Large-scale neural networks implemented with non-volatile memory as the synaptic weight element: impact of conductance response

Severin Sidler*‡, Irem Boybat*‡, Robert M. Shelby*, Pritish Narayanan*, Junwoo Jang*†, Alessandro Fumarola*‡, Kibong Moon†, Yusuf Leblebici‡, Hyunsang Hwang†, and Geoffrey W. Burr*

*IBM Research–Almaden, 650 Harry Road, San Jose, CA 95120, Tel: (408) 927–1512, Email: *gwburr@us.ibm.com*
†Department of Material Science and Engineering, Pohang University of Science and Technology, Pohang 790-784, Korea
‡EPFL, Lausanne, CH–1015 Switzerland

***Abstract*—We assess the impact of the conductance response of Non-Volatile Memory (NVM) devices employed as the synaptic weight element for on-chip acceleration of the training of large-scale artificial neural networks (ANN). We briefly review our previous work towards achieving competitive performance (classification accuracies) for such ANN with both Phase-Change Memory (PCM) [1], [2] and non-filamentary ReRAM based on PrCaMnO (PCMO) [3], and towards assessing the potential advantages for ML training over GPU–based hardware in terms of speed (up to 25× faster) and power (from 120–2850× lower power) [4]. We then discuss the "jump-table" concept, previously introduced to model real-world NVM such as PCM [1] or PCMO, to describe the full cumulative distribution function (CDF) of conductance-change at each device conductance value, for both potentiation (SET) and depression (RESET). Using several types of artificially–constructed jump-tables, we assess the relative importance of deviations from an ideal NVM with perfectly linear conductance response.**

I. Introduction

By performing computation at the location of data, non-Von Neumann (non–VN) computing *ought* to provide significant power and speed benefits (Fig. 1) on specific and assumably important tasks. For one such non–VN approach — on-chip training of large-scale ANN using NVM-based synapses [1]–[4] — viability will require several things. First, despite the inherent imperfections of NVM devices such as Phase Change Memory (PCM) [1], [2] or Resistive RAM (RRAM) [3], such NVM-based networks must achieve competitive performance levels (e.g., classification accuracies) when compared to ANN trained using CPUs or GPUs. Second, the benefits of performing computation at the data (Fig. 2) must confer a decided advantage in either training power or speed (or preferably, both). And finally, any on-chip accelerator should be applicable towards networks of different types (fully–connected "Deep" NN or Convolutional NN) and/or be reconfigurable for networks of different shapes (wide with many neurons, or deep with many layers).

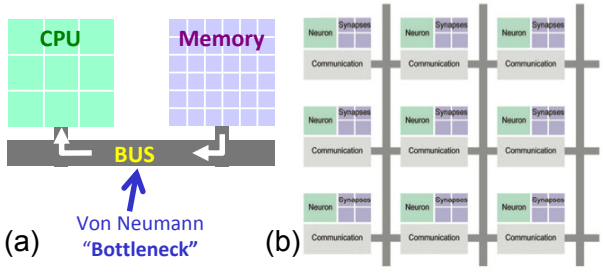

Fig. 1. In the Von Neumann architecture (a), data (both operations and operands) must move to and from the dedicated Central Processing Unit (CPU) along a bus. In contrast, in a Non–Von Neumann architecture, distributed computations take place at the location of the data, reducing the time and energy spent moving data around [1].

We briefly review our work [1]–[4] in assessing the accuracy, speed and power potential of on-chip NVM–based ML.

A. Potential for competitive classification accuracies

Using 2 phase-change memory (PCM) devices per synapse, we demonstrated a 3–layer perceptron with 164,885 synapses [1], trained with backpropagation [5] on a subset (5000 examples) of the MNIST database of handwritten digits [6] (Fig 3), using a modified weight-update rule compatible with NVM+selector crossbar arrays [1]. We proved that this modification does not degrade the high "test" (generalization) accuracies such a 3–layer network inherently delivers on this problem when trained in software [1]. However, nonlinearity and asymmetry in PCM conductance response limited both "training" and "test" accuracy in these initial experiments to 82–83% [1] (Fig. 4).

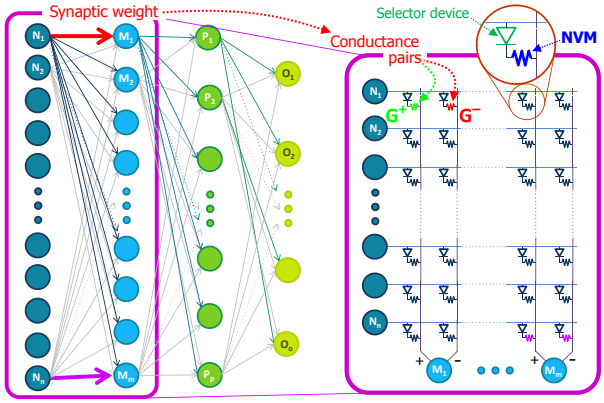

Fig. 2. Neuro-inspired non-Von Neumann computing [1]–[4], in which neurons activate each other through dense networks of programmable synaptic weights, can be implemented using dense crossbar arrays of nonvolatile memory (NVM) and selector device-pairs [1].

Asymmetry (between the gentle conductance increases of PCM partial–SET and the abruptness of PCM RESET) was mitigated by an occasional RESET strategy, which could be both infrequent and inaccurate [1]. While in these initial experiments, network parameters such as learning rate η had to be tuned very carefully, a modified 'LG' algorithm offered wider tolerance to η, higher classification accuracies, and lower training energy [4].

Tolerancing results showed that all NVM-based ANN can be expected to be **highly resilient to random effects** (NVM variability, yield, and stochasticity), but **highly sensitive to "gradient" effects that act to steer all synaptic weights** [1]. We showed that a bidirectional NVM with a symmetric, linear conductance response of finite but large dynamic range (e.g., each conductance step is relatively small) can deliver the same high classification accuracies on the MNIST digits as a conventional, software-based implementation (Fig. 5). One key observation is the importance of avoiding constraints on weight magnitude that arise when the two conductances are either both small or both large — e.g., synapses should remain in the center stripe of the "G-diamond" [2].

In this paper, we extend upon this observation to explore the impact of specific deviations from such an idealized linear conductance response.

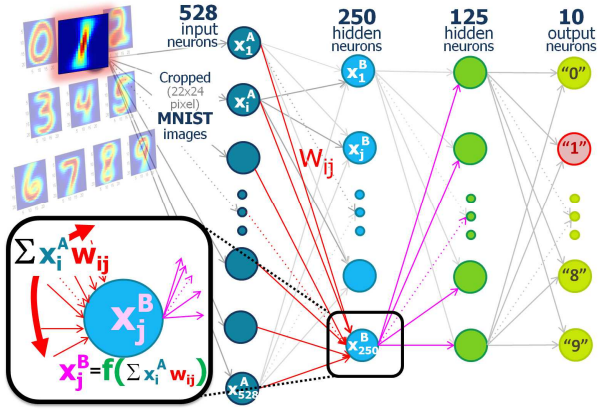

Fig. 3. In forward evaluation of a multilayer perceptron, each layer's neurons drive the next layer through weights w_{ij} and a nonlinearity $f()$. Input neurons are driven by input (for instance, pixels from successive MNIST images (cropped to 22×24)); the 10 output neurons classify which digit was presented [1].

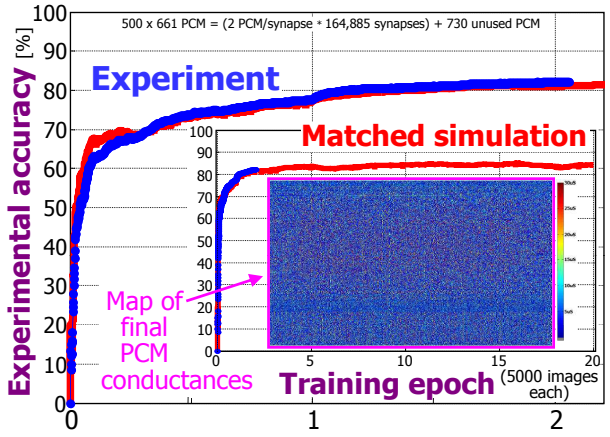

Fig. 4. Training accuracy for a 3–layer perceptron of 164,885 hardware-synapses [1], with all weight operations taking place on a 500 × 661 array of mushroom-cell PCM devices. Also shown is a matched computer simulation of this NN, using parameters extracted from the experiment [1].

B. Comparative analysis of speed and power

We have also assessed the potential advantages, in terms of speed and power, of on-chip machine learning (ML) of large-scale artificial neural networks (ANN) using Non-Volatile Memory (NVM)-based synapses, in comparison to conventional GPU–based hardware [4].

Under moderately-aggressive assumptions for parallel–read and –write speed, **PCM-based on-chip machine learning can potentially offer lower power and faster training (per ANN example) than GPU-based training** for both large and small networks (Fig. 6), even with the time and energy required for occasional RESET (forced by the large asymmetry between gentle partial-SET and abrupt RESET in PCM). Critical here is the design of area-efficient read/write circuitry, so that many copies of this circuitry operate in parallel (each handling a small number of columns (rows), $\mathbf{c}_s$).

II. Jump-table concept

A useful concept in modeling the behavior of real NVM devices for neuromorphic applications is the concept of a

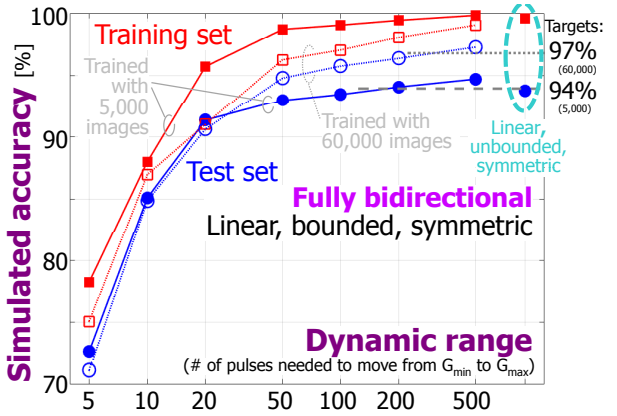

Fig. 5. When the dynamic range of the linear response is large, the classification accuracy can now reach that of the original network (a *test* accuracy of 94% when trained with 5,000 images; of 97% when trained with all 60,000 images) [2].

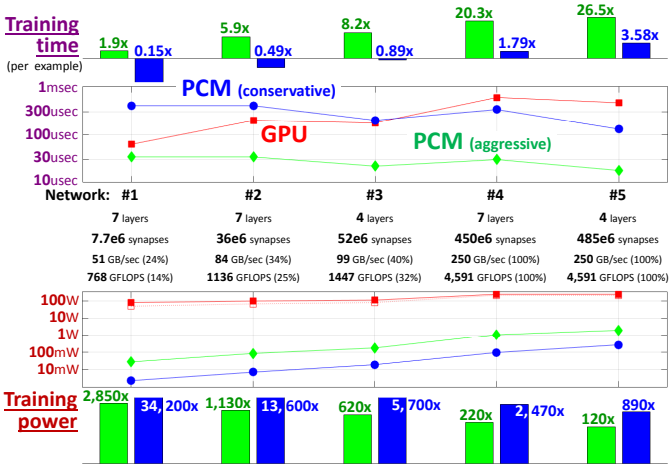

Fig. 6. Predicted training time (per ANN example) and power for 5 ANNs, ranging from 0.2GB to nearly 6GB [4]. Under moderately-aggressive assumptions for parallel–read and –write speed, PCM-based on-chip machine learning can offer lower power and faster training for both large and small networks [4].

"jump-table." For backpropagation training, where one or more copies of the same programming pulse are applied to the NVM for adjusting the weights [1], we simply need one jump-table for potentiation (SET) and one for depression (RESET).

With a pair of such jump-tables, we can capture the nonlinearity of conductance response as a function of conductance (e.g., the same pulse might create a large "jump" at low conductance, but a much smaller jump at high conductance), the asymmetry between positive (SET) and negative (RESET) conductance changes, and the inherent stochastic nature of each jump. Fig. 7(a) plots median conductance change for potentiation (blue) together with the $\pm 1\sigma$ stochastic variation about this median change (red). Fig. 7(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 at any given initial conductance G. This table is ideal for computer simulation because a random number r (uniform deviate, between 0.0 and 1.0) can be converted to a resulting ΔG produced by a single pulse by scanning along the row associated with the conductance G (of the device before the pulse is applied) to find the point at which the table entry just exceeds r.

III. Impact of Nonlinear Conductance Response

We have previously used a measured jump-table to simulate the SET response of PCM devices [1], and are currently exploring the use of similarly measured jump-tables for PCMO. In order to develop an intuitive understanding of the impact that various features of such jump-tables have on the classification performance in the ANN application, we study various artificially-constructed jump-tables. Except for the specific jump-tables, these simulations are identical to those performed in Ref [1], spanning 20 epochs.

The first question we address is the impact of asymmetry in conductance response. Here we assume both conductance responses are linear (Fig. 8(a)), but RESET conductance response is much steeper than SET, so that the stepsize of the depression (RESET) jump-table is increased (Fig. 8(b)). As shown by the solid curves with filled symbols in Fig. 8(c), even a small degree of asymmetry can cause classification accuracy to fall steeply. However, each downstream neuron has knowledge of the sign of the backpropagated correction, δ, and

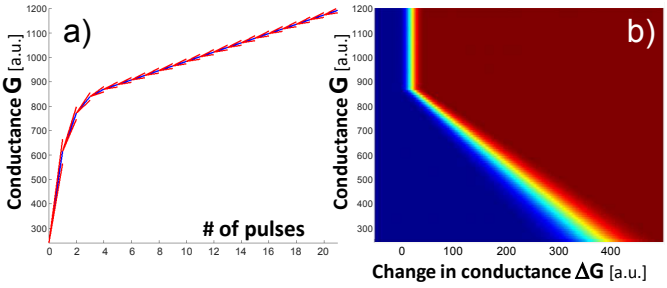

Fig. 7. (a) Example median (blue) and $\pm 1\sigma$ (red) conductance response for potentiation. (b) associated jump-table that fully captures this (artificially constructed in this case) conductance response, with cumulative probability plotted in color (from 0 to 100%) of any conductance change ΔG at any given initial conductance G.

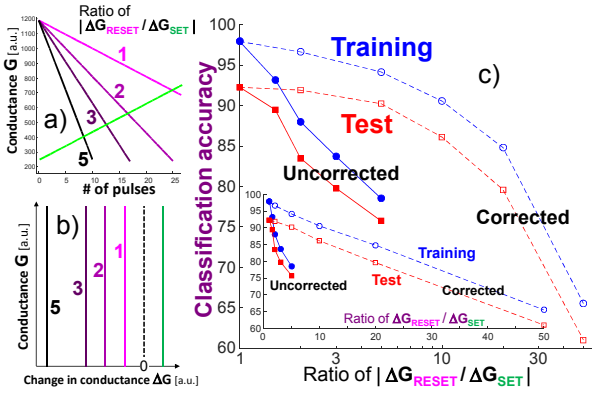

Fig. 8. (a) For a set of constructed linear conductance responses where the depression (RESET, magenta) response is steeper than the base potentiation (SET, green) response, the (b) resulting jump-table shows larger (but constant) steps for RESET. (For clarity, only median response is shown.) (c) Although 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.

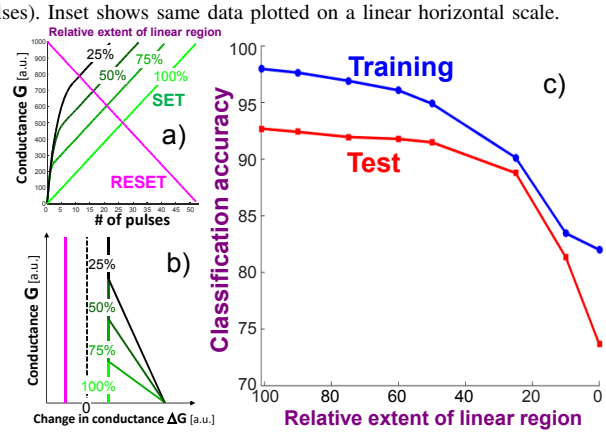

Fig. 9. Impact of relative extent of linear region on neural network performance. (RESET conductance response remains linear at all times). a) Conductance vs. number of pulses, b) hypothetical jump tables studied, and c) impact on training and test accuracy. A substantial non-linear conductance region (up to ~50%) can be accommodated without loss in application performance.

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. 8(c), this markedly expands the asymmetry that could potentially be accommodated.

Fig. 9 examines jump-tables that incorporate some degree of initial non-linearity in the SET conductance response (Fig. 9(a)). The relative extent of the linear region is varied from 100% (fully linear) down to near 0% (fully nonlinear). For this and all subsequent studies, we assume that RESET operations remain perfectly linear and symmetric to SET (Fig. 9(b)). We find that a substantial non-linear conductance region (up to ~50%) can be accommodated without a significant drop-off in the neural network performance (Fig. 9(c)).

Fig. 10 examines the impact of the strength of this initial non-linearity on the neural network performance. In these experiments, a stronger (weaker) non-linearity implies fewer (more) steps to traverse the extent of the non-linear region (representing 25% of the total conductance range, Fig. 10(a)).

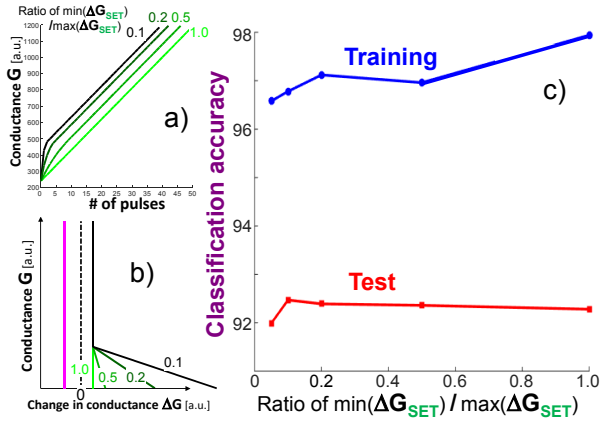

Fig. 10. Impact of the strength of an initial non-linearity on neural network performance. a) Conductance vs. number of pulses, b) hypothetical jump tables studied, and c) impact on training and test accuracy. Strength of an initial non-linearity does not impact test classification accuracy, so long as a sufficiently large linear region is available.

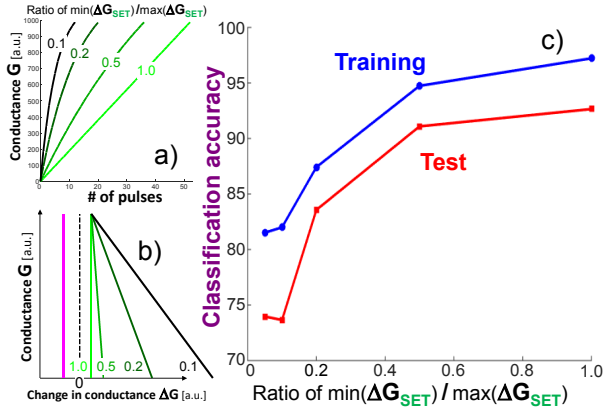

Fig. 11. Impact of fully non-linear conductance response. a) Conductance vs. number of pulses, b) hypothetical jump tables studied, and c) impact on training and test accuracy. Even in the absence of a linear region it is possible to achieve high performance – however, the ratio of minimum to maximum conductance change needs to be sufficiently large (>0.5).

The strength is defined as the ratio between the size of the final (minimum) conductance jump and the initial (maximum) conductance jump (Fig. 10(b)). Again, we find that the strength of the non-linearity has little impact on the test accuracy (Fig. 10(c)), so long as the linear region is sufficiently large.

We also investigate fully non-linear conductance responses of varying strengths (Figs. 11(a) and (b)). We find that it is still possible to achieve high classification accuracies (Fig. 11(c)), so long as the ratio of the minimum to maximum conductance jumps is >0.5. However, larger non-linearities cause a marked drop-off in network performance, as a large portion of the dynamic range can be used up by just a few training pulses.

IV. Conclusion

We have assessed the impact of the conductance response of Non-Volatile Memory (NVM) devices employed as the synaptic weight element for on-chip acceleration of the training of large-scale artificial neural networks (ANN). We briefly reviewed our previous work towards achieving competitive performance (classification accuracies) for such ANN with both Phase-Change Memory [1], [2] and non-filamentary ReRAM based on PrCaMnO (PCMO) [3], and towards assessing the potential advantages for ML training over GPU–based hardware in terms of speed (up to 25× faster) and power (from 120–2850× lower power) [4]. We discussed the "jump-table" concept, previously introduced to model real-world NVM such as PCM [1] or PCMO, to describe the full cumulative distribution function (CDF) of resulting conductance-change at each possible conductance value, for both potentiation (SET) and depression (RESET).

Using various artificially–constructed jump-tables, we assessed the relative importance of deviations from an ideal NVM with a linear conductance response. While even a small SET/RESET asymmetry between otherwise linear conductance responses can cause performance to fall off steeply, downstream neurons can partially compensate for this asymmetry by firing fewer RESET pulses (or more SET pulses), allowing reasonable performance even in the presence of a significant asymmetry. We also found that a substantial non-linear conductance region (up to ∼50%) can be accommodated, and that the strength of this initial non-linearity (ratio of minimum to maximum conductance change) can be significant, so long as a sufficiently large linear region is available. Even with fully nonlinear responses, it is possible to achieve high performance so long as the ratio of minimum to maximum conductance change is sufficiently close to unity (>0.5).

While the 'LG' algorithm, together with other approaches, should help a nonlinear, asymmetric NVM (such as PCM) act more like an ideal linear, bidirectional NVM, the identification of NVM devices and/or pulse-schemes that can offer a conductance response that is at least partly linear will help significantly in achieving high classification accuracies.

References

[1] G. W. Burr, R. M. Shelby, C. di Nolfo, J. W. Jang, R. S. Shenoy, P. Narayanan, K. Virwani, E. U. Giacometti, B. Kurdi, and H. Hwang, "Experimental demonstration and tolerancing of a large-scale neural network (165,000 synapses), using phase-change memory as the synaptic weight element," in *IEDM*, 2014, p. 29.5.

[2] G. W. Burr, R. M. Shelby, S. Sidler, C. di Nolfo, J. Jang, I. Boybat, R. S. Shenoy, P. Narayanan, K. Virwani, E. U. Giacometti, B. Kurdi, and H. Hwang, "Experimental demonstration and tolerancing of a large–scale neural network (165,000 synapses), using phase–change memory as the synaptic weight element," *IEEE Trans. Electr. Dev.*, vol. 62, no. 11, pp. 3498–3507, 2015.

[3] 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 Letters*, vol. 36, no. 5, pp. 457–459, 2015.

[4] G. W. Burr, P.Narayanan, R. M. Shelby, S. Sidler, I. Boybat, C. di Nolfo, and Y. Leblebici, "Large–scale neural networks implemented with nonvolatile memory as the synaptic weight element: comparative performance analysis (accuracy, speed, and power)," in *IEDM Technical Digest*, 2015, p. 4.4.

[5] D. Rumelhart, G. E. Hinton, and J. L. McClelland, "A general framework for parallel distributed processing," in *Parallel Distributed Processing*. MIT Press, 1986.

[6] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, p. 2278, 1998.