Received 30 August 2017; revised 21 November 2017; accepted 7 December 2017. Date of publication 11 December 2017; date of current version 15 January 2018. The review of this paper was arranged by Editor S. Moshkalev.

Digital Object Identifier 10.1109/JEDS.2017.2782184

Bidirectional Non-Filamentary RRAM as an Analog Neuromorphic Synapse, Part II: Impact of Al/Mo/$Pr_{0.7}Ca_{0.3}MnO_3$ Device Characteristics on Neural Network Training Accuracy

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

1 IBM Research–Almaden, San Jose, CA 95120, USA
2 Max Planck Institute for Microstructure Physics, 06120 Halle, Germany
3 École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
4 Department of Material Science and Engineering, Pohang University of Science and Technology, Pohang 790-784, South Korea
5 Samsung, Suwon 443-803, South Korea

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

ABSTRACT Neuromorphic computing embraces the “device history” offered by many analog non-volatile memory (NVM) devices to implement the small weight changes computed by a gradient-descent learning algorithm such as backpropagation. Deterministic and stochastic imperfections in the conductance response of real NVM devices can be encapsulated for modeling within a pair of “jump-tables.” Such jump-tables describe the full cumulative distribution function of conductance-change at each device conductance value, for both weight potentiation (SET) and depression (RESET). First, using several types of artificially constructed jump-tables, we revisit the relative importance of deviations from an ideal NVM with perfectly linear conductance response. Then, using jump-tables measured on improved non-filamentary resistive RAM devices based on $Pr_{0.7}Ca_{0.3}MnO_3$[see companion paper], we simulate the effects of their nonlinear conductance response on the training of a three-layer fully connected neural network. We find that, despite the relatively large conductance changes exhibited by any $Pr_{0.7}Ca_{0.3}MnO_3$device when either potentiating from its lowest conductance state or depressing from its highest conductance states, neural network training accuracies of >90% can be achieved. Highest accuracies are achieved by programming both conductances on each timestep (“fully bidirectional”), with the improved conductance on/off ratio of Al/Mo/PCMO resulting in marked improvements in training and test accuracy. Further accuracy improvements can be obtained by tuning the relative learning rate for potentiation (SET) by a factor of 1.66× with respect to depression (RESET), to offset the slight asymmetry between the average size of the associated SET and RESET conductance changes. Finally, we show that the bidirectional programming of Al/Mo/PCMO can be used to implement high-density neuromorphic systems with a single conductance per synapse, at only a slight degradation to accuracy.

INDEX TERMS Multi-layer neural network, neural network hardware, nonvolatile memory.

I. INTRODUCTION

Neuromorphic systems offer strong potential for fault-tolerant, massively parallel, energy-efficient computation [8]. Tasks involving image classification, speech recognition, machine translation and other pattern recognition tasks can potentially be more efficiently implemented

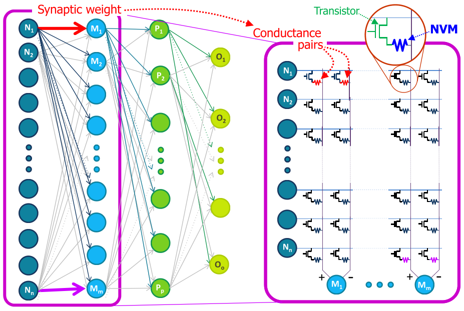


FIGURE 1. **Non-Von Neumann computing [1], [4]–[6] for implementing brain-inspired algorithms calls for multi-layer networks, in which each layer of neurons drives the next through dense networks of programmable synaptic weights. Dense crossbar arrays of nonvolatile memory (NVM) and transistor device-pairs, or potentially two-terminal selectors [7], can efficiently implement such neuromorphic networks [1].**

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 [8], [9] such as Spike-Timing-Dependent-Plasticity [10], to older and quite mature algorithms such as backpropagation [11] for Deep Neural Networks (DNNs) [12] (Fig. 1). While binary or trinary weights have been shown to be sufficient for forward-evaluation of DNNs [13], [14], weight update during training seems to require multiple bits of precision [15] or analog weights [4].

Numerous analog memory devices have been discussed for such neuromorphic applications, ranging from filamentary Resistive RAM (RRAM) [16]–[18], Phase Change Memory (PCM) [4], [19], [20], Conductive-Bridging RAM (CBRAM) [21], [22], and even Ferroelectric RAM [23]. 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 [8].

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 [24]. 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 [16].

In contrast, non-filamentary RRAM, such as devices based on $Pr_{0.7}Ca_{0.3}MnO_3$, show many of the desired characteristics of analog synaptic devices including bidirectional gradual conductance change. PCMO-based synapse devices with Mo electrodes exhibit bidirectional change but low conductance contrast [25]. 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 [26]. 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 excessive reactive oxidation at the metal/oxide interface after switching [27].

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 Part I, we described improved Al/Mo/PCMO devices offering improved retention and conductance contrast characteristics, and measured both switching energy as well as "jump-tables" describing both the median and stochastic device behavior to successive application of the same set of programming pulses (one pulse condition for potentiation (SET), and a second pulse condition for depression (RESET)).

In this Part II, we use these measured jump-tables to model the performance of PCMO-based devices on a typical multi-layer perceptron trained on MNIST. We revisit earlier work in which we artificially constructed jump-tables comprised of simple linear and non-linear functions of conductance change, in order to straightforwardly illustrate their impact on neural network classification performance. We study the difference between older Mo/PCMO devices [25] and these newer Al/Mo/PCMO devices [3], using two different weight-update schemes, "alternate bidirectional" and "fully bidirectional." These two schemes differ only in which conductance(s) get(s) programmed to achieve the weight change called for by backpropagation through the neural network. By use of an additional artificial jump-table, we prove that the accuracy issues of $Pr_{0.7}Ca_{0.3}MnO_3$devices cannot be simply assigned to the very large conductance changes exhibited when either potentiating from the lowest conductance states or depressing from the highest conductance states. We show that, by offsetting the slight asymmetry between the average size of the SET and RESET conductance changes, tuning of the relative learning rate for potentiation (SET) with respect to depression (RESET) can provide further accuracy improvements. And finally, we show that the bidirectional programming of Al/Mo/PCMO can be used to implement high-density neuromorphic systems with only a single conductance per synapse, at only a slight degradation to classification accuracy.

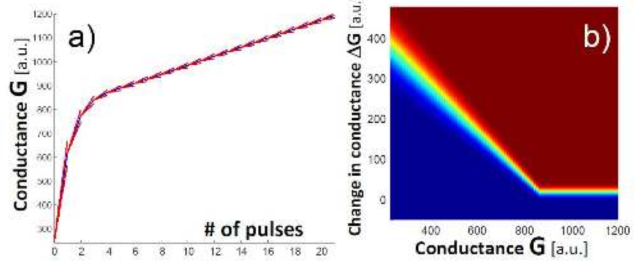


FIGURE 2. (a) Artificial conductance response (for illustration of the jump-table concept, not representing any particular PCMO device) as a function of pulse (median response in blue, $\pm 1\sigma$ response in red), and (b) corresponding "jump-table" plotting likelihood of conductance-change (cumulative probability in color from dark blue (0%) to dark red (100%)).

II. THE "JUMP-TABLE" CONCEPT

In [6], TiN/PCMO RRAMs were used as synaptic devices in a simulated neuromorphic system with a measure-then-fire programming scheme. The intrinsic non-linearity of the resistive switching was nearly completely compensated by choosing programming voltage carefully for each initial conductance value. It was observed that, when parameters were tuned so that the effective conductance response was linear, very high classification accuracy (up to $\approx$ 90.55%) could be reached for a three-layer perceptron architecture when trained on MNIST [6]. However, for real VLSI systems where millions of artificial synapses will need to be implemented [28], it will be highly impractical to sequentially measure every device before programming. Thus it is critical to determine if the native response of these NVM devices will be sufficient to deliver similarly high accuracy when programmed in a true "open loop" fashion.

Even when each programming pulse uses the same predefined amplitude, shape and duration, the conductance response of an array of NVM devices can exhibit significant non-idealities. Conductance response — e.g., the size of the conductance change induced by each pulse — can be nonlinear as a function of conductance, with identical SET pulses causing large jumps at low conductance, but much smaller jumps at high conductance. There can be asymmetry between positive (SET) and negative (RESET) conductance changes, even after the amplitude, shape and duration of the single pulse used for SET, and of the single pulse for RESET, are independently optimized for best match. And finally, in many real NVM devices, each conductance jump is inherently stochastic in nature.

All these behaviors can be compactly captured in the form of a pair of jump-tables, one for SET and one for RESET. Fig. 2(a) plots median conductance change just for SET (potentiation, blue curve), together with the $\pm 1\sigma$ stochastic variation about this median change (red lines). Fig. 2(b) shows the jump-table that fully captures this conductance response, plotting the cumulative probability (in color, from 0 to 100%) of any conductance change ΔG-per-pulse at any given initial conductance G. This number represents, for devices starting at the conductance G, the likelihood that a single programming pulse induces a conductance change *smaller* than the given ΔG.

Jump-tables can be empirically constructed by using constant-amplitude and duration pulses to explore the full range of conductances exhibited by the device [3]. Each measured programming pulse results in an increment of one and only one integer "bin" within an initially all-zero matrix spanning the expected range of (initial-conductance, conductance-change) space. After accumulating a large amount of data, this data is simply normalized within each initial-conductance-bin column (creating a probability density function (PDF) of conductance-change at a given initial conductance), and then integrated along the conductance-change axis to turn that PDF into a cumulative distribution function (CDF). By representing this data in terms of a monotonic CDF, a computer simulation need only sample a uniform random deviate for random number r, and then search the appropriate column of the jump-table (corresponding to the initial conductance before the pulse) to find the first ΔG entry which exceeds r.

III. BIDIRECTIONAL NON-FILAMENTARY RRAM: WHAT MAKES FOR A GOOD JUMP-TABLE

In addition to using a measured jump-table to simulate the SET response of PCM devices [1], we have previously studied various artificially-constructed jump-tables [2]. Because of their relevance in understanding the general role of different features found within jump-tables, we include these results here. These studies help provide an intuitive understanding of the impact that various features of such jump-tables have on the classification performance in the ANN application.

Except for the specific jump-tables, the ANN simulations performed here are identical to those in Ref [1]. A large-scale 3-layer perceptron (916 neurons and 164,885 synapses [1], [2], [4]–[6], [29]) was trained for 20 epochs on the first 5000 examples of the MNIST handwritten digit database, with images cropped to 22×24 pixels. The full "test" dataset of 10,000 different images was used for testing generalization after training was complete. Unless otherwise noted, two resistive switching devices are used to realize one synapse, with positive and negative weights encoded as the difference between the value of the two paired conductances.

Fig. 3 shows the network architecture used in the simulations. Every synapse is composed of two NVM devices, referred to as G^{+} and G^{-}, each with its own selection transistor. In our crossbar-compatible weight update scheme [1], both the upstream and downstream neurons fire between zero and four pulses towards their shared synapse, based on their internal knowledge of the most recent x (δ) value [28]. For bidirectional RRAM, unlike with PCM or unidirectional RRAM, the upstream neuron must change the voltage polarity of its pulses based on whether a partial SET or partial RESET operation is desired. Since the upstream neuron has no knowledge of the possible sign of the update, it must

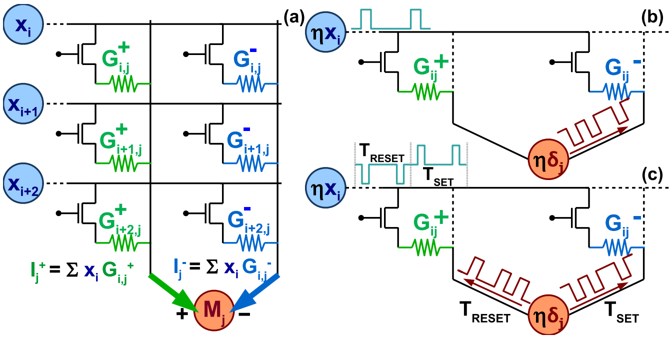


FIGURE 3. **(a) Circuit implementation of the analog multiply-accumulate operation (MAC) during the forward inference phase calls for highly–parallel current integration [28]. Weight update, in both (b) the 'alternate bidirectional' scheme and (c) the 'fully bidirectional' scheme, uses crossbar-compatible weight update [1] to program NVM devices only when both the upstream and downstream neurons have received a large *x* (δ) value during forward inference (reverse propagation). Learning rate (η) is used to scale from *x* (δ) to the actual number of programming pulses fired from the upstream (downstream) neuron.**

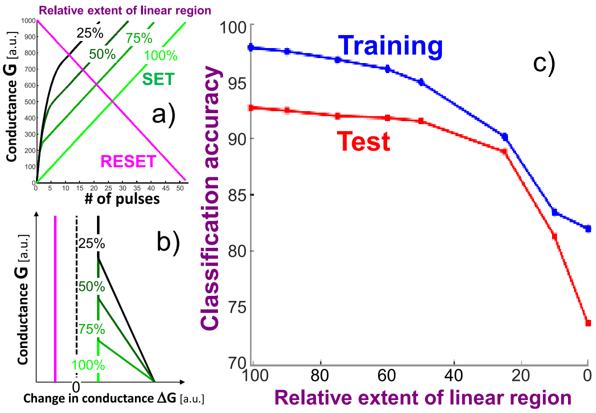


FIGURE 5. **Impact of the relative extent of the linear region of conductance change on neural network performance [2]. (RESET conductance response remains linear at all times). a) Conductance vs. number of pulses, b) hypothetical jump-tables studied, and c) impact on training and test accuracy. A substantial non-linear conductance region (up to ~50%) could be accommodated without loss in application performance.**

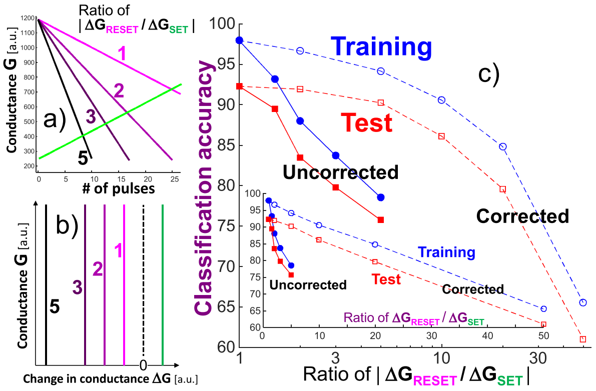


FIGURE 4. **(a) For a set of constructed linear conductance responses where the depression (RESET, magenta) response is steeper than the base potentiation (SET, green) response, the (b) resulting jump-table shows larger (but constant) steps for RESET. (For clarity, only median response is shown.) (c) Although even a small SET/RESET asymmetry causes performance to fall off steeply (solid curves with filled symbols), the downstream neuron can partially compensate for this asymmetry by firing fewer RESET pulses (or more SET pulses). Inset shows same data plotted on a linear horizontal scale.**

fire its pulses twice, in two separate time intervals. In contrast, the downstream neuron with knowledge of the sign of the update through the sign of δ fires its pulses only during the appropriate time interval. Conductance update occurs when the synapse receives pulses from both neurons simultaneously, as mediated by the selection transistor.

Since RRAM devices such as Al/Mo/PCMO are themselves Back-End-Of-the-Line (BEOL)–compatible, integrated two-terminal bidirectional-capable access devices — such as the Mixed-Ionic-Electronic-Conduction (MIEC)–based devices presented in [30] — could also be used, in order to have more silicon area available for implementing the required peripheral circuitry [28]. For the purposes of this study, artificial neurons are considered here to be ideal, implementing a perfect tanh() nonlinear activation function, although imperfect neurons can have their own effect on accuracy [29], [31].

In this paper, two variants on this weight update scheme will be discussed, as shown in Figs. 3(b) and (c). In the 'alternate bidirectional' scheme, only one conductance is programmed for every update step, alternating on each example. For instance, for an even-numbered example, only the positive conductance G^+ might be adjusted, either potentiated if the backpropagation algorithm calls for a weight increase, or depressed if the backpropagation algorithm requests a weight decrease. On odd-numbered examples, the weight change would be implemented using only changes to the G^- conductance. In contrast, in the 'fully bidirectional' scheme, both synapses are adjusted for each example, with both G^+ increased and G^- decreased, or vice-versa. We have previously shown that, for ideal bidirectional conductances, the 'fully bidirectional' scheme is preferable – and this scheme will be used throughout this first section on artificial deviations from such ideal jump-tables.

The first question we addressed was the impact of asymmetry in conductance response [2]. Here we assumed that both conductance responses were linear (Fig. 4(a)), but that RESET conductance response was much steeper than SET, so that the stepsize of the depression (RESET) jump-table was increased (Fig. 4(b)). As shown by the solid curves with filled symbols in Fig. 4(c), even a small degree of asymmetry tended to make classification accuracy fall steeply. However, each downstream neuron has knowledge of the sign of the backpropagated correction, δ, and thus knows whether it is attempting a SET or RESET. This implies that asymmetry can be partly offset by "correcting" a steeper RESET response by firing commensurately fewer RESET pulses (or more SET pulses). As shown by the dotted curves with open symbols in Fig. 4(c), this markedly expanded the asymmetry that could potentially be accommodated [2].

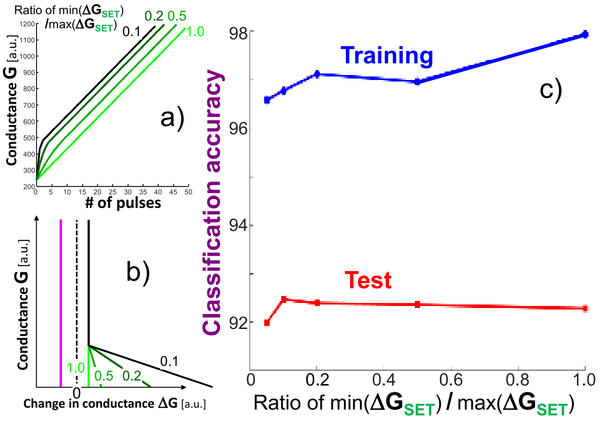


FIGURE 6. Impact of the "strength" of an initial non-linearity on neural network performance [2]. a) Conductance vs. number of pulses, b) hypothetical jump-tables studied, and c) impact on training and test accuracy. Strength of an initial non-linearity does not impact test classification accuracy, so long as a sufficiently large linear region is available.

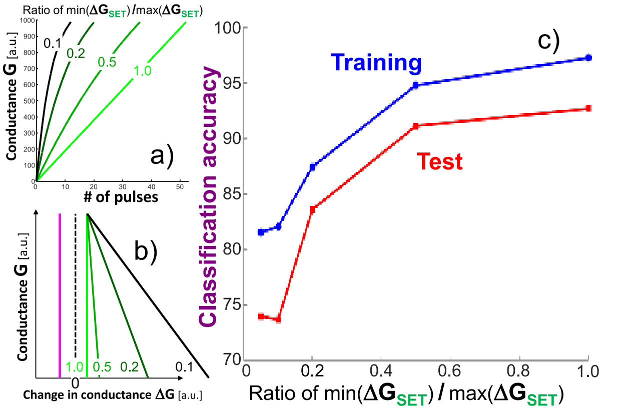


FIGURE 7. Impact of fully non-linear conductance response [2]. a) Conductance vs. number of pulses, b) hypothetical jump-tables studied, and c) impact on training and test accuracy. Even in the absence of a linear region it is possible to achieve high performance – however, the ratio of minimum to maximum conductance change needs to be sufficiently large (>0.5) [2].

Fig. 5 examines jump-tables that incorporated some degree of initial non-linearity in the SET conductance response (Fig. 5(a)) [2]. The relative extent of the linear region was varied from 100% (fully linear) down to near 0% (fully non-linear). For this and all subsequent studies in this section, we assumed that RESET operations were perfectly linear and symmetric to SET (Fig. 5(b)). We found that a substantial non-linear conductance region (up to ∼50%) could be accommodated without a significant drop-off in the neural network performance (Fig. 5(c)) [2].

Fig. 6 examines the impact of the strength of this initial non-linearity on the neural network performance [2]. In these experiments, a stronger (weaker) non-linearity implied fewer (more) steps to traverse the extent of the non-linear region (representing 25% of the total conductance range, Fig. 6(a)). The strength was defined as the ratio between the size of the final (minimum) conductance jump and the initial (maximum) conductance jump (Fig. 6(b)). Again, we found that the strength of the non-linearity had little impact on the test accuracy (Fig. 6(c)), so long as the linear region was sufficiently large [2].

We also investigated fully non-linear conductance responses of varying strengths (Figs. 7(a) and (b)). We found that it was still possible to achieve high classification accuracies (Fig. 7(c)), so long as the ratio of the minimum to maximum conductance jumps was >0.5. However, larger non-linearities caused a marked drop-off in network performance, as a large portion of the dynamic range could be used up by just a few training pulses [2].

IV. AL/MO/PCMO: SIMULATED NETWORK PERFORMANCE

In order to model the rate and variability of PCMO-based weight elements, jump-tables from measured device characteristics were constructed [3]. In Fig. 8, data from 50,000 SET ($-4.0V$, $10ms$) and RESET pulses ($3.5V$, $10ms$) applied to three 200nm-sized devices is plotted, for both Mo/PCMO and Al/Mo/PCMO devices. Due to intrinsic non-linearity, jump sizes at the extremes of both tables appear very large for both Mo/PCMO and Al/Mo/PCMO. Moreover, the average RESET jump size for intermediate conductance values is slightly larger than the SET. As we will see in this section of the paper, these two characteristics can each have an adverse effect on neural network classification performance.

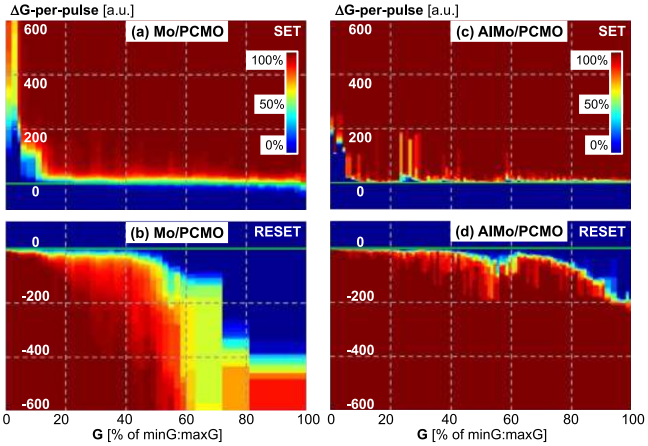


FIGURE 8. (a), (b) SET and RESET jump-tables for Mo/PCMO devices. (c), (d) SET and RESET jump-tables for Al/Mo/PCMO devices. The colormap indicates the cumulative distribution function of conductance-change-per-pulse (ΔG-per-pulse), as a function of the conductance (G). To synthesize the first SET-RESET pair of jump-tables (a), (b), data from initial conductance and resulting conductance change across 50000 total switching pulses applied across three different Mo/PCMO resistive switching memories with via-hole size of 200nm were measured. Similarly, three different Al/Mo/PCMO devices, also of 200nm via size, and a similar number of programming pulses were used to generate (c), (d). All SET pulses were (−4.0 V, 10 ms), all RESET pulses were (3.5V, 10 ms). See companion article [3] for more details.

Fig. 9 shows training and testing accuracy with varying values of the global learning rate, η. In the context of the

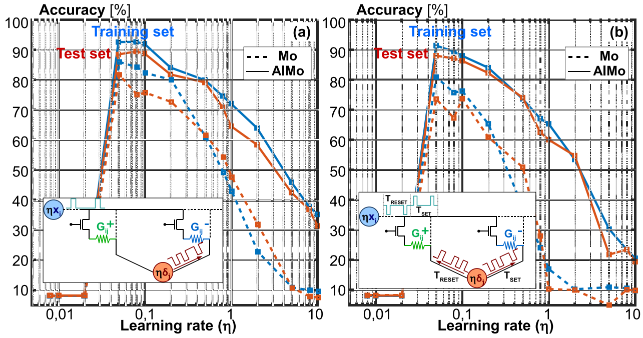


FIGURE 9. Training and test accuracy comparison (vs. global learning rate η) for Mo/PCMO and Al/Mo/PCMO with (a) 'alternate bidirectional' and (b) 'fully bidirectional' weight update schemes. Al/Mo/PCMO-based devices consistently show better neural network training accuracies.

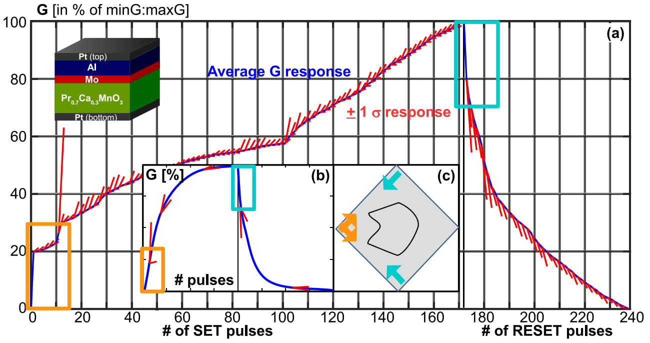


FIGURE 10. (a) Resulting median and $\pm 1\sigma$ conductance response of Al/Mo/PCMO devices under successive 10ms-long programming pulses, as computed from the measured jump-tables (Fig. 8). The median switching characteristic is plotted in blue, with the response at $\pm 1\sigma$ of the C.D.F. drawn in red. (b) Insets show the correspondence between the approximate conductance response of a PCMO-like NVM and (c) the expected effects within the G-diamond. The large conductance increases upon potentiating any low conductance state (orange rectangles) and the large conductance decreases upon depressing any high conductance state (light blue rectangles) leave only a small portion in the center of the G-diamond where synapses can be somewhat protected from large, abrupt weight changes.

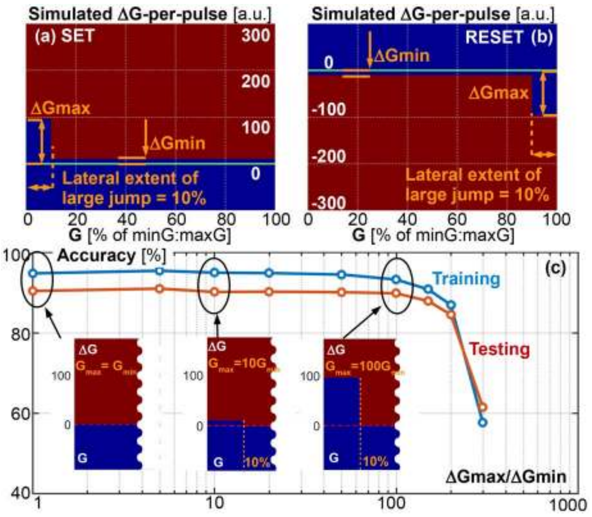


FIGURE 11. (a), (b) Artificial jump-tables of conductance-change-per-pulse (ΔG) constructed for both SET and RESET to simulate the effect of large jumps at the edges of the conductance range. Large jumps (ΔG_{max}) are produced over the ≈ 10% of the total conductance range at the left edge of the SET table (potentiating low conductances) and over a similar extent of the right edge of the RESET jump-table (depressing high conductances). The rest of the table represents a much smaller (but here symmetric) jump, ΔG_{min}. (c) Training and test accuracy while varying the amplitude of the large ΔG_{max} jump with respect to the small fixed ΔG_{min}, using the 'alternate bidirectional' configuration. The system is capable of achieving good performance up to values of $\Delta G_{max}/\Delta G_{min} \approx 100$. For comparison, the measured value for Al/Mo/PCMO is ≈ 189.

crossbar-compatible weight update, the learning rate represents the average number of programming pulses fired per update, and modulates the conversion of the local information within the neuron of x or δ to the number of programming pulses to fire. If the learning rate is very low, for example, no programming occurs at all, since pulses are only fired for exceedingly large x or δ values. Once learning rate is increased beyond some 'critical' threshold at which the system starts learning, classification accuracy tends to decrease as the learning rate increases, affected strongly by the discrepancy between desired and actual weight-changes as implemented within the imperfect memory devices. Because of the more limited conductance range over which similar SET and RESET characteristics are achieved, Mo/PCMO devices exhibit lower overall neural network accuracies than with Al/Mo/PCMO devices. Some of the characteristics highlighted in Part I of this paper [3], such as higher ON/OFF ratio, do in fact help make Al/Mo/PCMO a more promising candidate as a neuromorphic synapse.

Despite this improved performance, Al/Mo/PCMO-based RRAM still exhibit some undesirable characteristics. Fig. 10 shows the Al/Mo/PCMO conductance response (median and stochastic variance) produced by integrating the jump-table from minimum to maximum conductance and back. Although switching characteristics are somewhat linear over a portion of the conductance range, two very steep regions are present at the edges (as expected from inspection of Fig. 8). This causes identical programming pulses to induce very different conductance changes (and thus weight updates that fail to achieve those requested by the neural network) depending on the particular conductance state before programming. Unfortunately, prior measurement of the device initial conductance before each weight-update would introduce impractically-large time and energy demands to the neural network training procedure.

Fig. 11 shows another artificial jump-table experiment, to assess the impact of very large jumps at the lower and upper edges of the conductance range. This study, carried out using the 'alternate bidirectional' weight update scheme, shows that good performance can be maintained for very high jump-size disparity. This makes sense since, according to Fig. 5, large nonlinearities can be tolerated if the portion of the conductance regime over which they occur is small.

Another feature of Al/Mo/PCMO devices is that, because of the structure and switching asymmetry, the size of the conductance change for a single RESET pulse applied to an Al/Mo/PCMO device in an intermediate conductance state

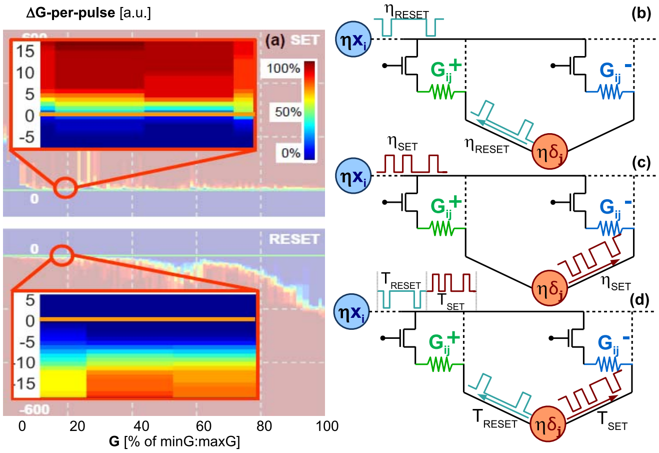


FIGURE 12. NVMs often show asymmetric conductance response due to the differing physics controlling potentiation and depression. Al/Mo/PCMO -based RRAM tend to exhibit a RESET transition which is steeper than the SET transition. This characteristic must to be taken into account when building large scale neuromorphic systems. (b) and (c) show how differentiated learning rates can be chosen in the 'alternate bidirectional' weight update, and (d) in the 'fully bidirectional' scheme.

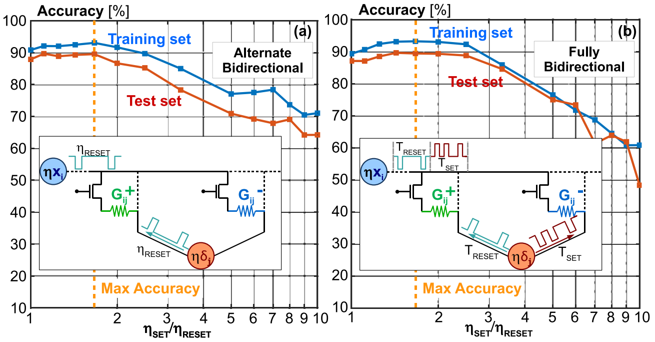


FIGURE 13. Training and test accuracy plotted vs. 'asymmetry correction factor', defined as the ratio between the learning rate for SET and RESET operation (η_{SET}/η_{RESET}) for (a) 'alternate bidirectional' and (b) 'fully bidirectional' weight update schemes. By decreasing the relative RESET learning rate ($\eta_{SET}/\eta_{RESET} > 1$), it is possible to increase classification accuracy, which peaks for $\eta_{SET}/\eta_{RESET} \approx 5/3$. However, over-correcting can cause performance to decline steeply.

is larger than the conductance change induced by a single SET pulse. In Fig. 12, we show a zoomed version of the Al/Mo/PCMO jump-table, clearly showing two different jump-sizes for SET and RESET at intermediate conductance values. Nevertheless, as was already shown in Fig. 4, it is possible to use peripheral circuitry to compensate for this phenomenon — by defining two different learning rates for SET and RESET operation. As a result, the downstream neuron — which knows whether it is inducing a SET or RESET through its knowledge of the sign of δ – can choose to intentionally fire more pulses when performing a SET operation.

The result of this correction is plotted in Fig. 13. For both update algorithms, it is possible to increase the accuracy and find an optimum correction factor (defined as the ratio η_{SET} $/\eta_{RESET}$) which, in both cases, is equal to ≈ 1.66.

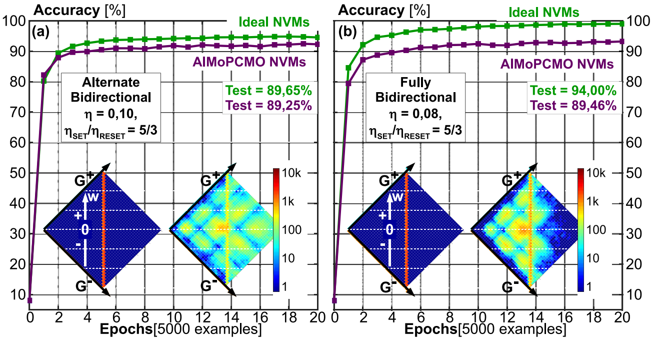


FIGURE 14. Optimized training evolution for Al/Mo/PCMO-RRAMs with the (a) 'alternate bidirectional' and (b) 'fully bidirectional' weight update schemes (purple curves), compared to ideal bidirectional NVMs (green curves). The inset shows the G-diamond plot of synapse distributions before and after training for the Al/Mo/PCMO-RRAMs.

In Fig. 14, training accuracy seems to converge very rapidly to high values ($> 90\%$) after the first few epochs. However, accuracy improvements then seem to saturate after this point within the training. It appears as if the network finds it impossible to tune the weights carefully, due to the occasional disruptions introduced by large jumps induced when potentiating a device from a low conductance value or depressing a device from a high conductance value. By looking at the G-diamond plots (inset), we note that the 'alternate bidirectional' weight update scheme tends to push synapse states to the lateral extents of the G-diamond (e.g., far to the left, towards $G^{+} \sim G^{-} \sim G_{min}$, or far to the right, towards $G^{+} \sim G^{-} \sim G_{max}$). As we have observed before [4], such synaptic states in which low weight is represented by two similarly-low or two similarly-high conductances tend to make it more difficult for the network to move this synapse to a state with a large magnitude should it wish to do so, which then seems to reduce the maximum accuracy achievable. Even with ideal synaptic devices, training accuracy with the 'alternate bidirectional' scheme can never exceed a maximum value of $\approx 95\%$ [4]. For ideal linear bidirectional NVMs, the 'fully bidirectional' scheme works better than the 'alternate bidirectional' scheme [4]. For the highly imperfect Mo/PCMO device, we observe that the 'alternate bidirectional' scheme is preferable (see Fig. 9), as if the adverse G-diamond effects of this scheme were outweighed by the benefits of only having to fire one potentially-problematic programming pulse instead of two. In contrast, with the improved characteristics and higher usable conductance range of Al/Mo/PCMO, and with the mild asymmetry between SET and RESET pulses compensated by tuning η_{SET}/η_{RESET}, then the 'fully bidirectional' scheme becomes slightly preferable, producing a final test accuracy of just under 90% (Fig. 14).

Two-NVM synapses are always needed in order to represent negative weights using only-positive physical quantities (conductances). However, there is no reason that one of these two NVM devices could not be shared amongst many different synapses, with all programming taking place on

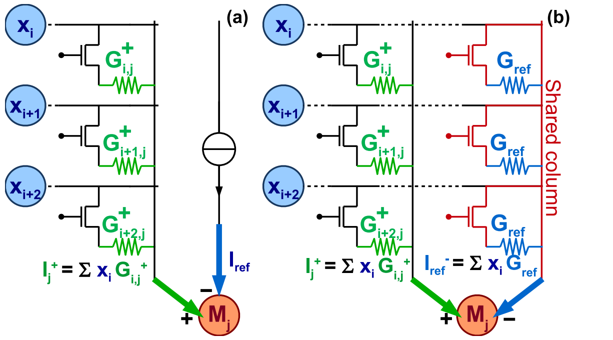


FIGURE 15. (a) Sketched circuit implementation of the 'single bidirectional' weight-update scheme with ideal reference current source. (b) Implementation of such a scheme using a shared conductance column. Note that when the reference current is implemented with reference conductances, it will be critical to compensate for any long-term time-dependent changes in their programmed conductance values, whether due to drift, decay, or any other thermal-switching effect.

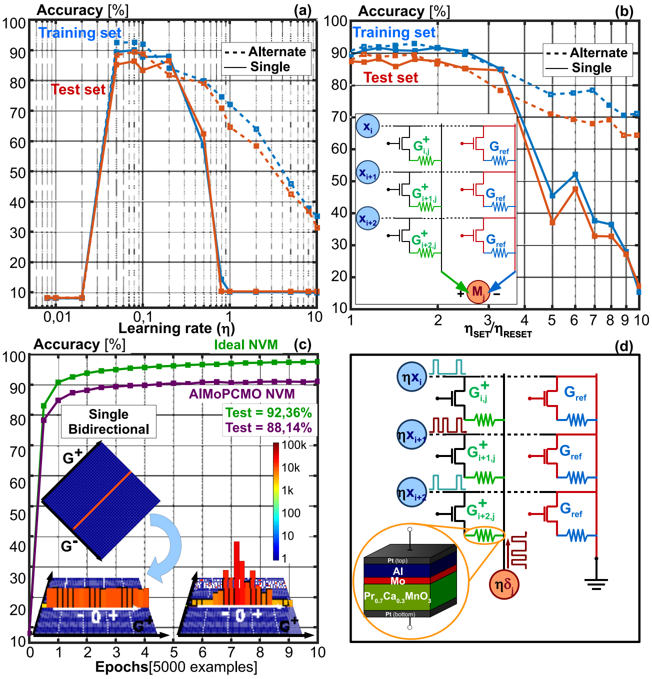


FIGURE 16. (a) Training and test accuracy (vs. relative learning rate) for Al/Mo/PCMO with 'single bidirectional' weight update scheme compared to the 'alternate bidirectional' weight update scheme. Despite having the same peak accuracy, the 'single bidirectional' scheme offers a smaller window for high-accuracy operation. (b) Training and test accuracy for the 'single bidirectional' scheme plotted vs. the 'asymmetry correction factor.' (c) Optimized training for image recognition with inset showing the distribution of conductances before and after the training was performed. (d) Circuit implementation of the 'single bidirectional' weight update scheme.

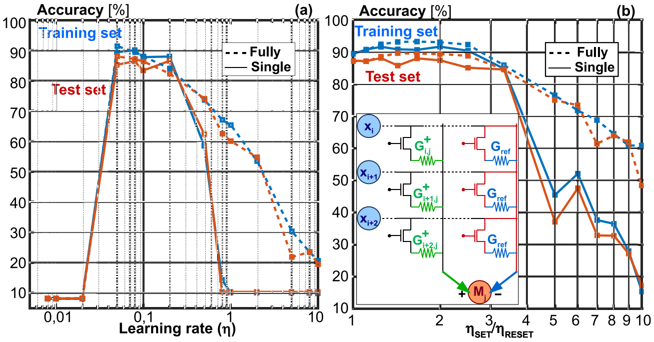


FIGURE 17. (a) Training and test accuracy (vs. relative learning rate) for Al/Mo/PCMO with 'single bidirectional' weight update compared to the 'fully bidirectional' weight update scheme. As with the 'alternate bidirectional' results, the 'single bidirectional' scheme can offer the same peak accuracy but a smaller operating window for learning rates. (b) Training and test accuracy plotted vs. the 'asymmetry correction factor.'

just G^+. This allows implementation of a neuromorphic system in which the density of synapses is very close to the density of conductances. However, for devices such as PCM or filamentary-RRAM, an explicit two-NVM scheme is essential in order to achieve smooth bidirectional weight change [4], since each device only supports gradual conductance change in only one direction (SET for PCM, RESET for filamentary-RRAM).

As shown in Fig. 10, Al/Mo/PCMO offers a truly bidirectional analog conductance behavior, making it possible to use just one device to represent each synaptic weight, together with a shared reference current. The value of such a reference should be

$$I_{ref} = \frac{G_{max}+G_{min}}{2}\cdot\sum_{1}^{n} x_i = G_{ref}\cdot\sum_{1}^{n} x_i, \qquad (1)$$

which can be generated easily either by adding a shared reference column where a large number of conductances are programmed one-time to specific values, or with circuit techniques for generating a specified reference current (Fig. 15).

In Figs. 16 and 17, high accuracy and tolerancing to learning-rate is shown for single-Al/Mo/PCMO synapses. As with the two-PCMO schemes already discussed, training accuracy can be maximized by tuning the ratio of the learning rates to compensate for the slight SET-RESET asymmetry in intermediate conductance states exhibited in the jump-tables (Fig. 8). The peak value ($\approx 91.62\%$ for training and $\approx 88.14\%$ for testing) is reached for $\eta_{SET}/\eta_{RESET} \approx 5/4$. In this single-Al/Mo/PCMO configuration, there is a one-to-one correspondence between conductance and weights, so the single diagonal line left across the center of the G-diamond can be plotted as a histogram. We note that, starting from a uniform distribution, most of the weights (conductances) become zero (e.g., their conductance approaches G_{ref}) during training. During programming, the reference column is disconnected from the firing neuron to prevent accidental programming of the reference conductances. One of the consequences of the 'single bidirectional' configuration is that the weight range available for every synapse is inherently smaller by a factor of two.

V. CONCLUSION

We have used measured jump-tables — which describe the full cumulative distribution function (CDF) of conductance-change at each device conductance value, for both weight potentiation (SET) and depression (RESET) — to model the performance of PCMO-based devices on a typical multi-layer perceptron trained on MNIST. Artificially constructed jump-tables illustrated the impact of simple deviations from linear and symmetric conductance change on neural network classification performance [2]. We studied the difference between older Mo/PCMO devices and Al/Mo/PCMO devices [3], using two different weight-update schemes, "alternate bidirectional" and "fully bidirectional." We showed that the accuracy issues of $Pr_{0.7}Ca_{0.3}MnO_3$devices are not simply due to the very large conductance changes exhibited when either potentiating from the lowest conductance states or depressing from the highest conductance states. By offsetting the slight asymmetry between the average size of the SET and RESET conductance changes, tuning of the relative learning rate for potentiation (SET) with respect to depression (RESET) was shown to provide further accuracy improvements. When compared with phase-change memory (PCM), PCMO synapses represent a potential improvement over previous work [1], [4] in terms of raw training accuracy and support for bidirectional conductance response. We showed that this can potentially support a more compact synaptic cell (1T1R instead of 2T2R) while avoiding the need to perform an "Occasional RESET" step. Thus the bidirectional programming of Al/Mo/PCMO can be used to implement high-density neuromorphic systems with only a single conductance per synapse, at only a slight degradation to classification accuracy.

REFERENCES

[1] 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.*, San Francisco, CA, USA, 2014, pp. 29.5.1–29.5.4.

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

[3] K. Moon *et al.*, "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," *IEEE J. Electron Devices Soc.*, to be published.

[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 non-volatile memory as the synaptic weight element: Comparative performance analysis (accuracy, speed, and power)," in *IEEE 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, 2015.

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

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

[9] 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.

[10] 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.

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

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

[13] 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.

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

[15] 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.

[16] 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.

[17] 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 *IEEE IEDM Tech. Dig.*, Washington, DC, USA, 2015, pp. 17.4.1–17.4.4.

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

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

[20] 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 *IEEE IEDM Tech. Dig.*, Washington, DC, USA, 2015, pp. 17.1.1–17.1.4.

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

[22] 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.

[23] 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.

[24] G. W. Burr *et al.*, "Phase change memory technology," *J. Vac. Sci. Technol. B, Nanotechnol. Microelectron. Mater. Process. Meas. Phenomena*, vol. 28, no. 2, pp. 223–262, 2010.

[25] 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 *IEEE IEDM Tech. Dig.*, Baltimore, MD, USA, 2009, pp. 1–4.

[26] 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. 1–10.

[27] 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 *IEEE IEDM Tech. Dig.*, Washington, DC, USA, 2015, pp. 17.6.1–17.6.4.

[28] P. Narayanan *et al.*, "Toward on-chip acceleration of the backpropagation algorithm using nonvolatile memory," *IBM J. Res Develop.*, vol. 61, no. 4, pp. 1–11, Jul./Sep. 2017.

[29] 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.

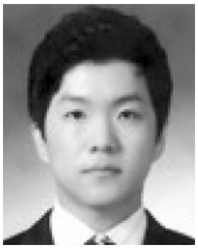
[30] G. W. Burr *et al.*, "Large-scale (512kbit) integration of multilayer-ready access-devices based on mixed-ionic-electronic-conduction (MIEC) at 100% yield," in *Proc. IEEE Symp. VLSI Technol. (VLSIT)*, Honolulu, HI, USA, 2012, pp. 41–42.

[31] P. Narayanan *et al.*, "Reducing circuit design complexity for neuromorphic machine learning systems based on non-volatile memory arrays," in *Proc. ISCAS*, Baltimore, MD, USA, 2017, pp. 1–4.

ALESSANDRO FUMAROLA received the M.S. degree in nanotechnology jointly from the Polytechnic of Turin, Italy, INP Grenoble, France, and EPFL Lausanne, Switzerland. He was with IBM Research, Almaden, San Jose, CA, USA, for six months. He is currently pursuing the Ph.D. degree with the Max Planck Institute for Microstructure Physics, Halle, Germany. 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 the master's degree work with the IBM Zurich Research Laboratory, Rüschlikon, Switzerland.

His current research interests include cognitive computing and systems engineering.



KIBONG MOON received the B.S. degree with 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.

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. In 2016, he joined Samsung Electronics.

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.

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 the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS, the IEEE TRANSACTIONS ON NANOTECHNOLOGY, the *ACM Journal on Emerging Technologies in Computing Systems*, and several IEEE conferences.

YUSUF LEBLEBICI (M'90–SM'98–F'10) received the B.Sc. and M.Sc. degrees in electrical engineering from Istanbul Technical University, Turkey, in 1984 and 1986, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana–Champaign, in 1990.

In 2002, he joined the Swiss Federal Institute of Technology, Lausanne, as a Professor. His research interests include design of high-speed CMOS digital and mixed-signal integrated circuits, computer-aided design, and reliability of VLSI systems.

HYUNSANG HWANG received the Ph.D. degree in materials science from the University of Texas at Austin, Austin, TX, USA, in 1992. After five years with LG Semiconductor Corporation, 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, Pohang University of Science and Technology, South Korea.

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.