JOURNAL OF SEMICONDUCTOR TECHNOLOGY AND SCIENCE, VOL. 22, NO. 5, OCTOBER, 2022 ISSN(Print) 1598-1657
https://doi.org/10.5573/JSTS.2022.22.5.341 ISSN(Online) 2233-4866

Impact of 3D NAND Current Variation on Inference Accuracy for In-memory Computing

Wonbo Shim

***Abstract*—3D NAND Flash has been proposed and investigated as a memory device candidate for the energy-efficient and ultra-high density compute-in-memory system. To achieve the acceptable accuracy for the inference applications, 3D NAND string current must be controlled precisely. However, there exist many challenging points which bothers the precise current control such as retention, temperature, pattern dependency in the cells of the 3D NAND string. In this work, we investigated the causes and effects of the 3D NAND string current variation and the resulted inference accuracy drop. The current variation drops the accuracy significantly so that the compensating design schemes must be implemented for the practical designs.**

***Index Terms*—Compute-in-memory, deep neural network, 3D NAND, variation**

I. INTRODUCTION

Recently, artificial intelligence has been emerging to successfully process the complex tasks such as image classification and language translation. The deep learning algorithms are using larger and deeper networks, so called deep neural networks (DNNs), to achieve higher accuracy. Such trend requires tremendous number of computation which may incur lots of power consumption for the computing hardware. Specifically, processing the DNN on the conventional von-Neumann computing architecture may suffer from the energy consumption due to the large number of data movements between the dynamic random access memory (DRAM) and the processing units.

Therefore, many kinds of novel computing architectures have been proposed to replace the von-Neumann architecture. Compute-in-memory (CIM) has been proposed as a powerful candidate with its high energy efficiency and throughput owing to its parallel computation in memory. CIM utilizes the nonvolatile memory (NVM) array for the computation. The weights of the DNNs are stored as a conductance of the memory cell. Then, multiply-and-accumulation (MAC) operations are conducted by activating the multiple rows of the array and reading out the summed current along the column. The results of the digitized analog current represent the output of the MAC computation.

Every nonvolatile memory device has been investigated as a candidate for the CIM applications. Emerging NVMs such as resistive random access memory (RRAM) [1, 2] and phase change memory (PCM) [3] have advantages in their multi-level characteristics and logic compatibility. But their small on and off current ratio make difficult to configuring large array size which may result in the low array efficiency of the chip. NOR Flash [4] could be another candidate thanks to its large on and off current ratio, however, recent NOR Flash technology has stopped at 28 nm technology node because of lots of fabrication challenges piled up.

Recently, 3D NAND Flash memory has been emerged as a CIM device candidate due to its ultra-high density and low on current [5, 6]. Incomparable high density of

Manuscript received Jun. 7, 2022; reviewed Aug. 31, 2022; accepted Oct. 1, 2022
Department of Electrical and Information Engineering, Seoul National University of Science and Technology, Seoul, Korea
E-mail : wbshim@seoultech.ac.kr

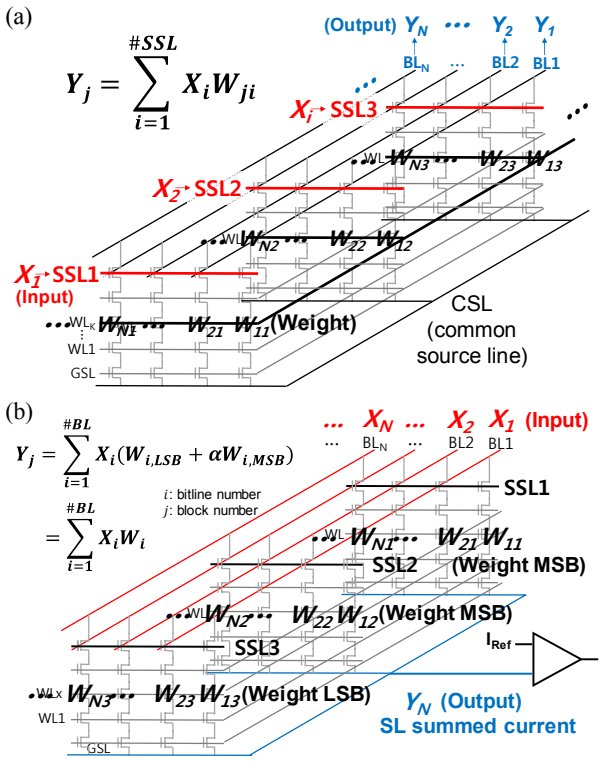

Fig. 1. 3D NAND based in-memory MAC operation scheme with (a) SSL input scheme; (b) BL input scheme.

3D NAND enables storing and processing huge DNNs within a single 3D NAND die [7]. The fabrication technology of 3D NAND is already matured and further developments are on-going. Moreover, its inherent low on current is advantageous for low power and parallel computation.

In this paper, we investigate the current variation of the 3D NAND and its impact on the accuracy of the inference application system. We first develop the current variation from the various sources of the 3D NAND in detail. Then, the electrical parameters are incorporated into the software simulation to evaluate the inference accuracy degradation.

II. 3D NAND BASED CIM ARCHITECTURE AND NONIDEAL EFFECTS ON 3D NAND STRING CURRENT

Fig. 1 shows the MAC operation schemes on 3D NAND architecture. The string select line (SSL) input scheme [8] is depicted in Fig. 1(a), where the input voltages corresponding to the input data are applied to

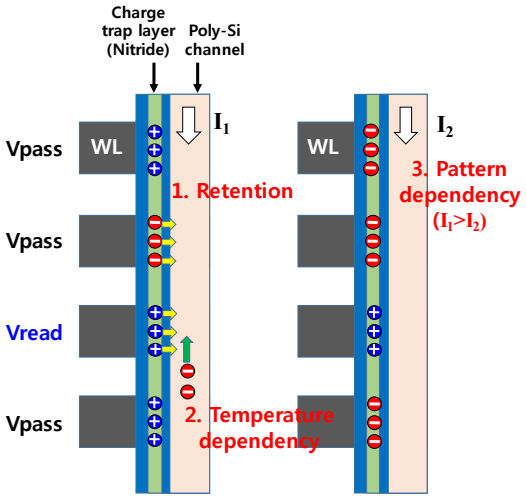

Fig. 2. Various current variation sources on 3D NAND string.

the string select lines (SSLs) while the string currents from the multiple cells are summed up along the bitline (BL). The analog-to-digital circuit (ADC) sense the aggregated currents and digitized results represent the output of MAC. Meanwhile, the BL input scheme [6] applies input voltage to the BLs and sense the output currents at source line (SL) as shown in Fig. 1(b). The significant bits (i.e. MSBs) of weights are duplicated to multiple SSLs. According to the applied input voltage at the bitlines, multi-bit weights can be processed at once. In terms of latency and energy efficiency, BL input scheme can take huge advantage when the network requires frequent input fetch. According to the target applications, the proper scheme must be exploited to achieve higher energy efficiency.

In spite of the advantages of the 3D NAND based CIM architecture, the inaccurate current of the memory cells may induce accuracy drop of the computation. Not different from other types of NVMs, 3D NAND string current would vary with various variation sources. The three critical factors that could affect the on-state current are shown in Fig. 2.

The retention characteristic of the charge trap flash (CTF) cells may cause serious impact on the on-state current. Due to the loss or gain of the trapped charges, i.e. electrons and holes, in the charge trap layer, threshold voltage of the individual cell may change over time. Not only the cells in the selected wordline to be read, but also the cells in the unselected wordlines that must be turned on during the read operation lose the charges over time

as well. Therefore, the string current amount would decrease or increase compared to immediately after being programmed.

The operating temperature at read operation may also affect the on-state current. Because the grain boundaries of the polysilicon channel incur the temperature dependency, the temperature rise would cause the on-state current increase of the 3D NAND channel.

The pattern dependency of the NAND string is another critical factor that affect the on-state current. Although the cells at different NAND strings which are sharing the same wordline are erased (holes trapped in the charge trap layer) as shown in Fig. 2, the programmed states on the unselected wordlines cause on-state current variation, i.e. I_1 is higher than I_2. The current variation could be mitigated with high pass voltage (Vpass), but it cannot be completely eliminated.

III. Simulation Results

In this section, we show the quantitative array-level simulation result of the 3D NAND string current variation and the respective inference accuracy degradation. To conduct the array-level simulation, we used SPICE simulation with the BSIM parameters extracted from the measured 3D NAND device characteristics [6] as indicated in Fig. 3. The conventional BSIM parameters was modified to describe the I-V characteristics and threshold voltage window of the programmed and erased CTF cells. Fig. 4 shows the I-V curve of on and off state 3D NAND string current with modified BSIM parameters which was used throughout this work. To evaluate the string current variation induced by various device level nonideal effects (i.e. Vt shift, temperature dependent mobility variation, program and erase pattern in NAND string), variations were reflected on the electron mobility and flat band voltage of the extracted BSIM parameters. Then, to evaluate the NAND string current variation from these nonideal effects with SPICE simulation, the netlist for the NAND array built with 32 WL NAND string was written up. A single array is composed of 128 NAND strings to be able to extract the standard deviation of the current variation with randomly written data, which means 32 CTF cells of every NAND string are randomly programmed (0 or 1). The string current variation can be

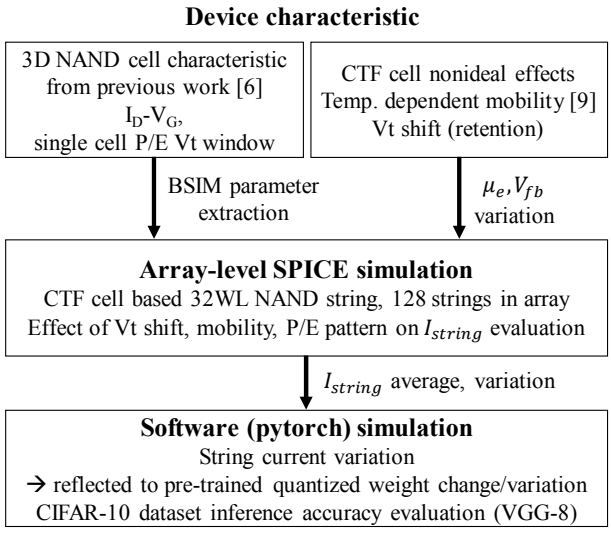

Fig. 3. Inference accuracy simulation flow of this work from device-level characterization to software simulation.

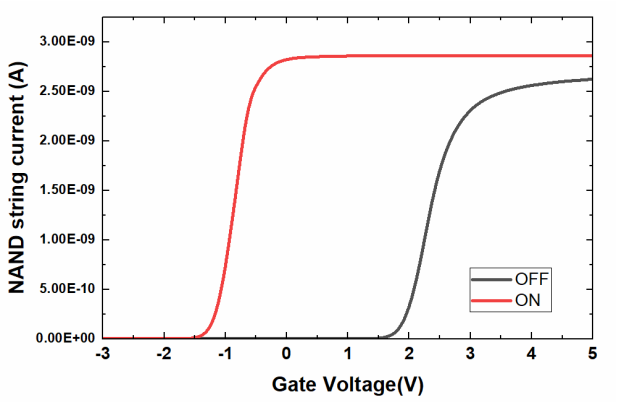

Fig. 4. Simulated I-V curve of on and off state 32 WL 3D NAND string current with modified BSIM parameters.

transferred to the change or variation of the pre-trained weight values. By incorporating the string current variation from SPICE simulation result into the weights of pre-trained model which are already quantized integer numbers, we evaluated the inference accuracy for the CIFAR-10 dataset on VGG-8 model. In the VGG-8 model, ReLU function was used as a nonlinear activation function. Throughout this work, the two-level input voltages (0 V and Vread) were used to represent binary input signals.

As mentioned in the previous section, charge loss or charge gain in the charge trapping layer over time change the threshold voltage (V_t) of the CTF cells. For the MAC operation on 3D NAND array, the on-state (erased) current plays dominant role for the accurate computation.

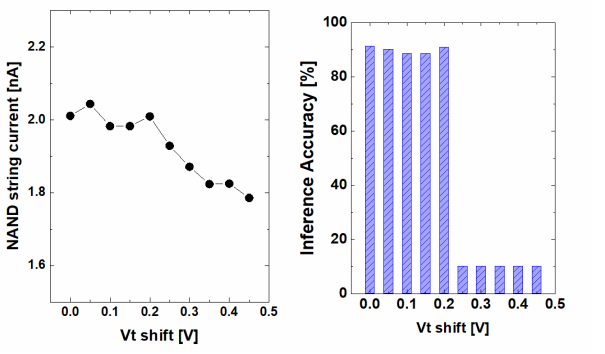

Fig. 5. (a) String current change with the threshold voltage shift of the erased cell; (b) CIFAR-10 inference accuracy with respect to the threshold voltage shift.

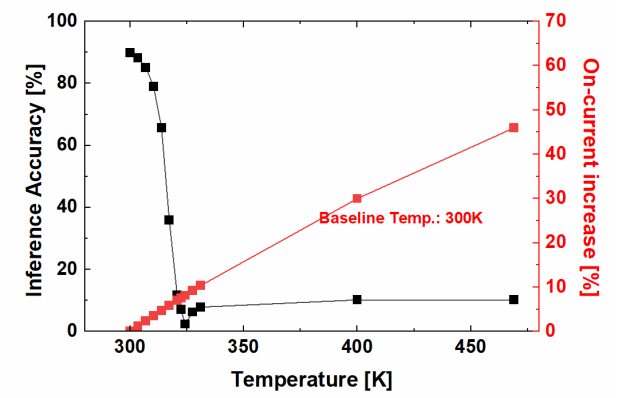

Fig. 6. Temperature dependency of the inference accuracy of 3D NAND based compute-in-memory system.

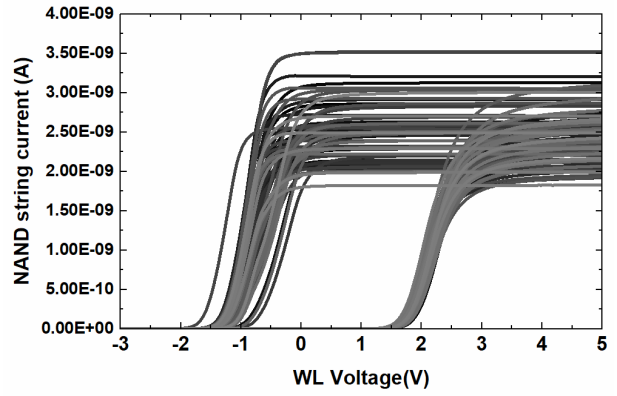

Fig. 7. Programmed pattern dependency at the 32 WL 3D NAND string current. 128 string currents are shown. The 64 cells on the selected WL are programmed and the other 64 cells are erased.

Therefore, the effect of the electron gain (or hole loss) of the on-state cells should be investigated. We assumed that all the erased cells gain or lose the same amount of charge at specific time which means the threshold voltage changes are identical for every erased cell. First, the 32 WL 3D NAND string current change with respect to the positive shift of the threshold voltage have been simulated as shown in Fig. 5(a). As the threshold voltage shift increases, the string current decrease gradually. 0.4 V of threshold voltage shift induces approximately 10 % current drop. We incorporated this current drop into the software simulation as weight mean value change to evaluate the inference accuracy drop of the CIFAR-10 dataset. Relatively high accuracy was maintained until 0.2 V of threshold voltage shift. However, accuracy drops to 10 % drastically with threshold voltage shift larger than 0.25 V. It can be noted that the threshold voltage shift must be carefully monitored not to exceed ~0.2 V.

The operating temperature dependency to the inference accuracy is shown in Fig. 6. Within the polysilicon channel, the grain boundaries impede the electron drift. The higher operating temperature helps to jump over this electrical barrier. From [9], the electron mobility at high temperature increases drastically, e.g., at 400 K, it is 30 % higher than at room temperature. If the sensing circuits are designed to operate at the room temperature near 300 K, the inference accuracy significantly drops at a temperature above 310 K if any temperature compensation schemes are not applied. It implies that the temperature compensation in the sensing circuits such as temperature dependent reference voltage or current must be implemented for the practical design.

The program and erase pattern dependency at the 32 WL 3D NAND string is shown in Fig. 7. Even though the 64 programmed cells and 64 erased cells have exactly same threshold voltage respectively, the randomly programmed pattern at the cells of unselected WLs affect the 3D NAND string current. It can be seen that both the threshold variation (-1 V to 0 V for erased cells) and the on-state current variation (2 nA to 3.5 nA for erased cells) exists. Because the on-state currents do not vary with higher than 1 V of WL voltage, the threshold voltage may not affect the total summed currents from multiple NAND strings. However, the on-state current variation directly affects the summed current amount so that the inference accuracy may be degraded. By applying the weight value variation with the on current standard variation results above, 84.5 % of inference accuracy for CIFAR-10 dataset was achieved which is

around 5 % lower than the case without on-state current variation.

IV. CONCLUSION

The NAND string current variation induced by various sources was simulated. Accordingly, inference accuracy would be significantly degraded as the impact of the variation source increases. From this work, we provide the guideline of NAND Flash based CIM hardware with accuracy point of view. Noticeable degradation was shown when the threshold voltage shift of erased cells in NAND string shifts larger than 0.25 V, or temperature increases 10 degrees. The randomly programmed data pattern also degrades the accuracy as well. In conclusion, further device engineering and circuit design techniques must be developed to compensate the inference accuracy degradation of the NAND Flash based CIM engine.

ACKNOWLEDGMENTS

This study was financially supported by Seoul National University of Science and Technology.

REFERENCES

[1] W. Wu et al, “A methodology to improve linearity of analog RRAM for neuromorphic computing,” Symposium on VLSI Technology, June, 2018, art. No. 8510690, pp. 103-104.

[2] P. Lin et al, “Three-dimensional memristor circuits as complex neural networks” Nature Electronics 3, 225-232(2020).

[3] S. Ambrogio et al, “Equivalent-accuracy accelerated neural-network training using analogue memory,” Nature 558, pp. 60-67, 2018.

[4] X. Guo, et al, "Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology," *IEEE International Electron Devices Meeting (IEDM),* San Francisco, CA, 2017, pp. 6.5.1-6.5.4.

[5] P. Wang et al, “Three-dimensional NAND Flash for vector-matrix multiplication,” *IEEE Trans. VLSI Systems*, vol. 27, no. 4, pp. 988-991, 2019

[6] H. -T. Lue et al, ″Optimal design methods to transform 3D NAND Flash into a high-density, high-bandwidth and low-power nonvolatile computing in memory (nvCIM) accelerator for deep-learning neural networks (DNN),” *IEEE International Electron Devices Meeting (IEDM),* San Francisco, CA, 2019, pp. 38.1.1-38.1.4.

[7] W. Shim et al, “Architectural design of 3D NAND Flash based compute-in-memory for inference engine,” *ACM/IEEE International Symposium on Memory Systems (MEMSYS),* pp.77-85. Oct. 2020, virtual.

[8] W. Shim et al, "Technological Design of 3D NAND-Based Compute-in-Memory Architecture for GB-Scale Deep Neural Network," in IEEE Electron Device Letters, vol. 42, no. 2, pp. 160-163, Feb. 2021.

[9] D. Resnati et al, “Characterization and Modeling of Temperature Effects in 3-D NAND Flash Arrays—Part I: Polysilicon-Induced Variability,” IEEE Transactions on Electron Devices, vol. 65, no. 8, pp. 3199-3206, Aug. 2018.

Wonbo Shim received the B.S., Ph.D. degrees in electrical engineering from the Seoul National University, Korea in 2007 and 2013, respectively. He is currently an assistant professor of Department of Electrical and Information Engineering with Seoul National University of Science and Technology. From 2013 to 2019, he was a Senior Engineer with the Samsung Electronics, Hwaseong, Korea, in flash design team at memory division. From 2019 to 2021, he was a Postdoctoral Research Fellow in Georgia Institute of Technology. His current research interest includes energy efficient deep learning architecture design, nonvolatile memory device modeling for synaptic applications, and 3D NAND Flash design.