Reliability Perspective of Resistive Synaptic Devices on the Neuromorphic System Performance

Pai-Yu Chen, and Shimeng Yu*
School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, USA, *Email: shimengy@asu.edu

***Abstract*—Emerging non-volatile memory (eNVM) based synaptic devices are attractive for the replacement of SRAM in the hardware implementation of artificial neural networks (ANNs). However, one of the critical challenges for eNVM is the reliability concerns due to data retention and write endurance failures. This paper investigates the impact of these two failures in the multilayer perceptron (MLP) using our developed NeuroSim+ simulator. For the retention failure in offline classification, we consider various possible conductance drift scenarios and the reported physical model based on conductance variation. The results confirm that faster degradation on the classification accuracy is highly correlated with larger deviation in the weighted sum. For the endurance failure in online learning, the strength of conductance tuning is assumed to become weaker over write pulse cycles. The analysis suggests that the learning accuracy is less impacted because the network is able to adapt itself and activate more synapses to participate in the weight update when the tuning capability of synapses are degraded.**

***Index Terms*—Emerging non-volatile memory, endurance, artificial neural network, reliability, retention, synaptic devices**

I. Introduction

Neuromorphic computing based on artificial neural networks (ANNs) has attracted considerable attention owing to its great success in various intelligence applications such as speech and image recognition. Traditional implementation of ANN relies on CPUs/GPUs and/or FPGAs to speed up matrix operations by making effective use of their parallel processing capabilities. However, these platforms are still inadequate for real-time/low-power training with large-scale dataset that poses a high requirement on the computation and memory bandwidth. In recent years, several custom CMOS ASIC hardware accelerators have been developed (e.g. MIT's Eyeriss [1]) to further improve the computation and power efficiency, where SRAM is used to implement the synapses. But SRAM is area inefficient (with cell size $100F^2$~$200F^2$, F is the lithography feature size) thus part of the weights may have to be stored off-chip (i.e. in DRAM), introducing the bottleneck of off-chip memory access. To replace SRAM, emerging non-volatile memory (eNVM) based resistive synaptic devices, such as resistive random access memory (RRAM) [2-4] and phase change memory (PCM) [5, 6], are considered as promising candidates due to their compact device structure (with cell size $4F^2$~$12F^2$) and the ability to store "analog" weight in multi-level conductance states. At architecture level, the entire weight matrix is represented by a resistive synaptic array with crossbar structure (Fig. 1(b)) that enables the weighted sum (matrix-vector multiplication) to be performed in a parallel fashion.

Despite that the shift from digital to analog computing domain offers a significant improvement in the area, power and computation speed, synaptic devices usually suffer from non-ideal device effects, including nonlinear and noisy conductance tuning, limited precision and finite ON/OFF ratio, etc. Degradation of learning accuracy associated with these properties has been analyzed thoroughly using NeuroSim+ simulator in our prior work [7], but the reliability issues such as data retention and write endurance are unexplored. In fact, the reliability soft errors in HfO_x based RRAM caused by its stochastic nature of oxygen vacancies have been reported to be harmful to the learning performance in a winner-take-all ANN [8]. Degradation of learning accuracy is also observed with the retention-induced conductance variation in HfO_x based analog RRAM with a thermal enhanced layer [9]. Therefore, it is crucial to perform a comprehensive analysis of the reliability issues on the learning performance of ANN. In this work, we aim at investigating the impact of data retention and write endurance with generic assumptions of all possible failure mechanisms. The retention model presented in [9] will also be taken into account and its impact will be re-evaluated with our NeuroSim+ simulation framework.

II. Neruosim+ Simulation Framework

To study the feasibility of synaptic devices as analog weights on ANN, we have developed a simulation framework named NeuroSim+ for a 2-layer multilayer perceptron (MLP) NN with synaptic device properties incorporated into the weights [7]. As shown in Fig. 1(a), we use MNIST handwritten digits [10] as the training and testing dataset to implement online learning and offline classification. The MLP network topology is 400(input layer)-100(hidden layer)-10(output layer). 400 neurons of input layer correspond to 20×20 MNIST image (converted to black/white and edge cropped), and 10 neurons of output layer correspond to 10 classes of digits. Such a simple 2-layer MLP can achieve 96~97% in the software baseline.

The simulator can emulate hardware by mapping the weight matrixes to resistive synaptic arrays, as shown in Fig. 1(b). In this work, each synaptic device model has the conductance (G) incremental tuning as well as the retention and endurance properties. It should be noted that the synaptic devices can only represent positive weights, thus a mapping from the algorithm's weight (-1~1) to device's weight (0~1) is required. In neuron peripheral circuits, the array's weighted sum result will be mapped back to the algorithm's weighted sum result by subtracting the sum of input vector elements. For the learning modes, the simulator can perform online learning and offline

This work is supported in part by NSF-CCF-1552687, NSF-CCF-1715443, NSF/SRC E2CDA, and ASCENT (one of the six SRC/DARPA JUMP Centers).

978-1-5386-5479-8/18/$31.00 ©2018 IEEE

classification. In online learning, the MLP simulator takes into account the synaptic device properties in training the network with images randomly picked from the training dataset (60k images) and classifying the testing dataset (10k images). In offline classification, the network is pre-trained by software, and the MLP simulator only performs the inference with synaptic device properties.

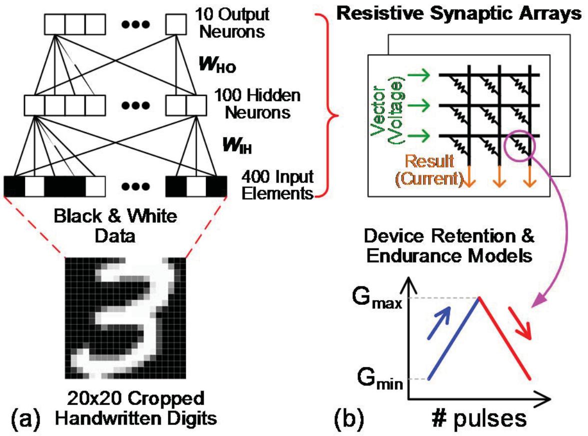

Figure 1. (a) The 2-layer multilayer perceptron (MLP) neural network (NN). The input MNIST images are cropped and encoded into black/white data for simplification. (b) In the simulator, the weights W_{IH} and W_{HO} are implemented with resistive synaptic arrays, where each synaptic device model includes the linear conductance (G) tuning with number of pulses as well as the retention and endurance properties.

III. Reliability Analysis

For memory application, the data retention and write endurance are the key metrics for the reliability evaluation of eNVM. In this section, we incorporate the retention and endurance models into the developed simulator to study these two issues. Since the emphasis is on the reliability, we set the synaptic weight to be 6-bit (64 levels) and assumes linear conductance tuning without variation in all the simulations.

A. Data Retention

Data retention refers to the ability of memory device to retain its programmed state over a long period of time. Typical retention specification for NVM in memory application is more than 10 years at 85ºC. Many binary eNVM devices have been able to meet this requirement. However, there are no reported data for analog eNVM that shows such retention, which can be attributed to the instability of intermediate conductance states [9]. To be general, we consider four scenarios of conductance drift for the retention analysis. As shown in Fig. 2(a)-(c), the conductance can either drift toward its maximum, minimum or intermediate states. These three scenarios have ever been reported in the retention measurement of binary eNVMs [3, 11, 12]. In addition, we also consider random conductance drift towards its maximum or minimum state with equal probability, as shown in Fig. 2(d). The formula for modeling the conductance drift behavior is assumed to follow the one that is widely used in PCM [6, 13], which can be described as

$$G=G_0\left(\frac{t}{t_0}\right)^{v} \qquad (1)$$

where G_0 is the initial conductance, t is the retention time, v is the drift coefficient and t_0 is the time constant which is assumed to be 1 second in this work. In the retention analyses, the offline classification is used with the conductance ON/OFF ratio set to be 50, which is a sufficiently large ratio, in order to still capture the conductance drift at the lowest conductance state.

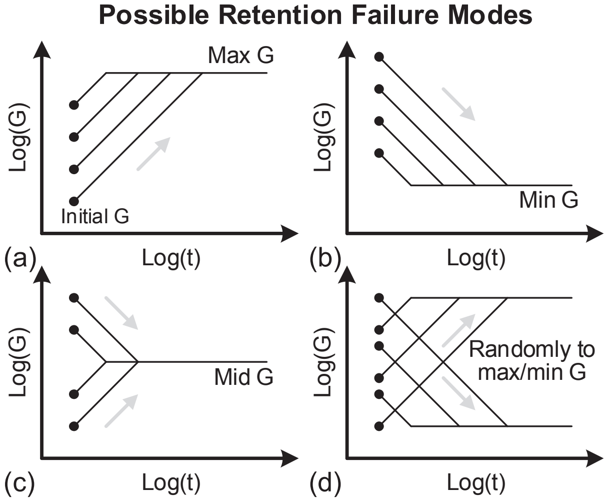

Figure 2. General assumptions of retention failure modes: conductance drifting towards its (a) maximum state, (b) minimum state, (c) intermediate state, or (d) maximum/minimum state with randomness.

Fig. 3(a) shows the degradation of classification accuracy over retention time at a fixed drift coefficient of 0.01 with different final weight states that the conductance drifts to. It can be simply calculated that the conductance change is ~20% over 10 years under such drift coefficient, and it leads to degradation of accuracy <90% for all final weight states. On the other hand, the result suggests that the final state either be at the maximum or minimum conductance has the poorest accuracy. To have a quantitative comparison between different final weight states, we measure the maximum drift coefficient of all states that still give an accuracy >90% at a retention time of 10 years. As shown in Fig. 3(b), the final weight at 0.6 can tolerate up to a maximum drift coefficient of ~0.012, which corresponds to ~25% of the conductance change at 10 years.

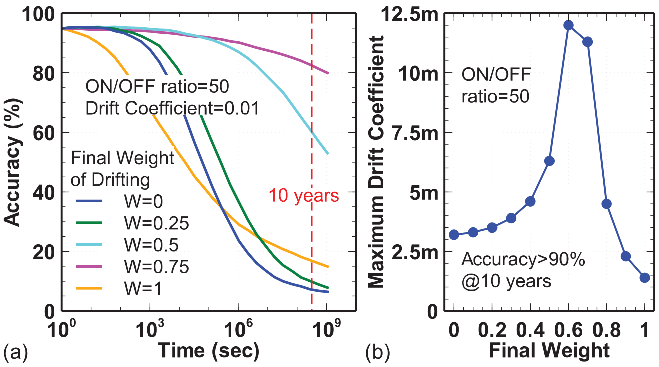

Figure 3. (a) Classification accuracy as a function of retention time with conductance drifting toward different final weight states. (b) The maximum drift coefficient as a function of final weights for achieving >90% accuracy at 10 years.

The reason why intermediate final weight states (Fig. 2(c)) have less accuracy degradation than either the maximum or minimum ones (Fig. 2(a)-(b)) can be largely attributed to the deviation of weighted sum after retention degradation. This can be simply observed from the distribution of the absolute difference of column conductance sum before and after retention degradation, as shown in Fig. 4 for the first and second layer of MLP NN. The difference (ΔW) is measured between the array conductance patterns before and after a retention of 10 years, and a small drift coefficient of 0.001 is used to ensure that most of the conductance have not reached their final states at 10 years. As all the conductance will drift in the same direction to the maximum or minimum final weight state, a larger deviation of weighted sum is expected, and the high inverse correlation between Fig. 4 and Fig. 3(b) confirms that the accuracy degradation is strongly affected by the amount of weighted sum deviation.

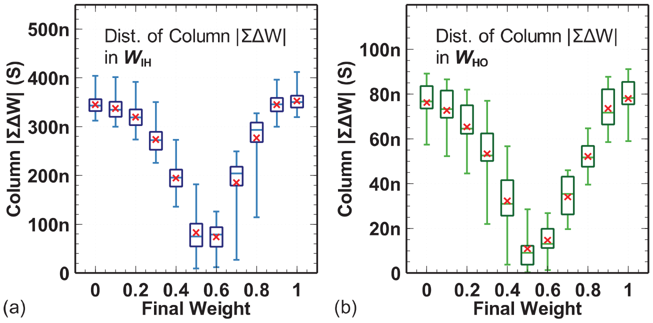

Figure 4. Distribution of the absolute difference of column conductance sum before and after 10 years (drift coefficient=0.001) in the (a) first and (b) second layer of MLP NN. Both results are highly correlated with Fig. 3(b).

The above argument can be further substantiated by the analysis of random conductance drift in Fig. 2(d), where its impact on the classification accuracy is shown in Fig. 5. With the same drift coefficient of 0.01, the accuracy degradation is much less severe than the ones in other drift scenarios (Fig. 3(a)), even we select the worst result in Fig. 5 for comparison. The reason is because the weighted sum deviation will be averaged out by this randomness. It can be expected that if either drifting towards maximum or minimum conductance is much more probable, the accuracy degradation will be as severe as that of W=0 or W=1 in Fig. 3(a).

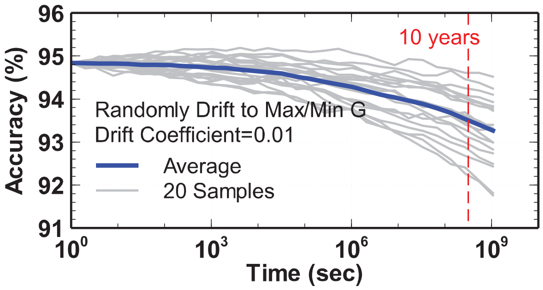

Figure 5. Monte carlo simulation on the accuracy with conductance randomly drifting toward its maximum or minimum states. Under the same drift coefficient, the randomness behavior does not lead to radical change in weighted sum thus the impact on the accuracy is much smaller compared to other conductance drift senarios.

In fact, the only experimental work so far that reported the retention properties in analog RRAM suggests that its behavior can be due to multiple hops of oxygen vacancies over long retention time [9], which is analogous to Brownian Motion. It also shows that the read current distribution of each conductance level follows a normal distribution, where its standard deviation (σ) increases with retention time. In other words, the retention behavior can be modeled as an increasing conductance variation over time, which is illustrated in Fig. 6(a). From [9], its σ is described as

$$\sigma=\lambda\sqrt{t}+\theta \qquad (2)$$

where λ and θ are fitting parameters. Since these fitting parameters can vary in different devices, conductance states and even temperatures, we rather evaluate the impact of this retention behavior based on σ. As shown in Fig. 6(b), a σ of ~0.2 will lead to a significant degradation on the accuracy. It can be calculated that given θ=0, λ should be smaller than ~7e-6 for the accuracy to remain >90% at 10 years.

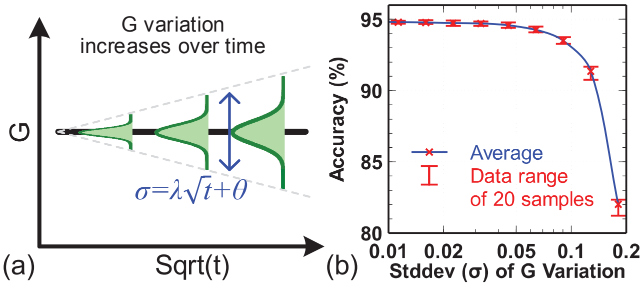

Figure 6. (a) The retention model proposed in [9] suggesting the an increasing conductance variation over time. (b) The impact of conductance variation on the classification accuracy.

B. Write Endurance

In memory application, the write endurance specifies the number of times that a memory device can be programmed (written) before the write failure occurs. Typical binary eNVM devices can achieve >10^6 write cycles (between the highest and lowest conductance states). However, the analog eNVM endurance definition should be different as it has only incremental conductance change by each write pulse. So far, there is no prior work discussing the endurance behavior of analog eNVM for neuromorphic computing. To study the endurance effect in this work, we assume that the strength of conductance tuning (ΔG) decreases over write pulse cycles, which is expressed as

$$\Delta G=\Delta G_0(1\text{-}r)^{(\#pulses)} \qquad (3)$$

where ΔG_0 is the ideal conductance change without considering endurance degradation, r is the reduction ratio, #pulses means the cumulative number of pulses that has been applied to the device. As illustrated in Fig. 7(a), the conductance will be eventually unchangeable after an excessive number of cycles. To analyze its impact, we apply the endurance property in the online learning of the MLP NN. As shown in Fig. 7(b), the learning accuracy degradation begins to be noticeable as we gradually increase r to be >0.01. We also apply variations of 10% and 20% on the ratio, and it does not really either significantly alleviate or worsen the degradation.

In the endurance analysis, we assume the maximum conductance of the device is 100 nS. It can be calculated that the required cumulative number of pulses to reduce the strength of conductance tuning by 50% and 90% are ~70 and ~230 under r=0.01, respectively. Fig. 8(a)-(b) shows the distribution of the sum of absolute conductance change in the first and second layer of MLP NN without endurance effect to achieve the targeted learning accuracy. The conductance changes with 70 and 230 write pulses are also labeled. Given only the results of Fig. 8(a)-(b), we may easily conclude that r=0.01 is too large thus there will be a significant accuracy degradation, because most of the devices require far more pulses than these two numbers to achieve >90% accuracy. However, the accuracy with r=0.01 in Fig. 7(b) disproves this argument.

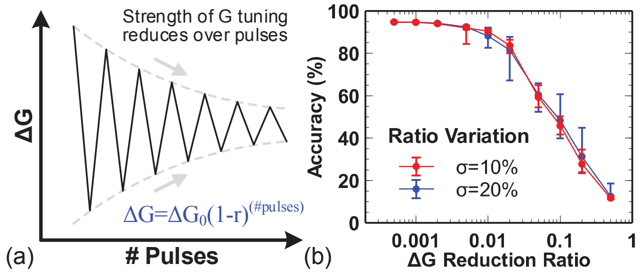

Figure 7. (a) Endurance degradation in weight update of synaptic devices. Strength of conductance tuning decreases over pulse cycles. (b) The impact of ΔG reduction ratio (with 10% and 20% variation) on the learning accuracy. 10 device samples are measured for each data point.

In fact, the network has the ability to adapt itself to this endurance degradation by relying on other devices whose conductance is still tunable. As shown in Fig. 8(c)-(d), the conductance cannot be further tuned beyond a certain amount of total conductance change (~150 nS), and the network will keep activating other inactive devices to take over the responsibility of learning during the entire learning process. Besides, analog eNVM devices with $>10^3$ write pulses of conductance tuning were also demonstrated [2, 5]. Therefore, the endurance issue may not be as critical as estimated.

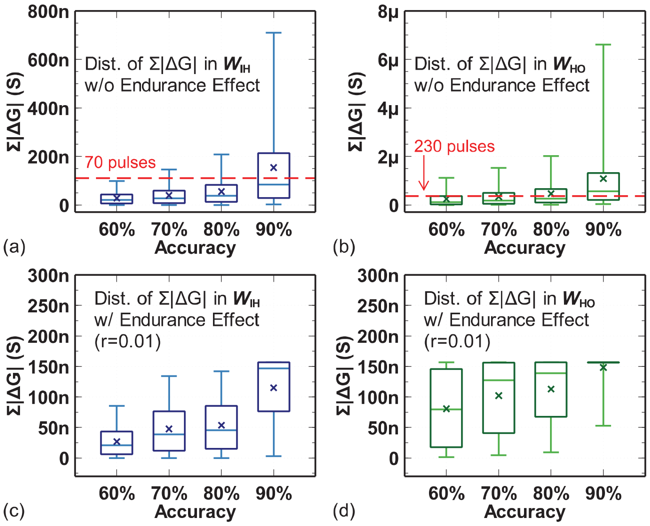

Figure 8. Distribution of the sum of absolute conductance change in the (a) first and (b) second layer without endurance effect, and (c) first and (d) second layer of MLP NN with endurance effect (r=0.01). The network can adapt itself to this endurance degradation by activating other synaptic devices whose conductance are still tunable.

IV. Conclusion

Data retention and write endurance are important reliability properties of synaptic devices. We have investigated the impact of these two properties on a 2-layer MLP NN using our developed NeuroSim+ simulation framework. It is observed that there is a strong correlation between the degradation of offline classification accuracy and the weighted sum deviation, thus retention behaviors which causes less deviation will have smaller impact on the accuracy. The analysis also includes the existing retention model based on conductance variation, enabling estimation of the model parameters based on targeted performance. In contrast, the endurance issue defined in this work is considered to be less critical than estimated because the network is able to alleviate it during online learning by making use of other devices whose conductance are still tunable.

V. Acknowlegment

The authors acknowledge Dr. Bin Gao from Tsinghua University for helpful discussions in this work.

VI. Reference

[1] Y.-H. Chen, et al., "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 262-263, 2016.

[2] L. Gao, et al., "Fully parallel write/read in resistive synaptic array for accelerating on-chip learning," *Nanotechnology,* vol. 26, no. 45, pp. 455204, 2015.

[3] S. H. Jo, et al., "Si memristive devices applied to memory and neuromorphic circuits," *IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 13-16, 2010.

[4] M. Prezioso, et al., "Training and operation of an integrated neuromorphic network based on metal-oxide memristors," *Nature,* vol. 521, no. 7550, pp. 61-64, 2015.

[5] D. Kuzum, et al., "Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing," *Nano letters,* vol. 12, no. 5, pp. 2179-2186, 2011.

[6] S. Kim, et al., "Resistance and threshold switching voltage drift behavior in phase-change memory and their temperature dependence at microsecond time scales studied using a micro-thermal stage," *IEEE Transactions on Electron Devices,* vol. 58, no. 3, pp. 584-592, 2011.

[7] P.-Y. Chen, et al., "NeuroSim+: an integrated device-to-algorithm framework for benchmarking synaptic devices and array architectures," *IEEE International Electron Devices Meeting (IEDM)*, pp. 135-138, 2017.

[8] A. M. Tosson, et al., "A Study of the Effect of RRAM Reliability Soft Errors on the Performance of RRAM-Based Neuromorphic Systems," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* vol. 25, no. 11, pp. 3125-3137, 2017.

[9] M. Zhao, et al., "Investigation of statistical retention of filamentary analog RRAM for neuromophic computing," *IEEE International Electron Devices Meeting (IEDM)*, pp. 872-875, 2017.

[10] MNIST handwritten digits, http://yann.lecun.com/exdb/mnist/

[11] Y. Y. Chen, et al., "Improvement of data retention in HfO_2/Hf 1T1R RRAM cell under low operating current," *IEEE International Electron Devices Meeting (IEDM)*, pp. 252-255, 2013.

[12] A. Prakash, et al., "TaO_x-based resistive switching memories: prospective and challenges," *Nanoscale research letters,* vol. 8, no. 1, pp. 418, 2013.

[13] R. A. Cobley, et al., "A model for multilevel phase-change memories incorporating resistance drift effects," *IEEE Journal of the Electron Devices Society,* vol. 3, no. 1, pp. 15-23, 2015.