MLC PCM Techniques to Improve Nerual Network Inference Retention Time by 10^5X and Reduce Accuracy Degradation by 10.8X

W. S. Khwa, K. Akarvardar*, Y. S. Chen, Y. C. Chiu, J. C. Liu, J. J. Wu, H. Y. Lee, S. M. Yu, C. H. Lee, T. C. Chen, Y. C. Lin, C. F. Hsu, T. Y. Lee, T. K. Ku, C. H. Kuo, J. Y. Wu, X. Y. Bao*, C. S. Chang, Y. D. Chih, H.-S. P. Wong*, M. F. Chang
Taiwan Semiconductor Manufacturing Company Ltd., Corporate Research, Hsinchu, Taiwan, R.O.C.
* Taiwan Semiconductor Manufacturing Company Ltd., Corporate Research, San Jose, CA, USA
Tel: 886-3-5636688 Ext: 7223792, email: wskhwa@tsmc.com

Abstract **—**We present three novel MLC PCM techniques – (1) device requirement balancing, (2) prediction-based MSB-biased referencing, and (3) bit-prioritized placement to address the MLC device challenges in neural network applications. Using measured MLC bit error rates, the proposed techniques can improve the MLC PCM retention time by 10^5 times while keeping the ResNet-20 inference accuracy degradation within 3% and reduce the accuracy degradation by 91% (10.8X) for CIFAR-100 dataset in the presence of temporal resistance drift.

I. Introduction

Phase Change Memory (PCM) has promising multi-level-cell (MLC) capability [1-3] to address the increasing on-chip memory capacity needs (Fig. 1) in neuromorphic and in-memory computing applications [3-7]. In these studies, overlapping MLC states are tolerated by the error-resilient nature of neural networks. However, the accumulation of information losses could start to affect the accuracy for complex tasks [8, 9]. Precise computation with distinguishable MLC states eliminates the information loss but imposes stringent device requirements. This work shows how bit error rate may inaccurately represent the robustness of a neural network and demonstrates three techniques to improve inference accuracy and retention time of a MLC PCM system.

II. Challenges of MLC Devices

Figure 2 details the PCM device with dual metal concept [3] and its MLC programmability. Fig. 2 (a) shows the bottom electrode (BE) and inner metal critical dimensions (CD) are 20 nm and ~10 nm, respectively. Fig. 2 (b) insect demonstrates the smooth SET transition of cell resistance (R_{CELL}) needed in MLC programming between 2MΩ and 20KΩ. Fig. 2 (b) presents the normalized MLC R_{CELL} variability from 31,000 devices using program-verify approach.

Figure 3 depicts three MLC device challenges in neural network– (1) unbalanced device requirement, (2) asymmetric MLC distribution, and (3) asymmetric MLC retention capability. The unbalanced device requirement arises from the fact that upper and lower weight bits affect the accuracy by different amount. Fig. 4 (a) shows the top-1 CIFAR-100 accuracy degradation of a ResNet-20 network (IN/W/OUT = 8b/8b/19b) as a function of injected errors. Fig. 4 (b) extracts the allowed error tolerance of each bit for a 2% accuracy degradation and shows a difference of over 10^3X between W[0] and W[7]. The MLC device responsible for the two uppermost bit faces the most stringent requirement. Second, the trained weight distribution is typically non-uniform (Fig. 5). Similar trend has also been observed on AlexNet [10] and VGG-16 [11]. This leads to an asymmetric number of devices among MLC states (i.e. "00/11" have higher probability than "01/10"). Lastly, the programmed PCM cell resistance drifts upward in a power-law relation with time ratio (t/t_0) [12, 13]. Fig. 6 shows the measured drift coefficient (γ) increases with initial cell resistance (R_0) and implies higher resistance states are more susceptible to retention error. Without considering these factors, a MLC PCM system targeting to achieve the lowest overall bit error rate is suboptimal for neural networks.

III. MLC Techniques and Results

Figure 7 illustrates the three proposed schemes – (1) device requirement balancing (DRB), (2) prediction-based MSB-biased referencing (PMR), and (3) bit-prioritized placement (BPP). The DRB pairs the less error-tolerant bit with the more error-tolerant bit (i.e. W[7]+W[0]) to relax the overall device requirement and balance the number of devices in each MLC states (Fig. 8). The PMR leverages the fact that the W[7] distribution can be viewed as two W[7] subsets with W[6]=0 and W[6]=1. Fig. 9 shows the probability of W[6]=W[7] is as high as 94.28% in ResNet-20. This motivates us to adjust W[7]'s sensing reference based on W[6] value to lower W[7]'s bit error rate (BER). Lastly, the BPP reduces upper-bit BER by modulating the two intermediate resistance distributions outward. This allocates more margin to the upper-bit and enhance the resilient against resistance drift.

Figure 10 (a) and (b) show the measured MLC BER of 8196 devices using the baseline variation-aware placement scheme [14] and (b) the BPP, respectively. The former minimizes the overall BER, while the latter purposely trades the lower-bit BER to reduce the upper-bit BER. Fig. 10 (c) shows PMR further suppresses the BER of W[7]. Based on the measured MLC BER, we simulate the accuracy degradation of ResNet-20 on CIFAR-100 (Fig. 11). The MLC PCM retention time to maintain the inference accuracy degradation within 3% is improved by 10^5X and the accuracy degradation is reduced by 91% (10.8X). Noted the overall BER of the baseline is lower than that of the proposed techniques. This shows evaluating the robustness of a MLC PCM neural network system from its overall BER alone may lead to inaccurate results. The proposed techniques are also applicable to other memory devices [15, 16], algorithms [17], and beyond MLC.

IV. Conclusion

Three neural network aware techniques were implemented to drastically improve neural network inference accuracy and retention time. The demonstration of how an MLC PCM system can be modulated for better reliability despite having a higher overall bit error rate favors an end-to-end neural network aware approach to evaluate MLC efficacy for the best inference accuracy.

Acknowledgement

The authors would like to thank Carlos H. Diaz for the insightful discussion.

References

[1] F. Bedeschi *et al.*, *JSSC*, 2009. [2] W.S. Khwa *et al.*, *JSSC*, 2017. [3] J. Y. Wu *et al.*, *IEDM*, 2018. [4] I. Boybat *et al.*, *Nat Commun*, 2018. [5] I. Giannopoulos *et al.*, IEDM, 2018, [6] C.X. Xue *et al.*, *Nat. Electron* 4, 81-90 (2021), [7] C.X. Xue, *ISSCC*, 2021, [8] A. S. Rekhi *et al.*, *DAC*, 2019, [9] B. Reagen *et al.*, *DAC*, 2018, [10] E. Lee *et al.*, *Electronics*, 2021, [11] E. Park *et al.*, *ISCA*, 2018, [12] M. Boniardi *et al.*, *TED*, 2010. [13] W. S. Khwa *et al.*, *IMW*, 2015. [14] B. Q. Le *et al.*, *TED*, 2019. [15] I. Giannopoulos *et al.*, *IEDM*, 2018. [16] S. Kim *et al.*, *IEDM*, 2013. [17] S. Ambrogio *et al.*, *IEDM*, 2019.

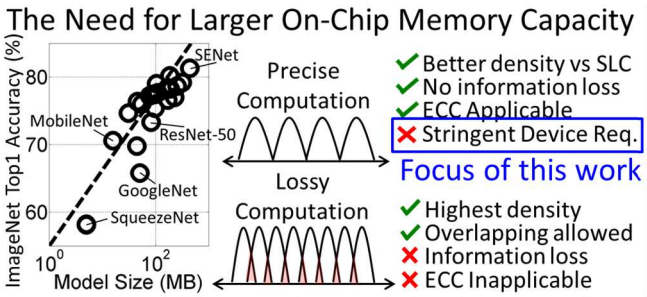

Fig. 1 MLC addresses the need for larger on-chip memory capacity. Distinguishable MLC states offers no information loss but faces stringent device requirement.

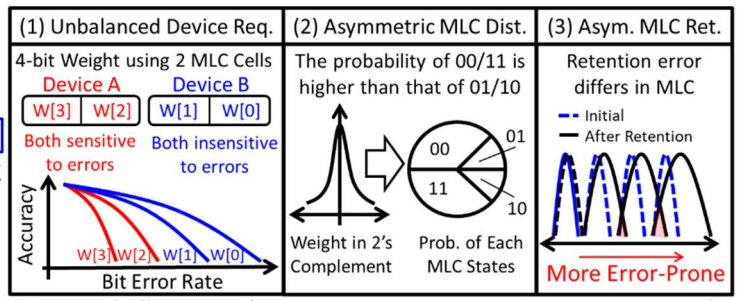

Fig. 3 Challenges of using MLC memory devices in neural networks include (1) unbalanced device requirement from different error sensitivity of each weight bits, (2) asymmetric MLC state probability from non-uniform weight distribution, and (3) asymmetric MLC retention capability from resistance drift.

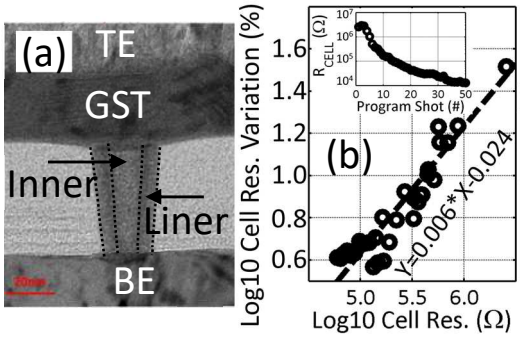

Fig. 2 (a) TEM of the PCM device with dual metal concept. Fig. 2 (b) shows smooth program curve (insect) and the measured cell resistance variation using program-verify.

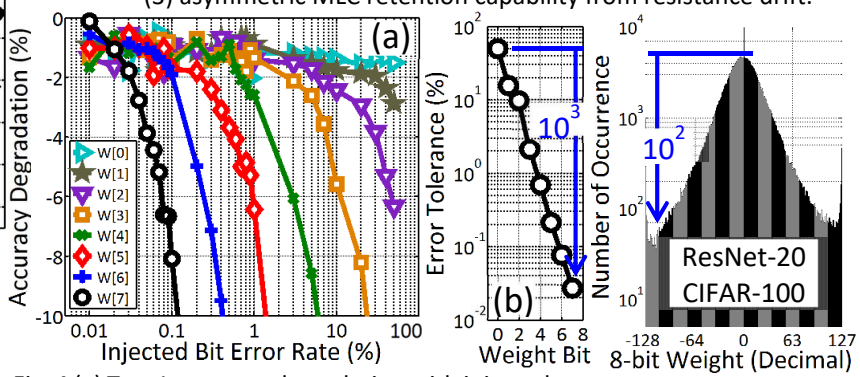

Fig. 4 (a) Top-1 accuracy degradation with injected bit errors. (b) The extracted error tolerance of each bit allowed for a 2% accuracy degradation.

Fig. 5 Trained 8-bit weight shows a 100X difference between center and edge.

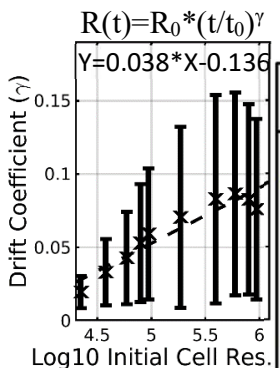

Fig. 6 Measured PCM drift coefficient (γ) vs initial resistance (R_0).

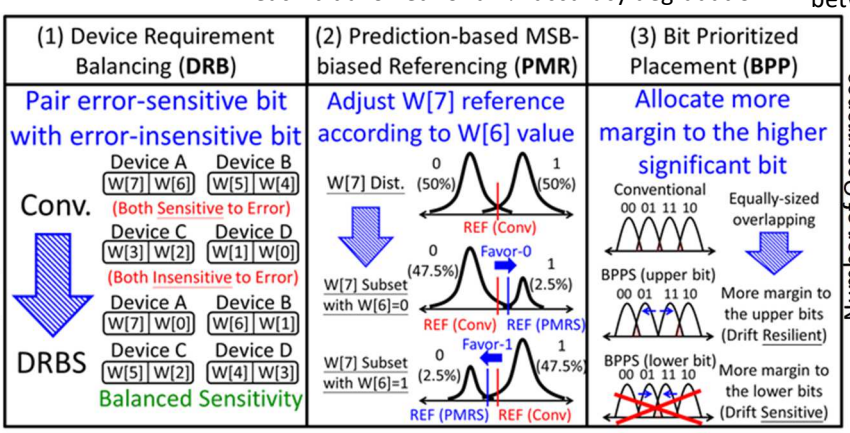

Fig. 7 Illustration of the three proposed schemes – (1) device requirement balancing (DRB), (2) prediction-based MSB-biased referencing (PMR), and (3) bit-prioritized placement (BPP).

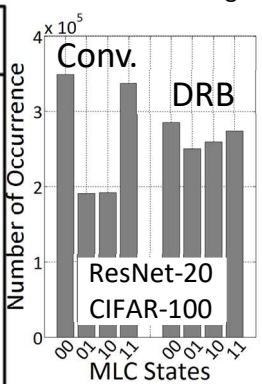

Fig. 8 DRB balances the number of devices in each MLC states.

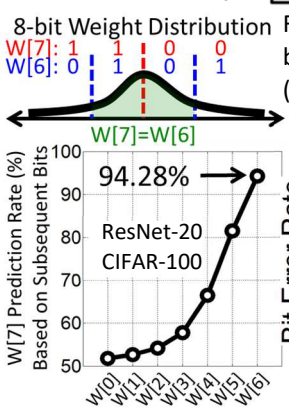

Fig. 9 PMR exploits the high probability of W[7]=W[6] to adjust W[7] sensing reference based on W[6] value.

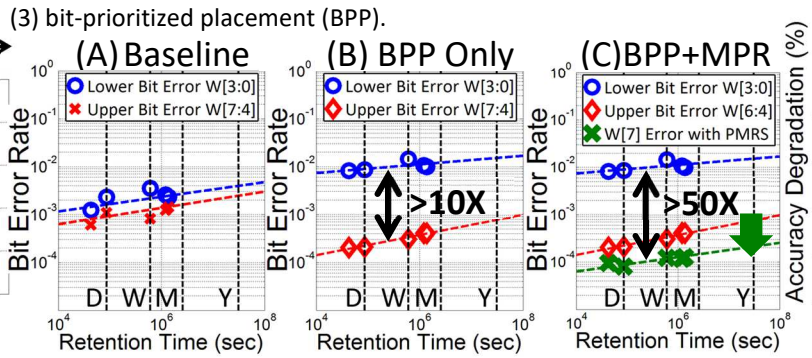

Fig. 10 Measured bit error rates from using (a) the baseline variation-aware placement scheme, (b) the proposed BPP scheme, and (c) the proposed BPP+MPR schemes. Dashed lines are added to indicate time for Day (D), Week (W), Month (M), and Year (Y).

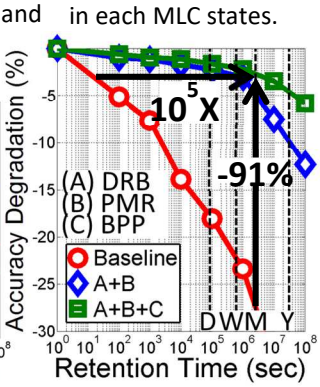

Fig. 11 Simulation for CIFAR-100 accuracy degradation vs time based on measured BER.