IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 30, 2020, accepted August 10, 2020, date of publication August 20, 2020, date of current version September 1, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3018226

Low-Power Binary Neuron Circuit With Adjustable Threshold for Binary Neural Networks Using NAND Flash Memory

SUNG-TAE LEE, SUNG YUN WOO, (Member, IEEE), AND JONG-HO LEE, (Fellow, IEEE)
Department of Electrical and Computer Engineering and ISRC, Seoul National University, Seoul 08826, South Korea

Corresponding author: Jong-Ho Lee (jhl@snu.ac.kr)

This work was supported in part by the National Research Foundation of Korea under Grant NRF-2016M3A7B4909604, and in part by the Brain Korea 21 Plus Project in 2020.

ABSTRACT Recent studies have demonstrated that binary neural networks (BNN) could achieve a satisfying inference accuracy on representative image datasets. BNN conducts XNOR and bit-counting operations instead of high-precision vector-matrix multiplication (VMM), significantly reducing the memory storage. In this work, an analog bit-counting scheme is proposed to decrease the burden of neuron circuits with a synaptic architecture utilizing NAND flash memory. A novel binary neuron circuit with a double-gate positive feedback (PF) device is demonstrated to replace the sense amplifier, adder, and comparator, thereby reducing the burden of the complementary metal-oxide semiconductor (CMOS) circuits and power consumption. By using the double-gate PF device, the threshold voltage of the neuron circuits can be adaptively matched to the threshold value in the algorithms eliminating the accuracy degradation introduced by the process variation. Thanks to the super-steep *SS* characteristics of the PF device, the proposed neuron circuit with the PF device has an off-state current of 1 pA, representing 10^5 times improvement compared to the neuron circuit with a conventional metal-oxide-semiconductor field effect transistor (MOSFET) device. A system simulation of a hardware-based BNN shows that the low-variance conductance distribution (8.4 %) of the synaptic device and the adjustable threshold of the neuron circuit implement a highly efficient BNN with a high inference accuracy.

INDEX TERMS Neuromorphic, in-memory computing, hardware neural networks, neuron circuits, synaptic device, NAND flash memory.

I. INTRODUCTION

Recently, neuromorphic computing inspired by brain architecture has gained much interest because of its extremely low-power and massively parallel operations [1], [2]. In the von Neumann architecture, vector-matrix multiplication (VMM) causes enormous energy consumption due to the memory wall problem of data movement between memory and arithmetic units. On the other hand, neuromorphic computing resolves this problem, by computing vector-matrix multiplication (VMM) with a nonvolatile memory array in a single pulse step, overcoming the limit of the von Neumann bottleneck. To implement neuromorphic computing with nonvolatile memory, researchers have proposed implementing the analog conductance of synaptic devices [3], [4]. However, it is challenging to implement an accurate analog conductance state in a memory device due to the non-ideal analog conductance characteristics of the memory device [4], [5].

Recently, researchers have demonstrated that BNN can obtain a comparable inference fidelity to high-precision neural networks on various datasets, such as MNIST, CIFAR-10, and ImageNet [6]–[8]. The BNN dramatically reduces the memory storage and computing resource by binarized activation and weight [6]–[14]. Instead of a high-precision analog state, it allows a binary state of the memory device, which provides a practical way for the implementation of a hardware neural network system [8], [14].

In a neuromorphic system, 2T2R (two select transistors with two RRAMs) was mainly studied as a binary

The associate editor coordinating the review of this manuscript and approving it for publication was Junxiu Liu.

synapse [8], [14]. Recent high-performance DNN algorithms typically demand a large parameter size. Therefore, NAND flash memory can be a promising candidate for synaptic devices to meet this requirement. NAND flash memory offers ultra-high bit density for ample data storage and low fabrication cost per bit, and it has been well known as a mature technology [15]–[17]. In previous research, we reported neuromorphic systems utilizing NAND flash memory as a multi-level synapse for on-chip learning [18] and as a binary synaptic device for BNN digitally [19].

First, in this study, we propose an analog bit-counting scheme with a synaptic architecture utilizing NAND flash memory. The proposed analog bit-counting scheme replaces the digital sense amplifier, adder, and digital comparator with a binary neuron circuit, significantly reducing the CMOS overhead in the neuron circuits compared to a digital bit-counting scheme. A one-bit current sense amplifier (CSA) can serve as a neuron circuit to produce a binary neuron output in an ideal case. However, it may cause considerable inference accuracy degradation because the threshold of the binary neuron circuit can be different from the threshold value in the algorithms due to the process variation [8], [20]. In a previous study, an ADC-like multi-level sense amplifier (MLSA) was employed instead of a one-bit CSA to minimize the accuracy degradation [8]. However, the ADC-like MLSA requires a large CMOS overhead.

Second, for the first time, we propose a low-power binary neuron circuit with a PF device that has an adaptive-threshold to resolve the above problem. Note that the proposed binary neuron circuit serves as a low-power comparator with an adaptive-threshold function, which is different from the conventional integrate-and-fire neuron circuits. We demonstrate that the threshold voltage of the neuron circuits can be adaptively changed by the gate bias or program/erase pulse. Therefore, the proposed neuron circuit can eliminate the accuracy degradation introduced by the process variation without any CMOS overhead. In addition, the PF device based on a gated-thyristor has a super-steep subthreshold swing (*SS*) [21]–[23]. Finally, we show that the proposed neuron circuit with the PF device significantly reduces the off-state current of a neuron circuit compared to a neuron circuit with the conventional MOSFET device, thanks to the super-steep *SS* characteristics of the PF device.

II. BINARY SYNAPTIC ARCHITECTURE BASED ON NAND FLASH MEMORY

Fig. 1 (a) and (b) show a synaptic string array architecture consisting of a 2T2S synaptic string structure in a digital bit-counting scheme and an operating voltage scheme, respectively [19]. The 2T2S synaptic string consisting of two input transistors and two NAND strings is capable of XNOR operation. Two input voltages (V_{in1}, V_{in2}) are applied to each gate of the two input transistors which are reused for all synaptic devices in one synaptic string, therefore the number of input transistors is significantly decreased compared to the 2T2R scheme in a previous work [8]. A synaptic device consists

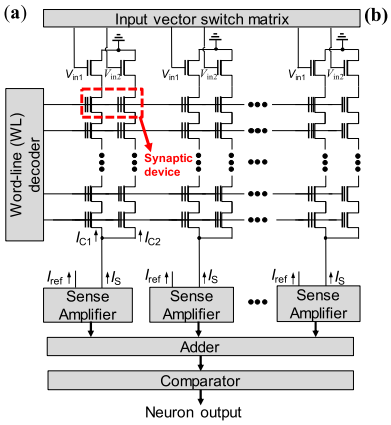

FIGURE 1. (a) Schematic diagram of the synaptic string array architecture based on a 2T2S synaptic string, which calculates bit-counting in a digital fashion. (b) Operating voltage scheme.

of two NAND cells whose complementary state defines the synaptic weight. As shown in Fig. 1 (b), a weight of $+1$ can be defined as the state where the right cell has a high threshold voltage ($V_{t,high}$), and the left cell has a low threshold voltage ($V_{t,low}$). In contrast, a weight of -1 can be defined as the reverse state of the two NAND cells. In addition, a complementary state of two input voltages (V_{in1}, V_{in2}) defines the input value. The state of (V_{on}, V_{off}) and (V_{off}, V_{on}) can represent an input value of $+1$ and -1, respectively, shown in Fig. 1 (b). By using the above scheme, the string current (I_S), which represents the XNOR output, is determined by the combination of the complementary input voltages and the state of the two adjacent NAND flash cells. The Fixed reference current (I_{REF}) of the sense amplifier is set to a value which is between the on-current (I_{on}) and off-current (I_{off}) of the NAND flash cells. In this scheme, the current sense amplifier compares the fixed reference current (I_{REF}) with a string current (I_S) which is the sum of the currents of the two NAND cells (I_{C1}, I_{C2}) to generate an XNOR output.

The word-line (WL) decoder applies the read bias (V_{read}) and pass bias (V_{PASS}) to a selected WL and unselected WLs, respectively. The input vector switch matrix applies input pulses to the input transistor. The adder sums the XNOR operation outputs, and the summed result goes through a binary comparator to produce a binary output. When V_{read} is imposed on the WL sequentially along the synaptic string, the output of each post-synaptic neuron is sequentially generated. Thus, the output of the k^{th} neuron in the post-synaptic neuron layer is generated when V_{read} is applied to the k^{th} WL. Because multipliers are not required, power consumption is enormously reduced.

We propose an analog bit-counting scheme with a 2T2S synaptic string by using a binary neuron circuit. When multiple currents are summed, the I_{on} dominates the total current (I_T), because the on/off resistance ratio of the NAND cells is sufficiently large [19]. For example, when the number of the synaptic string in the array is 256, the weighted sum of 0 can correspond to an I_T of 128 I_{on}s. When I_T is smaller than 128 I_{on}s, the binary neuron circuit generates an output

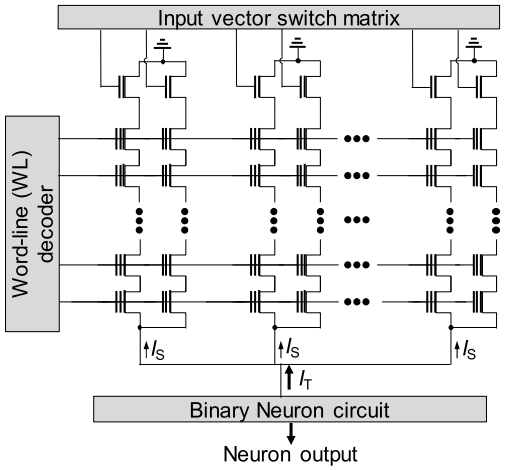

FIGURE 2. Schematic diagram of the synaptic string array architecture based on a 2T2S synaptic string, which calculates bit-counting in an analog fashion.

−1, which means there are more XNOR outputs of −1 than XNOR outputs of +1 in a row. In this scheme, the binary neuron circuit replaces the digital sense amplifiers, adder, and comparator shown in Fig. 2, which significantly decreases the power consumption and the burden of the circuits compared to the digital scheme in Fig. 1. Furthermore, the neuron circuit can be reused for all neurons in the neuron layer, therefore increasing the integration density compared to the previous work [8]. On the other hand, the process variation of the neuron circuit can reduce the inference accuracy in the proposed analog bit-counting scheme. In principle, a binary CSA can be used as a neuron to generate a binary output. However, process variation results in the intrinsic offset of the CSA, therefore the threshold of the neuron can be different from the target value in the algorithms. It makes sensing pass rate worse when the total current (I_T) from the synaptic array increases as the size of an array becomes large [20], which causes a significant accuracy degradation. In a previous study, an ADC-like multi-level sense amplifier (MLSA) was employed instead of a one-bit CSA to minimize the accuracy degradation [8]. However, the ADC-like MLSA requires an immense CMOS overhead. We propose a low-power binary neuron circuit with a double-gate PF device that adaptively changes the threshold voltage of the neuron circuit, which significantly reduces the accuracy degradation without the CMOS overhead.

III. BINARY NEURON CIRCUIT WITH THE PF DEVICE

Fig. 3 (a) and (b) show the 3-D schematic and top views of the fabricated PF device with a structure of a double-gate floating-body. The PF device has a cathode region (n^+-region), gated region (p-channel), ungated region (n-channel), and anode region (p^+-region) from the right shown in Fig. 3 (b). The n-channel and p-channel of which the doping concentration is $\sim 1 \times 10^{18}$ cm^{-3} serve as hole and electron injection barriers, respectively. The O/N/O stack of which the thickness is 2/4.2/9 nm is formed between the

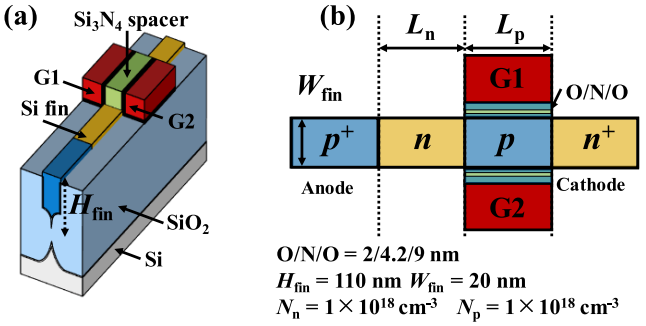

FIGURE 3. (a) 3-D schematic view and (b) Top view of the PF device.

p-channel and gate to store charge in the nitride (N) layer. Fig. 4 shows a TEM image of the fabricated PF device. The n^+ poly-Si double gates (G1, G2) are defined on the left and right sides of the Si_3N_4 spacer.

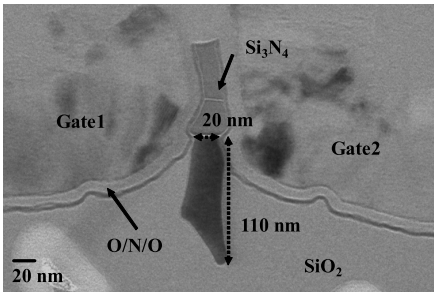

FIGURE 4. TEM image of the fabricated PF device.

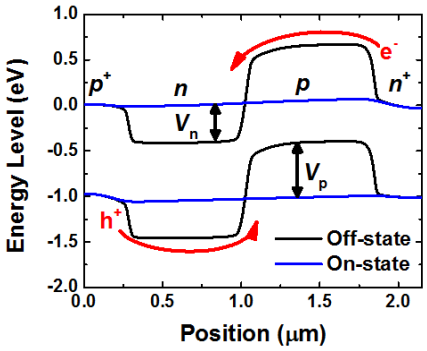

FIGURE 5. Energy band diagram of the PF device.

Fig. 5 represents the simulated energy band-diagram of the device to explain the positive feedback (PF) operation. The simulation is executed at a V_{G1} of 2 V and −1 V, which correspond to the turn-on and turn-off states, respectively, at a fixed V_{G2} of 0 V. In the turn-off state, the electron injection barrier (V_p) and hole injection barrier (V_n) impede the movement of the holes and electrons shown in Fig. 5. When the V_{G1} increases from −1 V to 2 V, the V_p decreases, which results in the injection of electrons from the n^+ region into the n region. It decreases V_n, which results in the injection of holes from the p^+ region into the p region further decreasing the V_p, and electrons flow into the n region again. When this PF process occurs, the device turns on rapidly with a steep SS.

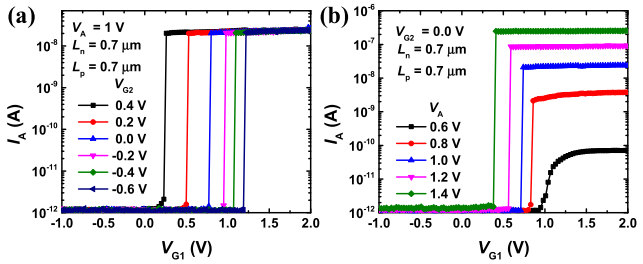

FIGURE 6. I_A-V_{G1} curves measured in the fabricated PF device as a parameter of (a) V_{G2} and (b) V_A. Here, the lengths of both the electron (L_n) and hole injection barriers (L_p) are 0.7 μm.

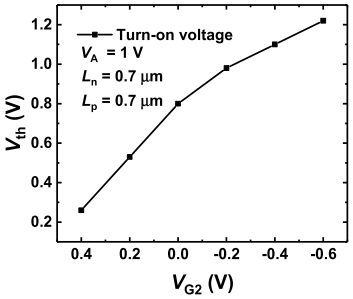

FIGURE 7. Change of V_{th} with V_{G2} at V_A of 1 V.

Fig. 6 (a) shows the anode current (I_A) versus G1 voltage (V_{G1}) curves measured in the fabricated PF device as a parameter of V_{G2}. As V_{G2} increases, the threshold voltage (V_{th}) decreases because the electron injection barrier (V_p) effectively decreases. Fig. 6 (b) shows the I_A-V_{G1} curves as a parameter of the anode bias (V_A). The built-in potential (V_{bi}) in the *p-n* junction impedes the current flow when the V_A is small. When V_A is larger than the V_{bi}, the I_{on} significantly increases. As V_A increases further, the carriers are generated in the reverse biased *p-n* junction and they accumulate in the *n* region and *p* region. As a result, the V_p and V_n decrease as the V_A increases, resulting in a decrease of V_{th}. Fig. 7 shows the change of the V_{th} with V_{G2} at a fixed V_A of 1 V. The threshold voltage can be modulated by changing the bias applied to the second gate (G2) shown in Fig. 7. Therefore, the threshold voltage of neuron circuits can be adaptively controlled, which significantly reduces the accuracy degradation introduced by the process variation.

Fig. 8 (a), (b) and (c) show the I_A-V_{G1} curves of the fabricated PF device as a parameter of the number of pulses when the device is programmed with V_{PGM}s of 7, 7.5 and 8 V, respectively. As the number of V_{PGM}s applied to G2 increases, more electrons are trapped in the charge trap layer of the PF device, which increases the concentration of holes in the *p* region. Then, the V_{th} of the PF device increases gradually. Fig. 8 (d) explains the changes of V_{th} with the number of pulses as a parameter of the pulse amplitude of V_{PGM}. As the pulse amplitude increases, V_{th} increases at the same number of pulses because more electrons are trapped in the charge trap layer. Therefore, the V_{th} of the neuron circuit can be changed by controlling the amplitude of V_{PGM} and the number of V_{PGM}s for the double-gate PF device.

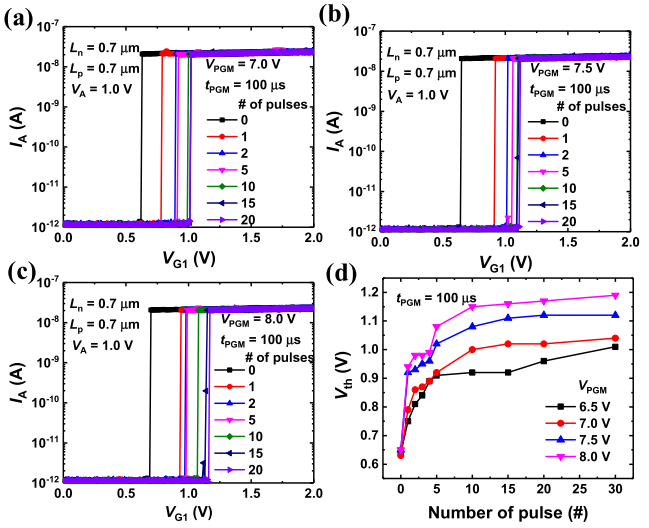

FIGURE 8. I_A-V_{G1} curves measured in the fabricated PF device as a parameter of the number of pulses when the V_{PGM} is (a) 7 V, (b) 7.5 V and (c) 8 V. (d) The change in V_{th} with the number of pulses as a parameter of the amplitude of V_{PGM}.

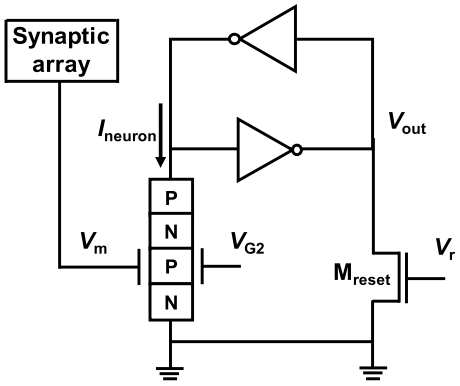

FIGURE 9. A schematic circuit diagram of the proposed binary neuron circuit using the double-gate PF device.

Fig. 9 shows a schematic diagram of the proposed binary neuron circuit using the double-gate PF device. The neuron circuit consists of the double-gate PF device, two invertors and one *n* MOSFET. The supply voltage (VDD) is 1.2 V. As current from synaptic array increases, the membrane voltage (V_m) increases. When the V_m exceeds the V_{th} of the neuron circuit, then the on-current flows in the neuron circuit and the output voltage (V_{out}) of the neuron circuit becomes VDD, which can be regarded as a binary output of +1. Then, V_{out} is initialized to 0 V by applying the reset pulse (V_r) to reset-MOSFET (M_{reset}).

Fig. 10 (a) and (b) show the transient waveforms of the neuron circuit as a parameter of the membrane voltage (V_m) and V_{th} of the neuron circuit, respectively. In Fig. 10 (a), the V_{th} of the double-gate PF device is fixed at 0.7 V by controlling V_{G2} or the program/erase operation. V_m increases from V_{m4} to V_{m1}. When V_{m1} exceeds the fixed V_{th} of 0.7 V, the V_{out1} becomes VDD. In Fig. 10 (b), the V_m is fixed at 0.55 V (V_{m1}). Then, V_{th} decreases from V_{th1} to V_{th3}. Because V_{th3} is lower than V_{m1}, V_{out} becomes VDD (V_{out} at V_{th3}).

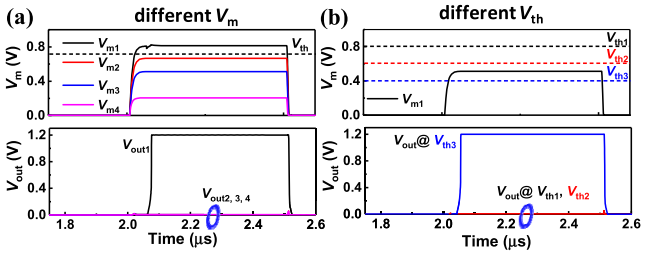

FIGURE 10. Simulated transient results of the neuron circuit as parameters of (a) the membrane voltage (V_m) and (b) V_{th} of the neuron circuit.

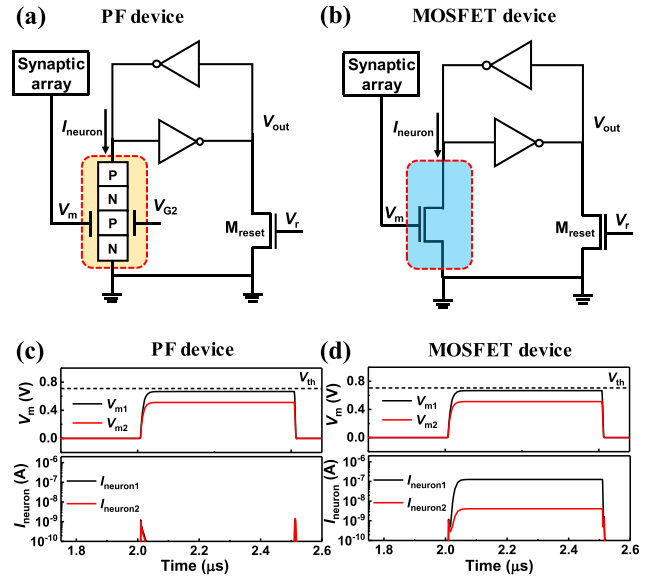

FIGURE 11. Schematic diagrams of the proposed neuron circuits using (a) double-gate PF device and (b) conventional MOSFET. Transient waveforms of the membrane potential, and current of the neuron device in the neuron circuits using (c) double-gate PF device and (d) conventional MOSFET to compare the power consumption in the off-state (V_{out} = 0 V).

Fig. 11 (a) and (b) show schematic diagrams of neuron circuits using the double-gate PF device and conventional MOSFET, respectively, for a comparison of the power consumption in the off-state (V_{out} = 0 V) of the neuron circuit. Fig. 11 (c) and (d) represent the membrane voltage (V_m), and the current of the neuron device in the neuron circuits using the double-gate PF device and conventional MOSFET, respectively. When V_m is lower than the V_{th}, the PF device having a steep switching characteristic shows a very low I_{off} (∼1 pA) during the read operation of the synaptic arrays shown in Fig. 11 (c). On the other hand, in the neuron circuit using the conventional MOSFET, the subthreshold current (∼ 100 nA) of the *n* MOSFET flows during the read operation of the synaptic arrays shown in Fig. 11 (d). Therefore, during the off-state of the neuron circuit, the neuron circuit with the PF device significantly reduces the power consumption compared to the neuron circuit with a conventional MOSFET.

The effect of synaptic device variation on the inference accuracy of a hardware-based BNN is investigated. Fig. 12 (a) and (b) show the inference accuracy with the weight variation of the synaptic devices on the MNIST and

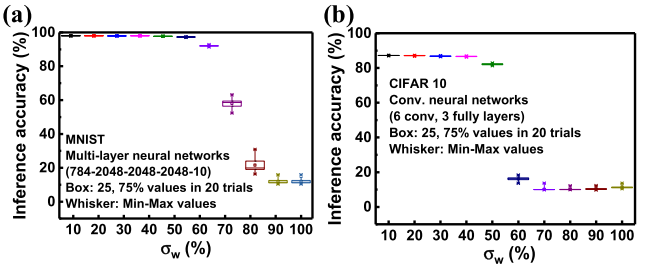

FIGURE 12. Simulated classification accuracy with respect to the conductance variation of synaptic devices in the case of (a) MNIST and (b) CIFAR 10.

CIFAR 10 images, respectively. A weight variation occurs when the weights obtained in the off-chip training are transferred to the synaptic devices. The variation of the conductance of the NAND cells is assumed to follow the Gaussian distribution [24]. The effect of device variation is more detrimental to convolution neural networks classifying CIFAR 10 than the multi-layer neural networks classifying MNIST. Note that little decrease in accuracy is observed when the sigma (σ_w) of the synaptic weight variation is within about 40%. As noted in previous work [19], when the V_{PGM} is 16 V, the sigma over mean (σ/μ) of the conductance of the NAND cells in an array is about 8.4 %. Therefore, BNN using NAND flash cells as synaptic devices is very robust to the effect of device variation.

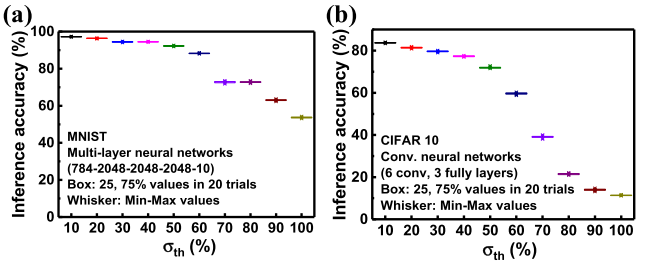

FIGURE 13. Simulated classification accuracy with respect to the threshold voltage variation of the neuron circuit for the (a) MNIST and (b) CIFAR 10 datasets.

The effect of the V_{th} variation in the neuron circuits on the inference accuracy is also investigated. Fig. 13 (a) and (b) show the effect of the variation of V_{th} on the inference accuracy of the MNIST and CIFAR 10 datasets, respectively. The threshold voltage variation (σ_{th}) of the neuron circuits is assumed to follow a Gaussian distribution. The classification accuracy for the MNIST and CIFAR10 datasets decreases significantly as the sigma (σ_{th}) of the threshold voltage in neuron circuits increases by ∼60% and ∼50%, respectively. In particular, the classification accuracy for the CIFAR10 dataset decreases more severely as σ_{th} increases above ∼50%. The threshold voltage transferred to the binary neuron circuit can be different from the threshold value in the algorithm due to process variation or variation in the transfer process. By using the double-gate PF device, the V_{th} can be matched to the threshold value in the algorithms by controlling the gate bias or the program/erase pulse shown in Figs. 6 and 8. Therefore, the proposed binary neuron circuit consisting of 6 transistors

can eliminate the accuracy degradation without CMOS overhead compared to the ADC-like MLSA [8]. Comparing the inference accuracy within 40% sigma (σ_w, σ_{th}) in Figs. 12 and 13 shows that the V_{th} variation in the neuron circuits has a greater effect on the inference accuracy than the weight variation in synaptic arrays in BNN. Therefore, in this work, it can be said that the proposed neuron circuit with the PF device capable of controlling the V_{th} accurately improves the inference accuracy while reducing the power consumption.

IV. CONCLUSION

An analog bit-counting scheme has been proposed to decrease the burden of neuron circuits in a binary neural network with a synaptic architecture utilizing NAND flash memory compared to the digital bit-counting scheme. A novel binary neuron circuit with a double-gate positive feedback (PF) device was proposed to replace the sense amplifier, adder, and comparator, thereby decreasing the power consumption and the burden of the CMOS circuits. The proposed neuron circuit consisting of 6 transistors, including the double-gate PF device, eliminates accuracy degradation without additional CMOS overhead compared to a multi-level sense amplifier. The V_{th} variation of the neuron circuits was more detrimental to the inference accuracy compared to the weight variation of the synaptic devices up to 40 % sigma (σ_w, σ_{th}). By controlling the gate bias or program/erase pulse for the double-gate PF device, we demonstrate that the threshold voltage of the neuron circuits can be adaptively matched to the threshold value in the algorithms. Thanks to the super-steep *SS* characteristics of the PF device, the proposed neuron circuit with the PF device significantly reduces the off-state current ($\sim$1 pA) of the neuron circuit compared to the neuron circuit with the conventional MOSFET device (I_{off} $\sim$100 nA). Note that, to accommodate a vast volume of parameters and a large network size required in recent neural networks, high-density NAND flash and the proposed neuron circuit are promising candidates for a neuromorphic system. Therefore, practical realization of hardware neural networks consisting of NAND flash memory and neuron circuits needs to be demonstrated and requires further study. The proposed binary neuron circuit with a synaptic device utilizing NAND flash memory in this work can show the feasibility of energy-efficient and high-density neuromorphic hardware with a high inference accuracy.

ACKNOWLEDGMENT

(Sung-Tae Lee and Sung Yun Woo contributed equally to this work.)

REFERENCES

[1] B. L. Jackson, B. Rajendran, G. S. Corrado, M. Breitwisch, G. W. Burr, R. Cheek, K. Gopalakrishnan, S. Raoux, C. T. Rettner, A. Padilla, A. G. Schrott, R. S. Shenoy, B. N. Kurdi, C. H. Lam, and D. S. Modha, "Nanoscale electronic synapses using phase change devices," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 9, no. 2, pp. 1–20, May 2013.

[2] M. Suri, O. Bichler, D. Querlioz, O. Cueto, L. Perniola, V. Sousa, D. Vuillaume, C. Gamrat, and B. DeSalvo, "Phase change memory as synapse for ultra-dense neuromorphic systems: Application to complex visual pattern extraction," in *IEDM Tech. Dig.*, Dec. 2011, pp. 4.4.1–4.4.4.

[3] S. Park, A. Sheri, J. Kim, J. Noh, J. Jang, M. Jeon, B. Lee, B. R. Lee, B. H. Lee, and H. Hwang, "Neuromorphic speech systems using advanced ReRAM-based synapse," in *IEDM Tech. Dig.*, Dec. 2013, pp. 6–25.

[4] G. W. Burr, R. M. Shelby, S. Sidler, C. di Nolfo, J. Jang, I. Boybat, R. S. Shenoy, P. Narayanan, K. Virwani, E. U. Giacometti, B. N. Kurdi, and H. Hwang, "Experimental demonstration and tolerancing of a large-scale neural network (165 000 synapses) using phase-change memory as the synaptic weight element," *IEEE Trans. Electron Devices*, vol. 62, no. 11, pp. 3498–3507, Nov. 2015.

[5] P.-Y. Chen, B. Lin, I.-T. Wang, T.-H. Hou, J. Ye, S. Vrudhula, J.-S. Seo, Y. Cao, and S. Yu, "Mitigating effects of non-ideal synaptic device characteristics for on-chip learning," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, Nov. 2015, pp. 194–199.

[6] L. Jiang, M. Kim, W. Wen, and D. Wang, "XNOR-POP: A processing-in-memory architecture for binary convolutional neural networks in wide-IO2 DRAMs," in *Proc. IEEE/ACM Int. Symp. Low Power Electron. Design (ISLPED)*, Jul. 2017, pp. 1–6.

[7] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks," in *Proc. 30th Conf. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 4107–4115.

[8] X. Sun, S. Yin, X. Peng, R. Liu, J.-S. Seo, and S. Yu, "XNOR-RRAM: A scalable and parallel resistive synaptic architecture for binary neural networks," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2018, pp. 1423–1428.

[9] R. Andri, L. Cavigelli, D. Rossi, and L. Benini, "YodaNN: An architecture for ultralow power binary-weight CNN acceleration," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 37, no. 1, pp. 48–60, Jan. 2018.

[10] S. Yu, "Neuro-inspired computing with emerging nonvolatile memorys," *Proc. IEEE*, vol. 106, no. 2, pp. 260–285, Feb. 2018.

[11] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or −1," Mar. 2016, *arXiv:1602.02830*. [Online]. Available: http://arxiv.org/abs/1602.02830

[12] T. Tang, L. Xia, B. Li, Y. Wang, and H. Yang, "Binary convolutional neural network on RRAM," in *Proc. 22nd Asia South Pacific Design Autom. Conf. (ASP-DAC)*, Jan. 2017, pp. 782–787.

[13] Y. Guan and T. Ohsawa, "Co-design of DNN model optimization for binary ReRAM array in-memory processing," in *Proc. IEEE 11th Int. Memory Workshop (IMW)*, May 2019, pp. 1–4.

[14] Z. Zhou, P. Huang, Y. C. Xiang, W. S. Shen, Y. D. Zhao, Y. L. Feng, B. Gao, H. Q. Wu, H. Qian, L. F. Liu, X. Zhang, X. Y. Liu, and J. F. Kang, "A new hardware implementation approach of BNNs based on nonlinear 2T2R synaptic cell," in *IEDM Tech. Dig.*, Dec. 2018, pp. 20.7.1–20.7.4.

[15] R. Yamashita *et al.*, "A 512 Gb 3b/cell flash memory on 64-word-line-layer BiCS technology," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2017, pp. 196–197.

[16] D. Kang *et al.*, "A 512 Gb 3-bit/cell 3D 6^{th}-generation V-NAND flash memory with 82 MB/s write throughput and 1.2 Gb/s interface," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, pp. 216–218, 2019.

[17] H. Huh *et al.*, "A 1 Tb 4b/cell 96-stacked-WL 3D NAND flash memory with 30 MB/s program throughput using peripheral circuit under memory cell array technique," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2020, pp. 220–221.

[18] S.-T. Lee, S. Lim, N. Choi, J.-H. Bae, C.-H. Kim, S. Lee, D. H. Lee, T. Lee, S. Chung, B.-G. Park, and J.-H. Lee, "Neuromorphic technology based on charge storage memory devices," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2018, pp. 169–170.

[19] S.-T. Lee, H. Kim, J.-H. Bae, H. Yoo, N. Y. Choi, D. Kwon, S. Lim, B.-G. Park, and J.-H. Lee, "High-density and highly-reliable binary neural networks using NAND flash memory cells as synaptic devices," in *IEDM Tech. Dig.*, Dec. 2019, pp. 38.4.1–38.4.4.

[20] C.-P. Lo, W.-Z. Lin, W.-Y. Lin, H.-T. Lin, T.-H. Yang, Y.-N. Chiang, Y.-C. King, C.-J. Lin, Y.-D. Chih, T.-Y. J. Chang, M.-S. Ho, and M.-F. Chang, "Embedded 2 Mb ReRAM macro with 2.6 ns read access time using dynamic-trip-point-mismatch sampling current-mode sense amplifier for IoE applications," in *Proc. Symp. VLSI Circuits*, 2017, pp. C164–C165.

[21] J. Wan, C. L. Royer, A. Zaslavsky, and S. Cristoloveanu, "A systematic study of the sharp-switching Z2-FET device: From mechanism to modeling and compact memory applications," *Solid-State Electron.*, vol. 90, pp. 2–11, Dec. 2013.
[22] Y. Jeon, M. Kim, D. Lim, and S. Kim, "Steep subthreshold swing n- and p-channel operation of bendable feedback field-effect transistors with $p^+ - i - n^+$ nanowires by dual-top-gate voltage modulation," *Nano Lett.*, vol. 15, no. 8, pp. 4905–4913, Aug. 2015.
[23] K. B. Choi, S. Y. Woo, W.-M. Kang, S. Lee, C.-H. Kim, J.-H. Bae, S. Lim, and J.-H. Lee, "A split-gate positive feedback device with an integrate-and-fire capability for a high-density low-power neuron circuit," *Frontiers Neurosci.*, vol. 12, no. 704, pp. 1–13, Oct. 2018.
[24] S.-T. Lee, S. Lim, N. Y. Choi, J.-H. Bae, D. Kwon, B.-G. Park, and J.-H. Lee, "Operation scheme of multi-layer neural networks using NAND flash memory as high-density synaptic devices," *IEEE J. Electron Devices Soc.*, vol. 7, pp. 1085–1093, 2019.

SUNG-TAE LEE received the B.S. degree in electrical and computer engineering from Seoul National University (SNU), Seoul, South Korea, in 2016, where he is currently pursuing the combined master's and Ph.D. degree. He is also with the Inter-University Semiconductor Research Center, SNU. His current research interests include neuromorphic systems and its application in computing.

SUNG YUN WOO (Member, IEEE) received the B.S. degree in electrical engineering from Kyungpook National University, in 2014. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Seoul National University (SNU), Seoul, South Korea. He is also with the Inter-University Semiconductor Research Center, SNU. His current research interests include neuromorphic systems and neural networks.

JONG-HO LEE (Fellow, IEEE) received the Ph.D. degree in electronic engineering from Seoul National University (SNU), Seoul, in 1993. He was a Postdoctoral Fellow with the Massachusetts Institute of Technology, Cambridge, MA, USA, from 1998 to 1999. He has been a Professor with the School of Electrical and Computer Engineering, SNU, since 2009.