A 40nm Low-Power Logic Compatible Phase Change Memory Technology

J.Y. Wu, Y.S. Chen, W.S. Khwa, S.M. Yu, T.Y. Wang, J.C. Tseng, Y.D. Chih and Carlos H. Diaz
Taiwan Semiconductor Manufacturing Company Ltd., Corporate Research, 168, Park Ave. 2,
Hsinchu Science Park, Taiwan
Tel: 886-3-5636688, email: jywuz@tsmc.com

Abstract

An embedded phase change memory technology in 40nm low-power logic platform is demonstrated with minimal added process complexity - two non-critical additional masks over standard logic. Specially designed hard mask and etching process was used to achieve 50% shrinkage of the memory cell bottom electrode dimension with same lithography tooling as the 40nm logic platform. Bottom electrode CD shrinkage along with optimization of the electrode materials in terms of electrical and thermal conductivity enabled significant (~4x) write current reduction attaining competitive levels of ~300 μA at 40nm BE CD. Embedded PCM cells reported in this work demonstrated over 100x memory window - (RESET/SET resistance switching ratio), over 200k cycling endurance with extrapolated 10 year retention at 120 ℃. In this work not only large switching resistance ratios but also highly-controllable resistance values that are almost independent of the PCM starting resistance state are presented along with the corresponding programing pulse requirements. The switching resistance ratio and resistance value controllability are key features for neural network and compute-in-memory applications. In this work, their benefits on design margins for energy efficient high-density binary neural network for inference applications aiming accuracy levels of well over 90% is asserted over an MNIST dataset.

INTRODUCTION

Phase Change Memory (PCM) has been reported as a good candidate for non-volatile memory applications [1-5]. Regardless of the application, low power operation, high density, and low process complexity – cost are key requirements to be met; for embedded memory applications compatibility with underlying logic process is also a upmost importance. Tackling the power challenge at fundamental memory cell level relates to minimizing the current required to promote phase change and the overall voltage drop across the memory element and its access device. Minimizing operating current entails PCM material optimization along with the memory cell structure electro-thermal design. Primary cell structures can be categorized as confined and non-confined [6-7]. Confined cells have the potential of smallest footprint at expense of challenging phase change material bottom-up gap-fill inside small holes or requiring complex patterning schemes [7-9]. In this work we focus on non-confined structures for 1T1R embedded memory arrays where transistor as access device enables write/reads at lower operating power supply as compared to ovonic threshold switches (OTS) as selector devices [10-11]. We report on operating current reduction while preserving solid memory switching windows as demonstrated on 1Mb test arrays by cost-effective bottom electrode dimension scaling and electro-thermal design, memory cell PCM material and etch optimization (un-doped to doped GST); results show good scalability potential and compare very well with published work from other groups [7,12-13]. This work also reports on our findings pertaining memory cell programing schemes aimed attain dependable and well controllable resistance switching independent of the initial state of the memory cells. Finally, models based on this work's experimental results are utilized to study the design space and guidelines the PCM technology potential benefits for energy-efficient binary neural network applications. These results complement well other reports on PCM as a key technology candidate for artificial intelligence (AI) applications [14-15].

DEVICE STRUCTURE AND THERMAL SIMULATION

Figure 1 shows the TEM image of mushroom structure PCM device fabricated between M4 and M5 of 40nm CMOS logic technology. Figure 2 shows the BE process flow for shrinking 50% BE CD from original logic design rule. Dual metal-1 and metal-2 used in BE process flow are capable of filling in 40nm BE CD. CMP process is performed to decide final BE height. Doped GST is then deposited onto small scale BE, followed by a top electrode (TE) layer. TE/GST are patterned and connected to M5 through top Via. In order to optimize etching condition for TE and GST individually, two step etching process is compared in Fig. 3. Composition change induced by etch process is observed on TE#1 (high Cl_2 flow) and TE#2 (low Cl_2 flow) is the key to achieve uniform GST composition. Optimized halogen gases for GST etching is also important to minimize etch damage as shown in Fig. 4 [16-17]. In Fig. 5, dual metal thicknesses ratio (BE#1 with Met-1/Met-2=1/8 and BE#2 with Met-1/Met-2=8/1) are designed for engineering BE structure through TCAD thermal simulation [18]. For BE#2, more heat can be generated and stay in the GST/BE interface for reducing write current requirement. In addition, melting volume for BE#2 is smaller as comparing to BE#1. In Fig. 6 and Fig. 7, electrical resistivity of Met-2 and thermal conductivity of Met-1 are studied on two different BE structures. From simulation result, BE#2 can combine both advantages for dramatically write current reduction as comparing to BE#1.

DEVICE PERFORMANCE

In Fig. 8, BE#2 can demonstrate 300 μA write current for full amorphous state that is ~4x reduction comparing to

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

BE#1. In Fig. 9, RESET current reduction with BE area scaling is studied and experimental result follows the trend of $I \sim A^{0.8}$, better than literature's results ($I \sim A^{0.65}$) [19]. Fig. 10 shows RT and 150℃ resistance drift behavior and estimated RESET R drift coefficient is 0.1. Designed 1T1R PCM 1Mb test-chip on 40nm low-power logic platform is shown in Fig. 11. In Fig. 12, good HRS and LRS switching result is demonstrated with over 100x switching ratio.

Figure 13 shows the resistance distribution of HRS and LRS with various drain voltages (Vd). There is a great improvement of the chip operated to higher resistance by larger Vd due to larger amorphous volume formed while LRS resistance slightly increases owing to some residual amorphous regions existed after SET. In Fig. 14, SET speed is characterized with programming the cell back to the HRS before applying each SET pulse. The 100ns SET speed is defined by reaching 100x resistance switching window on doped GST. Through waveform modulation by increasing falling tail pulse width, symmetry of RESET/SET conductance as a function of write voltage can be demonstrated in Fig. 15. PCM program characteristic features advantage for precisely resistance control and it is critical for improving inference accuracy in computing-in-memory architecture. Fig. 16 demonstrated good cycling endurance up to 200K with applying 200ns pulse width without read verification. Fig. 17 is retention result for doped GST material. Calculated Ea is ~2.9 eV for predicting 10 year data retention at 120 ℃. Through write-and-verification scheme, standard deviation of LRS and HRS resistance can be improved to < 3.5%.

COMPUTING IN MEMORY

To improve the data utilization efficiency in CMOS-based deep neural network (DNN) accelerators, parallelized computation across multiple processing-elements (PE) are highly preferred [20-21]. Computing-in-memory (CIM) is an alternative a possible approach that integrates the computation into the memory array. In CIM, the matrix multiplication could be performed efficiently by activating multiple wordlines (WL) simultaneously and comparing the accumulated bitline current (I_{BL}) or voltage (V_{BL}) [22-23]. From the characterization result on 40nm PCM test-chip, the impact of cell resistance standard deviation (std) and resistance ratio (R_{ratio}) on inference accuracy are simulated using a LeNet-5 binary neural network (BNN) with MNIST dataset [24-25]. We focused on the last two fully-connected (FC) layers, because they are technically more challenging due to more WL activations.

Fig. 18 shows the product-sum result distributions of the last two FC layers, with number of positive one and number of negative one on the x-axis and y-axis, respectively. A {x=18, y=13} point would indicate this MNIST test image yields a +5 product-sum result. The average product-sum result from FC-1 and FC-2 are +4.37 and +8.2, respectively. This implies that FC-1 would face more technical challenge due to its narrower margin. Fig. 19 shows the impact analysis of std and R_{ratio} on inference accuracy. To achieve a >90% accuracy on a two-layer FC (64x64x10), a minimal R_{ratio} of 10 is required with 3% and 4% std on FC-1 and FC-2, respectively. Inference accuracy is very sensitive to std. Increasing the std from 3% to 4%, with a R_{ratio} of 10, on FC-1 increases the error rate by >2.9X. If R_{ratio} could be increased to 143, then the std can be relaxed to 5%. Larger R_{ratio} is highly favorable and could relax the std requirement. Fig. 20 shows a case analysis of PCM versus two other emerging memory candidates. Even for BNN, in which both the weights and neuron activations are binarized to +1/-1, it is still critical to have a sufficiently large R_{ratio} to compensate for the intrinsic memory device variation. In this regard, PCM benefits tremendously from its ability to fine-tune cell resistance and its high R_{ratio}.

CONCLUSION

A 40nm CMOS-compatible PCM technology is demonstrated with cost-effective BE reduction method and damage-free TE/GST patterning processes. 300 μA write current is achieved with 4x reduction through optimizing dual metal thicknesses ratio. Over 100x resistance ratio, good resistance controllability, reliable cycling and good high temperature data retention are also demonstrated. Proposed PCM technology is a promising candidate of DNN hardware accelerator for handwritten MNIST accuracy over 90% with binary neural network.

Acknowledgements

The authors would like to thank the support of Cheng-Chung Chien and Julian R. Lee in process modules as well as Jeff Wu and Pike Chang for TCAD simulation.

References

[1] G.Servalli, "A 45nm Generation Phase Change Memory Technology", IEDM Tech. Dig., p113, 2009.

[2] F. Pellizzer1, et. al., "A 90nm Phase Change Memory Technology for Stand-Alone Non-Volatile Memory Applications", VLSI Tech. Symp., 2006.

[3] Daewon Ha, et. al. "Recent Advances in High Density Phase Change Memory (PRAM)", IEDM Tech. Dig., 2007.

[4] Scott W. Fong., et. al., "Phase-Change Memory—Towards a Storage-Class Memory", IEEE TED, Vol 64, No.11, pp.4374-4385, 2017.

[5] H.-S. Philip Wong, et. al., "Phase Change Memory", IEDM Tech. Dig., p2201, 2010.

[6] Y.J. Song, et. al., "Highly Reliable 256Mb PRAM with Advanced Ring Contact Technology and Novel Encapsulating Technology", VLSI Tech. Symp., 2006.

[7] M. J. Kang, et. al., "PRAM cell technology and characterization in 20nm node size", IEDM Tech. Dig., p39, 2011.

[8] M. Breitwisch, et. al.,"Novel Lithography-Independent Pore Phase Change Memory", Symp. VLSI Tech., p100, 2007.

[9] J.I. Lee, et. al., "Highly Scalable Phase Change Memory with CVD GeSbTe for Sub 50nm Generation ", Symp. VLSI Tech., p102, 2007.

[10] DerChang Kau, et. al., "A stackable cross point phase change memory", IEDM Tech. Dig., p617, 2009.

[11] H. Y. Cheng, et. al., "An Ultra High Endurance and Thermally Stable Selector based on TeAsGeSiSe Chalcogenides Compatible with BEOL IC Integration for Cross-Point PCM", IEDM Tech. Dig., p28, 2016.

[12] Y. H. Ha, et. al., "An Edge Contact Type Cell for Phase Change RAM Featuring Very Low Power Consumption", Symp. VLSI Tech., p175, 2003.

[13] W.C. Chien, et. al., "Reliability Study of a 128Mb Phase Change Memory Chip Implemented with Doped Ga-Sb-Ge with Extraordinary Thermal Stability", IEDM Tech. Dig., p662, 2016.

[14] S. Yu, et. al., "Neuro-inspired computing with emerging nonvolatile memorys," in Proceedings of the IEEE, vol. 106, no. 2, pp. 260-285, 2018.

[15] G. W. Burr, et. al., "Experimental Demonstration and Tolerancing of a Large-Scale Neural Network (165000 Synapses) Using Phase-Change Memory as the Synaptic Weight Element", IEEE TED, vol. 62, no. 11, pp. 3498-3507, 2015.

[16] E. A. Josepht, et. al., "Patterning of N:Ge2Sb2Te5 Films and the Characterization of Etch Induced Modification for Non-Volatile Phase Change Memory Applications", VLSI-TSA, 2008.

[17] Se-Koo Kang, et. al., "Etch Damage of Ge2Sb2Te5 for Different Halogen Gases", Jpn. J. Appl. Phys., 086501-1, 2011

[18] J.Y. Wu, et. al., "A Low Power Phase Change Memory Using Thermally Confined TaN/TiN Bottom Electrode", IEDM Tech. Dig., p43, 2011.

[19] D. Ielmini, et. al., "Phase change materials in non-volatile storage", Materialstoday, vol. 14, pp. 600-607, 2011.

[20] Y.H. Chen, et. al., "Eyeriss: an energy-efficient reconfigurable accelerator for deep convolutional neural networks," IEEE ISSCC, p262, 2016.

[21] N. P. Jouppi, et. al., "In-datacenter performance analysis of a Tensor Processing Unit," ACM/IEEE ISCA, p1, 2017.

[22] A. Biswas, et. al., "Conv-RAM: An energy-efficient SRAM with embedded convolution computation for low-power CNN-based machine learning applications," IEEE ISSCC, pp. 488-490, 2018.

[23] S. K. Gonugondla, et. al., "A 42pJ/decision 3.12TOPS/W robust in-memory machine learning classifier with on-chip training," IEEE ISSCC, pp. 490-492, 2018.

[24] W. S. Khwa, et. al., "A 65nm 4Kb algorithm-dependent computing-in-memory SRAM unit-macro with 2.3ns and 55.8TOPS/W fully parallel product-sum operation for binary DNN edge processors," IEEE ISSCC, pp. 496-498, 2018.

[25] R. Liu, et. al., "Parallelizing SRAM Arrays with Customized Bit-cell for Binary Neural Networks", Design Automation Conference (DAC), pp. 21.1-21.6, 2018.

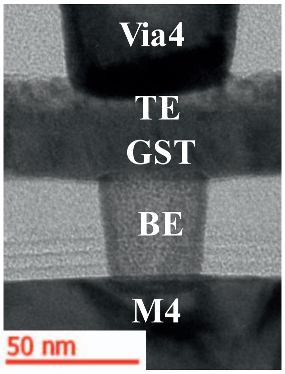


Fig. 1 TEM image for intergraded PCM on 40nm low-power logic platform. PCM is fabricated between M4 and M5.

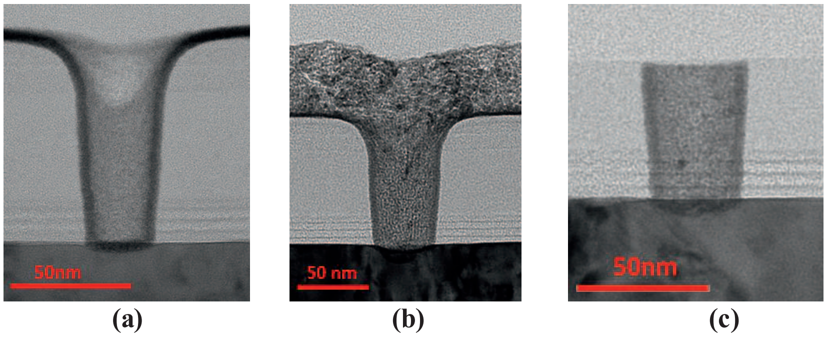


Fig. 2 Post BE etch TEM images with shrinking 50% BE CD from logic design rule after (a) metal-1 deposition, (b) metal-2 deposition and (c) CMP.

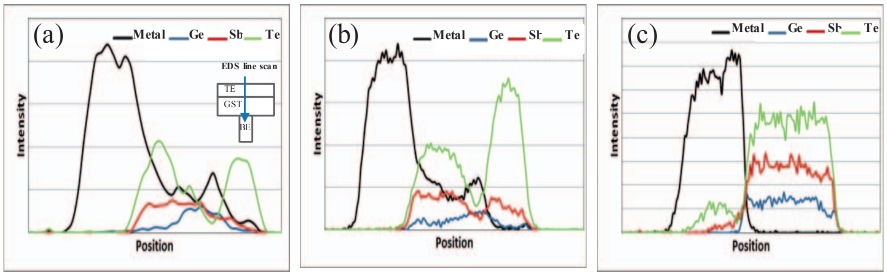


Fig. 3 TE-Metal/Ge/Sb/Te composition uniformity is studied by EDS line scan (vertical) with different TE and GST etching conditions that includes (a) TE#1/GST#1, (b) TE#1/GST#2 and (c) TE#2/GST#2. TE metal etch is as important as GST etch since both etch conditions can affect Ge/Sb/Te composition uniformity.

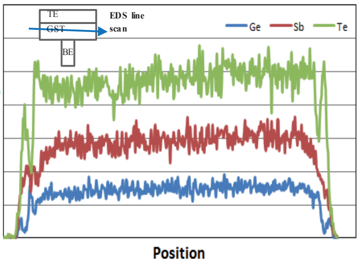


Fig. 4 Ge/Sb/Te composition uniformity in GST is studied by EDS line scan (horizontal) with optimized TE/GST etching conditions.

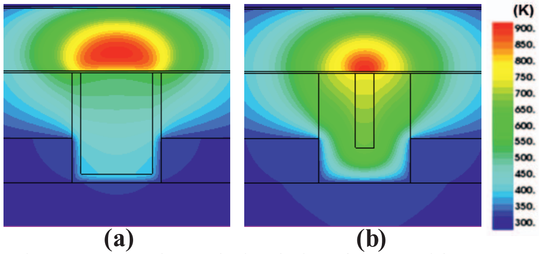


Fig. 5 TCAD thermal simulation for BE with engineering dual metal thicknesses ratio. Met-1 and Met-2 are outer and inner layers, respectively. (a) BE#1 with Met-1/Met-2=1/8 and (b) BE#2 with Met-1/Met-2=8/1.

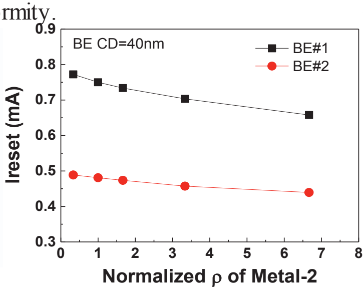


Fig. 6 BE engineering by changing electrical resistivity of Met-2 (inner layer). Higher electrical resistivity Met-2 can reduce RESET current.

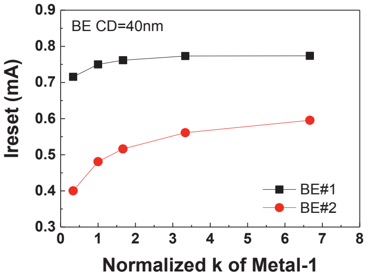


Fig. 7 BE engineering by changing thermal conductivity of Met-1 (outer layer). Low thermal conductivity Met-1 can reduce RESET current.

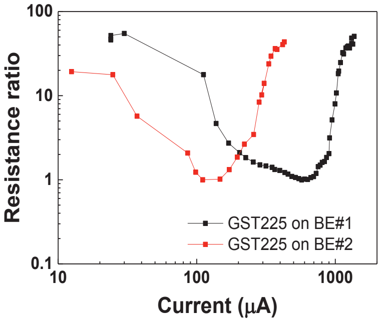


Fig. 8 Undoped GST225 R-I curves are measured on different BE structures. RESET current as low as 300uA is demonstrated on BE#2.

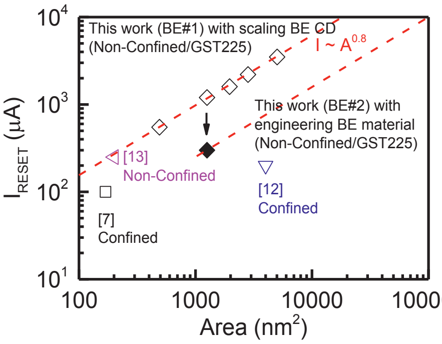


Fig. 9 RESET current as a function of BE area is plotted from experiment results. Similar to thermal simulation result, write current can be further reduced on BE#2.

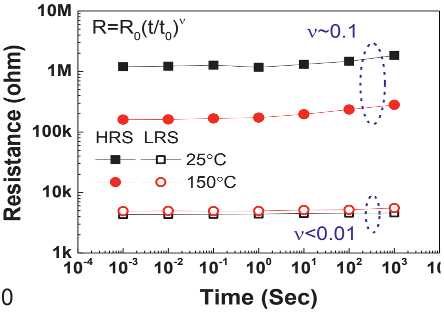


Fig. 10 Data retention result at RT and 150 ℃. LRS is stable with time (ν<0.01) while HRS starts to drift after 1 sec with ν~0.1. ν is defined as drift coefficient.

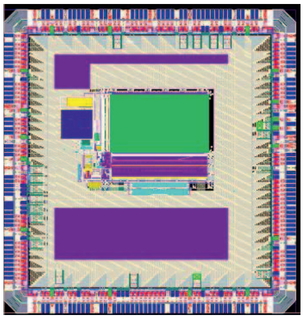


Fig.11 The 1Mb 1T1R PCM test-chip on 40nm low-power logic platform.

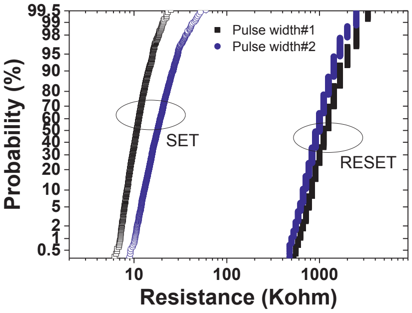


Fig.12 Resistance distribution of HRS and LRS are compared with different write speeds. Longer PW#2 can achieve larger memory window than PW#1.

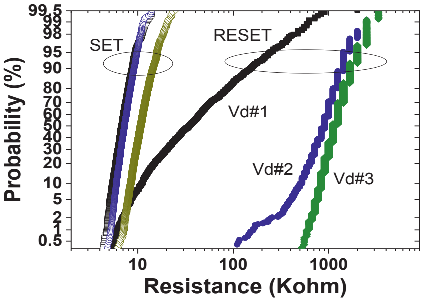


Fig.13 Resistance distribution of HRS and LRS with various drain voltages (Vd). A great improvement of HRS by a larger Vd while LRS resistance only slightly increases.

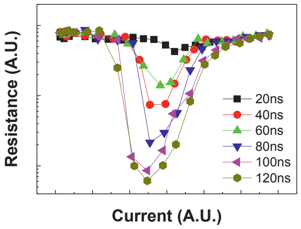


Fig.14 R-I curve is measured for RESET and SET operation. Estimated SET speed are 100ns for reaching 100x R switching ratio.

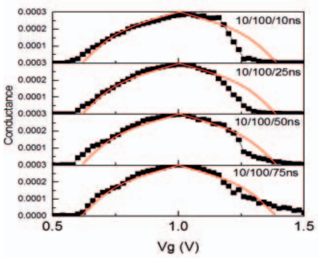


Fig.15 Write waveform modulation can achieve symmetry of RESET/SET conductance as a function of write voltage.

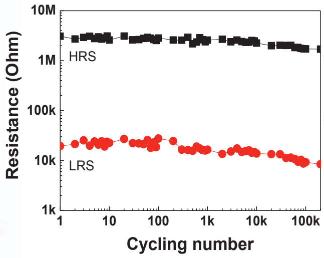


Fig. 16 200K cycling endurance of 1T1R PCM is demonstrated. RESET and SET are operated with 200ns pulse width.

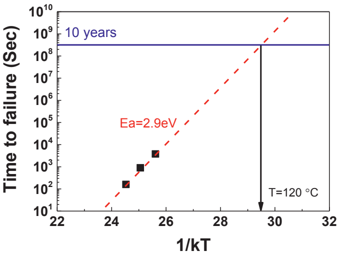
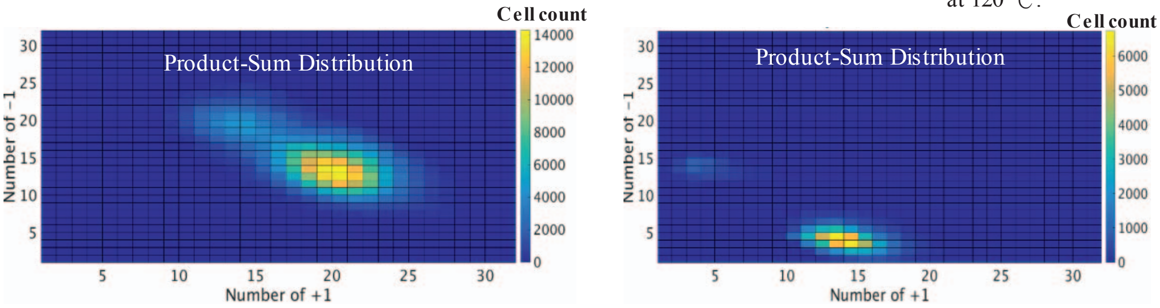


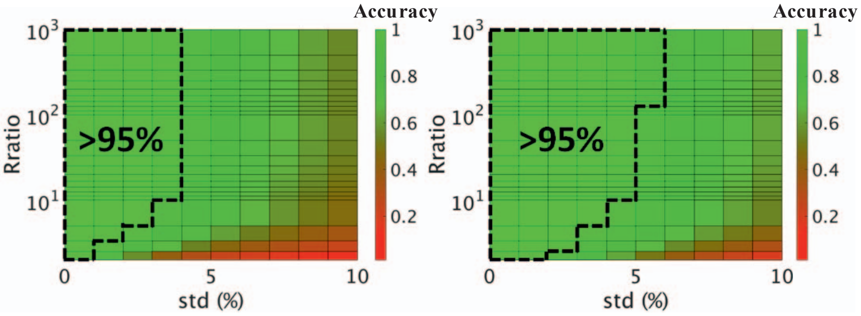
Fig. 17 Data retention result for extracting Ea on 1T1R PCM device for predicting 10 year lifetime at 120 ℃.



(a) FC-1

(b) FC-2

Fig. 18 Simulated product-sum distributions from MNIST dataset show the last two FC layers will have different requirement. The average product-sum result for (a) FC-1 and (b) FC-2 are +4.37 and +8.2, respectively. FC-1 layer faces more stringent requirement due to its narrower margin from having a smaller difference between the number of +1 and -1.



(a) FC-1

(b) FC-2

Fig. 19 Simulation analysis shows both standard deviation (std) and resistance ratio (R_{ratio}) are critical for MNIST accuracy. In particular, for a FC size of 64x64x10, a minimum R_{ratio} of 10 is suggested with 3% and 4% std on FC-1 and FC2, respectively. If R_{ratio} is increased to 143, the std could be relaxed to 5%.

PCM: Rratio = 100, std = 3.5%
Memory 1: Rratio = 10, std = 6%
Memory 2: Rratio = 2, std = 5%

	Accuracy %
PCM (FC-1)	95.303
Memory 1 (FC-1)	51.780
Memory 2 (FC-1)	0.013
PCM (FC-2)	99.122
Memory 1 (FC-2)	77.494
Memory 2 (FC-2)	22.274

Fig. 20 MNIST accuracy simulation analysis shows that even for binary neural network, a large R_{ratio} is still critical to compensate the intrinsic memory device variation.