IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 27, 2020, accepted November 13, 2020, date of publication December 1, 2020,
date of current version December 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3041734

Double-Gated Asymmetric Floating-Gate-Based Synaptic Device for Effective Performance Enhancement Through Online Learning

DONGHYUN RYU, TAE-HYUNG KIM, (Graduate Student Member, IEEE), TAEJIN JANG, (Graduate Student Member, IEEE), JUNSU YU, (Student Member, IEEE), JONG-HO LEE, (Fellow, IEEE), AND BYUNG-GOOK PARK, (Fellow, IEEE)
Inter-University Semiconductor Research Center, Department of Electrical and Computer Engineering, Seoul National University, Seoul 08826, South Korea

Corresponding author: Byung-Gook Park (bgpark@snu.ac.kr)

This work was supported in part by the Brain Korea 21 Plus Project in 2020, and in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) funded by the Korea government (MSIT) under Grant 2020-0-01294.

ABSTRACT In this paper, we propose a floating-gate-based synaptic transistor with two independent control gates that implement both offline and online learning. Unlike previous research on double-gated synaptic transistors, the proposed device is capable of online learning without facing a fan-out problem. Basic operation of the device was verified and a program/erase scheme based on Fowler-Northeim tunneling is suggested for the multi-conductance utilization of the synaptic device. With the proposed P/E scheme, an offline-trained single-layered hardware-based spiking neural network was simulated for MNIST classification, resulting in 87.37% classification accuracy with 10% conductance variation. To alleviate this performance degradation, the online learning method is applied on the offline-trained SNN by reusing 3,000 training images. The effectiveness of the proposed method is also verified under existence of the synaptic weight variance. As a result, up to 86.89% of the performance degradation is alleviated.

INDEX TERMS CMOS, flash memory, synaptic device, neuromorphic system, offline learning, online learning.

I. INTRODUCTION

Neuromorphic systems are rising candidates for the next generation computing system due to their massively parallel data processing capability and minimal power consumption [1]–[8]. Various researchers have implemented neuromorphic systems using their unique methods and performing machine learning tasks such as pattern recognition or image denoising [9]–[13]. Neuromorphic systems consist of neuron circuits and synaptic devices, and their implementation differ depending on the specific combination of incorporated circuits and devices. The most widely used neuron circuits include integrate-and-fire (IF) neurons, a simplified model of a biological neuron that integrates current in a membrane capacitor and generates an action potential when the membrane voltage exceeds the threshold [14]–[19]. IF neurons receive and transmit a signal in various forms such as left-justified encoding or Poisson encoding [20]–[22]. Behavior of IF neurons are proven to be equivalent to rectified linear unit (ReLU) activation function of non-SNNs, making offline learning possible by weight transfer from weights that are calculated from an external computer. Candidates for synaptic devices include a flash memory as well as emerging memory devices such as resistive random-access memory (RRAM), phase change random access memory (PCRAM), and ferroelectric tunnel junction (FTJ) [23]–[30]. Both gradual switching devices and abrupt switching devices are used as synapses. Although it is more effective to use gradual switching synaptic devices, which represent continuous synaptic weight in single memory cell, it is also possible to implement one synaptic weight with multiple single-bit devices [31]–[33]. There were studies that implemented a neuromorphic system with widely used flash memory as synaptic devices [34]–[37]. Such studies include fabricating a hardware-based neural network, which conducts vector-matrix-multiplication in a NOR flash array or implementing a binarized neural network by conducting an XNOR operation on a NAND flash [38], [39]. However, there is

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

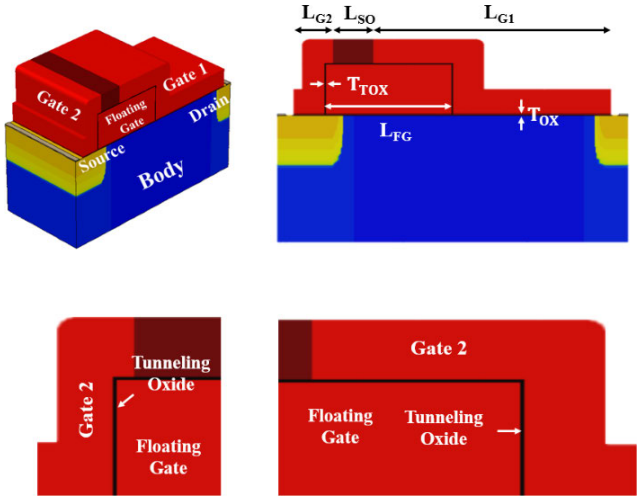

FIGURE 1. (a) 3-dimensional structure of the proposed double gate synaptic device and its (b) cross-sectional view, (c) enlarged view of gate2 region and (d) enlarged view of gate1 region.

a downside in using conventional flash arrays since require complex external controllers to implement online learning. For this reason, some researchers devised a double-gated synaptic transistor to easily implement online learning such as spike timing-dependent plasticity [15], [40]. The double-gated synaptic transistor mentioned above can realize online learning and lifelong learning, which are the major advantages of hardware-based neuromorphic systems. However, they induce problems in system operation. Conventional double-gated synaptic devices use output pulse of presynaptic neuron as current source to prevent drain current flowing when only teaching signal is given. However, this method makes the presynaptic neuron to drive an excessive amount of current, causing a fan-out problem [37]–[42]. For example, transmitting spike to one thousand synaptic devices which operate at 1uA each, will require the presynaptic neuron to drive 1mA of current.

Therefore, in this research, we propose a double-gated synaptic transistor capable of online learning without facing a fan-out problem. We handle the fan-out problem by connecting the drain to voltage source. We also prevent the drain current flowing at teaching signal pulse by adopting asymmetrically shaped gates. In the following sections, we analyze the basic operations of the proposed device and the results for neuromorphic system operations when the device is used as a synapse. Moreover, a new learning technique which utilize both online and offline learning is proposed to minimize performance degradation of the neuromorphic system.

II. SIMULATION RESULTS FOR SYNAPTIC DEVICE

A. DEVICE STRUCTURE

Fig. 1 presents three-dimensional and cross-sectional views of proposed double-gated synaptic transistor. Gate 1 length (L_{G1}), gate 2 length (L_{G2}), and floating-gate length (L_{FG}) are set to be 1.5 μm, 0.25 μm, and 0.8 μm, respectively. Gate1 and gate2 are separated by 0.25 μm of separating oxide

TABLE 1. Parameters for the proposed synaptic device.

	Value		Value
Gate 1 Length (L_{G1})	1.5 [um]	Overlap Length (L_{OV})	60 [nm]
Gate 2 Length (L_{G2})	0.25 [um]	Drain junction Length (L_{DOV})	0.36 [um]
Thickness of Tunneling Oxide (T_{TOX})	5 [nm]	Source junction Length (L_{SOV})	0.16 [um]
Thickness of gate dielectric Oxide (T_{OX})	15 [nm]	Separating Oxide Length (L_{SO})	0.25 [um]
Floating Gate Length (L_{FG})	0.8 [um]		

(L_{SO}). The gate-to-S/D overlap length (L_{OV}) is 0.36 um for the drain region and 0.16 μm for the source region, which ensures effective channel control. The thickness of the tunneling-oxide (T_{ox}) and the gate dielectric oxide are fixed to 5 nm and 15 nm, respectively. Doped poly-silicon is used for gate1 and gate2 for effective program/erase and online learning. The doping concentration is 5×10^{20} cm^{-3} of phosphorus in the S/D region and 1×10^{17} cm^{-3} of boron in the silicon box, and the graded doping profile is adopted to account for the realistic fabrication conditions. All of the structural parameters are specified in Table 1.

B. SIMULATION CONDITION

We analyzed the device characteristics through simulation. To increase the accuracy of simulation, we carefully calibrated the entire simulation conditions using measurement data from previous studies. Device simulation was conducted using Synopsys Sentaurus 3D technology computer-aided design (TCAD) simulation. Referring [47], all of the physics and parameters are calibrated to implement the actual program/erase (PGM/ERS) of a memory device and the operation of a complementary metal–oxide–semiconductor (CMOS) device (Fig. 2). Bandgap narrowing as well as Shockley-Read-Hall (SRH) recombination models are used and the mobility properties are also considered by using Philips and Lombardi models. Quantum potential and Fermi-Dirac models are included to consider the density gradient quantization and carrier density. The hydrodynamic carrier transport model is adopted as a carrier transport model. We can check that the simulation results, performed with the above physics, fits well with the data from [48]. To simulate the memory characteristics, we calibrated the electron tunneling mass of 5 nm tunneling-oxide using [49]–[51] as reference. The electron tunneling mass is tuned to 0.35 m_0, which shows a well matched threshold voltage (V_{th}) shift.

C. DEVICE OPERATION AND PGM/ERS CHARACTERISTICS

Synaptic transistors require two or more gates to perform online learning. Online learning is implemented using potential difference between the input signal and the feedback

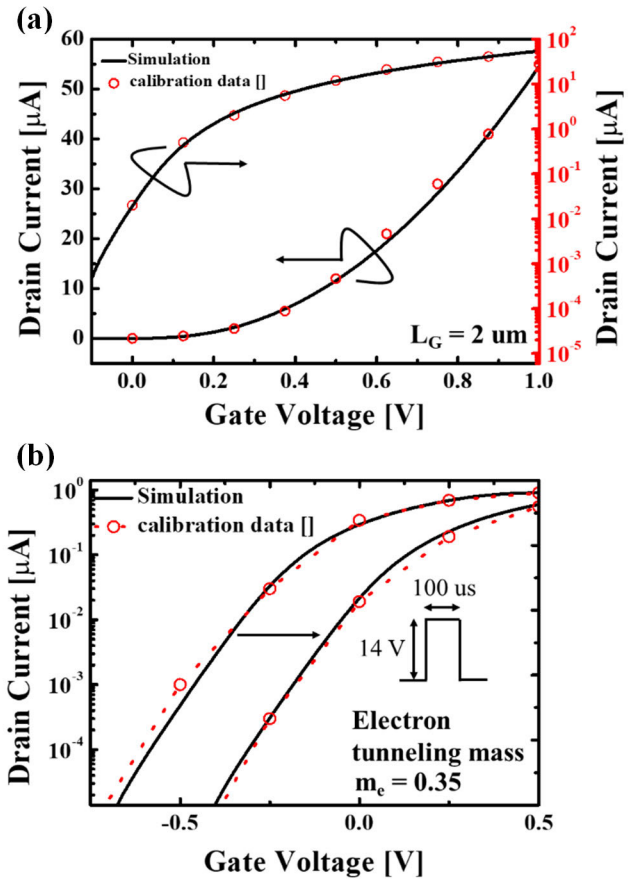

FIGURE 2. Simulated transfer curves of parameters calibrated using measurement data. (a) Parameter and physics calibration process data from [44]. (b) Memory characteristic calibration process data from [45]–[47].

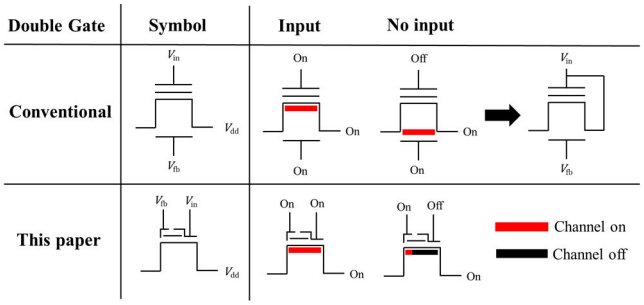

FIGURE 3. Diagram illustrating operation of conventional and proposed double-gated synaptic transistor.

signal. However, malfunction may occur if a channel is formed by a feedback signal. Therefore, V_{in} and V_{dd} are connected to ensure that drain current flow only when input signal is given as shown in Fig. 3. However, in this case, output spikes of presynaptic neurons operate as current sources inducing extreme fan-out problems. In neuromorphic systems, this problem becomes even more serious as numerous synaptic devices are connected in parallel since R_L becomes far lower than R_{out} (Fig. 4).

In the proposed device, the role of gate 1 and gate 2 are different. Coupling ratio between gate 1 and the floating-gate

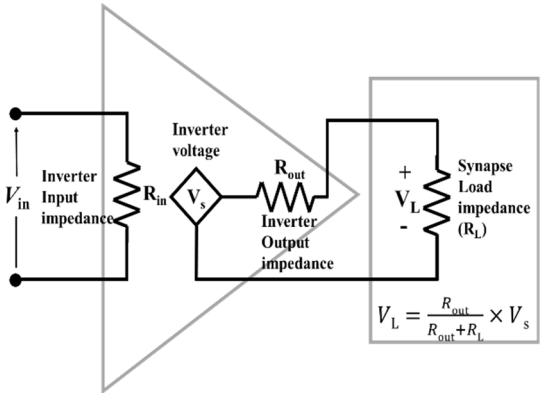

FIGURE 4. Schematic of neuron circuit considering each impedance.

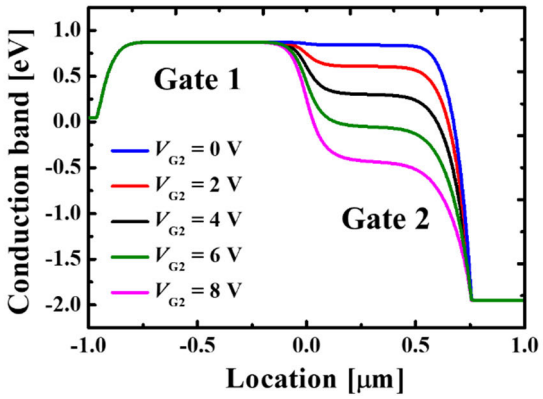

FIGURE 5. Energy band diagrams along the source-channel-drain with respect to different gate2 bias. 0V is applied to gate1 and 2V is applied to drain. Note that source-side energy barrier remains constant disregarding changes in gate 2 bias.

is larger than that of gate 2 and the floating-gate. Therefore, gate 1 dominantly controls the on/off operation of the device. Due to the different coupling ratio, the voltage across silicon dioxide between gate 2 and the floating-gate is larger than that of gate 1 and the floating-gate. Therefore, gate 2 becomes the charge source for programming on the floating-gate by Fowler-Nordheim (FN) tunneling. One important characteristic of the proposed device is that the drain current does not flow with a feedback signal even if the drain is connected to a voltage source. The reason for this is the existence of an area under gate 1, which maintains a source side energy barrier regardless of the voltage applied to gate 2 (Fig. 3). It can be verified in Fig. 5, which is the energy band diagrams of the channel with respect to different gate 2 voltages. The characteristics explained above play an important role in preventing a malfunction in online learning, where a signal is given to both gate 1 and gate 2. Without such a characteristic, the neuron will fire at an excessive rate when a large feedback signal is given to gate 2.

The PGM/ERS operation of the proposed device utilizes FN tunneling across silicon oxide between gate 2 and the floating gate. Before diving into online learning, we first analyze basic PGM/ERS characteristics for utilizing the proposed device as a multi-conductance synaptic device

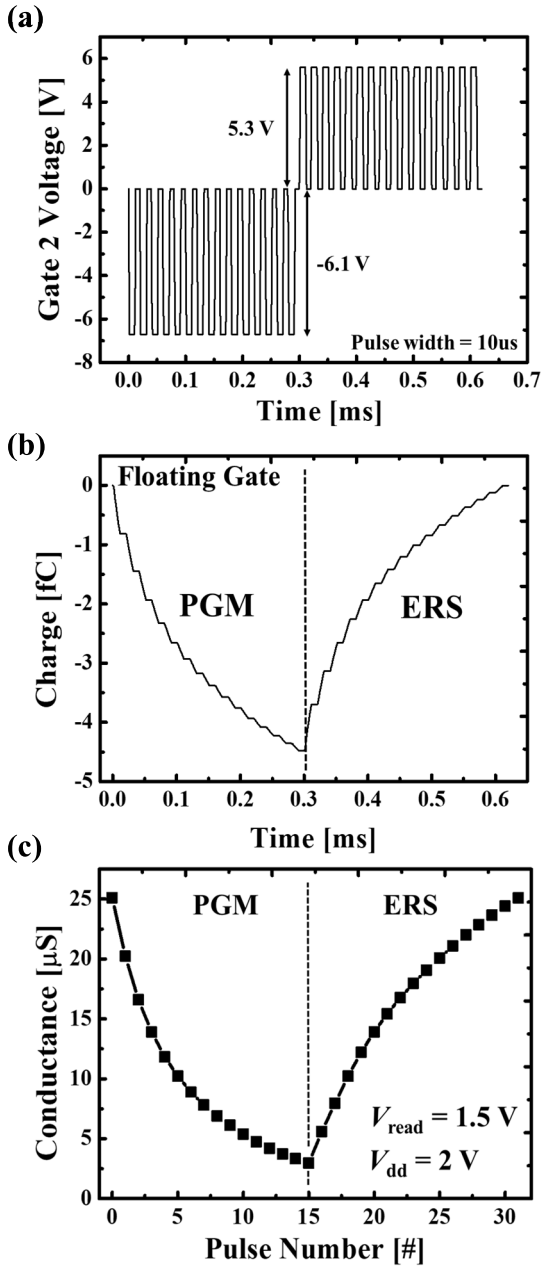

FIGURE 6. Program/Erase characteristics of the proposed device. (a) Pulse P/E bias scheme, (b) changes in floating-gate charge and (c) conductance are presented.

for offline learning. Fig. 6(a) presents the bias condition of PGM/ERS for offline learning. Rectangular pulses are applied to gate 2 of the device, and gate 1 and the drain are biased as 0 V during the learning process. For PGM, an individual PGM pulse is applied to gate 2 for 10 μs with the magnitude of -6.1 V. When the PGM pulse is applied to gate 2, most of the voltage is applied to the tunneling-oxide formed between gate 2 and the floating-gate since the coupling ratio between gate 1 and the floating-gate is larger than that of between gate 2 and the floating-gate. For this reason, the FN tunneling of electron between gate 2 and the floating-gate occurs, changing the charge stored in floating-gate. It can be seen that the amount of charge decreases as the pulse is applied. However, as the PGM pulse is applied repeatedly, the charge of the floating-gate lowers the potential of the floating-gate, which degrades the PGM efficiency. Therefore, the change rate of the charge reduces as the pulse number increases.

Transfer curves were verified for each of the PGM/ERS states. Gate 2 and drain are biased at 0 V and 2 V respectively. Then, gate 1 is set as the control gate as mentioned above. During gradual PGM, as the pulse number of PGM increases, the reduced charge in the floating-gate causes the threshold voltage (V_{th}) of the device to increase. In contrast, with a repeatedly applied ERS pulse, Vth decreases due to the accumulated charge in the floating-gate. This results in 0.65 V and 0.4 V of memory window (MW) for PGM and ERS, respectively. The conductance values are extracted in each of the states at a read voltage condition of 1.5 V, which ensures a stable operation of the proposed synaptic device. The amount of conductance decreases with the PGM states. However, the conductance change rate becomes smaller because the sufficiently discharged floating-gate depresses the tunneling of additional electrons. In the case of the ERS states, the opposite tendency of conductance is shown, but the same tendency of the conductance change rate is also verified due to the sufficiently charged floating-gate. This shows the same tendency with the charge change of the floating gate as illustrated in Fig. 6(c).

D. SYSTEM LEVEL SIMULATION OF SYNAPTIC DEVICE

In order to analyze the system-level performance of the proposed device, an offline-trained hardware-based SNN is simulated. A single-layered hardware-based SNN is trained to classify the modified National Institute of Standards and Technology (MNIST) dataset, which exhibited 92.06% of classification accuracy on ideal non-SNN.

Fig. 8(a) presents a schematic of a synaptic array using a proposed double-gated device. $V_{1\ldots i}^{pre}$ are input voltage to synaptic array, $V_{1\ldots j}^{post}$ are postsynaptic spikes, and $V_{1\ldots j}^{fb+}/V_{1\ldots j}^{fb-}$ are feedback signals for online learning. One synaptic weight is represented by a pair of two synaptic devices, denoted by G^+ and G^-, respectively [52]–[55]. G^+ injects a current to the membrane capacitor of a postsynaptic neuron, which is responsible to the positive part of a weight. On the other hand, G^- withdraws current from the membrane capacitor, which is responsible to the negative part of a weight.

Weight values are calculated externally from a non-spiking optimized neural network using a stochastic gradient descent with minimum square loss and transferred into our SNN after weight quantization. For a pair of synaptic devices representing a positive weight, G^- is programmed to its lowest conductance. For a pair representing a negative weight, G^+ is at its lowest conductance.

As illustrated in Fig. 8(b), pixel values of input images from MNIST was represented using a Poisson-encoded spike train, where the spiking rate is proportional to the input pixel value. The maximum of 255 spikes are used for a single pixel

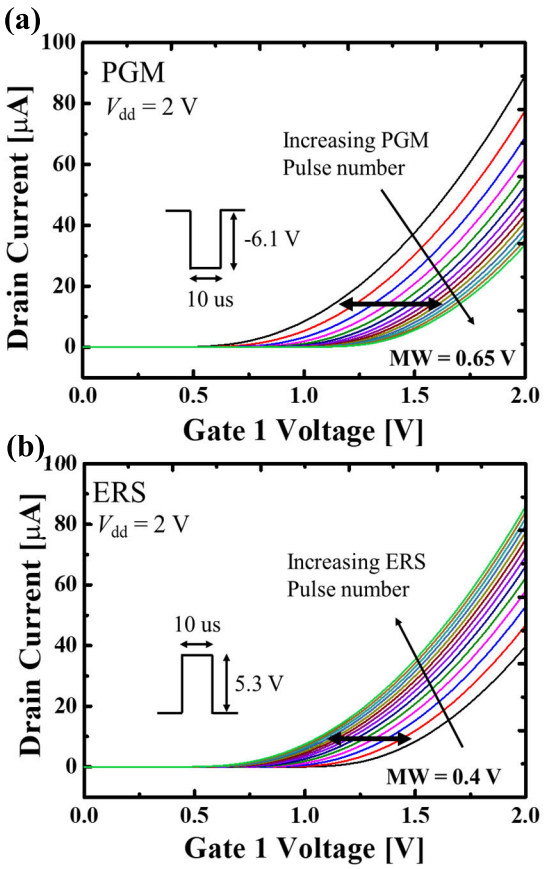

FIGURE 7. Transfer curves of the proposed device with (a) increasing PGM pulse and (b) increasing ERS pulse. Each transfer curve corresponds to a conductance presented in Fig. 6(c).

since the MNIST dataset are 8-bit grayscale images. Using the synaptic device as described above yielded a classification accuracy of 91.64%.

III. ONLINE LEARNING

A. IMPLEMENTATION OF ONLINE LEARNING ON DEVICE

In this session, we maximize the performance of a neuromorphic system by utilizing both offline and online learning. Offline learning is effective in the way that it adopts optimal synaptic weight values computed from an external computer. However, it is prone to performance degradation from problems including variance, limited fan-in, and overflow [56]–[58]. In other words, computer weights are optimal for the ideal neural network, but they are sub-optimal for hardware-based neural networks. Therefore, in this paper, we apply online learning on a neuromorphic system trained offline proceeding in order to maximize the performance. Similar to previously researched double-gated synaptic devices [14], [15], the proposed device utilizes the overlapping input and teaching signal for online learning. The teaching signal nor the input signal provide enough voltage for change in floating gate electron density. Therefore, changes of the floating gate charge and the threshold voltage upon the online programming scheme must be analyzed.

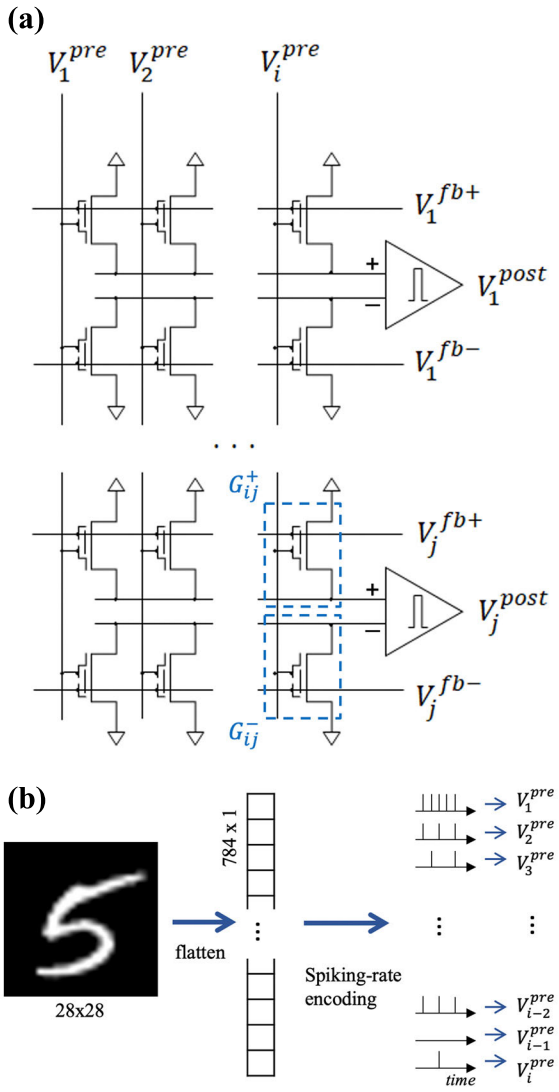

FIGURE 8. (a) Schematic of hardware-based single layer neural network with excitatory and inhibitory synapses. 7280 pairs of devices are used to formulate single-layered perceptron used in this paper. (b) Diagram of generating input signals for one MNIST image.

As presented in Fig. 9(a), a bipolar input pulse with an amplitude of 1.5 V and a duration of 600 ns is applied to gate 1 for the reading and current summation. The feedback pulse is applied to gate 2 for an online depression and potentiation, and the voltage and duration of which differs depending on the target polarity and magnitude of the conductance change. In the case of online depression, a −5.9 V feedback pulse is applied to gate 2, which causes a maximum potential difference between gate 2 and gate 1 to become −7.4 V. This is a sufficient potential difference for FN tunneling of the electron from gate 2 to the floating gate. For online potentiation, an applied gate 1 input pulse is the same as the online depression, and the feedback pulse is applied with a positive 5.1 V pulse as opposed to the online depression. In this case, the potential maximum difference between gate 2 and the floating gate becomes 6.6 V, which causes FN tunneling of the electron from the floating gate to gate 2. The resulting charge

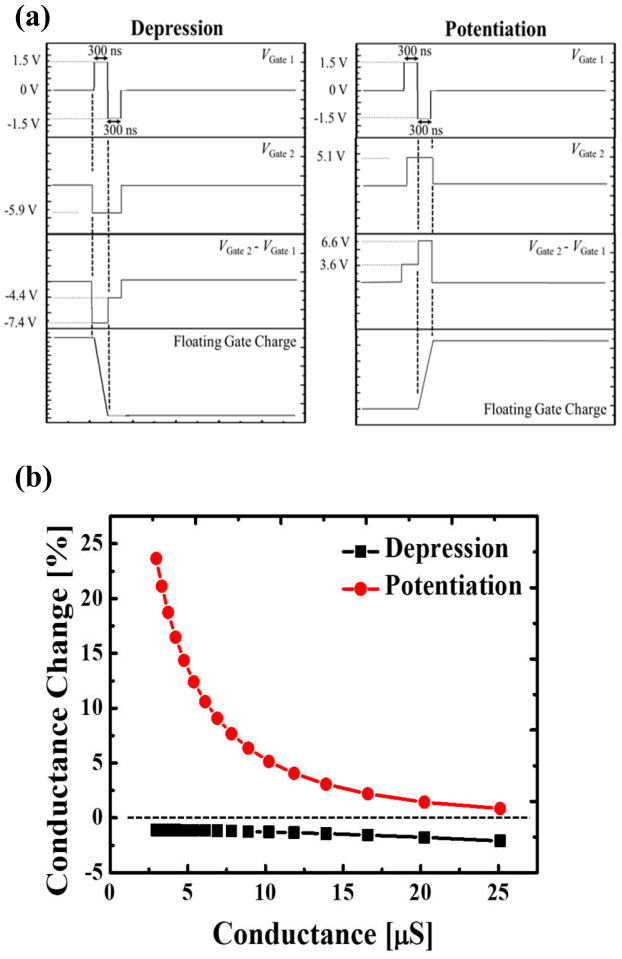

FIGURE 9. (a) Programming pulse scheme for on-line learning and (b) conductance change rate in each off-line trained state. Conductance change as response to potentiation/depression pulse of (a) differs depending on the initial device conductance.

change amount is less than 10% of the offline learning charge change amount. The amount of Vth shift differs depending on the conductance of the synaptic device before receiving a signal, and the conductance change (ΔS/S) on the online depression and potentiation are presented in Fig. 9(b).

B. SYSTEM-LEVEL SIMULATION FOR ONLINE LEARNING

In actual implementation of hardware neural networks, performance degradation occurs due to the device conductance variation. Therefore, the performance gap between the ideal non-SNN and the hardware-based SNN can be mitigated by fine conductance modulation, which can be achieved by applying online learning on offline-trained SNN of Fig. 8. By applying a teaching signal with a certain rule on the double-gated synaptic device, we can emulate a gradient descent in updating the synaptic weight. First, consider a single-layered perceptron:

$$Y = Relu(WX) \tag{1}$$
$$Y = [y_1, y_2, \cdots, y_n]^T \tag{2}$$
$$X = [x_1, x_2, \cdots, x_m]^T \tag{3}$$
$$L = \frac{1}{n}\sum_j (y_j - t_j)^2 \tag{4}$$

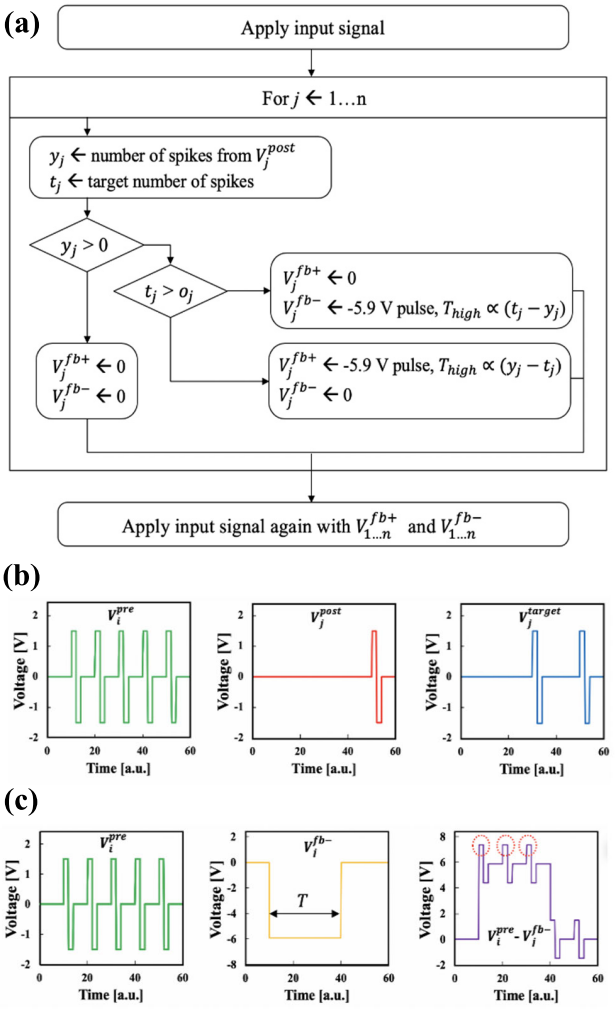

FIGURE 10. (a) Flowchart for online learning scheme for single-layered perceptron represented by Fig. 7 and equations (1)-(4). An example corresponding to flowchart (a) applied to the neural network of Fig. 7 is presented in (b) and (c). (b) presents voltages when an input signal is given, and (c) presents voltages when the same input signal is given again with feedback voltage. Pulse duration T is proportional to number of target spikes and output spikes. Conductance change occurs in time marked with red dotted circle.

where X is the input vector, Y is the output vector, W is the weight matric, and L is the loss function. Minimizing the mean square error (L) with the gradient descent yields the following equation:

$$\Delta w_{ij} = -\eta \frac{dL}{dw_{ij}} = -\eta sign(w_{ij}x_i)(y_j - t_j)x_i \tag{5}$$

To emulate this equation on hardware, we update the conductance of G^+ and G^- with teaching signals and input signals. Since decreasing G^- by ΔG yields the same result as increasing G^+ by ΔG, we only use the online depression upon online learning for reducing the overall current level and power consumption. For weights connected to the output neuron corresponding to the label, we decrease G^- since the weight must increase, and for the others, we decrease G^+. According to Fig. 8(b), conductance change of the

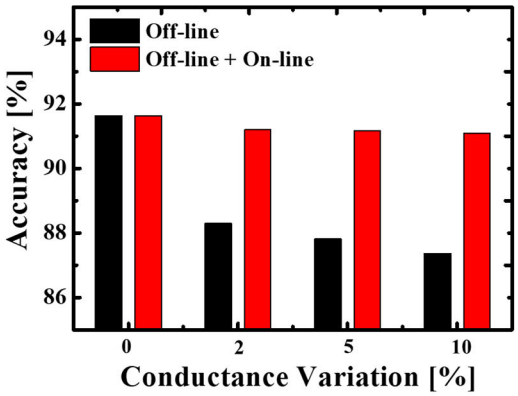

FIGURE 11. MNIST classification accuracy of SNN with synaptic weight variance.

synaptic transistor occurs at the presence of a presynaptic spike overlapping with the teaching signal. Therefore, with rate-coded data, Δw_{ij} is proportional to the expected number of input spikes, overlapping with the teaching signal. If the teaching signal with duty proportional to $|y_j - t_j|$ is given to gate 2, Δw_{ij} becomes proportional to $(y_j - t_j)x_i$, successfully emulating a weight update by a gradient descent. A step-by-step flowchart for online learning is presented in Fig. 10(a). During the online learning procedure, we further reduce training loss by reusing training images. If the hardware SNN operates as desired, only the output neuron corresponding to the target class should fire at the target speed and the others should remain silent. Therefore, if a neuron fired less than the target firing rate, the synaptic weights connected should increase, and if it fired more, synaptic weights should decrease. An example for node voltages is presented in Fig. 10(b)-(d). As illustrated in Fig. 10(d), a weight change occurs when $V^{pre} - V^{fb}$ becomes maximum. Since duration of teaching signal V^{fb} is proportional to the difference between the number of target output spikes and the actual number of output spikes, it can be seen that the larger the output error, the more the conductance change occurs.

The conductance variance of nonvolatile memory is inevitable. There always exists a mismatch between target conductance and programmed conductance. To verify the effect of conductance variation on SNN performance, we assumed an ideal synaptic device with continuous conductance states and a log-normal distribution of conductance variation is also considered. As the variance increases, the system accuracy becomes lower since the synaptic weights are not at its local optimum anymore. Applying the online learning method, we can set the synaptic weights to become closer to the local optimum, increasing the performance, reducing up to 87% of the performance degradation. The proposed online learning method shows significant performance enhancement and this additional online learning method should be performed without a fan-out problem. Classification accuracy of the hardware-based SNN before and after applying online learning with respect to weight variance is presented in Fig. 11.

IV. CONCLUSION

In this paper, we analyzed the device characteristics and the system level operation of a floating-gate-based synaptic device with two control gates. The proposed device is programmed to 16 different conductance levels by applying a pulse at one of two control gates, thus designating it as a qualified synaptic device candidate for a hardware-based spiking neural network. Through an effective online learning method, the performance of the hardware neural network is maximized and variation immunity is achieved. The proposed synaptic transistor and its training strategy enable efficient lifelong learning of a neuromorphic system.

ACKNOWLEDGMENT

(Donghyun Ryu and Tae-Hyung Kim contributed equally to this work.)

REFERENCES

[1] A.-D. Almási, S. Wozniak, V. Cristea, Y. Leblebici, and T. Engbersen, "Review of advances in neural networks: Neural design technology stack," *Neurocomputing*, vol. 174, pp. 31–41, Jan. 2016.

[2] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. K. Esser, R. Appuswamy, B. Taba, A. Amir, M. D. Flickner, W. P. Risk, R. Manohar, and D. S. Modha, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, Aug. 2014.

[3] B. Han, A. Sengupta, and K. Roy, "On the energy benefits of spiking deep neural networks: A case study," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 971–976.

[4] G. W. Burr, R. M. Shelby, A. Sebastian, S. Kim, S. Kim, S. Sidler, K. Virwani, M. Ishii, P. Narayanan, A. Fumarola, L. L. Sanches, I. Boybat, M. Le Gallo, K. Moon, J. Woo, H. Hwang, and Y. Leblebici, "Neuromorphic computing using non-volatile memory," *Adv. Phys.*, vol. 2, no. 1, pp. 89–124, Dec. 2016.

[5] C. Mead, "Neuromorphic electronic systems," *Proc. IEEE*, vol. 78, no. 10, pp. 1629–1636, Oct. 1990.

[6] J. Misra and I. Saha, "Artificial neural networks in hardware: A survey of two decades of progress," *Neurocomputing*, vol. 74, nos. 1–3, pp. 239–255, Dec. 2010.

[7] A. Pantazi, S. Wozniak, T. Tuma, and E. Eleftheriou, "All-memristive neuromorphic computing with level-tuned neurons," *Nanotechnology*, vol. 27, no. 35, Jul. 2016, Art. no. 355205.

[8] D. S. Jeong, K. M. Kim, S. Kim, B. J. Choi, and C. S. Hwang, "Memristors for energy-efficient new computing paradigms," *Adv. Electron. Mater*, vol. 2, no. 9, Aug. 2016, Art. no. 1600090.

[9] M. Chu, B. Kim, S. Park, H. Hwang, M. Jeon, B. H. Lee, and B.-G. Lee, "Neuromorphic hardware system for visual pattern recognition with memristor array and CMOS neuron," *IEEE Trans. Ind. Electron.*, vol. 63, no. 4, pp. 2410–2419, Sep. 2014.

[10] S. Park, H. Kim, M. Choo, J. Noh, A. Sheri, S. Jung, K. Seo, J. Park, S. Kim, W. Lee, and J. Shin, "RRAM-based synapse for neuromorphic system with pattern recognition function," in *IEDM Tech. Dig.*, Dec. 2012, pp. 10-1–10-2.

[11] H. Kim, S. Hwang, J. Park, S. Yun, J.-H. Lee, and B.-G. Park, "Spiking neural network using synaptic transistors and neuron circuits for pattern recognition with noisy images," *IEEE Electron Device Lett.*, vol. 39, no. 4, pp. 630–633, Apr. 2018.

[12] S. K. Bose, V. Mohan, and A. Basu, "A 75Kb SRAM in 65 nm CMOS for in-memory computing based neuromorphic image denoising," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Oct. 2020.

[13] S. Kim, B. Choi, J. Yoon, Y. Lee, H.-D. Kim, M.-H. Kang, and S.-J. Choi, "Binarized neural network with silicon nanosheet synaptic transistors for supervised pattern classification," *Sci. Rep.*, vol. 9, no. 1, pp. 1–7, Aug. 2019.

[14] J. Park, M.-W. Kwon, H. Kim, S. Hwang, J.-J. Lee, and B.-G. Park, "Compact neuromorphic system with four-terminal Si-based synaptic devices for spiking neural networks," *IEEE Trans. Electron Devices*, vol. 64, no. 5, pp. 2438–2444, May 2017.

[15] H. Kim, S. Hwang, J. Park, and B.-G. Park, "Silicon synaptic transistor for hardware-based spiking neural network and neuromorphic system," *Nanotechnology*, vol. 28, no. 40, Sep. 2017, Art. no. 405202.
[16] G. Indiveri, B. Linares-Barranco, T. J. Hamilton, A. V. Schaik, R. Etienne-Cummings, T. Delbruck, S.-C. Liu, P. Dudek, P. Häfliger, S. Renaud, J. Schemmel, G. Cauwenberghs, J. Arthur, K. Hynna, F. Folowosele, S. Saighi, T. Serrano-Gotarredona, J. Wijekoon, Y. Wang, and K. Boahen, "Neuromorphic silicon neuron circuits," *Frontiers Neurosci.*, vol. 5, p. 73, May 2011.
[17] C.-S. Poon and K. Zhou, "Neuromorphic silicon neurons and large-scale neural networks: Challenges and opportunities," *Frontiers Neurosci.*, vol. 5, p. 108, Sep. 2011.
[18] V. Kornijcuk, H. Lim, J. Y. Seok, G. Kim, S. K. Kim, I. Kim, B. J. Choi, and D. S. Jeong, "Leaky integrate- and-fire neuron circuit based on floating-gate integrator," *Frontiers Neurosci.*, vol. 10, p. 212, May 2016.
[19] C.-L. Chen, K. Kim, Q. Truong, A. Shen, Z. Li, and Y. Chen, "A spiking neuron circuit based on a carbon nanotube transistor," *Nanotechnology*, vol. 23, no. 27, Jun. 2012, Art. no. 275202.
[20] C. Johnson, S. Roychowdhury, and G. K. Venayagamoorthy, "A reversibility analysis of encoding methods for spiking neural networks," in *Proc. Int. Joint Conf. Neural Netw.*, Jul. 2011, pp. 1802–1809.
[21] A. Manwani and C. Koch, "Synaptic transmission: An information-theoretic perspective," in *Proc. Adv. Neural Inf. Process. Syst.*, 1997, pp. 201–207.
[22] J. A. Wall, L. J. McDaid, L. P. Maguire, and T. M. McGinnity, "Spiking neural network model of sound localization using the interaural intensity difference," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 4, pp. 574–586, Apr. 2012.
[23] G. W. Burr, R. M. Shelby, C. D. Nolfo, J. W. Jang, R. S. Shenoy, P. Narayanan, K. Virwani, E. U. Giacometti, B. 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," in *IEDM Tech. Dig.*, Dec. 2014, pp. 29-1–29-5.
[24] V. Milo, G. Pedretti, R. Carboni, A. Calderoni, N. Ramaswamy, S. Ambrogio, and D. Ielmini, "A 4-transistors/1-resistor hybrid synapse based on resistive switching memory (RRAM) capable of spike-rate-dependent plasticity (SRDP)," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 26, no. 12, pp. 2806–2815, Dec. 2018.
[25] C.-H. Kim, S. Lim, S. Y. Woo, W. M. Kang, Y. T. Seo, S. T. Lee, S. Lee, D. Kwon, S. Oh, Y. Noh, and H. Kim, "Emerging memory technologies for neuromorphic computing," *Nanotechnolgy*, vol. 30, no. 3, Nov. 2018, Art. no. 032001.
[26] C. Yoon, J. H. Lee, S. Lee, J. H. Jeon, J. T. Jang, D. H. Kim, Y. H. Kim, and B. H. Park, "Synaptic plasticity selectively activated by polarization-dependent energy-efficient ion migration in an ultrathin ferroelectric tunnel junction," *Nano Lett.*, vol. 17, no. 3, pp. 1949–1955, Feb. 2017.
[27] C. Sung, J. Song, D. Lee, S. Lim, M. Kwak, and H. Hwang, "Ultra-thin <10 nm) dual-oxide (Al_2O_3/TiO_2) hybrid device (memory/selector) with extremely low I_{off} < 1nA) and I_{reset} < 1nA) for 3D storage class memory," in *Proc. IEEE Symp. VLSI Technol.*, Jul. 2019, pp. 62–63.
[28] S. Ambrogio, A. Kumar, A. Chen, G. W. Burr, M. Gallot, K. Spoon, H. Tsai, C. Mackin, M. Wesson, S. Kariyappa, P. Narayanan, and C.-C. Liu, "Reducing the impact of phase-change memory conductance drift on the inference of large-scale hardware neural networks," in *IEDM Tech. Dig.*, Dec. 2019, p. 6.
[29] K. Ota, J. Deguchi, S. Fujii, M. Saitoh, M. Yamaguchi, R. Berdan, T. Marukame, Y. Nishi, K. Matsuo, K. Takahashi, Y. Kamiya, and S. Miyano, "Performance maximization of in-memory reinforcement learning with variability-controlled $Hf_{1-x}Zr_xO_2$ ferroelectric tunnel junctions," in *IEDM Tech. Dig.*, Dec. 2019, pp. 6-1–6-2.
[30] T.-Y. Wu, T.-S. Chang, H.-Y. Lee, S.-S. Sheu, W.-C. Lo, T.-H. Hou, H.-H. Huang, Y.-H. Chu, C.-C. Chang, M.-H. Wu, C.-H. Hsu, C.-T. Wu, M.-C. Wu, and W.-W. Wu, "Sub-nA low-current HZO ferroelectric tunnel junction for high-performance and accurate deep learning acceleration," in *IEDM Tech. Dig.*, Dec. 2019, pp. 6-1–6-3.
[31] D. Kuzum, R. G. D. Jeyasingh, and H.-S.-P. Wong, "Energy efficient programming of nanoelectronic synaptic devices for large-scale implementation of associative and temporal sequence learning," in *IEDM Tech. Dig.*, Dec. 2011, pp. 30-1–30-3.
[32] Y. Matveyev, R. Kirtaev, A. Fetisova, S. Zakharchenko, D. Negrov, and A. Zenkevich, "Crossbar nanoscale HfO_2-based electronic synapses," *Nanosc. Res. Lett.*, vol. 11, no. 1, pp. 1–6, Mar. 2016.
[33] M.-H. Kim, S. Kim, S. Bang, T.-H. Kim, D. K. Lee, S. Cho, J.-H. Lee, and B.-G. Park, "Pulse area dependent gradual resistance switching characteristics of CMOS compatible SiN_x-based resistive memory," *Solid-State Electron.*, vol. 132, pp. 109–114, Jun. 2017.
[34] W. Chen, R. Fang, M. B. Balaban, W. Yu, Y. Gonzalez-Velo, H. J. Barnaby, and M. N. Kozicki, "A CMOS-compatible electronic synapse device based on $Cu/SiO_2/W$ programmable metallization cells," *Nanotechnology*, vol. 27, no. 25, May 2016, Art. no. 255202.
[35] M. R. Azghadi, B. Linares-Barranco, D. Abbott, and P. H. W. Leong, "A hybrid CMOS-memristor neuromorphic synapse," *IEEE Trans. Biomed. Circuits Syst.*, vol. 11, no. 2, pp. 434–445, Apr. 2017.
[36] X. Gu, Z. Wan, and S. S. Iyer, "Charge-trap transistors for CMOS-only analog memory," *IEEE Trans. Electron Devices*, vol. 66, no. 10, pp. 4183–4187, Oct. 2019.
[37] C.-H. Kim, S. Lee, S. Y. Woo, W.-M. Kang, S. Lim, J.-H. Bae, J. Kim, and J.-H. Lee, "Demonstration of unsupervised learning with spike-timing-dependent plasticity using a TFT-type NOR flash memory array," *IEEE Trans. Electron Devices*, vol. 65, no. 5, pp. 1774–1780, May 2018.
[38] 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. 38-1–38-4.
[39] X. Guo, F. M. Bayat, M. Bavandpour, M. Klachko, M. R. Mahmoodi, M. Prezioso, K. K. Likharev, and D. B. Strukov, "Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology," in *IEDM Tech. Dig.*, Dec. 2017, pp. 6-1–6-5.
[40] S. Y. Woo, K.-B. Choi, J. Kim, W.-M. Kang, C.-H. Kim, Y.-T. Seo, J.-H. Bae, B.-G. Park, and J.-H. Lee, "Implementation of homeostasis functionality in neuron circuit using double-gate device for spiking neural network," *Solid-State Electron.*, vol. 165, Mar. 2020, Art. no. 107741.
[41] C. C. Wang and D. Markovic, "Delay estimation and sizing of CMOS logic using logical effort with slope correction," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 56, no. 8, pp. 634–638, Aug. 2009.
[42] S. B. Eryilmaz, D. Kuzum, S. Yu, and H.-S.-P. Wong, "Device and system level design considerations for analog-non-volatile-memory based neuromorphic architectures," in *IEDM Tech. Dig.*, Dec. 2015, p. 4-1.
[43] A. Sengupta, Y. Shim, and K. Roy, "Proposal for an all-spin artificial neural network: Emulating neural and synaptic functionalities through domain wall motion in ferromagnets," *IEEE Trans. Biomed. Circuits Syst.*, vol. 10, no. 6, pp. 1152–1160, Dec. 2016.
[44] S. B. Eryilmaz, S. Joshi, E. Neftci, W. Wan, G. Cauwenberghs, and H.-S.-P. Wong, "Neuromorphic architectures with electronic synapses," in *Proc. 17th Int. Symp. Qual. Electron. Design (ISQED)*, Mar. 2016, pp. 118–123.
[45] I. Chakraborty, D. Roy, and K. Roy, "Technology aware training in memristive neuromorphic systems for nonideal synaptic crossbars," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 5, pp. 335–344, Oct. 2018.
[46] M. M. Eshaghian-Wilner, A. Friesz, A. Khitun, S. Navab, A. C. Parker, K. L. Wang, and C. Zhou, "Emulation of neural networks on a nanoscale architecture," *J. Phys. Conf.*, vol. 61, no. 1, pp. 288–292, 2007.
[47] *TCAD Sentaurus User Manual*, Synopsys, Mountain View, CA, USA, Sep. 2014.
[48] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*. Cambridge, U.K.: Cambridge Univ. Press, 2013.
[49] J.-H. Ahn, D.-I. Moon, S.-W. Ko, C.-H. Kim, J.-Y. Kim, M.-S. Kim, M.-L. Seol, J.-B. Moon, J.-M. Choi, J.-S. Oh, S.-J. Choi, and Y.-K. Choi, "A SONOS device with a separated charge trapping layer for improvement of charge injection," *AIP Adv.*, vol. 7, no. 3, Mar. 2017, Art. no. 035205.
[50] M. Hirose, "Electron tunneling through ultrathin SiO_2," *Mater. Sci. Eng., B*, vol. 41, no. 1, pp. 35–38, Oct. 1996.
[51] M. Depas, B. Vermeire, P. W. Mertens, R. L. Van Meirhaeghe, and M. M. Heyns, "Determination of tunnelling parameters in ultra-thin oxide layer poly-$Si/SiO_2/Si$ structures," *Solid-State Electron.*, vol. 38, no. 8, pp. 1465–1471, Aug. 1995.
[52] K. Moon, M. Kwak, J. Park, D. Lee, and H. Hwang, "Improved conductance linearity and conductance ratio of 1T2R synapse device for neuromorphic systems," *IEEE Electron Device Lett.*, vol. 38, no. 8, pp. 1023–1026, Aug. 2017.
[53] H. Tian, X. Cao, Y. Xie, X. Yan, A. Kostelec, D. DiMarzio, C. Chang, L.-D. Zhao, W. Wu, J. Tice, J. J. Cha, J. Guo, and H. Wang, "Emulating bilingual synaptic response using a junction-based artificial synaptic device," *ACS Nano*, vol. 11, no. 7, pp. 7156–7163, Jun. 2017.

[54] N. Duan, Y. Li, H.-C. Chiang, S.-P. Huang, K.-S. Yin, J. Chen, C.-I. Yang, T.-C. Chang, and X.-S. Miao, "Gate modulation of excitatory and inhibitory synaptic plasticity in a low-temperature polysilicon thin film synaptic transistor," *ACS Appl. Electron. Mater.*, vol. 1, no. 1, pp. 132–140, Dec. 2018.
[55] G. Lecerf, J. Tomas, and S. Saighi, "Excitatory and inhibitory memristive synapses for spiking neural networks," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2013, pp. 1616–1619.
[56] S. Kim, M. Lim, Y. Kim, H.-D. Kim, and S.-J. Choi, "Impact of synaptic device variations on pattern recognition accuracy in a hardware neural network," *Sci. Rep.*, vol. 8, no. 1, pp. 1–7, Feb. 2018.
[57] C. Frenkel, M. Lefebvre, J.-D. Legat, and D. Bol, "A 0.086-mm^2 12.7-pJ/SOP 64k-synapse 256-neuron online-learning digital spiking neuromorphic processor in 28-nm CMOS," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 1, pp. 145–158, Feb. 2019.
[58] Y. Li, P. Saratchandran, and N. Sundararajan, "Analysis of minimal radial basis function network algorithm for real-time identification of nonlinear dynamic systems," *IEE Proc.-Control Theory Appl.*, vol. 147, no. 4, pp. 476–484, Jul. 2000.

DONGHYUN RYU received the B.S. degree in electrical and computer engineering from Soongsil University, Seoul, South Korea, in 2017, and the M.S. degree in electrical and computer engineering from Seoul National University (SNU), Seoul, in 2019, where he is currently pursuing the Ph.D. degree in electrical and computer engineering.

TAE-HYUNG KIM (Graduate Student Member, IEEE) received the B.S. degree in electrical and computer engineering from Seoul National University (SNU), Seoul, South Korea, in 2017, and the M.S. degree, in 2019.

TAEJIN JANG (Graduate Student Member, IEEE) received the B.S. degree from Seoul National University (SNU), in 2016, where he is currently pursuing the Ph.D. degree in electrical engineering.

JUNSU YU (Student Member, IEEE) received the B.S. degree in electrical and computer engineering from Seoul National University (SNU), in 2019, where he is currently pursuing the M.S. degree in electrical engineering.

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

BYUNG-GOOK PARK (Fellow, IEEE) received the B.S. and M.S. degrees in electronic engineering from Seoul National University (SNU), Seoul, South Korea, in 1982 and 1984, respectively, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 1990. In 1994, he joined the School of Electrical Engineering, SNU, as an Assistant Professor, where he is currently a Professor.