Received 12 September 2019; revised 27 September 2019; accepted 2 October 2019. Date of publication 15 October 2019; date of current version 30 October 2019. The review of this article was arranged by Editor M. Liu.

Digital Object Identifier 10.1109/JEDS.2019.2947316

Operation Scheme of Multi-Layer Neural Networks Using NAND Flash Memory as High-Density Synaptic Devices

SUNG-TAE LEE [1,2], SUHWAN LIM [1,2], NAG YONG CHOI [1,2], JONG-HO BAE [1,2], DONGSEOK KWON [1,2], BYUNG-GOOK PARK [1,2] (Member, IEEE), AND JONG-HO LEE [1,2] (Fellow, IEEE)

1 School of EECS, Seoul National University, Seoul 151-742, South Korea
2 Inter-University Semiconductor Research Center, Seoul National University, Seoul 151-742, South Korea

CORRESPONDING AUTHOR: J.-H. LEE (e-mail: 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 2019.

ABSTRACT We propose a designing of multi-layer neural networks using 2D NAND flash memory cell as a high-density and reliable synaptic device. Our operation scheme eliminates the waste of NAND flash cells and allows analogue input values. A 3-layer perceptron network with 40,545 synapses is trained on a MNIST database set using an adaptive weight update method for hardware-based multi-layer neural networks. The conductance response of NAND flash cells is measured and it is shown that the unidirectional conductance response is suitable for implementing multi-layer neural networks using NAND flash memory cells as synaptic devices. Using an online-learning, we obtained higher learning accuracy with NAND synaptic devices compared to that with a memristor-based synapse regardless of weight update methods. Using an adaptive weight update method based on a unidirectional conductance response, we obtained a 94.19% learning accuracy with NAND synaptic devices. This accuracy is comparable to 94.69% obtained by synapses based on the ideal perfect linear device. Therefore, NAND flash memory which is mature technology and has great advantage in cell density can be a promising synaptic device for implementing high-density multi-layer neural networks.

INDEX TERMS Neuromorphic, NAND flash memory, deep neural networks (DNNs), synaptic device, deep learning, multi-layer neural networks, hardware-based neural network.

I. INTRODUCTION

The neuromorphic computing that mimics neuro-biological architectures present in the nervous system has been emerged as an attractive field of research because of its power efficiency [1]. Until now, spike-timing-dependent plasticity (STDP) algorithm [2], [3] motivated by learning process of real brains has actively been researched. However, STDP learning algorithm is still improving but not yet mature, resulting in poor performance compared to backpropagation algorithm [4].

Unlike STDP, backpropagation is a widely used, well-studied method in training deep neural networks (DNNs), offering outstanding performance on datasets such as hand-written digits (MNIST) [5]. Multi-layer neural networks based on synaptic devices can reduce power consumption greatly by replacing the vector-by-matrix multiplication with a dense crossbar array of analog devices such as PCM, RRAM [6], and NOR flash memory [7], [8].

However, several problems need to be addressed before memristive crossbar arrays can be widely adopted, such as high device variability, absence of precise device models and stochastic behavior of devices [9].

In order to evade above problems, we can use Si-based devices such as NOR flash memory and SRAM [10]. However, these memories have limitation of density because of bit lines and word lines contact in each cell device. On the other hand, NAND flash memory reduces ground wires

and bit lines considerably, which allows a denser layout and greater storage capacity per chip than those devices [11].

In the meantime, most of the works have been studied about synaptic devices using PCM, RRAM [6] and NOR flash memory [7], [8]. Another group used the measured characteristics of a single device to implement vector matrix multiplication in the NAND flash memory architecture [12]. However, all cells in NAND flash memory cannot be fully used as synaptic devices because the size of all synapse layers is determined by the number of word lines and the number of bit lines [12]. In addition, in the scheme of applying input voltages to word-lines [12], it is very difficult to allow analogue input values because of nonlinearity of $I_{\mathrm{BL}} - V_{\mathrm{WL}}$ characteristics. In this work, we propose a new operation scheme for implementing multi-layer neural networks using 2-D NAND flash memory cells as high-density, reliable synaptic devices. NAND flash memory is a mature technology and has great advantages in cell density and large storage capacity per chip because the cell string in the array can be fabricated vertically (vertical NAND flash) [13] and each cell string has many cells connected in series between the bit line and the source line. Unlike [12], we apply input values (voltage) into bit-line to allow analogue input value satisfying weighted sum output equation. In addition, the operation scheme in this work eliminates the waste of NAND flash cells. Negative synaptic weight can be represented using the difference in conductance (synaptic weight, $W_{ij} = \mathrm{G}_{ij}^{+} - \mathrm{G}_{ij}^{-}$) between a pair of adjacent cells. In our operation scheme, the current subtractor subtracting the current from two adjacent synaptic strings can be reused for all synapses in the synapse string, which reduces the burden of circuits. We measured floating-gate 2-D (planar) NAND flash cell strings fabricated with 26 nm technology. We also investigated the device variation by measuring NAND flash cells and checked the reliability of NAND flash cells by measuring endurance and retention characteristics. Using a matched computer simulation, a 3-layer perceptron network with 40,545 synapses is trained using the weight update method in [14] appropriate for our device and the MNIST data set.

II. OPERATION SCHEME OF MULTI-LAYER NEURAL NETWORKS

To implement multi-layer neural networks using a synaptic device array, adaptive learning rule for hardware-based multi-layer neural networks different from software-based algorithm is needed as shown in Table 1. The input signal ($a_i^{(l-1)}$ for the i^{th} neuron in the $l-1$ layer) and the weight (W_{ij} for the weight of the synapse between the i^{th} neuron in $l-1$ layer and the j^{th} neuron in l layer) can be represented by voltage ($V_i^{(l-1)}$) and the conductance difference of a pair of synaptic devices ($G_{ij}^{+} - \mathrm{G}_{ij}^{-}$), respectively. In forward evaluation of a multi-layer perceptron, each layer's inputs (V_i) drive the next layer's neurons through weights W_{ij} and activation function f. For backward propagation, each layer's error values (V_j^l) drive the preceding layer's error value using gradient descent method. By using sign of ΔW_{ij}, we can update the conductance of synaptic devices. Weight (W_{ij}) of synaptic device can be modified by one step ($|\Delta G_{ij}^{-}|, -|\Delta G_{ij}^{+}|$) at each iteration according to sign of ΔW_{ij} to reduce the burden of periphery circuit [14].

TABLE 1. Learning rule of software-based and hardware-based neural networks.

Target	Software-based	Hardware-based
Weights W_{ij}	W_{ij}	$G_{ij}^{+} - G_{ij}^{-}$
Forward propagation $S_j^{(l)}$	$\sum_i^N W_{ij} a_i^{(l-1)}$	$a_i^{(l-1)} \rightarrow V_i^{(l-1)}$ $\sum_i^N (G_{ij}^{+} - G_{ij}^{-}) V_i^{(l-1)}$
Backward propagation $\delta_i^{(l-1)}$	$\sum_j^M W_{ij} \delta_j^{(l)} \cdot f'(s_i^{(l-1)})$	$\delta_j^{(l)} \rightarrow V_j^{(l)}$ $\sum_j^M (G_{ij}^{+} - G_{ij}^{-}) V_j^{(l)} \cdot f'(V_i^{(l-1)})$
Weight updates ΔW_{ij}	$-\eta \cdot \delta_j^{(l)} \cdot f(s_i^{(l-1)})$	$\begin{cases} \lvert\Delta G_{ij}^{-}\rvert & \text{if } \Delta W_{ij} > 0 \\ -\lvert\Delta G_{ij}^{+}\rvert & \text{if } \Delta W_{ij} < 0 \end{cases}$

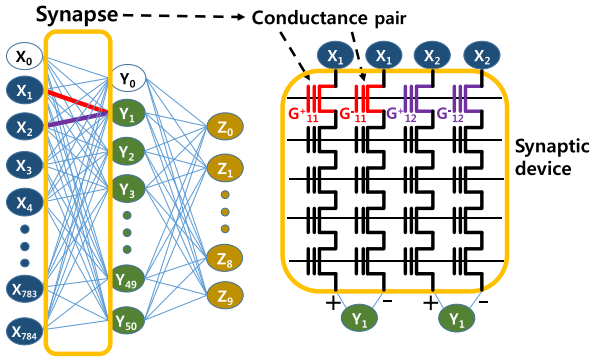

FIGURE 1. 3-layer perceptron in which synapse can be implemented using NAND flash memory. Synaptic weight is encoded by the conductance difference between a pair of adjacent NAND cells.

As shown in Fig. 1, to use NAND flash memory cells as synaptic devices, we apply input values (voltages) to the bit-lines for the following reasons. The scheme which applies input to the bit-lines allows analogue input values satisfying weighted sum output equation, $I = \sum(G^{+} - G^{-})V$, because output current is zero when the bit-line voltage is zero and bit-line current increases linearly with increasing bit-line bias in linear region. However, in the scheme of applying input voltage to the word-lines [12], it is very difficult to allow analog input values because output current may not be zero when the word-line voltage is zero and bit-line current increases exponentially with increasing word-line bias. In addition, as shown in Fig. 1, by using conductance

difference between a pair of adjacent cells to represent synaptic weight ($W_{ij} = G_{ij}^{+} - G_{ij}^{-}$), negative synaptic weight can be represented [4]. Note two cells in the same position in two adjacent cell strings have nearly identical device characteristics, so we can minimize the mismatch between two cells that represent one synapse.

Forward propagation simply subtracts read current between a pair of bit lines and sums the total currents through electronic devices such as capacitor as shown in Fig. 2 (a). Fig. 2 (a) represents 785 × 51 × 2 synapse array between the l-1th neuron layer and the l^{th} neuron layer. The output currents for all neurons in l^{th} layer are produced sequentially when the read pulse sequentially enters the word line as shown in Fig. 2 (b). When the k^{th} pulse is applied to the k^{th} word-line, the output current flows for the kth neuron in the l^{th} layer. During this process pass bias is applied to the unselected word-lines to read the k^{th} cell current as shown in Fig. 2 (c) and (d). Since the resistance of the selected cell is always much greater than that of the unselected cells with large pass bias applied, the output current primarily depends on the threshold voltage of the k^{th} word-line cell. Then, the overall current for the k^{th} row is summed as shown in Fig. 2 (c). When the overall output current for all neurons in l^{th} layer is sequentially produced, the output current stored in computing system is passed to l layer neurons.

Another research group has proposed an operation scheme to implement vector matrix multiplication (VMM) in the NAND flash memory architecture [12]. In [12], the size of all synapse layers is determined by the number of word-lines and the number of bit-lines, so that NAND flash cells cannot be fully used as synaptic devices. On the other hand, there is no waste of NAND flash cells in our scheme. In addition, using this scheme, the current subtractor subtracting the synaptic string can be reused for all synapse in synapse string as shown in Fig. 2 (a), which reduces the burden of circuits. In addition, program inhibition by boosting the channel potential is used to program only one cell in a row by applying a high bias to the unselected bit-lines and a low bias to the selected bit-line [15].

For an M x N synapse array, the time complexity of VMM using NAND flash memory is O(N) and it is larger than O (1) which is the time complexity of VMM using memristor array. However, recent state-of-the-art DNN algorithms typically require enormous parameter size. As a way to accommodate this, NAND flash memory which has great advantages in cell density can be a promising candidate for synaptic device. Because NAND flash architecture reduces ground wires and bit-lines considerably, which allows a denser layout and greater storage capacity per chip than other memory devices. In addition, it can be fabricated vertically, which allows great density [16]–[18]. Furthermore, by using sequential reading method, the current subtractor subtracting the current of synaptic string which consists of two adjacent cell strings can be reused for all synapses in the synapse string, which significantly reduces the burden of circuits. In addition, 3-D NAND flash

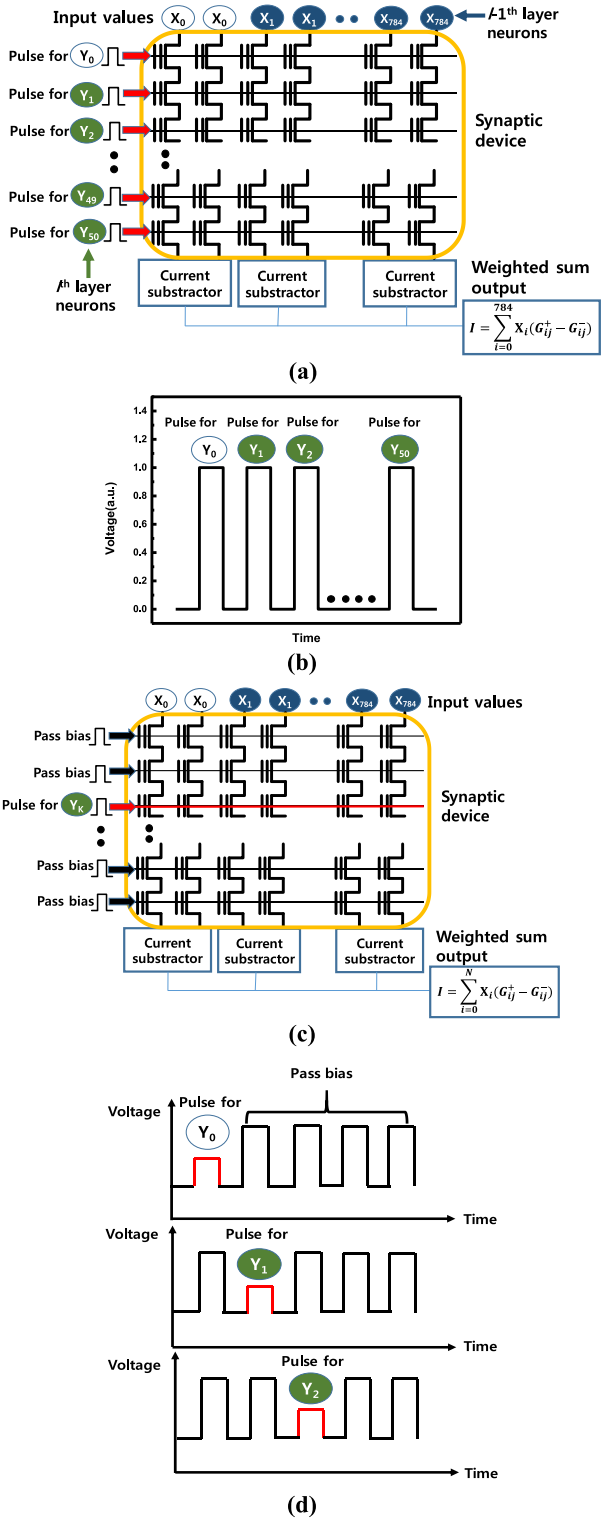

FIGURE 2. (a) Schematic of forward propagation of multi-layer neural networks using NAND flash memory. (b) Pulse-timing diagraph which is applied to word-lines. (c) Schematic of Forward propagation for producing the output current for k^{th} neuron in l^{th} layer. (d) Pulse-timing diagraph for producing the output current for each neuron in l^{th} layer.

memory has been demonstrated as technologically mature and cost-competitive technology among the various non-volatile memory technologies [16]–[18]. Therefore, NAND

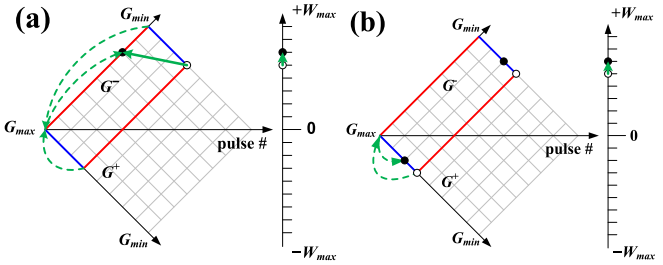

FIGURE 3. (a), (b) Weight update method when G^- reaches its minimum value.

flash can be a promising synaptic device for implementing high-density multi-layer neural networks.

As the synaptic device has discrete and finite conductance, we need adaptive weight update method. Conductance of synaptic device can be modified by one step at each iteration according to sign of error value to reduce the burden of periphery circuit which updates the conductance of synaptic devices by applying programming/erasing pulses. In software-based multi-layer neural networks, weight can be modified to exact target value according to the calculated error values. However, in hardware-based multi-layer neural networks, as the synaptic weight has discrete values, the weight can be modified to approximate target values. If multiple pulses are required for updating synaptic weight to approximate target value, we need to check the current conductance of the device and calculate the number of pulses required to reach the target conductance. It imposes big burden to periphery circuit. Therefore, updating the synaptic weight value by one step at each iteration reduces burden of external circuit.

As two synaptic devices are required to represent negative weight value ($W_{ij} = G_{ij}^{+} - G_{ij}^{-}$), there are several ways to update synaptic weight. In other words, both increasing G^+ and/or decreasing G^- result in increasing the weight. As NAND flash memory has higher learning accuracy when the conductance response is unidirectional as shown in Fig. 4(c) than when the conductance response is bidirectional, we can only decrease the G^- to increase the weight value (W_{ij}). However, as devices have finite conductance response, there is a case when weight needs to be increased ($\Delta W_{ij} > 0$) but G^- reaches its minimum conductance value (G_{min}), and G^- can no longer be decreased. In this case, there are two ways to update weight values. First, it is possible to initialize both G^+ and G^-, with a subsequent decrease in G^- as shown in Fig. 3 (a) [19]. Second, G^+ should be reset to G_{max} and decreased to the target value by applying a series of program pulses sequentially as shown in Fig. 3 (b) [14].

III. EXPERIMENTAL MEASUREMENT

In this work, we used floating gate 2D (planar) NAND flash cell strings fabricated with the 26 nm technology. One cell string consists of 64 cells, two dummy cells, a drain select line (DSL) transistor, and a source select line (SSL) transistor. The channel length and width are 26 and 20 nm, respectively [20].

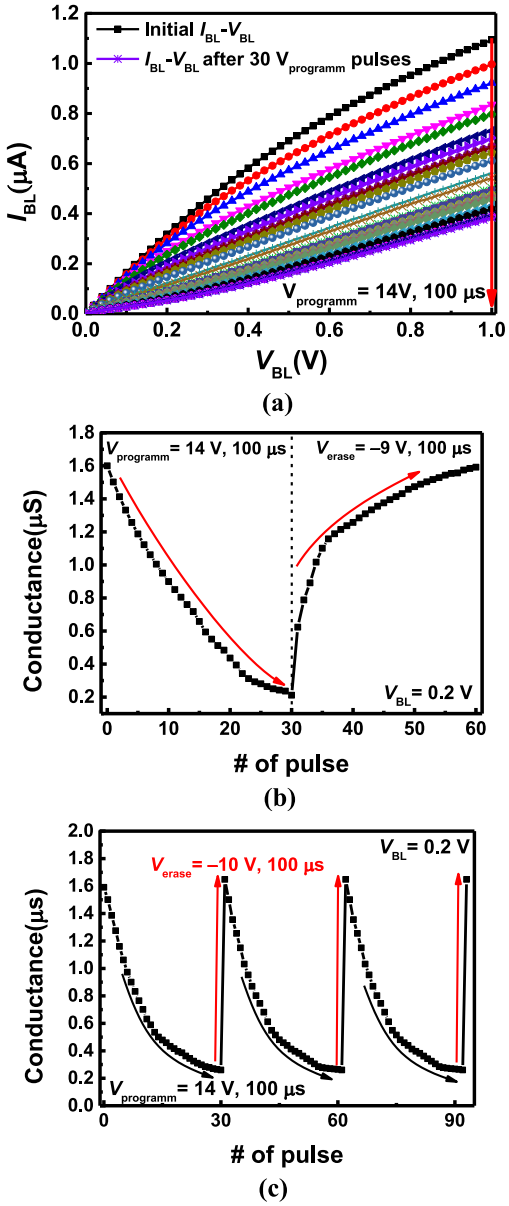

FIGURE 4. (a) Measured I_{BL}-V_{BL} curves when selected cell is programmed 30 times. (b) Bidirectional conductance response when selected cell is programmed 30 times and erased 30 times. (c) Unidirectional conductance response for 3 cycles when selected cell is programmed 30 times and erased at once.

Fig. 4 (a) shows the decreasing bit-line current (I_{BL}) curves when selected cell is programmed 30 times in a V_{BL} range of 0 V 1 V at a pass bias of 6.5 V. As device has higher accuracy when it has large dynamic range [21], we used voltage pulse of 14 V which is minimum voltage for programming for a given program time of 100 μs. Electrons are emitted from channel and injected into floating gate by applying programming pulse, which increase the threshold

voltage and decrease the bit-line current (I_{BL}). Using identical programming pulse (14 V, 100 μs) reduces the burden of periphery circuit, because different pulse size according to present conductance state needs enormous amount of calculation.

Fig. 4 (b) and (c) show measured bidirectional conductance response and unidirectional conductance response of selected cell, respectively. Conductance is measured as I_{BL}/V_{BL} at V_{BL} of 0.2 V. The selected cell is programmed (14 V, 100 μs) 30 times and erased (−9 V, 100 μs) 30 times to represent bidirectional conductance response. On the other hand, selected cell is programmed (14 V, 100 μs) 30 times and erased (−10 V, 100 μs) once to represent unidirectional conductance response as shown in Fig. 4 (c).

For comparing our synaptic device with other devices reported up to date, we use the behavior model for NAND flash cell, ideal perfect linear device, and memristive device. Fig. 5 shows normalized conductance versus the number of pulses in three devices, using behavior model [22] in equations (1) and (2)

$$\delta G_p = \alpha_p \exp\left(-\beta_p \frac{G - G_{\min}}{G_{\max} - G_{\min}}\right) \tag{1}$$

$$\delta G_d = \alpha_d \exp\left(-\beta_d \frac{G_{\max} - G}{G_{\max} - G_{\min}}\right) \tag{2}$$

where α_p is a fitting parameter and β_p is a nonlinearity factor of the potentiation characteristic, similarly α_d and β_d for the depression characteristic. In addition, G is the conductance of electronic synapse devices. $G_{\max}$ and $G_{\min}$ are the maximum and minimum conductance, respectively.

Equation (1) can be expressed as follows

$$\delta G_p = \frac{G(n+1) - G(n)}{1} = \frac{\Delta G}{\Delta n} = \alpha \exp\left(-\beta \frac{G(n) - G_{\min}}{G_{\max} - G_{\min}}\right) \tag{3}$$

where, n is the number of pulse, α and β represent α_p and β_p, respectively. We can approximate above equation as follows to be transformed into the derivative form

$$\frac{dG}{dn} = \alpha \exp\left(-\beta \frac{G(n) - G_{\min}}{G_{\max} - G_{\min}}\right). \tag{4}$$

Integrating the above equation yields the following equation

$$G_{LTP}(n) = G_{\min} + \frac{G_{\max} - G_{\min}}{\beta} \ln\left(\frac{\alpha\beta}{G_{\max} - G_{\min}}\right) + \frac{G_{\max} - G_{\min}}{\beta} \ln\left(n - 1 + \frac{G_{\max} - G_{\min}}{\alpha\beta}\right). \tag{5}$$

Therefore, the conductance logarithmically increases as the number of pulses increases in the behavior model.

In memory devices, the amount of the stored charge increases logarithmically as the number of potentiation pulses increases [23], because the previously stored charge reduces the amount of charge stored by the additional pulses by the Coulomb repulsion. The charge stored in the floating-gate below the gate acts as a gate bias to induce carriers (electrons or holes) in the channel. Thus, the effective gate bias increases logarithmically as the number of potentiation pulses increases. Furthermore, previous works using physical modeling have shown that the threshold voltage logarithmically increases as the time of program pulse increases [24], [25]. An increase in the program time corresponds to an increase in the number of pulses. In addition, we measured the conductance in the linear region. In the linear region, the current linearly decreases with increasing threshold voltage. Consequently, the current logarithmically increases as the number of erase pulses increases, and logarithmically decreases as the number of program pulses increases.

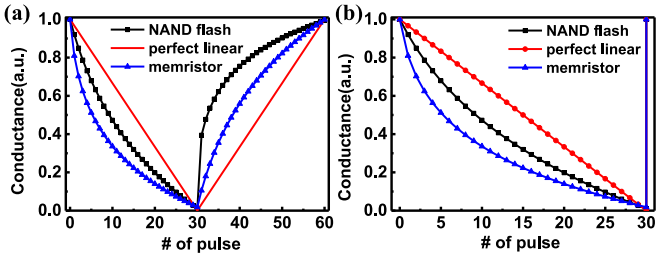

FIGURE 5. (a) Bidirectional conductance responses of NAND flash, ideal perfect linear device, and memristor [26]. (b) Unidirectional conductance responses of NAND flash, ideal perfect linear device, and memristor [26].

Since both the physical modeling of floating gate device and the behavior model in [22] mean that the conductance logarithmically increases as the number of erase pulses increases, we used the behavior model to fit the conductance behavior of NAND flash memory cells.

Because the maximum value of conductance for real devices is limited, the dynamic range of conductance is important for learning performance. In this case, we assumed that the conductance of each device reaches minimum conductance from the maximum conductance after 30 pulses. Fig. 5 (a) and (b) show bidirectional and unidirectional conductance response of devices, respectively.

IV. SIMULATION RESULT OF MNIST PATTERN RECOGNITION

We designed a 3-layer perceptron networks using NAND flash as synaptic devices and evaluated classification accuracy for MNIST hand written digit sets using matched computer simulation. Fig. 6 shows full learning procedure for designed neural networks. We adopt the online learning updating the weight of synaptic device at each training sample to reduce the burden of synapse array and peripheral circuits. In addition, weight (W) of synaptic device can be modified by one step ($|\Delta G_{ij}^{-}|, -|\Delta G_{ij}^{+}|$) at each iteration according to sign of ΔW to reduce the burden of periphery circuit. In this simulation work, the conductance response data from Fig. 5 are used for multi-layer perceptron. Fig. 7 represents activation function. Black and red line indicates hard-sigmoid function and differential value of it,

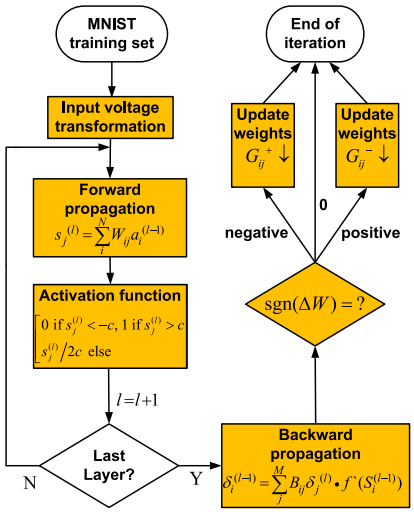

FIGURE 6. Online learning procedure for hardware-based multi-layer neural networks.

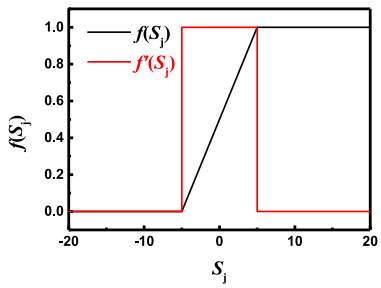

FIGURE 7. Activation function. Black line indicates hard-sigmoid function.

respectively. 60000 sets of MNIST are used for training and 10000 sets are used for testing accuracy. Input neurons are driven by pixels transformed to gray scale digital pulse (0 to 1 V). After forward propagation and computing error values, weight can be changed by one step at each iteration using identical pulse. The sign of weight (ΔW_{ij}) is used to determine whether the weight should be increased or decreased.

Fig. 8 shows simulated classification accuracy using conductance response in Fig. 5. Using bidirectional conductance response in Fig. 5 (a), the simulated accuracies for NAND flash, perfect linear, and memristor devices are 87.92%, 94.14% and 85.99% respectively as shown in Fig. 8 (a). In Fig. 5 (a), the NAND flash has more linear conductance response than memristor during programming, but has more nonlinear conductance during erasing. Therefore, in bidirectional conductance case, accuracy obtained by using NAND flash is similar to accuracy obtained with a memristor-based synapse.

On the other hand, using unidirectional conductance response in Fig. 5 (b) and weight update method in Fig. 3 (a), the simulated accuracies for NAND flash, perfect linear, and memristor devices are 86.14%, 93.89% and 72.58% respectively as shown in Fig. 8 (b).

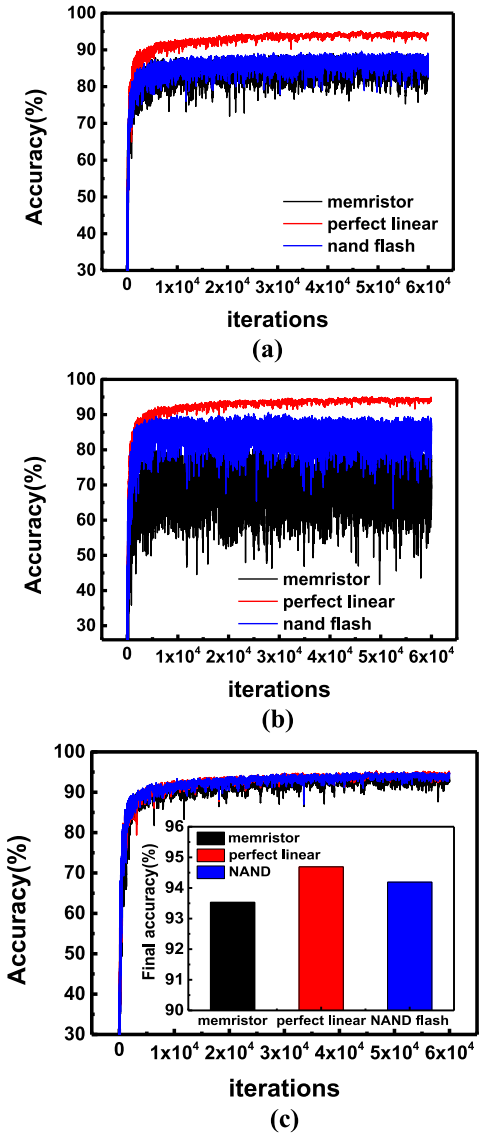

FIGURE 8. (a) Simulated classification accuracy obtained by using the bidirectional conductance response in Fig. 5 (a). (b) Simulated classification accuracy obtained by using the unidirectional conductance response in Fig. 5 (b) and weight update method in Fig. 3 (a). (c) Simulated classification accuracy obtained by using the unidirectional conductance response in Fig. 5 (b) and weight update method in Fig. 3 (b).

Even if the weight update method changes, the ideal synapse with a linear conductance response has almost the same accuracy. However, when the weight update method shown in Fig. 3 (a) is applied, a network composed of synapses with a large nonlinearity is greatly degraded in accuracy. Since the NAND synapse device has a more linear conductance response than the memristor device as shown in Fig. 5 (b), a network composed of NAND synapse devices has higher accuracy than that composed of memristor devices.

Fig. 8 (c) shows accuracy obtained by using unidirectional conductance response in Fig. 5 (b) and weight update method

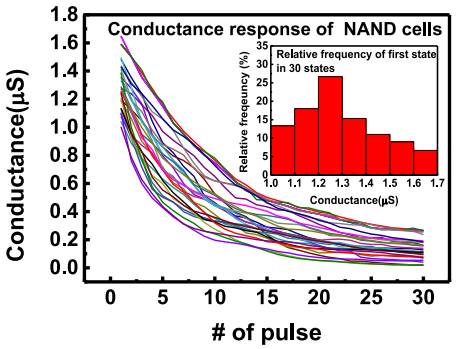

FIGURE 9. Conductance responses of 30 NAND flash cells for 30 states.

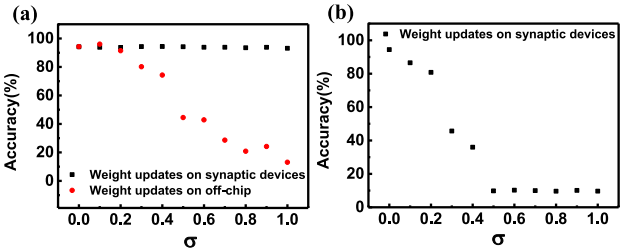

FIGURE 10. Simulated classification accuracy with respect to (a) device-to-device variation and (b) cycle-to-cycle variation. The variation is assumed to have a Gaussian distribution.

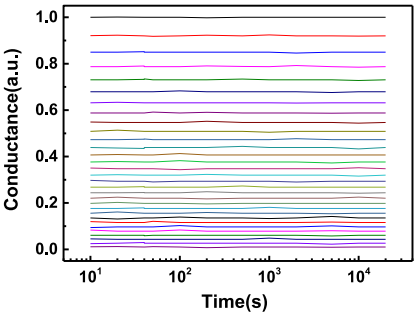

FIGURE 11. Retention characteristics of conductance states. NAND flash memory cells fabricated with 26 nm technology are measured.

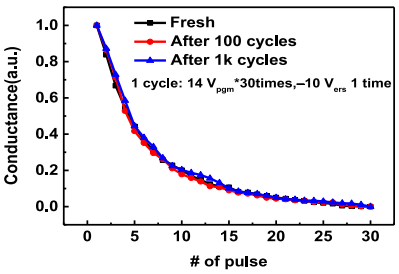

FIGURE 12. Conductance response of fresh, 100 and 1k cycled cell.

in Fig. 3 (b). The simulated accuracies for NAND flash, perfect linear, and memristor devices are 94.19%, 94.69% and 93.53% respectively. The learning accuracy obtained by the unidirectional conductance response and the weight update method in Fig. 3 (b) is higher than that obtained by the bidirectional conductance response. This is because the weight update method in Fig. 3 (b) reduces the asymmetry between weight increase and weight decrease, which is an important factor for high learning accuracy [14]. Therefore, when the weight update method in Fig. 3 (b) is applied, the accuracy obtained using NAND flash cells is similar to that obtained with ideal perfect linear devices. Thus, the unidirectional conductance response is suitable to implement multi-layer neural networks using NAND flash cells as synaptic devices.

NAND flash cells were measured for 30 conductance states to investigate the device variation as shown in Fig. 9. In addition, drain current has a linear relationship with threshold voltage in linear region and threshold voltage in NAND flash follows Gaussian distribution [27]. Therefore, we assumed the conductance distribution of NAND flash memory cells follows a Gaussian distribution, $X \sim N(1.25, 0.04)$. Learning accuracy with respect to device-to-device variation and cycle-to-cycle variation is investigated to check the effect of device variation on learning accuracy. Fig. 10 (a) shows the effect of device-to-device variation on learning accuracy when the weights are updated on off-chip and the weights are updated on synaptic devices. We assumed the distribution of the conductance of synaptic devices follows the Gaussian distribution $X(1, \sigma)$ and the standard deviation varies from 0 to 1. Learning accuracy is degraded from 94.3% to 13.11% when the weights are updated on off-chip. However, learning accuracy is negligibly degraded from 94.19% to 93.12% when the weights are updated on synaptic devices as the standard deviation increases from 0 to 1. In other words, neural network is robust to device-to-device variation when the weights are updated on synaptic devices. Fig. 10 (b) shows the effect of cycle-to-cycle variation on learning accuracy. As shown in Fig. 10 (b), the learning accuracy is degraded from 94.3% to 10.11% as the standard deviation increases from 0 to 1. Therefore, the cycle-to-cycle variation has a more detrimental effect on the learning accuracy than the device-to-device variation.

To check the reliability of NAND flash cells, endurance and retention properties are measured. Fig. 11 shows the retention characteristics of conductance states at 25°C. Compared to conventional NAND flash memory, a smaller program bias is used and the amount of electrons stored in the floating-gate is relatively smaller. Therefore, synaptic devices using NAND flash cells have excellent retention characteristics as shown in Fig. 11. We also investigated the cycle-to-cycle variation of NAND flash memory cells. As shown in Fig. 12, we can observe the conductance of the cell is almost the same up to 1k cycles. In one cycle, the cell is programmed 30 times by applying 30 pulses with a width of 100 μs and a voltage of 14 V and erased by applying 1 pulse with a width of 100 μs and a voltage of -10 V. The cycle-to-cycle variation is expressed as a percentage of the entire weight range. The calculated cycle-to-cycle variation is 1.7%. Fig. 13 illustrates the pass bias disturbance. One program pulse reduces conductance by 130 nS, while conductance is reduced by 4nS, after 10^5

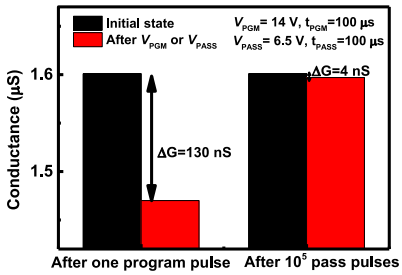

FIGURE 13. Conductance change after applying 10^5 pass bias pulses. When one pulse is applied for the program, the conductance changes significantly, whereas the pass bias pulses of 10^5 have a negligible effect on the conductance.

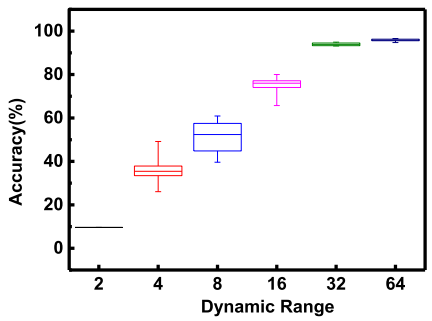

FIGURE 14. Learning accuracy with the dynamic range of synaptic device.

pass bias pulses are applied. Therefore, pass disturbance has a negligible effect on the conductance of synaptic devices.

We investigate the effect of dynamic range on learning accuracy. Fig. 14 shows the learning accuracy over the dynamic range of synaptic devices.When the dynamic range is above 32, the learning accuracy remains above 94% and the accuracy drops significantly when the dynamic range is less than about 30. When learning MNIST data, it can be said that 30 levels of dynamic range are sufficient. It is expected that higher dynamic range is required for learning more complex images.

V. CONCLUSION

In this paper, we have proposed an operation scheme of multi-layer neural networks using 2D NAND flash memory cell as a high-density and reliable synaptic device. Our scheme eliminates the waste of NAND flash cells and allows analogue input values satisfying weighted sum output equation. The conductance response of NAND flash cell is compared with those of memristor and perfect linear device. By using the conductance response and suitable weight update methods for hardware-based multi-layer neural networks, we implemented a 3-layer perceptron networks. A 3-layer perceptron network with 40545 synapses was trained on a MNIST database set. By comparing bidirectional with unidirectional conductance responses in terms of classification accuracy, it has been shown that unidirectional conductance response is suitable to implement multi-layer neural networks using NAND flash cells as synaptic devices with adaptive weight update method. Simulated classification accuracy using NAND flash cells is comparable to that obtained by perfect linear device. Finally, NAND flash memory which is cost-competitive, mature technology and has great advantage in cell density and large storage capacity can be a promising synaptic device for implementing multi-layer neural networks.

REFERENCES

[1] D. Kuzum, S. Yu, and H.-S. P. Wong, "Synaptic electronics: Materials, devices and applications," *Nanotechnology*, vol. 24, no. 38, Sep. 2013, Art. no. 382001.

[2] B. L. Jackson, "Nanoscale electronic synapses using phase change devices," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 9, no. 2, p. 12, May 2013.

[3] M. Suri, "Phase change memory as synapse for ultra-dense neuromorphic systems: Application to complex visual pattern extraction," in *Proc. IEEE Electron Devices Meeting*, Dec. 2011, pp. 1–4.

[4] G. W. Burr, "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] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[6] M. Prezioso, F. Merrikh-Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev, and D. B. Strukov, "Training and operation of an integrated neuromorphic network based on metal-oxide memristors," *Nature*, vol. 521, pp. 61–64, Mar. 2015.

[7] F. M. Bayat, "Redesigning commercial floating-gate memory for analog computing applications," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2015, pp. 1921–1924.

[8] F. Merrikh-Bayat, X. Guo, M. Klachko, M. Prezioso, K. K. Likharev, and D. B. Strukov, "High-performance mixed-signal neurocomputing with nanoscale floating-gate memory cell arrays," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4782–4790, Oct. 2018.

[9] P. Pouyan, E. Amat, and A. Rubio, "Reliability challenges in design of memristive memories," in *Proc. 5th Eur. Workshop CMOS Variability*, Oct. 2014, pp. 1–6.

[10] J.-S. Seo, "A 45nm CMOS neuromorphic chip with a scalable architecture for learning in networks of spiking neurons," in *Proc. IEEE Custom Integr. Circuits Conf.*, Sep. 2011, pp. 1–4.

[11] R. Bez and P. Cappelletti, "Flash memory and beyond," in *Proc. IEEE VLSI Technol.*, Apr. 2005, p. 84.

[12] P. Wang, "Three-dimensional NAND flash for vector–matrix multiplication," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 27, no. 4, pp. 988–991, Apr. 2019.

[13] J. Jang, "Vertical cell array using TCAT(terabit cell array transistor) technology for ultra high density NAND flash memory," in *Proc. IEEE VLSI Technol.*, Jun. 2009, pp. 192–193.

[14] S. Lim, "Adaptive learning rule for hardware-based deep neural networks using electronic synapse devices," in *Neural Computing and Applications*. London, U.K.: Springer, Jul. 2018.

[15] A. S. Spinelli, C. M. Compagnoni, and A. L. Lacaita, "Reliability of NAND flash memories: Planar cells and emerging issues in 3D devices," *Computers*, vol. 6, no. 2, p. 16, 2017.

[16] S.-H. Lee, "Technology scaling challenges and opportunities of memory devices," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, Dec. 2016, pp. 1.1.1–1.1.8.

[17] D. Kang, "256Gb 3b/cell V-NAND flash memory with 48 stacked WL layers," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Jan./Feb. 2016, pp. 130–131.

[18] C. Kim, "A 512Gb 3b/cell 64-stacked WL 3D V-NAND flash memory," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2017, pp. 202–203.

[19] O. Bichler, M. Suri, D. Querlioz, D. Vuillaume, B. DeSalvo, and C. Gamrat, "Visual pattern extraction using energy-efficient '2-PCM synapse' neuromorphic architecture," *IEEE Trans. Electron Devices*, vol. 59, no. 8, pp. 2206–2214, Aug. 2012.

[20] H. Shim, "Highly reliable 26nm 64Gb MLC E2NAND (embedded-ECC & enhanced-efficiency) flash memory with MSP (memory signal processing) controller," in *Proc. IEEE VLSI Technol.*, Jun. 2011, pp. 216–217.
[21] G. W. Burr, "Neuromorphic computing using non-volatile memory," *Adv. Phys. X*, vol. 2, no. 1, pp. 89–124, 2016.
[22] D. Querlioz, O. Bichler, P. Dollfus, and C. Gamrat, "Immunity to device variations in a spiking neural network with memristive nanodevices," *IEEE Trans. Nanotechnol.*, vol. 12, no. 3, pp. 288–295, May 2013.
[23] F. R. Libsch and M. H. White, "Charge transport and storage of low programming voltage SONOS/MONOS memory devices," *Solid-State Electron.*, vol. 33, no. 1, pp. 105–126, 1990.
[24] M. R. Zakaria and M. N. Hashim, "An overview and simulation study of conventional flash memory floating gate device using concept F-N tunnelling mechanism," in *Proc. 5th Int. Conf. Intell. Syst. Model. Simulat.*, 2014, pp. 775–780.
[25] D.-H. Kim, "Program/erase model of nitride-based NAND-type charge trap flash memories," *Jpn. J. Appl. Phys.*, vol. 49, no. 8, pp. 1–4, Aug. 2010.
[26] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu, "Nanoscale memristor device as synapse in neuromorphic systems," *Nano Lett.*, vol. 10, no. 4, pp. 1297–1301, Mar. 2010.
[27] Y. Cai, "Threshold voltage distribution in MLC NAND flash memory: Characterization, analysis, and modeling," in *Proc. Design Autom. Test Europe Conf. Exhibit.*, 2013, pp. 1285–1290.

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. degrees. He is with the Inter-University Semiconductor Research Center, SNU. His current research interests include neuromorphic system and its application in computing.

SUHWAN LIM received the B.S. and M.S. degrees in electrical engineering from the Pohang University of Science and Technology, Pohang, South Korea, in 2008 and 2010, respectively. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Seoul National University, Seoul, South Korea. His current research interests include neuromorphic system and neural networks.

NAG YONG CHOI received the B.S. degree from Seoul National University, Seoul, South Korea, in 2014, where he is currently pursuing the combined master's and Ph.D. degrees with the Department of Electrical and Computer Engineering.

JONG-HO BAE received the B.S. degree in electrical engineering from the Pohang University of Science and Technology, Pohang, South Korea, in 2011. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Seoul National University, Seoul, South Korea, where he is with the Inter-University Semiconductor Research Center.

DONGSEOK KWON received the B.S. degree in electrical engineering from the Pohang University of Science and Technology, Pohang, South Korea, in 2017. He is currently pursuing the M.S. degree with the Department of Electrical and Computer Engineering, Seoul National University, Seoul, South Korea. His current research interests include neuromorphic system and its application in computing.

BYUNG-GOOK PARK (M'90) 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.

JONG-HO LEE (F'16) received the Ph.D. degree in electronic engineering from Seoul National University (SNU), Seoul, in 1993.

He was a Post-Doctoral 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.