Reliability Characterization of Logic-Compatible NAND Flash Memory based Synapses with 3-bit per Cell Weights and 1μA Current Steps

Minsu Kim, Jeehwan Song, Chris H. Kim
Department of Electrical and Computer Engineering
University of Minnesota
200 Union Street SE, Minneapolis, MN 55455, USA
phone: +1-(612)-475-7432, e-mail: kimx4916@umn.edu

***Abstract* — A logic-compatible embedded NAND (eNAND) flash memory based synapse with 3 bit per cell weight storage and 1μA current steps was demonstrated in a standard 65nm CMOS process. The eNAND flash based neuromorphic core consists of 16stack eNAND strings. Each flash cell is composed of 3 transistors (2 PMOS and 1 NMOS) and each string is connected to the main bitline via 2 additional NMOS access transistors. In this work, we realized 3 bit weight based on 1μA current steps using the proposed back-pattern tolerant program-verify scheme. To evaluate the reliability of eNAND Flash based synapses, we measured the temperature dependence, read disturbance, and retention characteristics from a 65nm test chip with 3 bit per cell weight storage and 1μA current steps.**

Index Terms – Embedded NAND (eNAND); logic-compatible NAND flash memory; neuromorphic computing; read disturbance; retention characteristic.

I. INTRODUCTION

Neuromorphic computing based on synaptic memory arrays is gaining popularity as a compact, low energy, and high performance alternative to GPU, FPGA, and ASIC based neural net accelerators. In neuromorphic arrays, the vector multiplication is performed in the analog domain by storing the weights in the memory array, loading the input data onto multiple wordlines, and summing up the cell currents through the bitline (Fig. 1). In principle, the current summation method can enable neural networks with faster operating speeds and lower power consumption compared to digital implementations. However, analog circuits have higher sensitivity to process-voltage-temperature (PVT) variations which necessitates synaptic memory devices to generate cell currents corresponding to the precise weight and input data. Fig. 2 shows different memory device options for neuromorphic applications [1-4]. Single-poly embedded flash (eFlash), which is the focus of this paper, can be built in a standard logic process using 3-5 discrete transistors per cell. Despite the relatively large area compared to emerging devices based on a filament layer or phase change material, single-poly eFlash can store precise multi-level non-volatile weights via incremental programming making it viable candidate for neuromorphic arrays. Logic-compatible eFlash cells can also play an important role in studying various aspects of neuromorphic computing as it enables hardware demonstration in a mature logic technology. In this work, we consider a NAND type array rather than a NOR type array due to its higher integration density, lower bitline capacitance, and in light of the huge investments made by industry in 3D

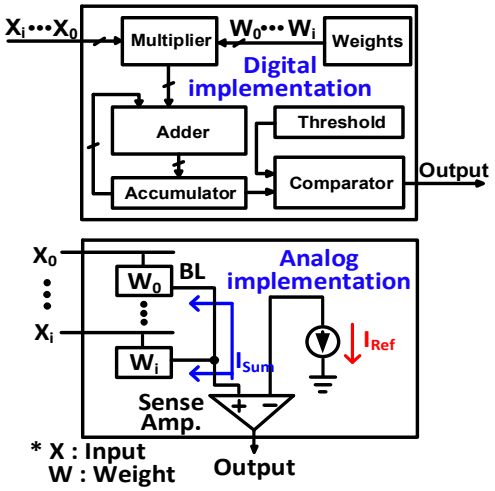


Fig. 1. Digital versus analog implementation of vector multiplication operation for neural network applications.

This work

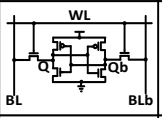
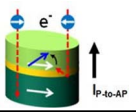
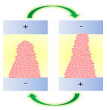
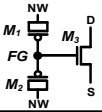
	SRAM [1]	MRAM [2]	RRAM [1]	PCRAM [3]	eFLASH [4]
Cell Configuration					
Nonvolatile?	No	Yes	Yes	Yes	Yes
Tunable?	No	No	No	Yes	Yes
Logic Compatible?	Yes	No	No	No	Yes
Muti level Weights?	No	No	No	Yes	Yes

Fig. 2. Memory device options for neuromorphic synapse array. Single poly eFlash is can be built in a mature CMOS technology, is non-volatile, and supports multiple level weights for accurate vector multiplication.

NAND flash technology [4-5]. To evaluate the reliability of eNAND flash based synapses, we measured the temperature dependence, read disturbance, and retention characteristics from a 65nm test chip with 3 bit per cell weight storage and 1μA current steps.

II. NEUROMORPHIC CORE & PROGRAMMING METHOD

A. *ENAND Flash based Neuromorphic Core*

The test chip consists of an eNAND flash array including 40 bitlines, high voltage control signals, bitline current sensing blocks and scan chains. The logic-compatible 3T eFlash cell

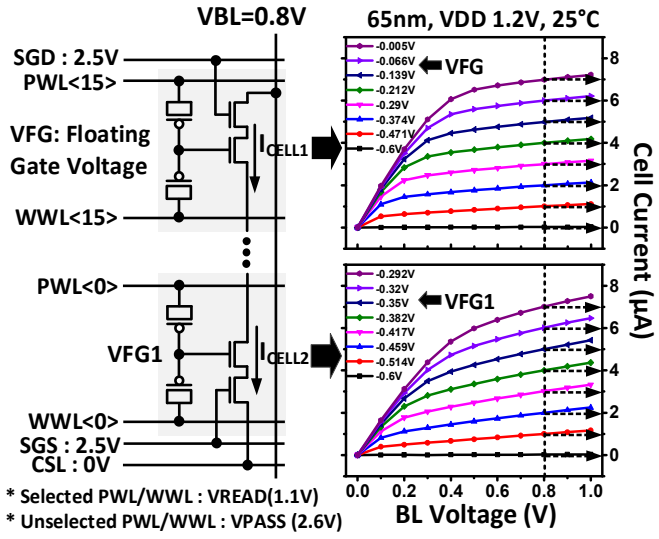


Fig. 3. Simulated I-V characteristics of top and bottom eFlash cells of a 16 stack NAND string. Series resistance varies depending on the location in the stack which can be compensated by incremental programming.

shown in Fig. 3 consists of two asymmetrically sized PMOS devices for efficient program and erase operation and an NMOS read device for the NAND string. Due to series resistance of the unselected devices, the I-V characteristics of the eNAND Flash cell vary depending on its location in the

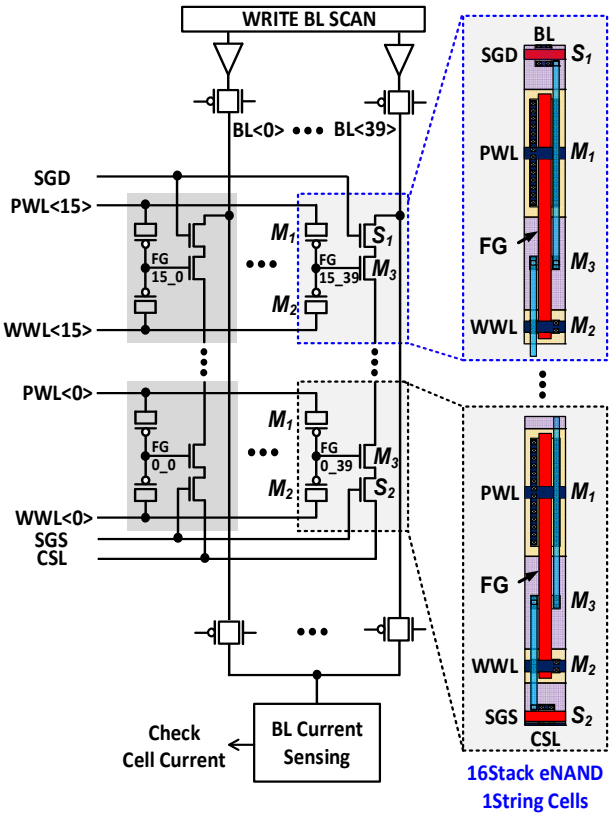


Fig. 4. Column slice of 16 stack eNAND flash based neuromorphic array. Control signals, BL current sensing block, and layout of the eNAND string are shown. IO devices in a standard logic process are used for the flash cell implementation.

stack. Simulation results in Fig. 3 (right) show that such location dependent variation can be cancelled out by fine-tuning the floating gate (FG) voltage. Fig. 4 shows the circuit diagram and layout of a 16 stack eNAND string. Each flash cell consists of 3 transistors and each string is connected to the main bitline via 2 additional NMOS access transistors. The wordline signal is controlled by the multi-story high voltage wordline driver that was previously proposed in [4]. To program the weights, we first erase the entire array and then

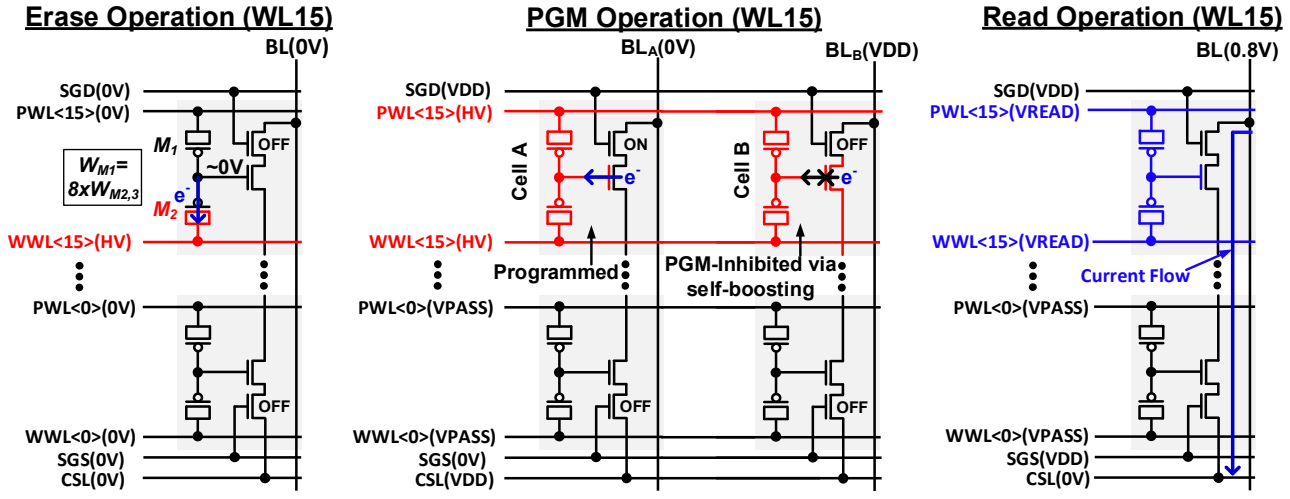


Fig. 5. Bias conditions of logic-compatible eNAND flash cell for erase, program and read modes. Example shown for WL #15, the top cell on the NAND string.

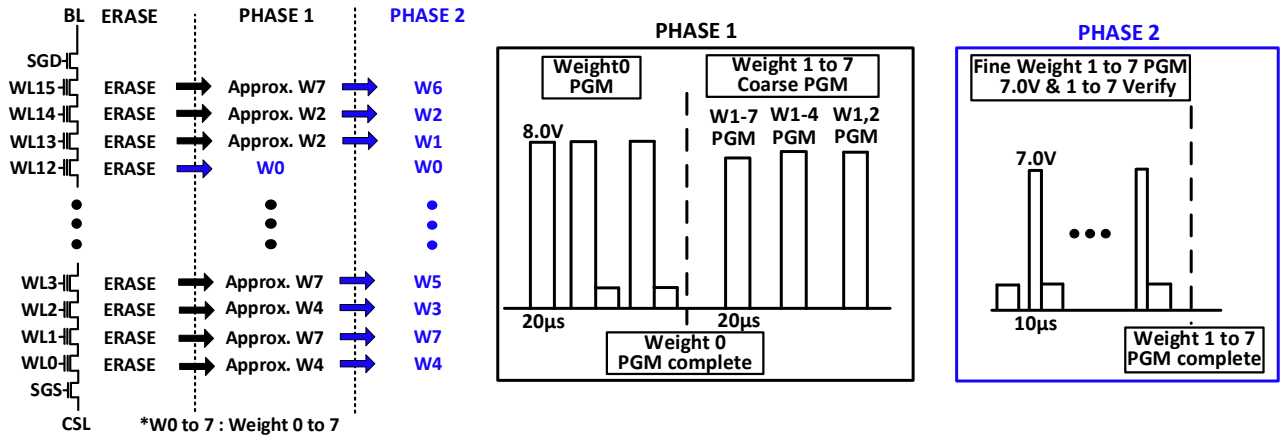


Fig. 6. Programming sequence for writing weights 0 to 7 into the 16 stack eNAND array. A 3 bit weight can be programmed into each cell using the proposed back-pattern tolerant program-verify scheme.

individually program each eNAND cell to obtain the desired cell current. The cell current value was verified after each program pulse for precise weight storage. Program and verify operations are repeated until the cell current reaches the target level. Program inhibition is achieved by cutting off the eNAND string from the main bitline and inducing self-boosting. Fig. 5 shows the bias conditions of the control signals during erase, program, program inhibition and read modes.

B. Programming 3bit Weight into eNAND cell

VPASS voltage is a critical operating parameter since a high VPASS voltage worsens disturbance while a low VPASS voltage leads to a stronger back-pattern dependence [5]. To suppress program disturbance, we used a 2.6V VPASS voltage in all our experiments. Due to the relatively low VPASS bias, the cell current may change depending on the weights stored in the unselected wordlines. To overcome this issue, we adopted the back-pattern tolerant program-verify scheme [5] shown in Fig. 6 which assures that the programmed cell current remains constant regardless of the cell status of the rest of the array. The operating sequence is as follows. First, we program the weight 0 cells on a given wordline while inhibiting the weight 1 to 7 cells. To ensure that the cell currents of all weight 0 cells are below 0.1 µA, we apply a high voltage (8.0V) for a long duration (20µs) until the cell current is practically zero. Once this is completed, we coarsely program the weight 1 and weight 2 cell currents close to the ideal weight 2 level. Similarly, weight 3 and weight 4 cell currents are programmed close to the ideal weight 4 level, and weight 5, 6, 7 cell currents ae programmed close to the weight 7 target using short program pulses as shown in Fig. 6. The first pulse is applied to weight 1 to 7 cells. The second pulse is applied to weight 1 to 4 cells and the third pulse is applied to weight 1 and 2 cells. The specific program voltage of each pulse was carefully chosen based on the program and program inhibition characteristics. The same sequence was repeated for the rest of the wordlines until the entire array is programmed. Finally, using smaller and shorter pulses (i.e., 7.0V, 10µs), we fine-tune the weight 1 to 7 cell currents to their target levels of 1µA to 7µA with a 1µA current step.

III. RELIABILITY MEASUREMENT RESULTS

We programmed random 3 bit weights into a eNAND Flash memory array under room temperature (25°C). A total of 160 strings × 16 stack cells/string = 2,560 cells were programmed. After programming the array, we measured the cell current at -10°C, 25°C and 70°C. As shown in Fig. 7, the cell current varies by less than 0.6µA across the different temperatures. We found that the cells closer to the bottom of

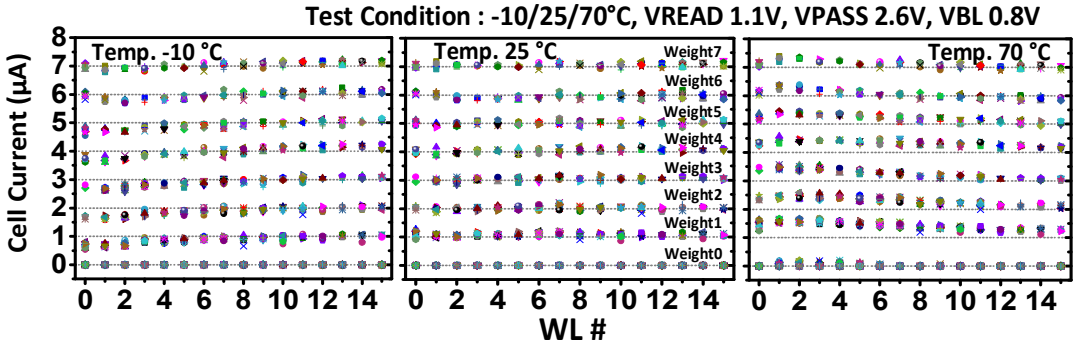


Fig. 7. Individual cell currents in wordline direction measured at -10, 25, and 70 °C. The cells were programmed with 3 bit resolution and 1µA current steps at room temperature (25°C). Cells closer to the bottom of the NAND string (i.e. WL #0) are more sensitive to temperature changes. The cell current marginally increases with temperature.

the NAND string have a higher temperature sensitivity. Measurements also show an increase in cell current at higher temperatures.

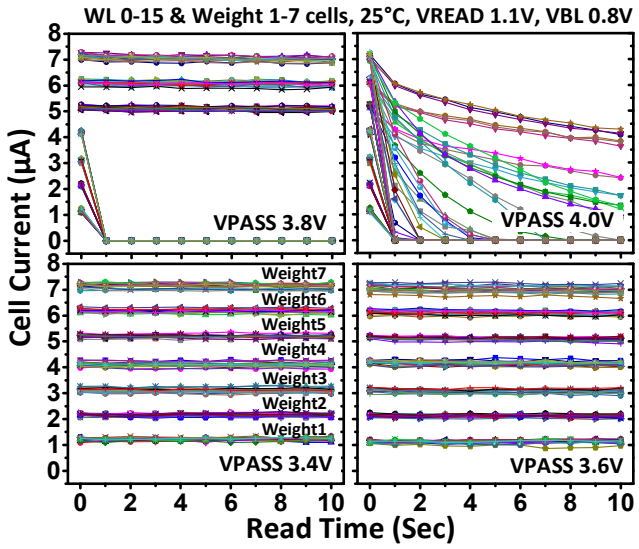


Fig. 8. VPASS disturbance characteristics of eNAND cells during read operation. Read currents of cells on different WLs are plotted for each weight level. The disturbance behavior is uniform across different WLs. Cell current remains constant for VPASS below 3.4V.

Fig. 8 shows the VPASS disturbance characteristics of the eNAND cells during read operation. Randomly selected weight 1 to 7 cells in each wordline were measured under different VPASS biases and read times. Test data shows negligible dependence on wordline location. The cell current remains constant below a VPASS voltage of 3.4V.

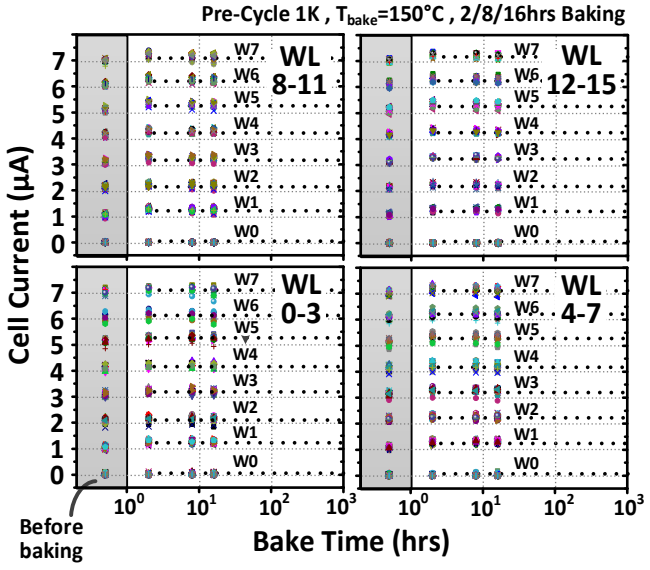


Fig. 9. Retention characteristics of weight 0 to weight 7 cells confirming minimal charge loss. Retention behavior is similar across different wordline groups suggesting good uniformity.

Fig. 9 shows the retention characteristics of weight 0 to 7 cells after baking the chip for 16 hours at 150°C. Before the baking test, all 2,560 cells were erased and programmed 1,000 times and then the cells were programmed with randomly chosen 3 bit weights. Test results show minimal charge loss in all wordline locations. The eNAND test chip was fabricated in a 65nm CMOS process, and the die microphotograph is shown in Fig.10.

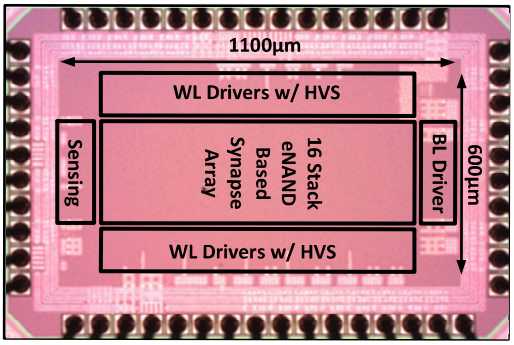


Fig. 10. 65nm logic-compatible eNAND test chip die photo

IV. Conclusion

This work presents reliability characterization results of a logic-compatible eNAND synapse array storing 3 bits per cell with 1µA current steps. A novel back pattern-tolerant verify scheme was employed to achieve accurate programming levels in a 16 stack eNAND string. Variation of the programmed cell current is less than 0.6µA over a temperature range of -10°C to 70°C, for all 2,560 measured cells. VPASS disturbance during read operation was not a concern for VPASS voltages below 3.4V. We verified that the cell currents remain unchanged after baking the chip for 16 hours at 150°C. Our extensive characterization results from the eNAND synapse array chip suggest that the proposed logic-compatible eNAND cells can achieve reliable 3 bit per cell weight storage with 1µA current steps.

References

[1] Z. Ye, R. Liu, H. Barnaby and S. Yu, "Evaluation of Single Event Effects in SRAM and RRAM based Neuromorphic Computing System for Inference," *2019 IEEE International Reliability Physics Symposium (IRPS)*, Monterey, CA, USA, 2019

[2] A. D. Patil, H. Hua, S. Gonugondla, M. Kang and N. R. Shanbhag, "An MRAM-based Deep In-Memory Architecture for Deep Neural Networks," *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, Sapporo, Japan, 2019

[3] H. Y. Cheng, W.C Chien, M. BrightSky, Y. H. Ho, Y. Zhu, A. Ray, R. Bruce, W. Kim, C. W. Yeh, "Novel fast-switching and high-data retention phase-change memory based on new Ga-Sb-Ge material," *2015 IEEE International Electron Devices Meeting (IEDM)*, Washington, DC, 2015, pp. 3.5.1-3.5.4.

[4] M. Kim, J. Kim, G. Park, L. Everson, H. Kim, S. Song, S. Lee and C. H. Kim, "A 68 Parallel Row Access Neuromorphic Core with 22K Multi-Level Synapses Based on Logic-Compatible Embedded Flash Memory Technology," *2018 IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, 2018, pp. 15.4.1-15.4.4.

[5] M. Kim, M. Liu, L. Everson, G. Park, Y. Jeon, S. Kim, S. Lee, S. Song and C. H. Kim, "A 3D NAND Flash Ready 8-Bit Convolutional Neural Network Core Demonstrated in a Standard Logic Process," *2018 IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, 2019, pp. 38.3.1-38.3.4.