm SiN_x layer was then deposited by plasma-enhanced chemical vapor deposition, and dry-etched by reactive-ion etching to form via-hole structures with various device diameters (ranging from 150 nm to 3 um). Finally, thin Mo and Al electrodes (3-4 nm and 10 nm, respectively) were sequentially deposited into this via-hole by PVD sputtering. An 80-nm-thick Pt layer was deposited as a top electrode and patterned by conventional lithography.

Electrical characteristics of the PCMO-based resistive memory devices were measured using an Agilent B1500A. Fig. 7 shows typical DC I-V characteristics of PCMO-based synapse devices with various reactive metals such as Mo, Al, and Al/Mo layers. The active diameter of each device was 500 nm. Under positive bias on the top electrode (TE), oxygen ions in the PCMO layer move to the metal, forming an interfacial oxide layer between the metal and PCMO layers and increasing device resistance. Conversely, under negative bias on the TE, oxygen ions move back to the PCMO layer, returning the device to low resistance. To avoid excessive current for the Mo/PCMO sample, we imposed an external compliance current of 100uA.

The SET current of the Al/Mo/PCMO device was lower than that of the Mo/PCMO device, which can be explained by the partial oxidation of the Al layer. Transmission electron microscopy (TEM) analysis (Fig. 8, top) revealed a 5nm-thick amorphous MoO_x interfacial layer at the Mo/PCMO interface due to the oxidation reaction. Energy Dispersive x-ray Spectroscopy (EDS) analysis showed that both the Mo and Al layers participate in the switching characteristics (Fig. 8, bottom), explaining the larger on/off ratio for the Al/Mo/PCMO device as compared to this Mo/PCMO device (Fig. 9(a)).

In addition to the higher ratio between ON and OFF resistance because of the high difference in resistivity between Al and AlO_x, there are noticeable improvements in conductance stability. While Mo electrode-based devices show good retention times for both the high and low resistance states, Al/PCMO exhibits poor stability in the LRS (Fig. 9(b)) because of the low oxidation free energy of Al. In a neuromorphic application, where the algorithm may be increasing the device conductance over time with infrequently applied programming pulses (Fig. 10(a)), the use of Al/PCMO devices means that much of this conductance change is being lost in the intervals between programming events (Fig. 10(b)).

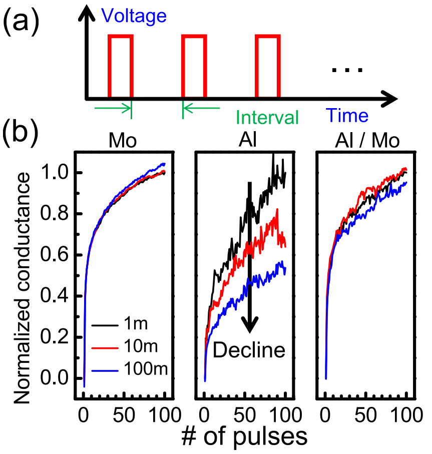


FIGURE 10. **(a) Proposed scheme for evaluating time-dependence of conductance response. (b) For the Mo/PCMO and Al/Mo variants, the changes in conductance state induced by constant-voltage pulses are mostly unaffected by longer time-delay *between* these pulses. In contrast, the Al/PCMO variant shows significant conductance loss as the time-delay between pulses is increased.**

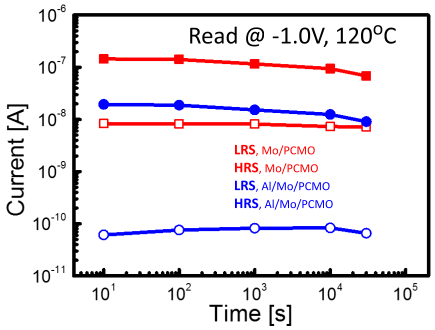


FIGURE 11. **Measured retention characteristics of 150nm diameter Mo/PCMO and Al/Mo/PCMO devices at 120°C.**

The combination of an Al electrode and Mo as diffusion barrier helps to maintain good retention (Fig. 9(b), Fig. 10(b)), even at high temperature (Fig. 11), without compromising too much of the on/off conductance contrast (Fig. 9(a)).

D. CONDUCTANCE RESPONSE OF AL/MO/PCMO

With the improved Al/Mo/PCMO device, we first evaluated synapse characteristics under DC programming conditions to confirm the potential for synapse applications. We confirmed that the Al/Mo/PCMO device can emulate synaptic behavior under both incremental and constant-amplitude voltage pulses as shown in Fig. 12. The conductance state was gradually increased/decreased under negative/positive bias, respectively, with both larger voltages (Fig. 13) and longer durations (Fig. 14) leading to a larger effect. Fig. 15 shows that the response of extensively-exercised devices is qualitatively quite similar to those of as-fabricated devices (Fig. 14), but with a reduced dynamic range.

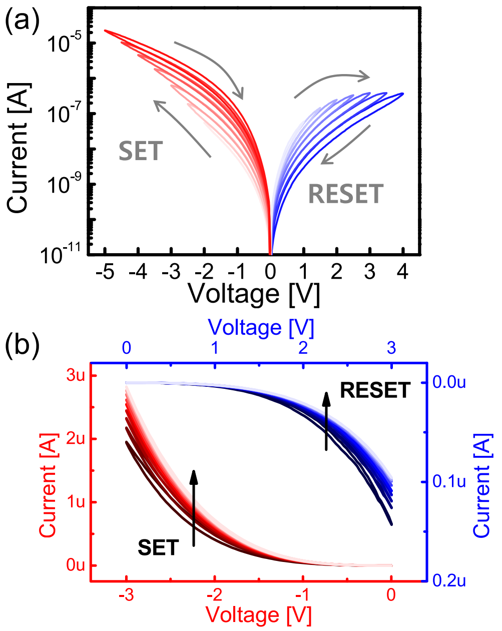


FIGURE 12. **(a) DC I-V characteristics of Al/Mo/PCMO device under consecutive (a) incremental and (b) constant-amplitude voltage pulse sweeps.**

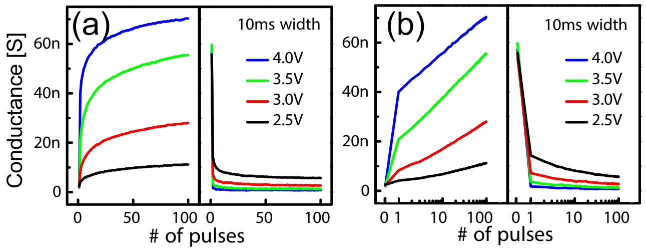


FIGURE 13. **(a) Potentiation and depression characteristics for fresh 500 nm Al/Mo/PCMO devices with equal pulses of various voltage amplitudes (100μs duration), using $V_{READ} = -1.0$ V. (b) Same data plotted on log scale. Note that, for completeness, the conductance before any pulses ("0") is indicated at an arbitrary position just to the left of the conductance after the first pulse ("1") – thus the slope of this first line segment cannot be assigned any physical meaning.**

III. JUMP-TABLE MEASUREMENTS FOR NEUROMORPHIC SIMULATION

We previously proposed a measure-then-fire programming scheme, and showed that this scheme led to improved neuromorphic training accuracy [6]. The intrinsic non-linearity of the resistive switching was compensated by proper choice of programming voltage, leading to a classification accuracy as high as ~90.55%. However, in a real VLSI system, it will not be practical to sequentially measure every device before performing the programming associated with the training of each individual neural network example.

Thus it is critical to assess whether the native devices, when programmed blindly with only two pulse conditions (one for SET and one for RESET), will prove to have a suitable response to attain the desired neuromorphic

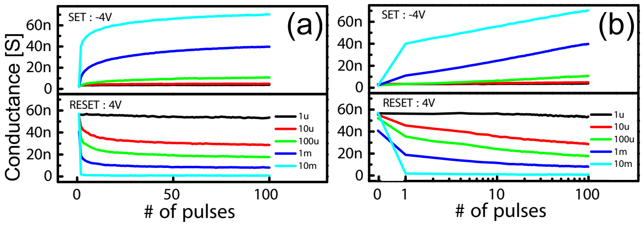


FIGURE 14. (a) Potentiation and Depression characteristics for fresh 500 nm Al/Mo/PCMO devices with equal pulses of various durations and constant amplitude, using $V_{READ} = -1.0$ V. (b) Same data plotted on log scale.

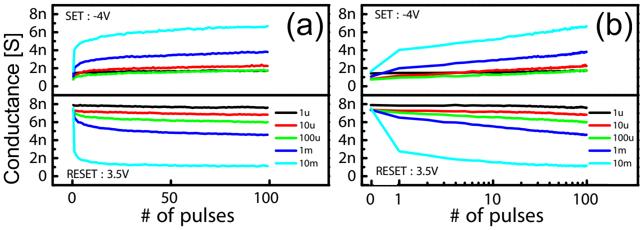


FIGURE 15. (a) Potentiation and depression characteristics for previously-exercised 500 nm Al/Mo/PCMO devices with equal pulses of various durations and constant amplitude, using $V_{READ} = -1.0$ V. (b) Same data plotted on log scale.

functionality or not. Previously, we used numerical formulae to fit curves like those in Figs. 13 and 14 in order to model device performance. Unfortunately, such formulae are often inadequate to accurately describe the device response, nor are they able to capture the stochastic nature of real device response.

A. JUMP-TABLE CONCEPT

We have previously introduced the concept of jump-tables [3], [29], which tabulate the cumulative probability of achieving a particular conductance change as a function of conductance. In contrast to conductance response curves that plot only median conductance as a function of the number of pulses (Fig. 16(a)), a jump-table can encode a complicated stochastic distribution (Fig. 16(b)). We can construct the required jump-tables — one for potentiating SET pulses, and another for depressing RESET pulses – from measurements performed with constant-amplitude and duration pulses. Such tables are capable of describing the relevant switching behaviour (including variability) of any uni- or bi-directional analog NVM device for this neuromorphic application, by allowing us to predict the statistical distribution of the "next" conductance state upon firing the particular voltage pulse (duration and amplitude) associated with the jump-table. Since DNN training with analog memories calls for only two types of pulses (one potentiation, one depression) [3], we need only two jump-tables. By definition, cumulative distribution functions are always monotonic, running from 0% to 100%. Thus comparison against a random uniform deviate makes it very straightforward to randomly sample the empirical distribution of conductance change for the purposes of simulation.

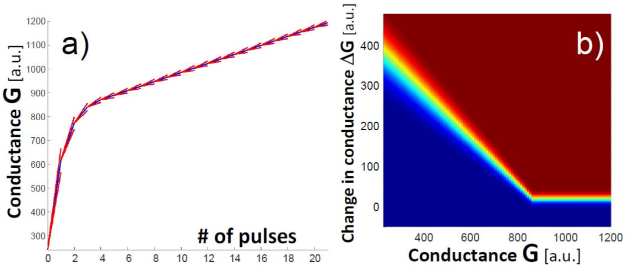


FIGURE 16. (a) Artificial conductance response (for illustration of the jump-table concept, not representing a particular PCMO device) as a function of pulse (median response in blue, ±1σ response in red), and (b) corresponding "jump-table" plotting likelihood of conductance-change (cumulative probability in color from dark blue (0%) to dark red (100%)).

B. JUMP-TABLE MEASUREMENTS

In order to reproduce closely the rate and variability of PCMO-based weight elements, jump-tables were constructed from measured device characteristics. In Fig. 17, data from 50,000 SET (-4.0 V, 10 ms) and RESET pulses (3.5 V, 10 ms) applied to three 200nm-sized devices is plotted in the form of jump-tables, for both Mo/PCMO and Al/Mo/PCMO. Conductances were measured at -1.0 V read voltage and are normalized to a conductance range of 0.5nS to 5nS (see Fig. 18). The magnitude of the largest jumps plotted here is significantly smaller than those shown in Fig. 14 for as-fabricated devices, but match the dynamic range for similarly extensively-exercised devices (Fig. 15).

An interesting realization is made by examining the raw data from which these jump-tables were synthesized (Fig. 18): there are clearly intermediate conductance values which the devices can reach either immediately after a polarity reversal (towards the left side of either plot in Fig. 18) or after many successive same-polarity pulses (right sides of the plots in Fig. 18). The next pulse induces a conductance change which appears to depend more on the number-of-pulses-since-reversal than on the particular absolute conductance value. This suggests that while absolute conductance is somewhat indicative of the device-state, conductance is insufficient to completely describe the device-state. Yet devices which experience a polarity-reversal at roughly the same conductance (e.g., the initial conductance of each trace in Fig. 18) exhibit nearly identical subsequent conductance evolutions as subsequent same-polarity pulses are applied. This implies that it should be possible, with additional measurements, to identify a complete state representation.

This could potentially be accomplished by specifying the conductance at which polarity was reversed and the number of successive pulses since that reversal. It might be possible, in future work on these materials, to connect this to the internal configuration of both the thickness of the oxidized layer and the asymmetry of the ion distribution within that

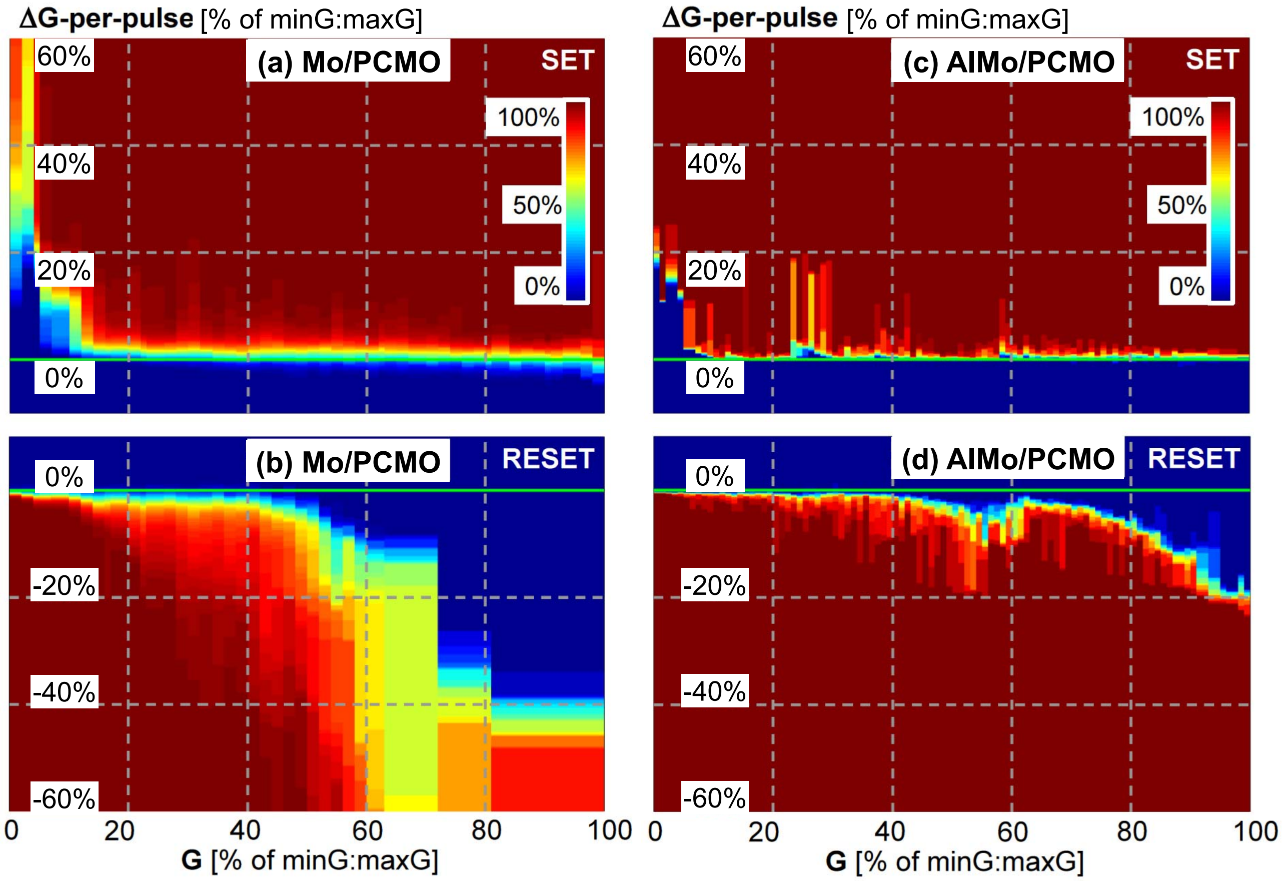


FIGURE 17. (a), (b) SET and RESET jump-tables for Al/Mo/PCMO devices. (c), (d) SET and RESET jump-tables for Al/Mo/PCMO devices. The cumulative distribution colormap of conductance-change-per-pulse (ΔG-per-pulse) as a function of the conductance (G) is plotted. 50,000 (a), (c) SET pulses (-4.0 V, 10 ms) and (b), (d) RESET pulses (3.5V, 10 ms) on three Al/Mo/PCMO, Mo/PCMO-based resistive switching memories with 200 nm vias were measured. Note that this normalized conductance range from 0% to 100% corresponds to absolute conductances ranging from minG∼0.5nS to maxG∼5.0nS (see Fig. 18).

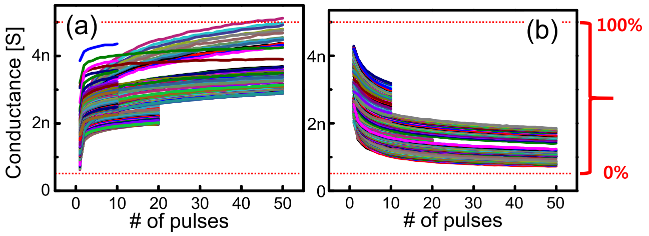


FIGURE 18. Raw conductance response data used to synthesize Fig. 17(c,d). Different measurements are shown in various colors in order to help distinguish individual measurement curves. Note that clearly there are intermediate conductance values which the devices can reach either immediately after a polarity reversal (towards the left side of either of these plots) or after many successive same-polarity pulses (right side of either plot), after which the conductance change will depend more on the number-of-pulses-since-reversal than on the particular absolute conductance value. This suggests that while absolute conductance is somewhat indicative of the device-state, it does not completely describe the device-state.

layer, given an accurate model for turning such an ionic distribution back into a device conductance. In the context of the present work, device conductance is the only available metric we have for specifying the device-state. We thus choose to treat the variability of subsequent conductance change at any given conductance as a random variable that we are unable to predict. As introduced into neural network simulations by the jump-tables, we can investigate how these conductance changes affect neural network performance [1].

Due to the intrinsic non-linearity of these devices, the extremes of both tables (low conductances in the SET tables, and high conductances in the RESET tables) show very large jumps for both Mo/PCMO and Al/Mo/PCMO. This is the equivalent of the rapid first step shown in Figs. 13 and 14. In addition to these large initial jumps, even in portions of the common conductance range where the changes induced by either SET or RESET pulses are both modest in size, the average RESET jump size is larger than the average SET jump size. These two characteristics can potentially compromise machine learning performance, and will be addressed in Part II of the paper [1].

In the remainder of this Part I, we turn to measurements of switching energy.

IV. SWITCHING ENERGY AND SCALING OF AL/MO/PCMO

To measure switching energy of Al/Mo/PCMO-based resistive memory devices, current response was recorded while

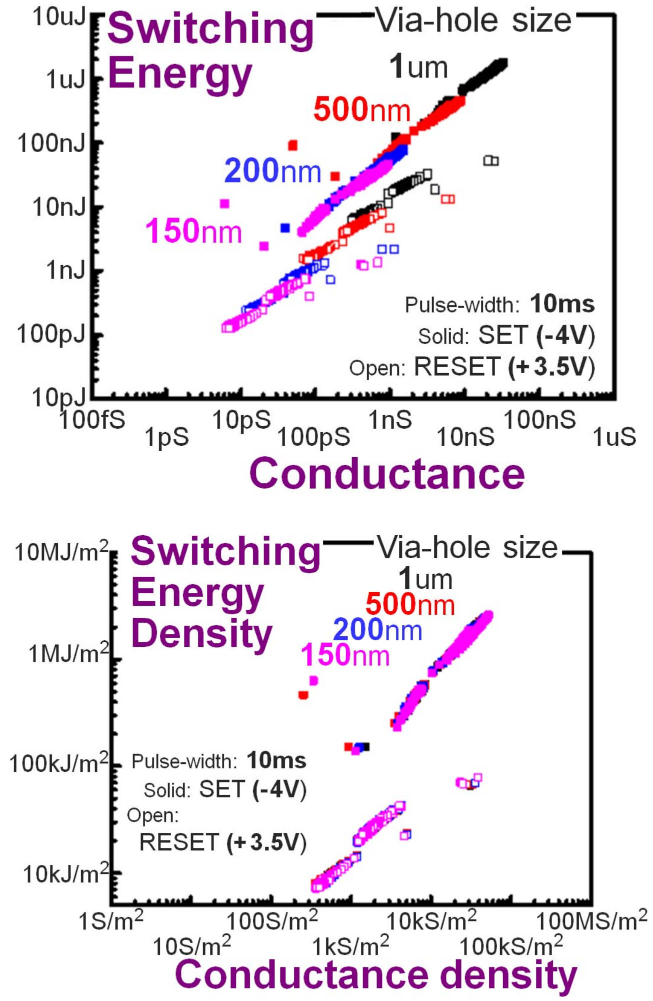


FIGURE 19. Switching (a) energy as a function of conductance and (b) energy density as a function of conductance density, measured for Al/Mo/PCMO,Mo/PCMO-based devices with -1V reading voltage.

switching pulses were applied. Switching-energy-per-pulse is calculated by integrating the applied voltage and the measured current, and plotted as a function of the current conductance before the pulse, as measured with -1.0 V read voltage.

A. AREA DEPENDENCE

In order to investigate the area dependency, SET pulses (−4.0 V, 10 ms) and RESET pulses (3.5 V, 10 ms) were applied on devices with various via-hole sizes (150 nm, 200 nm, 500 nm, and 1 um). For each cell, 50-cycle DC sweeps were performed for initialization, and 5-cycles of alternate 100 SET pulses and 100 RESET pulses were repeatedly applied. Care was taken to eliminate parasitic capacitances that otherwise masked low energy switching events. Device conductance was measured with read pulses of $-1V$.

Switching energy was found to increase with device conductance state, with SET energy larger than RESET energy because of the asymmetry between reduction and oxidation (Fig. 19(a)). Like other non-filamentary switching elements, PCMO-based memory devices show a dependence of programming energy on the active area. Measured switching energy, when normalized with respect to the active device area (Fig. 19(b)), shows the expected linear dependence between switching current and device hole-size.

It would be preferable to achieve femto-Joule level switching energy for efficient analog implementation of neuromorphic systems using resistive switching memory as synaptic devices [30]. Following the trend from 150nm down to 25nm, one can anticipate an improvement in switching energy by roughly 35× by such scaling. If the switching time could potentially be reduced from 10ms down to 10ns, then one would be able to achieve femto-Joule switching energy. Such aggressive scaling of both device area and switching time would be necessary in order to enable highly-parallelized weight update operations.

This of course requires that the full range of conductance would still be available, which might require higher voltage, thus risking permanent breakdown and damage of the devices. Potentially some further materials modification might provide fast, low-energy switching by trading back some of the retention improvements gained by combining Mo and Al layers.

V. CONCLUSION

We have shown improved non-filamentary RRAM devices based on $Pr_{0.7}Ca_{0.3}MnO_3$ by introducing a Mo buffer layer together with a reactive Al electrode to control the internal redox reaction and obtain stable analog conductance states offering high conductance contrast. Bidirectional analog synapse characteristics were observed and devices were measured to produce empirical "jump-tables" — likelihood of a given change of conductance at any given conductance state, for both potentiating SET and depressing RESET pulses. Such jump-tables are a critical component of accurate neural network simulations that can predict the effects of complex and stochastic device behavior (see companion paper [1]).

Future experimental work will be needed to identify a more suitable metric for device-state beyond absolute conductance. A clear dependence on both the device conductance at the point of polarity-reversal and the number of subsequent same-polarity pulses was observed. As part of such future work, it will be important to demonstrate that the loss of dynamic range illustrated here between fresh, as-fabricated devices (Fig. 14) and extensively-exercised devices (Fig. 15) is in fact an initial wear-in process and not indicative of a continual loss of dynamic range, representing an effective endurance-failure mechanism.

Finally, switching energy measurements were shown, illustrating the necessary path for future device research towards smaller devices and shorter pulses. However, in order to continue to obtain significant conductance changes with ultra-short pulses, it is likely that programming voltages will need to increase rather than decrease, slightly increasing power/energy but also increasing the risk of device breakdown and permanent damage.

REFERENCES

[1] A. Fumarola *et al.*, "Bidirectional non-filamentary RRAM as an analog neuromorphic synapse, part II: Impact of Al/Mo/Pr0.7Ca0.3MnO3 device characteristics on neural network training accuracy," *IEEE Trans. Electron Devices*, to be published.

[2] G. W. Burr *et al.*, "Neuromorphic computing using non-volatile memory," *Adv. Phys. X*, vol. 2, no. 1, pp. 89–124, 2017.

[3] 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 *IEDM Tech. Dig.*, San Francisco, CA, USA, 2014, pp. 29.5.1–29.5.4.

[4] 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," *IEEE Trans. Electron Devices*, vol. 62, no. 11, pp. 3498–3507, Nov. 2015.

[5] G. W. Burr *et al.*, "Large-scale neural networks implemented with nonvolatile memory as the synaptic weight element: Comparative performance analysis (accuracy, speed, and power)," in *IEDM Tech. Dig.*, Washington, DC, USA, 2015, pp. 4.4.1–4.4.4.

[6] J.-W. Jang, S. Park, G. W. Burr, H. Hwang, and Y.-H. Jeong, "Optimization of conductance change in $Pr_{1-x}Ca_xMnO_3$-based synaptic devices for neuromorphic systems," *IEEE Electron Device Lett.*, vol. 36, no. 5, pp. 457–459, May 2015.

[7] G. W. Burr *et al.*, "Access devices for 3D crosspoint memory," *J. Vac. Sci. Technol. B*, vol. 32, no. 4, 2014, Art. no. 040802.

[8] R. A. Nawrocki, R. M. Voyles, and S. E. Shaheen, "A mini review of neuromorphic architectures and implementations," *IEEE Trans. Electron Devices*, vol. 63, no. 10, pp. 3819–3829, Oct. 2016.

[9] S. Fusi, M. Annunziato, D. Badoni, A. Salamon, and D. J. Amit, "Spike-driven synaptic plasticity: Theory, simulation, VLSI implementation," *Neural Comput.*, vol. 12, no. 10, pp. 2227–2258, 2000.

[10] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, Oct. 1986.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[12] M. Courbariaux, Y. Bengio, and J.-P. David, "Binaryconnect: Training deep neural networks with binary weights during propagations," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2015, pp. 3123–3131.

[13] S. K. Esser *et al.*, "Convolutional networks for fast, energy-efficient neuromorphic computing," *Proc. Nat. Acad. Sci.*, vol. 113, no. 41, pp. 11441–11446, 2016.

[14] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, "Deep learning with limited numerical precision," in *Proc. Int. Conf. Mach. Learn.*, vol. 392. Lille, France, 2015, pp. 1737–1746.

[15] S. Yu, Y. Wu, R. Jeyasingh, D. G. Kuzum, and H. S. P. Wong, "An electronic synapse device based on metal oxide resistive switching memory for neuromorphic computation," *IEEE Trans. Electron Devices*, vol. 58, no. 8, pp. 2729–2737, Aug. 2011.

[16] M. Prezioso *et al.*, "Modeling and implementation of firing-rate neuromorphic-network classifiers with bilayer $Pt/Al_2O_3/TiO_{2-x}/Pt$ memristors," in *IEDM Tech. Dig.*, Washington, DC, USA, 2015, pp. 17.4.1–17.4.4.

[17] S. Yu *et al.*, "Scaling-up resistive synaptic arrays for neuro-inspired architecture: Challenges and prospect," in *IEDM Tech. Dig.*, Washington, DC, USA, 2015, pp. 17.3.1–17.3.4.

[18] B. L. Jackson *et al.*, "Nanoscale electronic synapses using phase change devices," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 9, no. 2, 2013, Art. no. 12.

[19] S. Kim *et al.*, "NVM neuromorphic core with 64k-cell (256-by-256) phase change memory synaptic array with on-chip neuron circuits for continuous in-situ learning," in *IEDM Tech. Dig.*, Washington, DC, USA, 2015, pp. 17.1.1–17.1.4.

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

[21] Y. J. Jeon, S. Kim, and W. D. Lu, "Utilizing multiple state variables to improve the dynamic range of analog switching in a memristor," *Appl. Phys. Lett.*, vol. 107, no. 17, 2015, Art. no. 173105.

[22] Y. Kaneko, Y. Nishitani, and M. Ueda, "Ferroelectric artificial synapses for recognition of a multishaded image," *IEEE Trans. Electron Devices*, vol. 61, no. 8, pp. 2827–2833, Aug. 2014.

[23] G. W. Burr *et al.*, "Phase change memory technology," *J. Vac. Sci. Technol. B*, vol. 28, no. 2, pp. 223–262, 2010.

[24] D.-J. Seong *et al.*, "Effect of oxygen migration and interface engineering on resistance switching behavior of reactive metal/polycrystalline $Pr_{0.7}Ca_{0.3}MnO_3$ device for nonvolatile memory applications," in *IEDM Tech. Dig.*, Baltimore, MD, USA, 2009, pp. 1–4.

[25] S. Park *et al.*, "RRAM-based synapse for neuromorphic system with pattern recognition function," in *IEDM Tech. Dig.*, vol. 10. San Francisco, CA, USA, 2012, pp. 10.2.1–10.2.4.

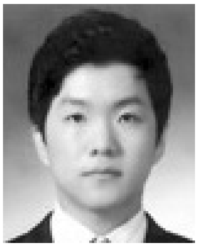
[26] K. Moon *et al.*, "High density neuromorphic system with $Mo/Pr_{0.7}Ca_{0.3}MnO_3$ synapse and NbO_2 IMT oscillator neuron," in *IEDM Tech. Dig.*, Washington, DC, USA, 2015, pp. 17.6.1–17.6.4.

[27] A. Fumarola *et al.*, "Accelerating machine learning with non-volatile memory: Exploring device and circuit tradeoffs," in *Proc. IEEE Int. Conf. Rebooting Comput. (ICRC)*, San Diego, CA, USA, 2016, pp. 1–8.

[28] Z. Li *et al.*, "Investigation on molybdenum and its conductive oxides as p-type metal gate candidates," *J. Electrochem. Soc.*, vol. 155, no. 7, pp. H481–H484, 2008.

[29] S. Sidler *et al.*, "Large-scale neural networks implemented with non-volatile memory as the synaptic weight element: Impact of conductance response," in *ESSDERC Techn. Dig.*, Lausanne, Switzerland, 2016, pp. 440–443.

[30] 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.

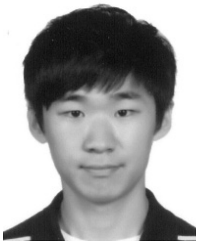


KIBONG MOON received the B.S. degree from the School of Electronics Engineering, Kyungpook National University, Daegu, South Korea, in 2013. He is currently pursuing the Ph.D. degree with the Department of Materials Science and Engineering, Pohang University of Science and Technology, Pohang, South Korea.

ALESSANDRO FUMAROLA received the M.S. degree in nanotechnology jointly from the Polytechnic of Turin, Italy, INP Grenoble, France, and EPFL Lausanne, Switzerland. He is currently pursuing the Ph.D. degree with the Max Planck Institute for Microstructure Physics, Halle, Germany. He was with IBM Research – Almaden, San Jose, CA, USA, for six months. His research interests include non-Von Neumann computing, resistive switching, and magnetic materials.

SEVERIN SIDLER received the B.S. and M.S. degrees in electrical engineering from EPFL, Lausanne, Switzerland. He has been a six-month intern with IBM Research – Almaden, San Jose, CA, USA, and performed his master's degree research with the IBM Zurich Research Laboratory, Rüschlikon, Switzerland.

His current research interests include cognitive computing and systems engineering.



JUNWOO JANG received the B.S. degree in electrical engineering from the Pohang University of Science and Technology, South Korea, in 2012 and the Ph.D. degree in 2016. He moved to the Department of Creative IT Engineering for his graduate studies, spent four months visiting IBM Research–Almaden, San Jose, CA, USA, in 2013. He joined Samsung Electronics in 2016.

PRITISH NARAYANAN received the Ph.D. degree in ECE from the University of Massachusetts Amherst in 2013, and joined IBM Research – Almaden, San Jose, CA, USA, as a Research Staff Member. His research interests include emerging technologies for logic, nonvolatile memory, and cognitive computing.

He was a recipient of the Best Paper Awards at ISVLSI 2008, IEEE DFT 2010, 2011, and NanoArch 2013, and has reviewed for IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS, the IEEE TRANSACTIONS ON NANOTECHNOLOGY, *ACM Journal on Emerging Technologies in Computing Systems*, and several IEEE conferences.

ROBERT M. SHELBY received the Ph.D. degree in chemistry from the University of California at Berkeley, Berkeley, CA, USA.

He joined IBM, Armonk, NY, USA, in 1978. He is currently a Research Staff Member with IBM Research–Almaden, San Jose, CA, USA.

Dr. Shelby is a fellow of the Optical Society of America.

GEOFFREY W. BURR (S'87–M'96–SM'13) received the Ph.D. degree in electrical engineering from the California Institute of Technology, Pasadena, CA, USA, in 1996.

He joined IBM Research–Almaden, San Jose, CA, USA, where he is currently a Principal Research Staff Member in 1996. His current research interests include nonvolatile memory and cognitive computing.

HYUNSANG HWANG received the Ph.D. degree in materials science from the University of Texas at Austin, Austin, Texas, USA, in 1992. He was with LG Semiconductor Corporation for five years. He became a Professor of materials science and engineering with the Gwangju Institute of Science and Technology, Gwangju, South Korea, in 1997. In 2012, he moved to the Materials Science and Engineering Department with Pohang University of Science and Technology, South Korea.