Stochastic weight updates in phase-change memory-based synapses and their influence on artificial neural networks

Irem Boybat^{*†§}, Manuel Le Gallo[†], Timoleon Moraitis[†], Yusuf Leblebici[§],
Abu Sebastian^{*†}, and Evangelos Eleftheriou[†]

[†]IBM Research – Zurich, 8803 Rüschlikon, Switzerland

^{*}Email: {ibo, ase}@zurich.ibm.com

[§]Ecole Polytechnique Federale de Lausanne (EPFL), 1015 Lausanne, Switzerland

Abstract—Artificial neural networks (ANN) have become a powerful tool for machine learning. Resistive memory devices can be used for the realization of a non-von Neumann computational platform for ANN training in an area-efficient way. For instance, the conductance values of phase-change memory (PCM) devices can be used to represent synaptic weights and can be updated in-situ according to learning rules. However, non-ideal device characteristics pose challenges to reach competitive classification accuracies. In this paper, we investigate the impact of granularity and stochasticity associated with the conductance changes on ANN performance. Using a PCM prototype chip fabricated in the 90 nm technology node, we present a detailed experimental characterization of the conductance changes. Simulations are done in order to quantify the effect of the experimentally observed conductance change granularity and stochasticity on classification accuracies in a fully connected ANN trained with backpropagation.

I. INTRODUCTION

Approaches based on artificial neural networks (ANNs) such as deep learning achieve unprecedented human-like performance in many real-world tasks. ANNs are commonly trained by a supervised training algorithm called backpropagation. After forward propagation of the input, the network output is compared to the input labels. The errors are back-propagated and all the synaptic weights are updated accordingly. However, training of large ANNs using conventional von Neumann computing systems is highly inefficient due to the physical separation between the memory and processing units. Recently, dense cross-bar arrays of resistive memory devices have been proposed as a non-von Neumann computational platform to perform the various steps involved in the training of ANNs [1], [2]. The conductance of resistive memory devices are used to represent the synaptic weights. The matrix-vector multiplications that are needed during the forward propagation of the network can then be achieved through Ohm's and Kirchhoff's laws. Weight updates can be achieved by modifying the conductance levels of the resistive memory devices by the application of appropriate electrical pulses.

Phase-change memory (PCM) is a type of resistive memory device where the synaptic weight can be stored in the phase configuration of the device. Non-volatile multi-level capabilities make PCM attractive for neural network applications. Burr et al. presented an experimental demonstration of a large-scale ANN using PCM devices fabricated in the 180 nm technology node, concluding that the performance of the neural network is quite robust to stochasticity of conductance changes [1]. However, the amount of device stochasticity may

be significantly larger for current and future devices fabricated at smaller technology nodes and with other material classes. In this paper, we investigate the impact of device stochasticity and its effect on network performance for PCM devices fabricated in the 90 nm technology node. First, we present a thorough experimental characterization of conductance changes in PCM devices fabricated in the 90 nm technology node. Secondly, we perform extensive simulations to explore how the statistics acquired from the experiments influence the performance of an ANN. Specifically, we focus on two aspects: the granularity of weight updates and the update randomness.

II. CONDUCTANCE CHANGE IN PHASE-CHANGE DEVICES

The PCM devices are based on doped-Ge₂Sb₂Te₅(GST) and were integrated into a prototype chip with 3 million devices fabricated in the 90 nm CMOS technology node [3]. The chip also integrates the circuitry for cell addressing, 8-bit on-chip ADC for cell readout, and voltage- or current-mode cell programming. An experimental hardware platform is built around the prototype. An FPGA board with embedded processor and Ethernet connection implements the overall system control and data management.

Fig. 1. Schematic illustration of a PCM device.

The PCM device consists of a tiny volume of phase-change material sandwiched between two electrodes (see Fig. 1) [4]. In an as-fabricated device, the material is in the crystalline phase with a high conductance. When a current pulse of sufficiently high amplitude (referred to as *RESET pulse*) is applied, a significant portion of the phase-change material melts due to Joule heating and when the pulse is stopped abruptly, the molten material quenches into the amorphous phase because of glass transition. The device will be in a low conductance state (initial conductance in Fig. 2(a)). When a current pulse (referred to as *SET pulse*) is applied to a PCM device in the low conductance state such that the temperature

Fig. 2. (a) Evolution of the conductance G as a function of the number of SET pulses. The error bar indicates the standard deviation over the 9,868 devices. (b) Histogram showing the distributions of conductance values after application of the third and the fourth SET pulses.

reached in the cell via Joule heating is high enough, but below the melting temperature, a part of the amorphous region crystallizes and the conductance increases again. With subsequent application of such pulses, one can progressively crystallize the amorphous region and thus progressively increase the device conductance. The crystallization dynamics are found to be mostly dominated by crystal growth at these length scales [5]. The extent of crystallization depends on the amplitude and duration of the SET pulse. Since the RESET pulse results in an abrupt decrease in conductance, two PCM devices in a differential configuration are typically employed to store the synaptic weights in an ANN [6], [1].

First, we present a thorough characterization of the change in conductance with the application of the SET pulses. The measurements are based on 9,868 devices. First, devices are initialized to a conductance close to zero. Thereafter, a SET pulse of $70 \mu\text{A}$ is applied. After the application of the pulse, the devices are read with a read voltage of approximately 0.3 V . The resulting current is converted to a voltage signal and digitized using the on-chip 8-bit ADC, calibrated using on-chip polysilicon resistors. The device conductance values are estimated based on the read voltage and the measured read current. This read operation is repeated 50 times with an interval of approx. 3 s and an average conductance value, G , is obtained. This averaging is required to eliminate the conductance variations arising from drift [7] and noise [3]. Figure 2(a) shows the evolution of the mean conductance of the 9,868 devices as a function of the number of pulses

Fig. 3. (a) Histogram showing the distribution of the conductance change ΔG between application of the third and the fourth SET pulses. (b) The mean behavior of ΔG as a function of the number of SET pulses. The error bar indicates the standard deviation over the 9,868 devices.

applied. It can be seen that G increases monotonically as a function of the number of pulses.

From this experimental data, it is also possible to estimate the conductance change per application of each SET pulse (denoted by ΔG). Figure 2(b) shows the distributions of the conductance values after application of the third and fourth SET pulses. From this data, one can calculate the change in conductance values corresponding to the application of the fourth SET pulse as shown in Fig. 3(a). The mean and standard deviation of ΔG as a function of the number of SET pulses are shown in Fig. 3(b).

In Fig. 4 we present the total conductance change as a function of the number of SET pulses. The mean conductance change is equal to the sum of the means of the ΔG random variables corresponding to the application of each SET pulse. It can also be seen that the standard deviation of the total conductance change is less than that obtained assuming independence of the ΔG random variables.

This characterization work highlights some of the key attributes associated with the conductance change with the application of SET pulses. First of all, the minimal mean conductance change that can be induced can be quite large especially for low values of device conductance. This lower bound characterizing the conductance updates implies a granularity of possible weight updates in a PCM synapse. Secondly, there is a significant randomness associated with ΔG . In fact it can be seen that the standard deviation of ΔG is comparable to or even larger than the mean for all the SET pulses. The ΔG

Fig. 4. Mean and standard deviation of the total conductance change as a function of the number of SET pulses. The measured standard deviation of the total conductance change is less than that obtained assuming that the ΔG random variable is independent.

granularity is comparable but the randomness associated with ΔG is larger compared to what is reported for 180 nm technology node [1]. However, note that the phase-change material used in our devices is not identical to that used in [1]. So a more thorough investigation of devices fabricated in different technology nodes but with identical phase-change materials is needed before we can draw conclusions on the influence of device dimensions on the ΔG randomness. This will be the subject of future work. The impact of the randomness and granularity in the weight change on the performance of ANNs will be investigated in Section III.

III. IMPACT ON THE PERFORMANCE OF ARTIFICIAL NEURAL NETWORKS

The ANN is trained using the backpropagation algorithm [8] on the common task of classifying the MNIST set [9] (Fig. 5(a)). The set contains 60,000 training and 10,000 test images of hand-written digits (all labeled), with each image consisting of 28x28 grey-scale pixels. Training the network with backpropagation takes place in two phases: the forward propagation and backpropagation. During forward propagation, the pointwise product of the activations in one layer with the weights connecting them to a neuron in the next layer are summed and passed through the nonlinear activation function of the latter neuron, creating the input to the subsequent layer. The input image propagates from all the layers to the output neuron layer. By computing the difference between the actual and the desired output vectors, the error of the network is calculated. The goal during training is to minimize the errors. The error is propagated from the output layer backwards to the input layer with a gradient-descent-based algorithm in the backpropagation phase. The synaptic weights of the network are updated afterwards. The learning rate parameter controls the size of the updates. The set of training images can be presented multiple times (*epochs*) to the network. During testing, the synaptic weights are fixed and the generalization accuracy of the network is measured as the percentage of correctly recognized digits using the test images. The network can reach a generalization accuracy of 98% on the MNIST dataset (see Fig. 5(b)).

An ANN of size 785x251x10 is used in the simulations. The last neuron of the input layer and the hidden layer are bias neurons. The synaptic weights $w \in [-1, 1]$ are randomly initialized in the interval $[-0.5, 0.5]$. A sigmoid function is

Fig. 5. (a) Artificial neural network (ANN) of size 785x251x10 used for the simulations. 1 bias neuron is added to both the input and the hidden layer. This simple network can reach reasonable test accuracies. (b) The ideal ANN simulator is trained using 20 epochs of all 60,000 training images and tested using all 10,000 test images from the MNIST set of handwritten digits. The network can classify more than 98% of the test images correctly. When trained using 5,000 training images, it achieves more than 94% generalization accuracy.

used as the activation function. For speed purposes, only the first 5,000 training images from the MNIST set are used for training, in 2 epochs. For testing, all 10,000 testing images from the set are shown to the network.

In our simulations, we model the synapse using two devices where one device represents a positive conductance G_+ and the other one represents a negative conductance G_- . The combined conductance of the synapse is calculated as $G = G_+ - G_-$ and re-scaled to represent the synaptic weight w . The maximum value of G_+ and G_- is $10 \mu S$ in accordance with the experimental results. During the forward propagation, the combined conductance G is read. During the weight update of the backpropagation phase, the conductance change ΔG is drawn from a Gaussian distribution with a fixed mean and a standard deviation. According to the weight change required by the network, multiple independent Gaussian random variables are added for a larger ΔG . ΔG is added to G_+ for potentiation and G_- for depression. If either G_+ or G_- is approaching its maximum value, the conductance value is reset and updated to keep the overall conductance G constant. A fixed learning rate is used for all simulations. Our PCM model differs from the characterization data in two aspects. First, each conductance change is modeled by a fixed mean value and does not capture the non-linear conductance behavior (see Fig. 6(a)). Secondly, the conductance changes are assumed to be independent. This assumption estimates a larger standard deviation than real device behavior (see Fig. 4).

First we investigate the impact of the finite-resolution conductance update on the network performance. Our experimental data shows that in PCM devices, the updates ΔG

Fig. 6. (a) The linear conductance model used in ANN simulations. (b) Generalization accuracy as a function of conductance step size. Simulations with no randomness (blue) and with a standard deviation equal to step size (red) are shown. Simulations are repeated 3 times and results are averaged. (c) Effect of randomness on conductance change. Mean and standard deviation of ΔG are equal on the black axis. Simulations are repeated 3 times and results are averaged.

are realized in finite steps and the conductance reaches a maximum value after a fixed number of pulses. The blue curve in Fig. 6(b) illustrates the impact of mean conductance step size on accuracy. Performance drops with increasing step size of the conductance update. In addition, we tested the effect of the presence of randomness in conductance updates that is observed in real devices, where the standard deviation of ΔG has similar magnitude as the mean. In Fig. 6(b), the comparison between $\sigma = 0$ (blue curve) and $\sigma = \mu$ (red curve) does not show any obvious effects introduced by the added variance of this size. To further understand the

effect of randomness in the conductance change, we also measure the performance of the network for varying degrees of randomness. The mean update $\mu(\Delta G)$ was kept fixed and we varied the standard deviation $\sigma(\Delta G)$ (see Fig. 6(c)). The network performance degrades significantly as the standard deviation to mean ratio is larger than 1.

IV. CONCLUSION

Cross-bar arrays comprising PCM devices have been proposed as a non-von Neumann computing platform to train ANNs efficiently. The synaptic weights are represented as the conductance states of PCM devices and the weight updates can be achieved by modifying these conductance values. In this paper, we presented a thorough experimental characterization of such conductance changes in 90 nm mushroom PCM cells with a specific focus on the granularity and randomness. We found that the standard deviation of the conductance change is comparable to the mean conductance change. An extensive study of the impact of this granularity and randomness on the performance of ANNs was conducted. This study showed that the classification accuracy degrades significantly with increased size of the mean conductance change. Moreover, when the standard deviation to mean ratio is larger than 1, the accuracy degrades further. This calls for innovations in device technology [10] or synaptic architectures that are more robust to these undesirable attributes [11]. It will be of interest to know how the conductance change granularity and stochasticity influences larger size networks and this will be the subject of future studies.

ACKNOWLEDGMENT

We thank N. Papandreou for technical assistance and S. R. Nandakumar and B. Rajendran for fruitful discussions. We acknowledge financial support from the Swiss National Science Foundation.

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," *IEEE Transactions on Electron Devices*, vol. 62, no. 11, pp. 3498–3507, 2015.
- [2] G. W. Burr *et al.*, "Neuromorphic computing using non-volatile memory," *Advances in Physics: X*, vol. 2, no. 1, pp. 89–124, 2017.
- [3] G. F. Close *et al.*, "Device, circuit and system-level analysis of noise in multi-bit phase-change memory," in *IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2010, pp. 29–5.
- [4] G. Burr *et al.*, "Recent progress in phase-change memory technology," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 6, no. 2, pp. 146–162, 2016.
- [5] A. Sebastian, M. Le Gallo, and D. Krebs, "Crystal growth within a phase change memory cell," *Nature Communications*, vol. 5, p. 4314, 2014.
- [6] M. Suri *et al.*, "Phase change memory as synapse for ultra-dense neuromorphic systems: Application to complex visual pattern extraction," in *Proc. IEEE International Electron Devices Meeting (IEDM)*, 2011, pp. 4.4.1–4.4.4.
- [7] A. Sebastian, D. Krebs, M. Le Gallo, H. Pozidis, and E. Eleftheriou, "A collective relaxation model for resistance drift in phase change memory cells," in *IEEE International Reliability Physics Symposium (IRPS)*. IEEE, 2015, pp. MY-5.
- [8] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, D. E. Rumelhart and J. L. McClelland, Eds. MIT Press, Cambridge, MA, 1986, vol. 1: Foundations, pp. 318–362.
- [9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, pp. 227–2324, 1998.
- [10] W. W. Koelmans *et al.*, "Projected phase-change memory devices," *Nature Communications*, vol. 6, p. 8181, 2015.
- [11] J. Bill and R. Legenstein, "A compound memristive synapse model for statistical learning through stdp in spiking neural networks," *Frontiers in Neuroscience*, vol. 8, p. 412, 2014.