Characterizing Endurance Degradation of Incremental Switching in Analog RRAM for Neuromorphic Systems

Meiran Zhao[1], Huaqiang Wu[1]*, Bin Gao[1], Xiaoyu Sun[2], Yuyi Liu[1], Peng Yao[1], Yue Xi[1], Xinyi Li[1], Qingtian Zhang[1], Kanwen Wang[3], Shimeng Yu[2], and He Qian[1]

[1]Institute of Microelectronics, Tsinghua University, Beijing 100084, China;
[2]Georgia Institute of Technology, Atlanta, GA 30332, USA; [3]Huawei Technologies CO., LTD.
Email: wuhq@tsinghua.edu.cn; gaob1@tsinghua.edu.cn;

***Abstract*—**Resistive random access memory (RRAM) is attractive for neuromorphic computing systems as synaptic weights. In the neural network training, incremental switching occurs between the analog conductance states, thus the analog RRAM devices have unique endurance degradation behaviors compared to the convention digital memory application. In this work, a fast measurement platform is developed to characterize the endurance of incremental switching in analog RRAM. It is found that under weak weight update pulses, the incremental switching cycles of RRAM can be increased for more than 5 orders of magnitude compared with full window switching under strong programming pulses. The 10^{11}-cycle endurance of analog RRAM is proved to be sufficient for training neural networks online for various datasets (from MNIST to ImageNet). However, the nonlinearity and dynamic range of analog RRAM degrade during cycling, which may influence the learning accuracy of the neural network when it re-trains with new datasets.

I. Introduction

To train a RRAM based neural network online, the conductance of RRAM, which represents synaptic weight (Fig. 1), is required to be tunable for a large number of updates, e.g. 10^5 for MNIST, and 10^7 for ImageNet (Table I) [1-4]. However, it is well known that RRAM for data storage application has a limited endurance, typically ranged between 10^5 to 10^7 [5]. It seems that such endurance cannot support online training of a neural network. On the other hand, it should be noted that the weight update of analog RRAM for neural network training is quite different from the write operation for memory application. For memory write operation, a large resistance window is typically achieved by strong programming pulses. Endurance failure occurs when the resistance window collapses (Fig. 2). Whereas, for neural network training, the conductance of analog RRAM only changes incrementally by a weak programming pulse (Fig. 3), which is called incremental switching or analog switching [6]. In this case, the damage of each pulse on the material properties is much less significant than the memory write, thus the lifetime is expected to be much longer. To our best knowledge, no prior work has studied the endurance of analog switching. It is not clear how to characterize the unique behaviors of endurance degradation on analog switching, and how much difference between the endurance of digital memory switching and analog switching.

There are two reasons why characterizing endurance of analog RRAM is challenging: 1) It is difficult to control the analog switching within a designated resistance window during repeated cycling, and it lacks of an effective evaluation criterion for endurance degradation; 2) Due to the large variability of RRAM, endurance test on single device is insufficient, but statistical measurement requires a long time. In this work, we developed a testing platform to characterize the endurance behaviors of analog switching. We found that the endurance degradation behaviors on analog switching are different from the memory switching. The impact of endurance degradation on neuromorphic system performance is also discussed.

II. Endurance Characterization Method

We fabricated 1T1R array where the RRAM cells are integrated on top of a transistor's drain contact. The transistor offers accurate control of various resistance states by tuning gate voltage during SET/RESET operations, which makes the characterization on analog switching possible. We also developed a fast and effective test platform for testing the 1T1R cell with Tektronix 4200 system. The test scheme of analog switching endurance is shown in Fig. 4. The main difference to the traditional endurance test scheme lies in setting the upper and lower boundaries of each state with variable voltage pulses.

The analog RRAM cell under test is TiN/ETML/HfO_x/TiN stack. HfO_x is switching layer, and electro-thermal-modulation-layer (ETML) contributes to excellent linear analog switching behaviors [7]. The devices are demonstrated good uniformity of analog switching from cycle to cycle and device to device (Fig. 5), which is important for studying endurance behaviors.

III. Endurance of Analog Switching

Firstly, we investigate how switching window influences endurance cycle. Three typical windows with different resistance levels show different endurance cycles (Fig. 6). The full window (high resistance state, HRS~1MΩ, low resistance state, LRS~20kΩ) shows the worst endurance. Most of the cells fail after 10^5 cycling. The high-R window case (HRS~1MΩ, LRS~100kΩ) also shows poor endurance, typically 10^6 cycles. The low-R window (HRS~100kΩ, LRS~20kΩ) shows much better endurance, greater than 10^8 cycles. When fixing low resistance state (LRS) around 20kΩ and changing high resistance state (HRS) from 1MΩ to 100kΩ, endurance shows improvement from 10^5 to 10^9 (Fig. 7). The endurance cycle increases significantly as the window reduces. However, when fixing HRS around 1MΩ or 200kΩ, and changing LRS from 20kΩ to 100kΩ, the endurance cycle changes less than 10 times (Fig. 8 & 9). Based on the statistical measurement for different

978-1-7281-1987-8/18/$31.00 ©2018 IEEE

switching windows (Fig. 10), it is concluded that higher resistance state influences endurance more significantly. This is attributed to the change of current mechanisms correlated with different morphology of oxygen vacancy (Vo) based conductive filament (CF) (Fig. 11). To get better endurance, the analog RRAM should be controlled within the relatively lower resistance range during training process. When the device fails, the resistance may be stuck randomly at LRS, HRS, or intermediate state (Fig. 12).

The above test is still similar as memory switching. To mimic the online training process in a neuromorphic system, the switching window is controlled within a very small range using weak programming pulse (update pulse), and the incremental conductance change after each pulse is recorded. With this test method, the analog switching does not fail even when >10^{11} update pulses are applied (Fig. 13). Here, we use update number instead of cycle number to distinguish this type of endurance test from conventional memory endurance test. It should be noticed that we stop at 10^{11} only because of the measurement time limit. 10^{11} is not the limit of update number.

Although the device does not fail, the analog switching performance degrades as update number increases. Different from memory switching, analog switching is usually evaluated with some new metrics, such as asymmetry/nonlinearity in weight update, dynamic range, variability, etc [8, 9]. We measure the analog switching in a full dynamic range after certain update numbers (Fig. 14), and we see that the dynamic range of analog switching decreases after 10^6 update pulses (Fig. 15). The nonlinearity of analog switching also becomes worse after 10^7 update pulses. The conductance change cannot keep a constant value throughout the dynamic range for both SET and RESET process (Fig. 16). In this case, although the neural network can still be trained, the learning accuracy will sacrifice.

Physical mechanism of the endurance degradation in analog switching is speculated (Fig. 17). Initially, multiple weak CFs are formed, contributing to the good linear analog switching [10]. When switching in low-R window, CF gap does not form, which is confirmed by the Ohmic current at resistance levels lower than 500kΩ (Fig. 11). In this case, electric field distributes uniformly in the switching layer. When switching in high-R/full window, CF gap forms, and large electric field concentrates at the gap region [10]. The large electric field damages the RRAM material properties seriously, resulting in poor endurance. After cycling, the multiple weak CFs morphology gradually transforms to a strong CF morphology, thus the nonlinearity and dynamic range degrade.

IV. Impact on Neural Network

A two-layer fully connected neural network for the MNIST dataset learning is used to investigate the impact of endurance degradation of analog switching (Fig. 18). The nonlinearity, dynamic range, endurance degradation, as well as variations, are all taken into consideration in the device behavioral model based on the measured results. When only considering the nonlinearity degradation, the learning accuracy degrades gradually by >6% at 10^9 update number (Fig. 19a). And when only considering the nonlinearity degradation, the learning accuracy remains unchanged until 10^7 update number, and then begins to degrade rapidly by >15% at 10^9 update number (Fig. 19b). When considering both nonlinearity and dynamic range degradation, the accuracy starts to decrease at 10^6 update number, but the degradation degree is similar (Fig. 19c). If we map the synaptic weight to different resistance window, by considering the endurance failure, the accuracy degradation behaviors are different (Fig. 20). It is found that the low-R window (10uS~50uS) is the best condition for online training thanks to its highest accuracy and longest endurance lifetime. Meanwhile, even though some of the cells in the network fail, the neural network can still keep a reasonable accuracy with different weight patterns (Fig. 21). This is due to the redundancy and error tolerance of neural network.

Here we define a new parameter, accumulated ΔG, to characterize the endurance of analog switching. Accumulated ΔG means the sum of the absolute value of conductance change under each update pulse. Assuming the dynamic range of RRAM is 5uS~50uS, the accumulated ΔG for a complete training of the MNIST dataset distributes around 250mS (Fig. 22). Although variation exists, most of the RRAM cells in the network show similar accumulated ΔG with this algorithm. Then we calculated the measured accumulated ΔG before obvious endurance degradation, which is around 6kS. With comparison of device measurements and network simulations (Fig. 23), it is proved that the analog RRAM can meet the requirement of online training. If we need to re-train the network for different datasets, the RRAM based network can re-train for >10^4 times. Even for a more complex dataset such as ImageNet shown in Table I, we can conclude that such device can also support re-training for many times.

V. Conclusion

For the first time, we investigated the endurance behaviors of incremental switching in analog RRAM for neuromorphic computing. Key achievements include: 1) We developed a test platform to characterize the endurance of analog switching; 2) We demonstrated the endurance of analog RRAM is sufficient for online training; 3) We found the nonlinearity and dynamic range of analog switching degrade with cycling, and the endurance cycle is dependent on the switching window. This work provides valuable guidelines for evaluation and optimization of RRAM based neuromorphic systems.

Acknowledgment

This work is supported in part by the MOST of China (2016YFA0201801), Beijing Innovation Center for Future Chips (ICFC), Beijing Municipal Science and Technology Project (D161100001716002, Z181100003218001), and NSFC (61674087, 61674089, 61674092, 61076115).

References

[1] A. Conneau *et al.*, *EACL* 2017, 1107.[2] Y. Li, *arXiv* 2017, 1701.07274.[3] K. Simonyan *et al.*, *Computer Science,* 2014. [4] D.-A. Clevert *et al.*, *Computer Science,* 2015.[5] G. Sassine *et al.*, *IRPS* 2018, P-MY.2-1. [6] P. Y. Chen *et al.*, *IRPS* 2018, 5C.4-1. [7] W. Wu *et al.*, *VLSI* 2018,103. [8] H. Wu *et al.*, *IEDM* 2017, 11.5.1. [9] H. Hwang *et al.*, *EDL* 2016. [10] M. Zhao *et al.*, *IEDM* 2017, 39.4.1.

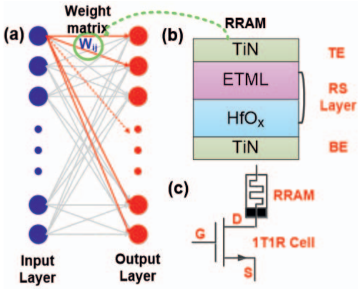


Fig. 1. Schematic diagram of (a) neural network; (b)RRAM; (c)1T1R cell. The conductance of RRAM defines weight in the network.

Task	DataSet	Update Number
Image Identification	MNIST	10^5
	CIFAR	10^5
	ImageNet	10^6-10^7
Natural language processing	-	10^7
Reinforcement learning	-	>10^8

Table. I. Typical requirement of weight update numbers for online training of represenative datasets.

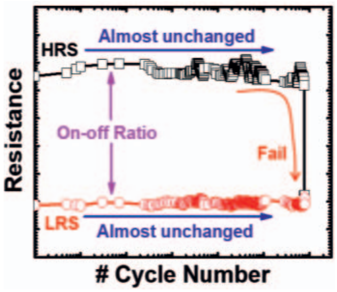


Fig. 2. Typical endurance failure behavior of RRAM for memory application. Resistance window closes after many cycles.

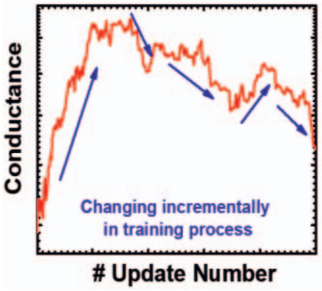


Fig. 3. Typical conductance change trace of analog RRAM during weight update process in a neural network.

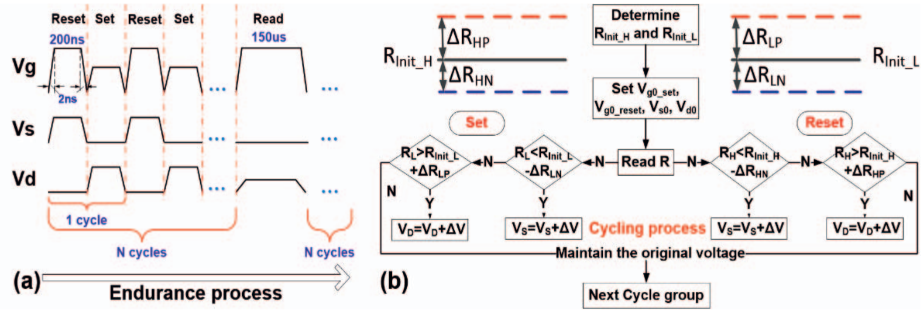


Fig. 4. Illustration of the analog switching endurance test process. (a) The waveform scheme and pulse condition of fast continuous cycling process; (b) The flow chart of determining the resistance window with verification opreation.

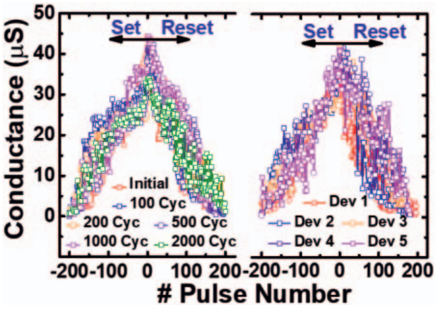


Fig. 5. The uniformity of analog switching behaviors of C2C and D2D. Pulse condition: width=50ns, Vset=1.4V, Vreset=-1.45V.

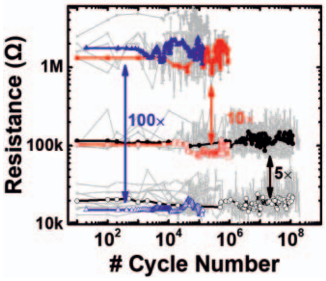


Fig. 6. Endurance failure bebavior of three windows. Grey lines are raw data. Colored lines are mean value.

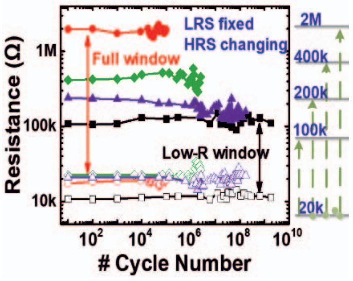


Fig. 7. Endurance failure behaviors with fixed LRS and various HRS (100k to 1MΩ). Measurement stops when the device fails.

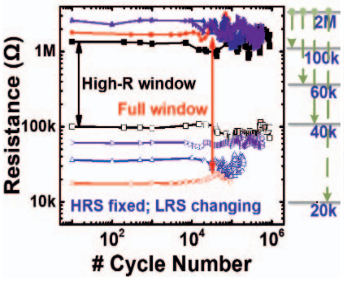


Fig. 8. Endurance failure behaviors with fixed HRS (1MΩ) and various LRS (20, 40, 60, 80 and 100kΩ).

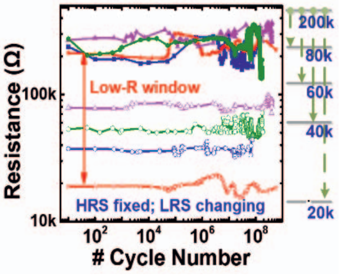


Fig. 9. Endurance failure behaviors with fixed HRS (200kΩ) and various LRS (20, 40, 60, and 80kΩ).

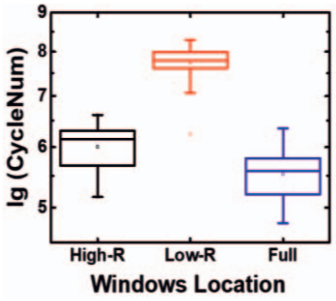


Fig. 10. Distribution of endurance lifetime for different switching windows.

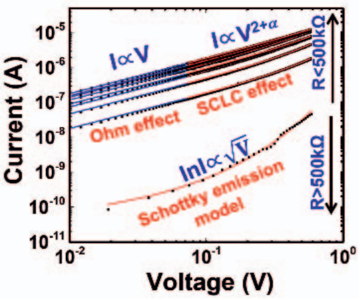


Fig. 11. I-V fitting for different resistance levels. Ohmic current is observed when R<500kΩ.

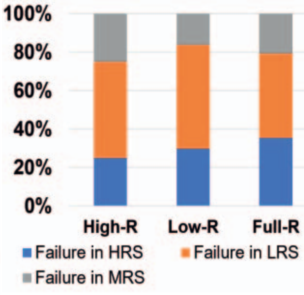


Fig. 12. Statistics of the stuck resistance levels when endurance failure occurs.

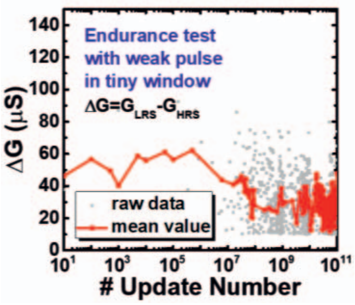
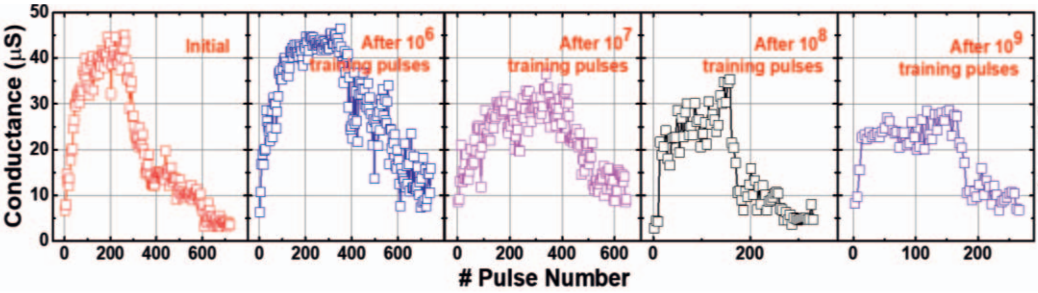


Fig. 13. Measured incremental conductance change within a small window during weak pulse cycling.



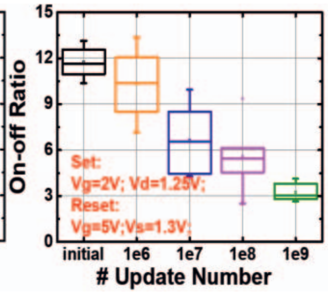
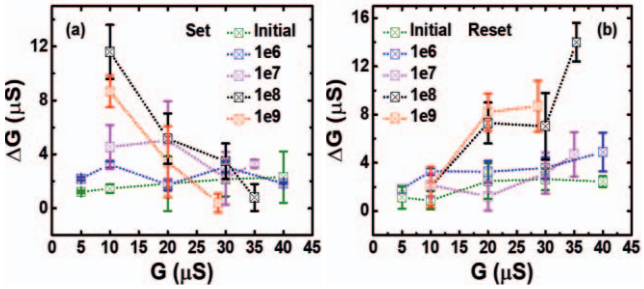


Fig. 14. Analog switching of full dynamic range with identical weight update pulses. The same RRAM device is measured after different numbers of update pulses, including initial, 10^6, 10^7, 10^8, and 10^9. Dynamic on/off ratio and nonlinearity of analog switching degrade as pulse number increasing.

Fig. 15. Distribution of dynamic on/off ratio of analog switching after different numbers of update pulse.



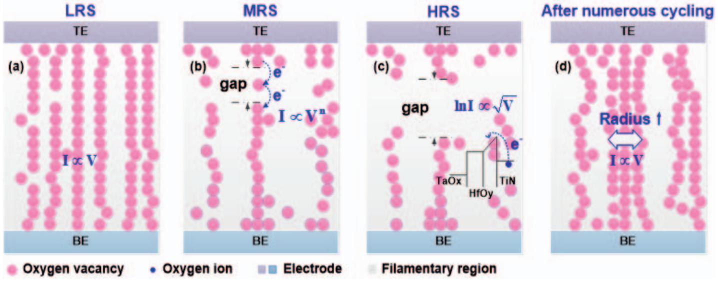
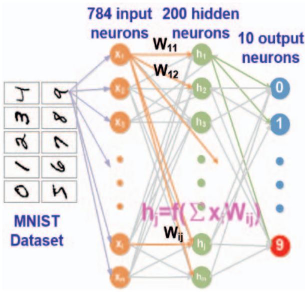
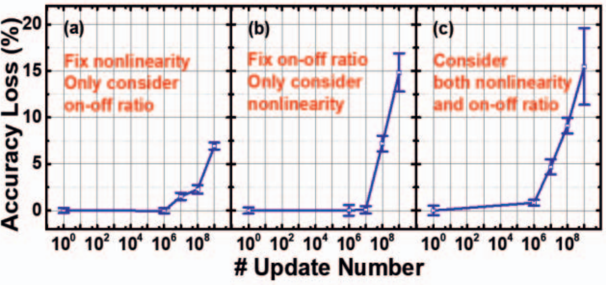


Fig. 16. Conductance change with one update pulse as a function of conductance level during (a) SET process and (b) RESET process. Closer to 0 means linear update, larger vaule means nonlinear update.

Fig. 17. Schematic of physical mechanism of endurance degradation for analog switching with different resistance windows.





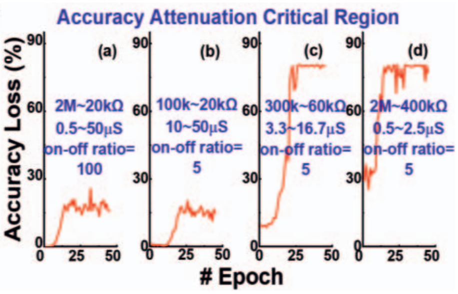
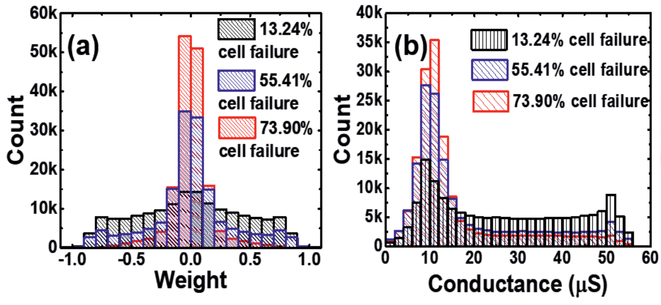
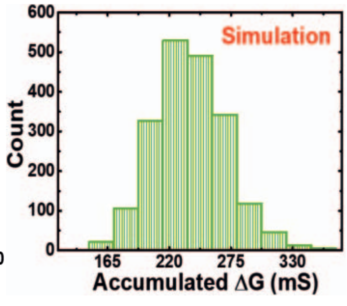


Fig. 18. Schematic of a two-layer fully-connected neural network for online learning and classifying MNIST dataset.

Fig. 19. Learning accuracy loss as a function of update number when only considering (a) on/off ratio; (b) nonlinearity and (c) both nonlinearity and on/off ratio.

Fig. 20. Learning accuracy loss with the degradation of device endurance. Different resistance ranges for weight mapping are compared.





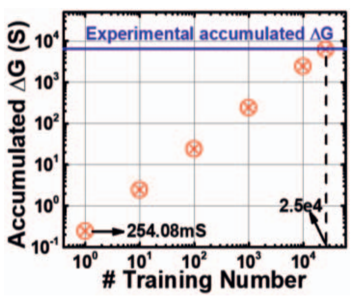


Fig. 21. Distribution of (a) algorithm weight and (b) RRAM conductance in the neural network after online training with different cell failure rate. Here, cell failure is only represented as resistance stuck at LRS, but the nonlinearity and ratio do not change. There is no obvious accuracy loss.

Fig. 22. Distribution of different RRAM's accumulated ΔG in the neural network after online training.

Fig. 23. Comparion of measured accumulated ΔG and simulated accumulated ΔG with different numbers of re-training of new datasets.