This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3083056, IEEE Access

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Spiking Neural Networks with Time-to-First-Spike Coding Using TFT-type Synaptic Device Model

SEONGBIN OH[1], (Graduate Student Member, IEEE), SOOCHANG LEE[1], (Graduate Student Member, IEEE), SUNG YUN WOO[1], (Member, IEEE), DONGSEOK KWON[1], (Graduate Student Member, IEEE), JISEONG IM[1], (Graduate Student Member, IEEE), JOON HWANG[1], (Member, IEEE), JONG-HO BAE[2], (Member, IEEE), BYUNG-GOOK PARK[1], (Fellow, IEEE), JONG-HO LEE[1], (Fellow, IEEE)

[1]Department of ECE and ISRC, Seoul National University, Seoul 08826, South Korea
[2]School of Electrical Engineering, Kookmin University, Seoul, 02707, Republic of Korea

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

This work was supported by the Brain Korea 21 Plus Project in 2021, and MOTE (Ministry of Trade, Industry & Energy (10080583) and KSRC (Korea Semiconductor Research Consortium) support program for the development of the future semiconductor device.

ABSTRACT In hardware-based spiking neural networks (SNNs), the conversion of analog input data into the arrival time of an input pulse is regarded as a good candidate for the encoding method due to its bio-plausibility and power-efficiency. In this work, we trained an SNN encoded by time to first spike (TTFS) and performed an inference process using the behavior of the fabricated TFT-type flash synaptic device. The exponentially decaying synaptic current model required in the inference process was implemented by reading devices in the subthreshold region using triangle pulses. In a high-level system simulation, the TTFS-SNN (two-layer MLP with 512 hidden neurons) reached a high accuracy of 97.94%. Compared to conventional rate-encoded SNNs, TTFS-SNN made 2.9 times faster judgment and consumed ~10 times less energy in the inference process. Additionally, to use the network in a more stable condition, we propose a method to operate it using a rectangle pulse in the saturation region of the synaptic device. The distortion caused by this approximation was minimized by shortening the pulse width. As a result, the modified inference system showed an accuracy of 97.36%, and the prediction time and energy consumption were reduced 3.97- and 83.04-times when compared to those of the rate-SNN. Finally, we analyzed the sensitivity of the network performance due to unexpected issues that may occur in the hardware system and thus explained the competitiveness of the proposed synaptic behavior in the saturation region.

INDEX TERMS Time-to-First-Spike (TTFS) Coding, Temporal Coding, Flash-based Synaptic Device, Hardware-based Spiking Neural Networks, Neuromorphic Systems

I. INTRODUCTION

Recently, SNNs have become regarded as a successful computing system and have been widely studied due to their compatibility with hardware implementation, enabling the parallel operation of massive data and low-power computing [1], [2]. However, it is difficult to train an SNN directly due to its nature of transmitting data by all-or-nothing discrete spikes. As one of the candidate methods for training SNNs, it was studied to transfer the weights trained by the ANN to the SNN [3], [4]. The conventional ReLU activation function can be approximated by a combination of the integrate and fire (I&F) neurons and the rate-encoding method that expresses the analog-valued input as the frequency of the input pulse in the SNN [5]. Due to this approximation, the SNN can achieve the performance of a highly advanced ANN without significant degradation and can greatly improve the learning speed by using GPU-accelerated training packages. However, the method of converting the analog input value into the frequency of the input pulse requires a very large number of spikes, which can lead to an increase in power consumption as the structure of the SNN becomes deeper and larger. These characteristics may not be suitable for edge computing in terms of power-efficiency and device endurance. In addition, many analog synaptic devices, including RRAM, show severe variation issues when operating with a small current density [6]. Thus, in order to build SNNs that are robust to variation, a synaptic device with a large current density should be used. Further, in order to obtain power-efficiency when using large current synaptic devices, it is necessary to study a network that uses only a small number of spikes for inference.

Another candidate for the encoding method is temporal encoding, in which the input values are converted to the arrival times of the spikes. An input neuron that accepts an input value larger than a small input value fires earlier. The input data is only represented by a single spike - regardless of the intensity of the input data; thus, sparse spikes are used in the inference process. In temporal coding, input data can be expressed in the order of firing time of the input nodes (rank-order) [7] or the arrival time of the spikes (time to first spike) [8]-[13].

To train the SNN encoded by the time-to-first-spike (TTFS), the conventional training method for the rate-encoded network is not suitable. The relationship between the input and output of each layer needs to be newly defined, and the modified method must be applied accordingly. Several previous works have been reported regarding training methods suitable for the TTFS-SNN [8]-[13]. However, the system of Rueckauer et al. (2018) [8] is more similar to the pulse width encoding system, because the integrated charge is affected by the pulse width rather than the arrival time of the input pulse. In Comsa et al. (2020) [9] and Zhang et al. (2020) [10], an alpha function was used as a synaptic current model, which may require more burden to be implemented in hardware. The network of Mostafa et al. (2017) [11] was successfully trained by defining a piece-wise linear relationship between the input and output of each layer, and a number of studies are followed, such as implementing a simple network in hardware by Billaudelle S et al. (2019) [12].

The contributions of our work are as follows:

1) A TFT-type flash device is fabricated and investigated as a synaptic device with nonvolatile memory function. Each device presents 32 states and is operated in the saturation region to ensure stable operation against drain-side noise or voltage changes.
2) We use the TTFS-encoded SNN model and training method proposed in prior work [11]. We newly implement the exponentially decaying synaptic current model in hardware by reading the device with triangle pulses in the subthreshold region.
3) We also propose a method of inference by operating a synaptic device in a saturation region to keep the network stable against unexpected variation. To do this, we use rectangular pulses instead of triangle pulses in the inference process and compare the accuracy, energy consumption, and latency with those of the rate-encoded network.
4) The sensitivity of the network's accuracy in regard to synaptic variability is analyzed by taking into account the operation region of the synapse for reading. We finally propose a stable and power-efficient method using an SNN.

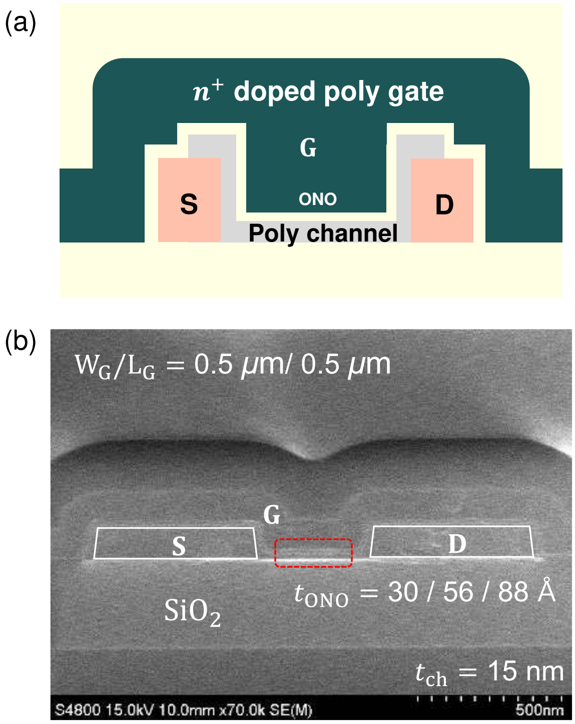

(a)
n^+ doped poly gate
G
ONO
S
D
Poly channel
(b)
W_G/L_G = 0.5 μm/ 0.5 μm
G
S
D
t_{ONO} = 30 / 56 / 88 Å
SiO_2
t_{ch} = 15 nm
S4800 15.0kV 10.0mm x70.0k SE(M)
500nm

FIGURE 1. (a) Schematic cross-sectional view of the TFT flash-based synaptic device. (b) Cross-sectional SEM image of the fabricated device.

II. Methods

A. Fabrication and Synaptic Operation of The TFT-type Flash Device

Fig. 1 shows a schematic view and TEM image of a TFT-type flash synapse device. The device has the structure of a FET with a charge trap layer and is applied to an AND-type array architecture. The thicknesses of the SiO_2 / Si_3N_4 / SiO_2 stack are 30 / 56 / 88 Å, respectively. The length (L), width (W), and thickness of the channel (t_{ch}) are 0.5 μm, 0.5 μm, and 15 nm, respectively. The device was fabricated with conventional CMOS process technology. First, an insulator layer and an n^+ doped poly-Si layer were sequentially deposited on the wafer, and then the poly-Si is patterned for the source and drain. Subsequently, a 15-nm-thick amorphous Si layer is deposited and then poly-crystallized through annealing. Finally, a charge trap layer and n^+ doped poly-Si are deposited, and the gate is formed.

The I_D-V_{GS} characteristics of the synapse device are shown in Fig. 2 (a). The V_{th} of the synaptic device increases when applying V_{PGM} (V_G = 0 V, V_S = -9.5 V, t_{pulse} = 100 μs) to the gate and source. The operation characteristics of the synaptic device are different depending on the read bias of the device. In this paper, the SNNs with synapses read in the subthreshold ($V_{GS} < V_{th}$) or saturation ($0 < V_{GS} - V_{th} < V_{DS}$) regions were analyzed. Fig. 2 (b) and (c) show the I_D-V_{DS} characteristics of the synaptic device when V_{GS} is 1.5 and 2.5 V, corresponding to the subthreshold and saturation region, respectively. For the

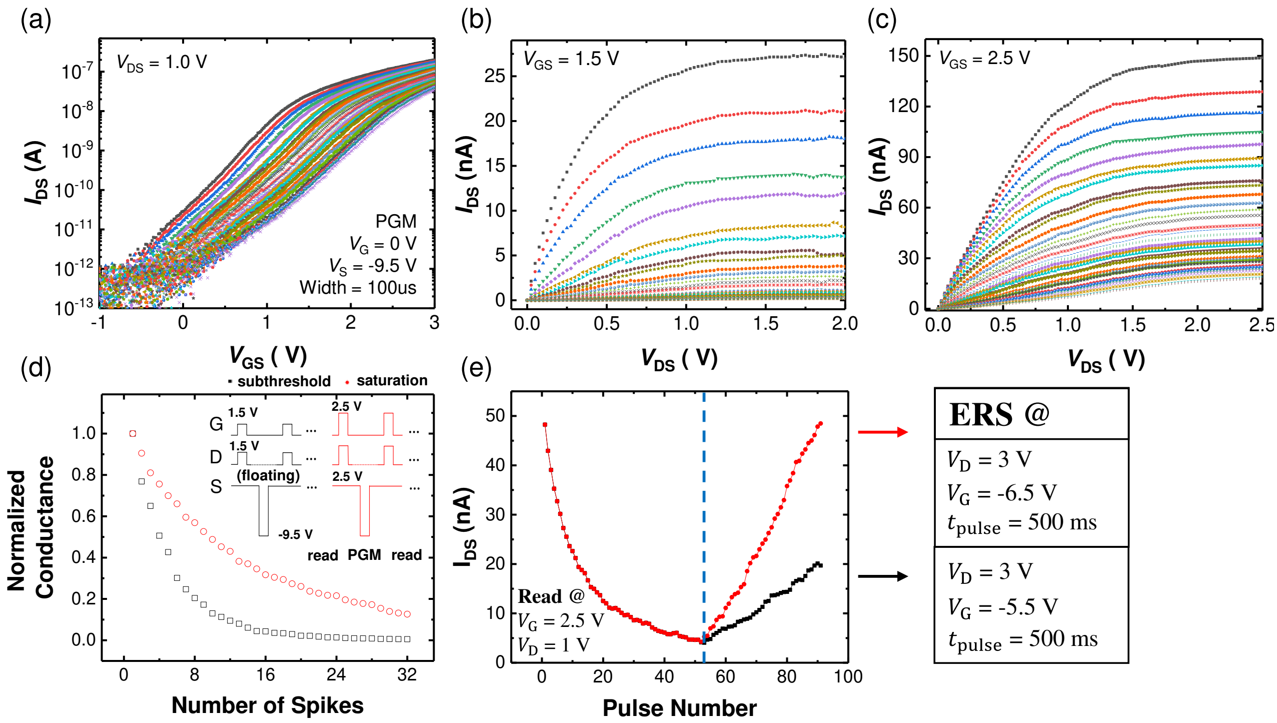


FIGURE 2. (a) Measured I_D-V_{GS} curve and program condition of the synaptic device. Measured I_D-V_{DS} characteristics of the synaptic device read (b) before threshold voltage (V_{GS}= 1.5 V) and (c) after threshold voltage (V_{GS} = 2.5 V). (d) The change of the normalized conductance as a parameter of the number of program pulses. The inset shows the voltage scheme for reading and programming the synaptic device in the subthreshold and saturation region. (e) Drain current change with respect to the number of pulses as a parameter of erase condition.

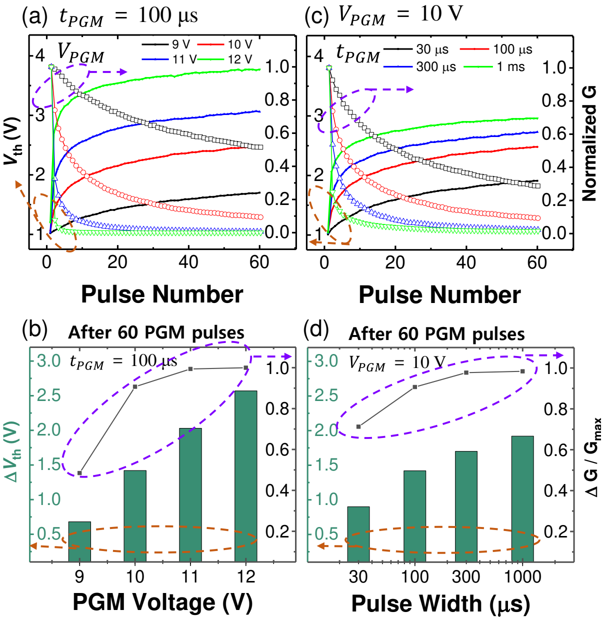


FIGURE 3. V_{th} and normalized conductance shift with respect to (a) the amplitude (the width is fixed at 100 μs) and (b) the width of the program pulses (the amplitude is fixed at 10 V). ΔV_{th} and ΔG / G_{max} as a function of (c) the amplitude and (d) the width of the program pulses.

reliable operation of the network, the synapse current should not change—even with unexpected changes in bias applied to the drain of the synapse device, such as noise [14] or an IR drop issue [15]. Hence, the drain bias is set to be larger than (V_{GS} - V_{th}) when reading the device. Fig. 2 (d) shows the change in the conductance of the device by the number of program spikes. The circle and square symbols are the results of reading in the subthreshold (V_{GS} = V_{DS} = 1.5 V) and saturation regions (V_{GS} = V_{DS} = 2.5 V), respectively. Each weight trained by the ANN is represented as the difference in conductance of the two synaptic devices [16]. The measurement results for the 32 states of one synaptic device are shown in Fig. 2 (a) to (c); so the weights can be quantized with 63 steps. Fig. 2 (e) shows the change in conductance as a function of the number of erase pulses. The device's conductance can be increased by increasing the number of applied erase pulses, so the weights of devices in the array can be updated.

Fig. 3 shows the V_{th} shift by program pulses. Obviously, as the amplitude or width of the program pulse increases, the value of ΔV_{th} also increases, which means that the device can be used over a larger dynamic range. Also, since the programmed charges interfere with the additional charge being programmed, the value of ΔV_{th} gradually decreases.

B. Training of Temporally Encoded Data

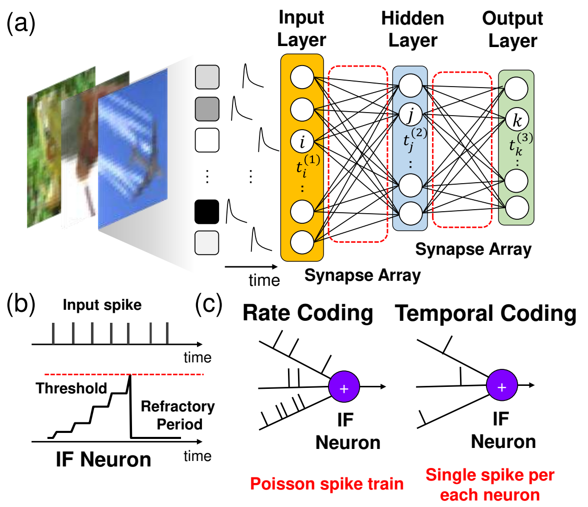


FIGURE 4. (a) Schematic of the TTFS-SNNs. (b) Integrate and fire behavior of neuron circuits. (c) Two methods (Rate coding and Temporal coding) for encoding analog data to SNN

The input data is encoded using the TTFS method. As shown in Fig. 4 (a), the arrival time of the spike of input neuron *i* is inversely proportional to the input value as follows:

$$t_{\text{i}}^{\text{input}} = \left[\frac{Y_{\max}-Y_{\text{i}}}{Y_{\max}} T_{\max}\right] \qquad (1)$$

where $Y_{\max}$ is the maximum value of input data, and Y_{i} is the input value of the i^{th} input neuron. $T_{\max}$ represents the number of time steps. In this paper, it is assumed that the synaptic current decays exponentially in the time-domain, as follows:

$$I_{\text{syn}}^{\text{ij}} = I_{\text{ij}} \exp(-(t-t_{\text{i}}^{(\text{l})}))\ H(t-t_{\text{i}}^{(\text{l})}) \qquad (2)$$

where $H(x)$ is a Heaviside step function, whose value is zero for negative arguments and one for positive arguments. In (2), I_{ij} represents the current value at the arrival time of the pulse between input *i* and output *j*. Also, $t_{\text{i}}^{(\text{l})}$ is a firing time of i^{th} input neuron in l^{th} layer.

As shown in Fig. 4 (b), the output neuron integrates the input spike train, and when the membrane voltage reaches the neuron threshold, the neuron sends a spike to the next layer. Since it is assumed that the TTFS-SNN has each neuron spike once at most, the already fired neurons enter the refractory period so that no more charge is integrated into the neuron. The membrane potential over time is given by:

$$V_{\text{mem}}^{\text{j}}(\text{t}) = \sum_{\text{i}} w_{\text{ij}} \left(1-\exp\left(-(t-t_{\text{i}}^{(\text{l})})\right)\right) H\left(t-t_{\text{i}}^{(\text{l})}\right) \qquad (3)$$

where w_{ij} represents the membrane voltage integrated by the j^{th} output neuron by the pulse of the i^{th} input neuron, which is obtained by dividing I_{ij} by the membrane capacitance. V_{mem} in (3) is implemented by the membrane charge of a neuron.

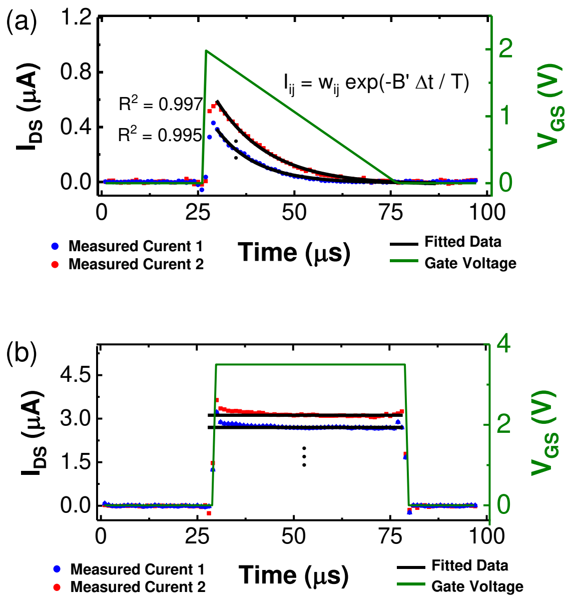


FIGURE 5. Measured I_D - t and V_{GS} - t plots of the synaptic device. (a) Exponentially decaying current by triangle pulse in subthreshold region and (b) rectangle current by step pulse in saturation region. t_{pulse}, t_{rise}, t_{fall} is 50 μs, 3 μs, 3 μs, respectively.

Since neurons integrate synaptic current by performing the integrate-and-fire (IF) operation, (3) can be obtained by integrating (2) over *t*. In previous studies, the neuron circuits for IF operation was implemented in hardware [17-19], and the operation of synapse array and neuron circuits was verified in circuit-level simulation [20]. Only the input neurons that have already fired at the firing time of the j^{th} output neuron influence the j^{th} membrane voltage; so, this is called a causal set C_{j}. The time when the j^{th} membrane voltage reaches the neuron threshold voltage (V_{th}) is expressed as $t_j^{(l+1)}$, which is the firing time of the j^{th} neuron. The membrane voltage at $t_j^{(l+1)}$ is expressed as:

$$V_{\text{th}} = \sum_{\text{i}\in C_{\text{j}}} w_{\text{ij}} \left(1-\exp\left(-(t_{\text{j}}^{(\text{l}+1)}-t_{\text{i}}^{(\text{l})})\right)\right). \qquad (4)$$

Hence, the relationship between input and output firing time is represented as:

$$\exp(t_{\text{j}}^{(\text{l}+1)}) = \frac{\sum_{\text{i}\in C_{\text{j}}} w_{\text{ij}} \exp(t_{\text{i}}^{(\text{l})})}{\sum_{\text{i}\in C_{\text{j}}} w_{\text{ij}} - V_{\text{th}}}. \qquad (5)$$

As expressed in (5), the exponential form of the input and output spike times of a layer has a piecewise-linear relation.

The most complex part of the above-mentioned SNN is implementing the exponentially decaying synaptic current model in hardware. In several previous works, TTFS data was trained on the network using the unconventional current model, but most models [21] were too complex to implement in hardware. In this paper, we propose a method for reading FET-type synaptic devices in the subthreshold region using triangle pulses expressed as:

$$V_{\mathrm{GS}}^{\mathrm{i}}(\mathrm{t}) = V_{\mathrm{o}}\,(1-\frac{t-t_{\mathrm{i}}}{T})\,H(t-t_{\mathrm{i}}) \qquad (if\ \ t_{\mathrm{i}} < t < t_{\mathrm{i}}+T) \quad (6)$$

where V_{o} is the maximum value of the read voltage, and T is the time width of triangle pulse. By the triangle gate voltage in the subthreshold region, the current has the form of

$$I_{\mathrm{sub}}^{\mathrm{ij}}(t) = A\exp(BV_{\mathrm{GS}}) = I_{\mathrm{ij}}\exp(-B'\frac{t-t_{\mathrm{i}}}{T})\,H(t-t_{\mathrm{i}}) \quad (7)$$

where A and B represent the coefficients of the current equation in the subthreshold region, and B' is a coefficient dependent on the subthreshold swing of the synapse device. When a triangle pulse that decreases linearly in the time-domain is applied to the gate of the device, the subthreshold current can effectively represent an exponentially decaying model. Fig. 5 shows the transient measurement results of a TFT type flash synaptic device in the time-domain. Even after programming the device, the synaptic current is fitted to the same equation. Then, by the equations (2) to (5), the relationship between input and output values is defined as follows:

$$\exp(B'\frac{t_{\mathrm{j}}^{(\mathrm{l}+1)}}{T}) = \frac{\sum_{\mathrm{i}\in C_{\mathrm{j}}} w_{\mathrm{ij}}\exp(B'\frac{t_{\mathrm{i}}^{(\mathrm{l})}}{T})}{\sum_{i\in C_{\mathrm{j}}} w_{\mathrm{ij}} - V_{\mathrm{th}}}. \quad (8)$$

When we set $\exp(B'\frac{t_{\mathrm{i}}^{(\mathrm{l})}}{T})$ to $z_{\mathrm{i}}^{(\mathrm{l})}$, the input and output values of the network are in the form of a piecewise-linear function in the z-domain as follows:

$$z_{\mathrm{j}}^{(\mathrm{l}+1)} = \frac{\sum_{\mathrm{i}\in C_{\mathrm{j}}} w_{\mathrm{ij}} z_{\mathrm{i}}^{(\mathrm{l})}}{\sum_{\mathrm{i}\in C_{\mathrm{j}}} w_{\mathrm{ij}} - V_{\mathrm{th}}}\,. \quad (9)$$

For training this network, we use the cross-entropy loss function, given by:

$$Loss[\mathrm{g}] = -\ln(\frac{(1/z[\mathrm{g}])}{\sum_{\mathrm{i}}(1/z[\mathrm{i}])}) \quad (10)$$

where g and i represent the indexes of the target neuron and other top neurons, respectively. The training aims to minimize the loss function, thus maximizing the difference between the firing times of the target class neuron and the other output neurons in the top layer. In order to use the architecture of the network more efficiently, it is recommended to remove useless neurons that do not participate in data propagation. In other words, it is desirable to eliminate dormant neurons that are not fired by any input data. Hence, we add the term expressed by:

$$\mathrm{Cost1} = K1 \times \sum_{\mathrm{j}} \max\left(0, V_{\mathrm{th}} - \sum_{\mathrm{i}\in C_{\mathrm{j}}} w_{\mathrm{ij}}\right) \quad (11)$$

where $K1$ is a hyper-parameter for network, and i and j represent the input and output indexes of each layer, respectively. This term ensures that an output neuron spikes if all input neurons spike. In addition, L2-norm regularization is used to avoid overfitting due to excessively large weight values. No other regularization skills are applied.

In this study, the arrival time of the spike is set linearly to facilitate hardware implementation. Therefore, if the distribution of input data set is expressed as:

$$Y_{\mathrm{x}} = \{ \quad \frac{0}{255} \quad , \quad \frac{1}{255} \quad , \ \dots \quad \frac{255}{255} \quad \}, \quad (12)$$

the firing time of each input neuron is expressed as:

$$t_{\mathrm{x}} = \{ \quad 255 \quad , \quad 254 \quad , \ \dots \quad 0 \quad \}. \quad (13)$$

Then, the input data set of ANN given by

$$z_{\mathrm{x}} = \{\exp(B'\times\frac{255}{T}),\ \exp(B'\times\frac{254}{T}),\ \dots\ \exp(B'\times\frac{0}{T})\} \quad (14)$$

is expressed as an exponential distribution.

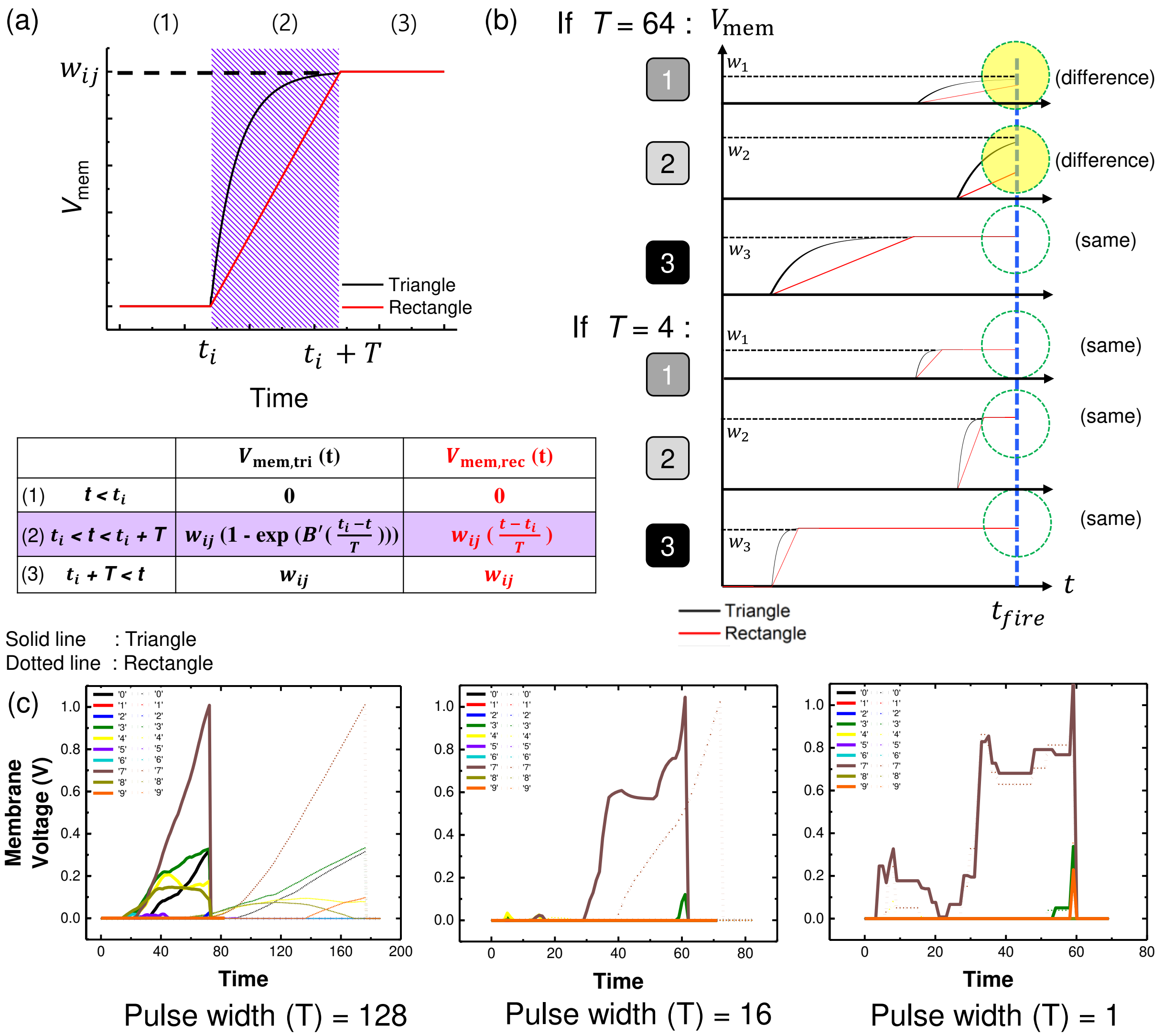
In this paper, the Adam optimizer from the PyTorch framework with a floating-point operation is used to train the ANN with the input-output relation shown in (9). Then, trained weights are transferred to the SNN, and the inference process is performed in the time-domain. The learning rate starts at 10^{-3} and exponentially decays to 10^{-6} as the epoch increases. The coefficient $K1$ in (11) is 100, and L2-norm regularization coefficient is 10^{-4}. The value of B', calculated from the subthreshold swing of our synaptic device shown in Fig. 2 (a) is 6.3. Lastly, the membrane threshold of the neuron is assumed to be 1 V.

III. Results

A. Approximation of Synaptic Current Model in Inference Process

For the inference process in hardware-based SNN, it is important to develop a training method for the ANN to achieve a good performance, but it is also necessary to ensure the inference process is performing reliably on chip. The SNN described in the previous section should be operated in the subthreshold region due to the exponentially decaying current model, which is very sensitive to unexpected device variations and noise. Besides, there is a hardware burden of adding the triangle pulse generator to each neuron model used in a conventional SNN. Hence, we propose an approximated inference process using rectangle pulses for a more stable operation in hardware. The use of a rectangle pulse allows operating the synaptic device in the saturation region, making it more robust against realistic problems such as noise or synaptic variation.

Fig. 6 (a) shows the integration of the membrane voltage (or membrane charge) of a neuron by one input voltage pulse.



	$V_{\text{mem,tri}}$ (t)	$V_{\text{mem,rec}}$ (t)
(1) $t < t_i$	0	0
(2) $t_i < t < t_i + T$	w_{ij} (1 - exp ($B'(\frac{t_i - t}{T})$))	w_{ij} ($\frac{t - t_i}{T}$)
(3) $t_i + T < t$	w_{ij}	w_{ij}

FIGURE 6. (a) Dynamics of V_{mem} in response to single triangle (black) and rectangle (red) pulses. The table shows the expression of the membrane voltage in time-domain. (b) Comparison of $V_{mem} - t$ plots by two-type pulses from input neurons 1,2,3. V_{mem} is compared for two-type networks (each of the pulse widths is 4 or 64 in 256 total time steps). (c) V_{mem} dynamics of the top layer neurons by randomly picked MNIST image. The solid line and dotted line represent the V_{mem} dynamics by the triangle and the rectangle pulses, respectively.

When the integration by two types of pulses with the same width is compared in the time-domain, the membrane voltage before the pulse arrives (region[1]) and after the pulse ends (region[3]) is the same. On the other hand, there is a difference when the membrane capacitor integrates the pulse (region(2)), which is the cause of an error in the approximation of replacing a triangle pulse with a rectangle pulse. In other words, input nodes with weighted sum values distorted by approximation are subsets of $D = \{i : \max(0, t - T) < t_i < t\}$ at time t. We call this set of input spikes the 'distorted set'. We propose a method to minimize the number of affected input neurons by reducing the pulse width compared to the total time step. For instance, if the pulse width decreases from 64 steps to 4 steps, the number of input neurons included in the distorted set $\{i : \max(0, t_{\text{fire}} - T) < t_i < t_{\text{fire}}\}$ of an output neuron decreases stochastically, as shown in Fig. 6 (b). As a result, the distortion of the membrane voltage due to the approximation can be reduced. Fig. 6 (c) shows the simulated membrane voltage of the top-layer neurons from the randomly picked MNIST image. The t_{fire} of the top neuron in the two types of inference becomes almost the same as the pulse width decreases sequentially with 128, 16, and 1 step. Therefore, it is shown that it can be approximated more accurately using a short pulse width.

Fig. 7 shows a simulated inference accuracy of the two-layer SNN using triangle and rectangle pulses in the inference process as a parameter of the pulse width. The accuracy of the SNN clearly increases as the number of hidden neurons increases. Furthermore, even if the inference process is performed using a triangle pulse, it is shown that the accuracy gradually drops as the pulse width decreases. This tendency is attributed to the dataset z_x shown in (14). If the pulse width is shortened, the deviation in the data set between the minimum and maximum values grows

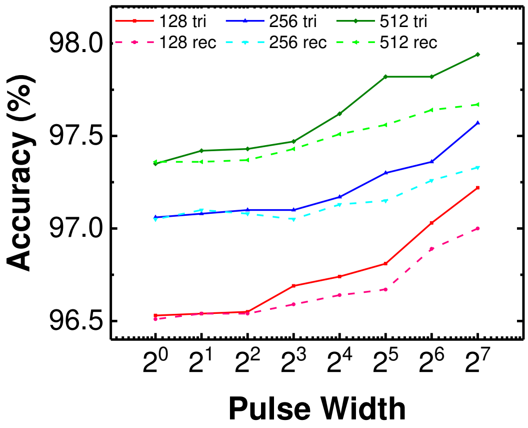


FIGURE 7. Accuracy - pulse width plots as a parameter of the number of hidden neurons. The solid and dotted lines indicate the performance of the network using triangle and rectangle pulses,

exponentially, making it more difficult to train an ANN of the same structure. In addition, if the pulse width is large, the inference with the rectangle pulse causes a large distortion, which decreases accuracy. However, as the pulse width becomes smaller and approaches 1 step, there is little difference in the performance of the two networks. Indeed, in a two-layer MLP with 512 hidden neurons, the accuracy decreases from 97.94 % to 97.36 % as the pulse width is reduced from 128 steps to 1 step. The gap between the accuracy of the SNN using the triangle and rectangle pulses was 0.27 % when the pulse width was 128 steps, but it reduced to 0 % as the pulse width decreases to 1 step. We also trained a simple SCNN structure of SCNN(5,32,2)-SCNN(5,16,2)-FC(10). Here, SCNN(5,32,2) means SCNN layer with 32 5 x 5 kernels and stride 2. FC(10) means FC layer with 10 output neurons. We obtained an accuracy of 98.88% on the SCNN, which is slightly lower than that of SCNN of the same size by Zhou et al [22]. This is because we set the input data set of ANN (z-domain) in an exponential (more difficult) form to facilitate hardware implementation in SNN (t-domain).

B. Comparison with Rate-Encoded Network

Fig. 8 shows a comparison of the characteristics between the conventional rate-encoded SNN and TTFS-SNN. The synaptic devices constructing the SNNs of Fig. 8 (a) and (b) operate in the subthreshold (V_{GS} = 1.5 V, V_{DS} = 1.5 V) and saturation (V_{GS} = 2.5 V, V_{DS} = 2.5 V) regions, respectively. In addition, it is assumed that the TTFS-SNNs shown in (a) and (b) use triangle and rectangle pulses, respectively.

The accuracy depicted by the solid lines in Fig. 8 is slightly lower for the TTFS-SNN compared to the rate-SNN. The slight difference in the accuracy of the two networks is due to the nature of TTFS networks making early judgments. The index of the top neuron fired first is the result predicted by the SNN, so input nodes with firing times later than t_{fire} cannot affect the prediction. Therefore, a relatively small number of neurons participate in the inference task, and the accuracy slightly decreases.

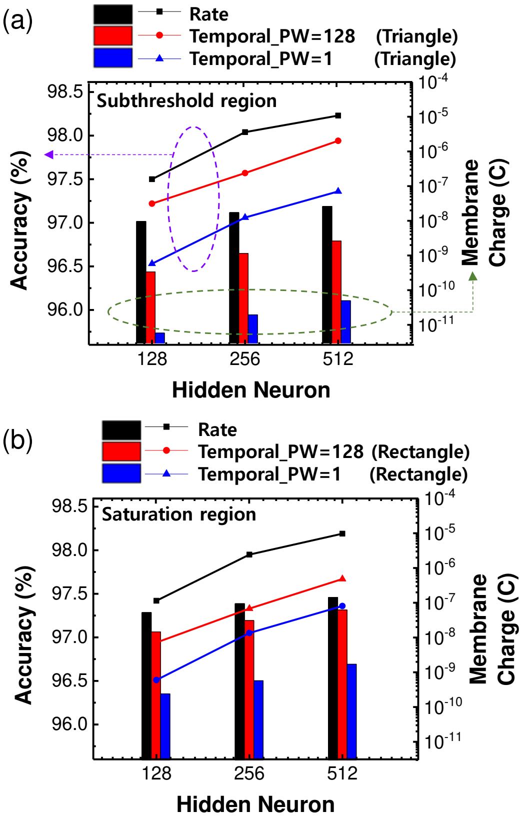


FIGURE 8. Comparison of accuracy (solid lines) and membrane charge (bars) as a parameter of the encoding rule. The synaptic device is operated in the (a) subthreshold region and (b) saturation region, respectively.

On the other hand, an integrated membrane charge, shown as bars in Fig. 8, is calculated by integrating the synaptic currents of the entire network during the time t. This parameter indicates the energy consumed by the synapse array, which accounts for a very high percentage of the consumed energy in the entire network [20]. One image is encoded with 256 time steps, and a single time step is assumed to be 1 μs. The TTFS-SNN consumes very little energy compared to the rate-SNN due to the smaller number of spikes and shorter prediction time. As the pulse width becomes shorter, the power-efficiency of the SNN is further improved, but the accuracy drops due to the wide deviation in the input data. The error rate, the latency, and the integrated charge values are compared in Tables I, II, and III. Indeed, in the 2-layer MLP (512 hidden neurons, operated in the saturation region, pulse width of 128), the TTFS-SNN has 0.61% less accuracy and 2.28 times less energy consumption than the rate-encoded SNN. When the pulse width is reduced to 1 time step, the accuracy drop reaches 0.93%, but the consumed energy decreases 83.0 times.

Another important advantage in the TTFS encoding method is low-latency. The rate-encoded network can only make a

TABLE I
THE SUMMARY OF SNN (784 – 128 – 10)

Operation Region	Encoding Rule	Error Rate (%)	Latency	Integrated Charge (× 10^{-10} C)
Subthreshold (Triangle)	Rate	2.56	256.0	96.2
	TTFS (PW = 128)	2.78	84.47	3.37
	TTFS (PW = 1)	3.47	73.95	0.0581
Saturation (Rectangle)	Rate	2.49	256.0	523
	TTFS (PW = 128)	3.0	187.7	146
	TTFS (PW = 1)	3.49	75.7	2.40

TABLE II
THE SUMMARY OF SNN (784 – 256 – 10)

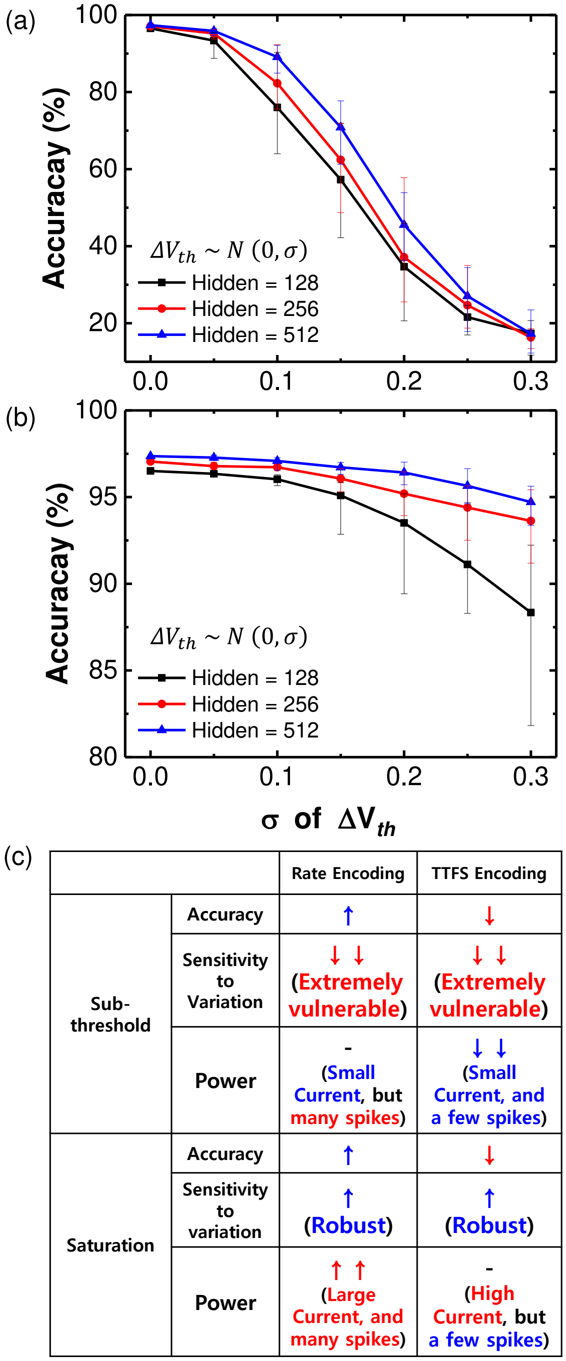
Operation Region	Encoding Rule	Error Rate (%)	Latency	Integrated Charge (× 10^{-10} C)
Subthreshold (Triangle)	Rate	1.95	256.0	173
	TTFS (PW = 128)	2.43	85.22	11.4
	TTFS (PW = 1)	2.94	68.84	0.195
Saturation (Rectangle)	Rate	1.95	256.0	941
	TTFS (PW = 128)	2.67	192.8	308
	TTFS (PW = 1)	2.95	70.74	5.73

TABLE III
THE SUMMARY OF SNN (784 – 512 – 10)

Operation Region	Encoding Rule	Error Rate (%)	Latency	Integrated Charge (× 10^{-10} C)
Subthreshold (Triangle)	Rate	1.85	256.0	261
	TTFS (PW = 128)	2.06	87.86	26.1
	TTFS (PW = 1)	2.64	62.13	0.497
Saturation (Rectangle)	Rate	1.72	256.0	1420
	TTFS (PW = 128)	2.33	173.6	622
	TTFS (PW = 1)	2.64	64.43	17.1

decision at the last time step after all input pulses arrive. On the other hand, the TTFS encoding method enables quick prediction because the inference process ends when one neuron in the top layer is fired. As shown in Table III, in the 2-layer MLP (512 hidden neurons, operated in the saturation region, pulse width of 1), TTFS coded network has improved latency by 3.9 times compared to the rate-coded network. Networks with triangle pulses show lower latency than those with rectangle pulses. This is because the triangle pulse takes a shorter time to reach the same membrane voltage as shown in Fig. 6 (a). The results depicted in Fig. 8 and Table I-III were obtained by system simulation.

C. Analysis on Variability of Synaptic Array



		Rate Encoding	TTFS Encoding
Sub-threshold	Accuracy	↑	↓
	Sensitivity to Variation	↓ ↓ (Extremely vulnerable)	↓ ↓ (Extremely vulnerable)
	Power	- (Small Current, but many spikes)	↓ ↓ (Small Current, and a few spikes)
Saturation	Accuracy	↑	↓
	Sensitivity to variation	↑ (Robust)	↑ (Robust)
	Power	↑ ↑ (Large Current, and many spikes)	- (High Current, but a few spikes)

FIGURE 9. Sensitivity of the network performance due to unexpected variation of V_{ov} of synaptic array operating in (a) subthreshold and (b) saturation regions. (c) Strength and weakness of SNN as a parameter of the data encoding method and the operation region of the synaptic device.

The biggest obstacle in inferring data on chip using pre-trained ANN weights is managing the variability of the conductance of the synaptic array [23], [24]. The weights of the SNN can change when it is transferred to the synaptic array or due to the V_{th} shift of the synaptic device over time. Also, noise caused by neurons may affect the operation of the synapse devices. Even though the synapse device is operated in the saturation region—so it is robust to the noise on the drain—the gate noise directly affects the synapse current. We investigate the sensitivity of the non-ideal characteristics of the synaptic array on the performance of the network. It is assumed that the change in threshold voltage has a normal distribution in the synaptic array [25], and the synaptic devices are operated in the subthreshold and saturation regions, respectively.

In the subthreshold region, the current has an exponential relationship to the overdrive voltage; so, the small changes of the threshold voltage affects the current exponentially as follows:

$$I_{\mathrm{D,real}} = I_{\mathrm{D,ideal}} \times \exp(B \times \Delta V_{\mathrm{th}}) \tag{15}$$

where $I_{\mathrm{D,real}}$ and $I_{\mathrm{D,ideal}}$ represent the device currents with and without synaptic variability, respectively. If the variation of the threshold voltage is assumed in the form:

$$\Delta V_{\mathrm{th}} \sim N(0, \sigma^2)\ , \tag{16}$$

the conductance considering the synaptic variability is expressed as:

$$G_{\mathrm{real}} = G_{\mathrm{ideal}} \times \exp(B \times N(0, \sigma^2)) \tag{17}$$

where G_{ideal} is the conductance value without any variation in the synaptic array. On the other hand, in the saturation region, the change in threshold voltage affects the synaptic current as follows:

$$I_{\mathrm{D,real}} = I_{\mathrm{D,ideal}} \times (1 - \frac{\Delta V_{\mathrm{th}}}{V_{\mathrm{ov}}})^2\ . \tag{18}$$

Hence, the conductance considering the synaptic variability is expressed as:

$$G_{\mathrm{real}} \approx G_{\mathrm{ideal}} \times N(1, \sigma^2)\ . \tag{19}$$

Fig. 9 (a) and (b) shows the simulated inference accuracy of the network applying the variation model described in (17) and (19). It is assumed that the pulse width used for inference is 1 step, and the result is expressed as error bars after 10 iterations.

As shown in Fig. 9 (a), if a FET-type synapse device is read in the subthreshold region, it is very vulnerable to variation. If the subthreshold swing of the device becomes steeper (larger B in (15)), the variation of the synaptic device is greater. In contrast, if a synaptic device with a large subthreshold swing is used to reduce the effect of this variation, the on/off ratio of the device is reduced, resulting in serious degradation in standby power. On the other hand, in the case of operating in the saturation region, the network is robust even with severe variations in the overdrive voltage. In addition, as the number of hidden neurons increases, the network becomes more resistant to variation.

Fig. 9 (c) shows the strength and weakness of the SNN as a parameter of the data encoding method and the operation region of the synaptic device. In summary, if a synaptic device is operated in a subthreshold region, it is very vulnerable to variation, which makes it difficult to construct an SNN. On the other hand, if the device is operated in the saturation region, the synaptic current is very large, so it is appropriate to use only a few spikes with the TTFS encoding method.

IV. Conclusion

In this paper, we have implemented an exponentially decaying synaptic current by reading a fabricated TFT-type synaptic device with a triangle pulse in the subthreshold region. An SNN was trained using TTFS-encoded data, and it reached an accuracy of 97.94% in a two-layer MLP (512 hidden neurons), which is higher than the results of previous work [11] encoded with TTFS data. We also proposed a method to read FET-type synaptic devices using a rectangular pulse in the saturation region (V_{GS} = 2.5 V, V_{DS} = 2.5 V) rather than the subthreshold region to keep the operating devices in a stable condition against unexpected external variability. The distortion of the weighted sum resulting from this approximation is minimized by using a short pulse width. As a result, an accuracy of 97.36% was obtained under stable operating conditions. The accuracy was reduced by 0.92% when compared to that of the rate-encoded SNN of the same size. However, the energy used for inference was reduced by 83.04 times, and the prediction time of the network was improved by 3.97 times. Finally, we investigated the sensitivity of the accuracy to synaptic variation as a parameter of the synaptic device operating region, and the FET-type synaptic device was found to be required to operate in the saturation region.

As the synapse current increases, the voltage drop across the parasitic resistance in the metal line increases, and the drain voltage of the synaptic devices in the array is position-dependent. In addition, if a large current flows through the current mirror in the neuron circuit, the summation of the current can be distorted due to fan-out issues. Hence, the TTFS encoding method, which uses only a small number of spikes to infer data, is foreseen to be a competitive candidate in neuromorphic systems targeting edge computing.

ACKNOWLEDGMENT

This work was supported by the Brain Korea 21 Plus Project in 2021, and MOTE (Ministry of Trade, Industry & Energy (10080583) and KSRC (Korea Semiconductor Research Consortium) support program for the development of the future semiconductor device.

REFERENCES

[1] 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.

[2] P. O'Connor, D. Neil, S. C. Liu, T. Delbruck, and M. Pfeiffer, "Real-time classification and sensor fusion with a spiking deep belief network," *Frontiers Neurosci.*, vol. 7, Oct. 2013.

[3] B. Rueckauer, I. A. Lungu, Y. Hu, M. Pfeiffer, and S. C. Liu, "Conversion of continuous-valued deep networks to efficient event-driven networks for image classification," *Frontiers Neurosci.*, vol. 11, Dec. 2017.

[4] P. U. Diehl, D. Neil, J. Binas, M. Cook, S. C. Liu, and M. Pfeiffer, "Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing," *IEEE Int. Jt. Conf. Neural Netw.*, Jul. 2015.

[5] S. Hwang, H. Kim, J. Park, M. W. Kwon, M. H. Baek, J. J. Lee, and B. G. Park, "System-level simulation of hardware spiking neural network based on synaptic transistors and I&F neuron circuits," *IEEE Electron Device Lett.* vol. 39, no. 9, pp. 1441-1444, Sep. 2018.

[6] J.-H. Lee et al., "Review of candidate devices for neuromorphic applications," in *Proc. 49th Eur. Solid-State Device Res. Conf. (ESSDERC)*, pp. 22–27, Sep. 2019.

[7] S. Thorpe, and G. Jacques, "Rank order coding," *Comput. Neurosci.,* pp. 113-118, 1998.

[8] B. Rueckauer, and S. C. Liu, "Conversion of analog to spiking neural networks using sparse temporal coding," *IEEE Int. Symp. Circuits Syst. (ISCAS)*, pp. 1-5, May. 2018.

[9] I. M. Comsa, T. Fischbacher, K. Potempa, A. Gesmundo, L. Versari, and J. Alakuijala, "Temporal coding in spiking neural networks with alpha synaptic function," *IEEE Int. Conf. Acoust. Speech Signal Process (ICASSP),* pp. 8529-8533, May. 2020.

[10] M. Zhang et al., "Rectified Linear Postsynaptic Potential Function for Backpropagation in Deep Spiking Neural Networks." *arXiv:2003.11837*, Nov. 2020.

[11] H. Mostafa, "Supervised learning based on temporal coding in spiking neural networks," *IEEE Trans. Neural Netw. Learn. Syst.,* vol. 29, no. 7, pp. 3227-3235, Aug. 2017.

[12] S. Billaudelle et al., "Versatile emulation of spiking neural networks on an accelerated neuromorphic substrate." *IEEE Int. Symp. Circuits Syst. (ISCAS)*, pp. 1-5, Oct. 2020.

[13] S. R. Kheradpisheh, M. Timothée, "S4NN: temporal backpropagation for spiking neural networks with one spike per neuron," *Int. J. Neural Syst.,* vol. 30, no. 6, 2020.

[14] Z. Chai et al., "Impact of RTN on pattern recognition accuracy of RRAM-based synaptic neural network," *IEEE Electron Device Lett.* vol. 39, no. 11, pp. 1652-1655, Nov. 2018.

[15] J. Liang, and H. S. P. Wong, "Cross-point memory array without cell selectors—Device characteristics and data storage pattern dependencies," *IEEE Trans. Electron Devices,* vol. 57, no. 10, pp. 2531-2538, Aug. 2010.

[16] G. W. Burr et al., "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. 106, no. 2, pp. 260-285, Jul. 2015.

[17] P. Livi et al., "A current-mode conductance-based silicon neuron for address-event neuromorphic systems," *IEEE Int. Symp. Circuits Syst. (ISCAS)*, pp. 2898-2901, May. 2009.

[18] W. Kang et al., "A Spiking Neural Network with a Global Self-Controller for Unsupervised Learning Based on Spike-Timing-Dependent Plasticity Using Flash Memory Synaptic Devices," *IEEE Int. Jt. Conf. Neural. Netw. (IJCNN)*, pp. 1-7, Jul. 2019.

[19] SY. Woo et al., "Low-Power and High-Density Neuron Device for Simultaneous Processing of Excitatory and Inhibitory Signals in Neuromorphic Systems," in *IEEE Access*, vol. 8, pp. 202639-202647, 2020.

[20] S. Oh et al., "Hardware Implementation of Spiking Neural Networks Using Time-To-First-Spike Encoding," *arXiv:2006.05033*, Jun. 2020.

[21] S. Park, S. Kim, B. Na, and S. Yoon, "T2FSNN: Deep Spiking Neural Networks with Time-to-first-spike Coding," *arXiv:2003.11741*, Mar. 2020.

[22] S. Zhou, L. Xiaohua, C. Ying, T. C. Sanjeev, S. Arindam, "Temporal-Coded Deep Spiking Neural Network with Easy Training and Robust Performance," *arXiv:1909.10837*, Aug. 2020.

[23] H. Kim et al., "Efficient precise weight tuning protocol considering variation of the synaptic devices and target accuracy," *Neurocomputing*, vol. 378, pp. 189-196, Feb. 2020.

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

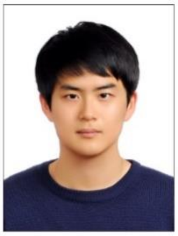
[25] S.-T. Lee, S. Lim, N. Y. Choi, J.-H. Bae, D. Kwon, B.-G. Park, and J.-H. Lee, ''Operation scheme of multi-layer neural networks using NAND flash memory as high-density synaptic devices,'' *IEEE J. Electron Devices Soc.*, vol. 7, pp. 1085–1093, 2019.

BIOGRAPHY

SEONGBIN OH (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering from Seoul National Univeresity (SNU), Seoul, South Korea, in 2017. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering. His current research interest includes neuromorphic systems and its application in computing.

SOOCHNAG LEE (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering from Seoul National Univeresity (SNU), Seoul, South Korea, in 2016. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering. His current research interest includes 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 ngineering, Seoul National University (SNU), Seoul, South Korea. His current research interest includes neuromorphic systems and its application in computing.



DONGSEOK KWON (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering from the Pohang University of Science and Technology (POSTECH), Pohang, South Korea, in 2017. He is currently pursuing the M.S. degree with the Department of Electrical and Computer Engineering, Seoul National University (SNU), Seoul, South Korea. His current research interest includes neuromorphic systems and its application in computing.

JISEONG IM (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering from Seoul National Univeresity (SNU), Seoul, South Korea, in 2020. He is currently pursuing the M.S. degree with the Department of Electrical and Computer Engineering. His current research interest includes neuromorphic systems and its application in computing.

JOON HWANG (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering from Seoul National Univeresity (SNU), Seoul, South Korea, in 2019. He is currently pursuing the M.S. degree with the Department of Electrical and Computer Engineering. His current research interest includes neuromorphic systems and its application in computing.

JONG-HO BAE (Member, IEEE) received the B.S. degree in electrical engineering from the Pohang University of Science and Technology (POSTECH), Pohang, South Korea, in 2011, and the Ph.D. degree from the Department of Electrical and Computer Engineering, Seoul National University (SNU), Seoul, South Korea, in 2018. He was a Postdoctoral Associate with the Inter-University Semiconductor Research Center (ISRC), SNU, from 2018 to 2019, and the University of California at Berkeley, Berkeley, CA, USA, from 2019 to 2020. In 2020, he joined the School of Electrical Engineering, Kookmin University, South Korea, as an Assistant Professor. His current research interests include charge trap memory, and hardware-based neural networks and its applications. B

BYUNG-GOOK PARK (Fellow, IEEE) received the B.S. and M.S. degrees in electronics 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 as an Assistant Professor with the Department of Electrical and Computer Engineering, SNU, where he is currently a Professor.

JONG-HO LEE (Fellow, IEEE) received the Ph.D. degree in electronic engineering from Seoul National University (SNU), Seoul, South Korea, 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. He is also a Lifetime Member of the Institute of Electronics Engineers of Korea (IEEK).