

Received 14 January 2024; revised 19 February 2024; accepted 28 February 2024. Date of publication 6 March 2024; date of current version 21 March 2024.
The review of this article was arranged by Editor P.-W. Li.

Digital Object Identifier 10.1109/JEDS.2024.3373889

Simulation and Optimization of IGZO-Based Neuromorphic System for Spiking Neural Networks

JUNHYEONG PARK^{1,2} , YUMIN YUN^{1,2}, MINJI KIM^{1,2}, AND SOO-YEON LEE^{1,2} 

¹ Department of Electrical and Computer Engineering, Seoul National University, Seoul 08826, South Korea
² Inter-University Semiconductor Research Center, Seoul National University, Seoul 08826, South Korea

CORRESPONDING AUTHOR: S.-Y. LEE (e-mail: sooyeon.lee@snu.ac.kr)

This work was supported by the BK21 FOUR Program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2024.

ABSTRACT In this paper, we conducted a simulation of an indium-gallium-zinc oxide (IGZO)-based neuromorphic system and proposed layer-by-layer membrane capacitor (C_{mem}) optimization for integrate-and-fire (I&F) neuron circuits to minimize the accuracy drop in spiking neural network (SNN). The fabricated synaptic transistor exhibited linear 32 synaptic weights with a large dynamic range (~ 846), and an n-type-only IGZO I&F neuron circuit was proposed and verified by HSPICE simulation. The network, consisting of three fully connected layers, was evaluated with an offline learning method employing synaptic transistor and I&F circuit models for three datasets: MNIST, Fashion-MNIST, and CIFAR-10. For offline learning, accuracy drop can occur due to information loss caused by overflow or underflow in neurons, which is largely affected by C_{mem} . To address this problem, we introduced a layer-by-layer C_{mem} optimization method that adjusts appropriate C_{mem} for each layer to minimize the information loss. As a result, high SNN accuracy was achieved for MNIST, Fashion-MNIST, and CIFAR-10 at 98.42%, 89.16%, and 48.06%, respectively. Furthermore, the optimized system showed minimal accuracy degradation under device-to-device variation.

INDEX TERMS Spiking neural network, neuromorphic system, IGZO, synaptic transistor, integrate-and-fire neuron.

I. INTRODUCTION

Neuromorphic computing systems for artificial neural networks (ANNs) and spiking neural networks (SNNs) have emerged as a promising solution to address memory bottleneck originating from von Neumann architecture with large-scale data [1], [2], [3]. Notably, SNN is well-suited for neuromorphic systems due to its biologically plausible mechanism mirroring the human brain and its potential for highly efficient parallel computing [4]. To realize a hardware SNN system, it is essential to implement: (1) artificial synapses that can mimic the synaptic plasticity and (2) artificial neurons that can emulate the spike behaviors [5], [6]. Recently, various studies on artificial synapses and neurons using indium-gallium-zinc oxide (IGZO), which has the advantages of a moderate mobility, a low leakage current, and a low-temperature process, have been actively proposed

to advance the SNN system [7], [8], [9]. Configuring an SNN system using both IGZO-based artificial synapses and neurons enables its fabrication through a low-temperature process, offering several advantages such as application on flexible or stretchable substrates and compatibility with back-end-of-line (BEOL) process [10], [11]. However, recent studies have primarily focused on developing either a synapse or a neuron individually, and there is still a lack of research on SNN systems that incorporate both synapses and neurons [12].

In order to achieve a high-performance SNN system, it is important to optimize the training method and system design that consider the properties of synapses and neurons [13]. In the training methods for SNNs, there are mainly two approaches: online and offline learning [14], [15]. For online learning, the spike-timing-dependent plasticity or

spike-rate-dependent plasticity are adopted as weight-update rules [16], [17]. Online learning can offer a robust SNN system by compensating for variations in synaptic devices during real-time learning. However, achieving high accuracy for complex datasets is still challenging [18], [19], [20], [21], [22]. Therefore, offline learning utilizing backpropagation algorithms has been studied as an alternative training method [23], [24], [25]. Offline learning is advantageous for achieving high accuracy for complex datasets because ANN weights trained using backpropagation algorithms are transferred to the SNN system [26]. However, the offline learning method suffers from accuracy drop during the ANN-to-SNN conversion due to the different neuron behaviors in ANN and SNN [27]. While ANN neurons pass inputs through an activation function and immediately produce an output, SNN neurons based on integration and fire (I&F) behavior must accumulate inputs until the membrane voltage (V_{mem}) exceeds a membrane threshold voltage (V_{MTH}) to generate a spike to the next layer [28]. Therefore, depending on the extent of V_{mem} change, information loss in SNNs can occur in two forms: overflow and underflow loss. Overflow loss occurs when voltage exceeding V_{MTH} is not utilized for subsequent spike generation. Conversely, underflow loss occurs when an increase in V_{mem} is insufficient to reach V_{MTH} , thereby preventing the propagation of information to the next layer [24], [29]. Especially, in deep SNNs with many layers, such information loss can accumulate as spikes pass through the layers, leading to error and performance degradation [27], [30]. Given that the extent of V_{mem} change is influenced by various factors including the conductance range of synaptic devices, neuron spike characteristics, and membrane capacitor (C_{mem}), it is imperative to design hardware SNN systems based on the properties of synapse and neuron.

In this paper, we simulated a hardware SNN system consisting of IGZO-based synaptic transistors and I&F neuron circuits. IGZO synaptic transistors were fabricated and measured to model synaptic weights and device-to-device variation. In addition, we proposed and verified an I&F neuron circuit by HSPICE simulation using the RPI polysilicon model, whose parameters were extracted by fitting the characteristics of fabricated IGZO thin-film transistors (TFTs). Subsequently, SNN simulations were conducted using MATLAB to evaluate the performance on three datasets: MNIST, Fashion-MNIST, and CIFAR-10. As the V_{mem} change is influenced by C_{mem} , we investigated the impact of C_{mem} on SNN performance and proposed a layer-by-layer C_{mem} optimization method to minimize information loss that occurs at each layer.

II. IGZO NEUROMORPHIC SYSTEM

A. IGZO SYNAPTIC TRANSISTOR

The neuromorphic system, which mimics the mechanism of the human brain, is composed of synaptic devices and neuron circuits. To implement the synaptic device, we fabricated IGZO synaptic transistors consisting of the

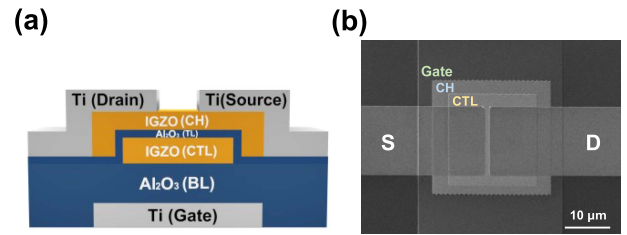


FIGURE 1. (a) Cross-sectional schematic and (b) SEM image of the IGZO synaptic transistor.

$\text{Al}_2\text{O}_3/\text{IGZO}/\text{Al}_2\text{O}_3$ gate stack and IGZO channel [31], as illustrated in Fig. 1(a). First, the buffer oxide (SiO_2) on Si substrate was cleaned using ultrasonication. A 70-nm titanium (Ti) layer was deposited using e-beam evaporator, and dry etched to form the gate electrode. A 70-nm Al_2O_3 blocking layer (BL) was deposited by atomic layer deposition (ALD) at 150 °C using the H_2O reactant and trimethylaluminum (TMA) precursor. A 40-nm IGZO charge trap layer (CTL) was deposited by RF sputtering using the IGZO target ($\text{In}:\text{Ga}:\text{Zn} = 1:1:1$ at%). The CTL was patterned by photolithography and wet etched. A 7-nm Al_2O_3 tunneling layer (TL) was deposited by ALD at 150 °C. A 40-nm IGZO channel was deposited by RF sputtering and defined by photolithography and wet etching. Then, 70-nm Ti drain/source were deposited by e-beam evaporation and formed by a lift-off process. Finally, the devices were annealed at 250 °C for 0.75 h in an air atmosphere. Each layer of the fabricated device was clearly defined with channel width/length of 15/0.9 μm , as shown in Fig. 1(b). All electrical measurements were performed using a semiconductor parameter analyzer (4200 SCS & 4225 PMU, Keithley) in a dark box at RT.

The synaptic transistor can modulate the threshold voltage (V_{TH}) by controlling the amount of charge in the CTL via Fowler-Nordheim tunneling between the channel and CTL. Under a positive gate bias (program), electrons from the channel can be trapped in the CTL. Conversely, under a negative gate bias (erase), electrons in the CTL are detrapped to the channel. To investigate the electrical behavior and uniformity of the synaptic transistor, we measured the transfer curves of 15 devices with a gate double sweep ranging from -20 V to 20 V, which exhibited clockwise hysteresis and uniform hysteresis window (Fig. 2(a)). As the representability of synaptic weights in the synaptic device is important for the neuromorphic system, we measured the synaptic weights of 15 devices using potentiation pulses. The condition of the potentiation pulse (-20 V, 200 μs) was chosen to achieve optimal linearity and dynamic range. There is potential to lower the voltage of the potentiation pulse by further optimizing the BL and TL in the synaptic transistor. Before the measurement, one depression pulse (20 V, 1 s) was applied to initialize the synaptic weight to the first level, and then a series of potentiation pulses were applied. All conductance values were read at V_{GS} of 0 V

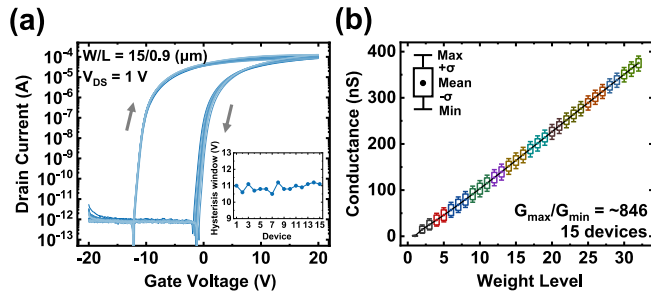


FIGURE 2. (a) Transfer curves under ± 20 V double sweep and hysteresis window size (inset) of 15 synaptic transistors. (b) Synaptic weight distribution (32 levels) of different 15 synaptic transistors.

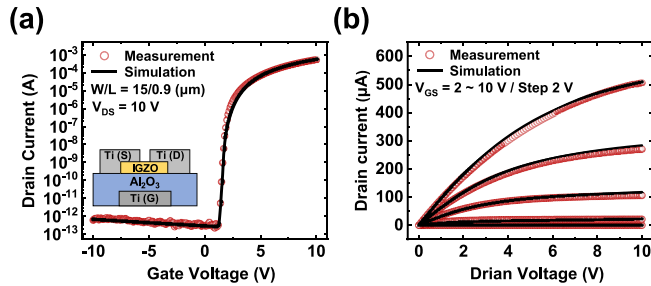


FIGURE 3. Measured and simulated (a) transfer and (b) output curves of the IGZO TFT.

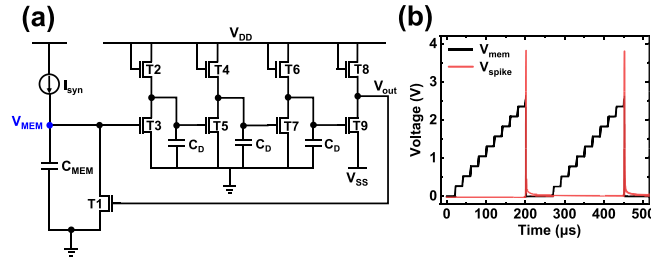


FIGURE 4. (a) Schematic of the proposed I&F neuron circuit based on IGZO TFTs. (b) Simulated output characteristic of the I&F neuron circuit.

and V_{DS} of 1 V. The devices show highly linear conductance modulation with 32 weight levels and a large dynamic range ($G_{\max}/G_{\min}$) of about 846, as shown in Fig. 2(b).

B. IGZO I&F NEURON CIRCUIT

We proposed an IGZO-based neuron circuit that emulates the I&F behavior widely used in SNNs [32]. The circuit consists of n-type-only TFTs because IGZO is inherently n-type material. There are four inverters, one reset TFT (T1), three delay capacitors (C_D), and one membrane capacitor (C_{mem}), as shown in Fig. 4(a). The circuit operation was verified by HSPICE simulation using the IGZO TFT library based on the fabricated device's electrical properties. The IGZO TFT was fabricated using the same process as the IGZO synaptic transistor, except for the deposition of IGZO CTL. The mobility, V_{TH} , and subthreshold swing of the IGZO TFT were $9.6 \text{ cm}^2/\text{V}\cdot\text{s}$, 1.9 V, and 220 mV/dec, respectively. We adjusted HSPICE RPI polysilicon model

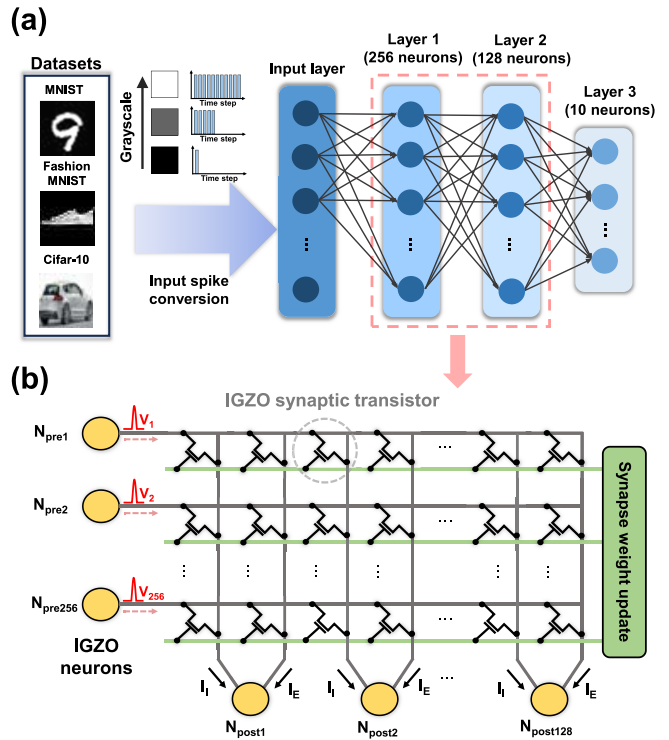


FIGURE 5. (a) Schematic of spiking neural network with three fully connected layers. (b) Hardware implementation of the SNN using IGZO-based synaptic transistors and I&F neurons.

parameters based on the measured data to simulate the circuit operation accurately. The measured and simulated electrical characteristics of the IGZO TFT are shown in Fig. 3. All transistors are $0.9 \text{ }\mu\text{m}$ in length. The widths of pull-up transistors (T2, T4, T6, and T8) and pull-down transistors (T3, T5, T7, and T9) in the inverters are $1 \text{ }\mu\text{m}$ and $35 \text{ }\mu\text{m}$, respectively. The large width of $60 \text{ }\mu\text{m}$ for T1 is used for the fast discharge of C_{mem} . The values of V_{DD} , V_{SS} , and C_D are 6 V, -0.2 V , and 500 fF, respectively. Fig. 4(b) shows the simulated V_{mem} and V_{spike} when the constant current spikes are periodically accumulated in C_{mem} . The neuron circuit successfully generates a spike and resets C_{mem} when V_{mem} exceeds the V_{MTH} of 2.45 V. The voltage and width of the spike are 3.8 V and 3.5 μs, respectively. The width of the spike can be modulated by changing C_D . When a spike is generated, V_{mem} is reset to zero voltage through T1.

C. SNN SYSTEM CONFIGURATION

For the SNN simulation, we utilized a network that consists of one input layer and three fully connected layers (256-128-10), as shown in Fig. 5(a). The size of the input layer is determined by the input image size of the dataset. The input image is converted to a series of spikes based on left-justified rate coding, where the number of spikes is proportional to the grayscale of the image [33]. The number of time steps is determined by the maximum grayscale of the input image because each input neuron propagates one spike per time step. The network layer can be configured in hardware

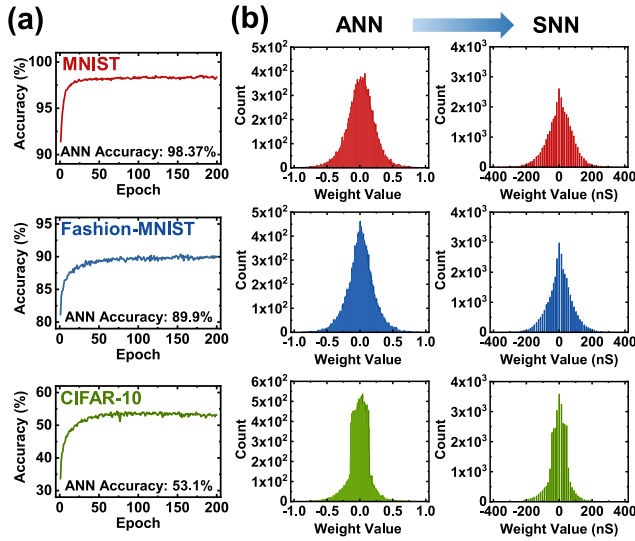


FIGURE 6. (a) ANN training accuracy of each dataset according to epoch. (b) Weight distribution change before and after synapse transfer.

utilizing IGZO-based synaptic transistors and I&F neurons, as shown in Fig. 5(b). The spikes generated by presynaptic neurons are converted to currents through the synapses, and then these currents accumulate in the C_{mem} of each postsynaptic neuron. It is assumed that each postsynaptic neuron is connected to both inhibitory and excitatory synapse lines to implement positive and negative weights.

III. RESULTS AND DISCUSSION

Fig. 6(a) shows the ANN training accuracy during 200 epochs for three datasets in the pre-defined software (Python). The training parameters were as follows: ReLU activation function, Adam optimization, a learning rate of 0.0005, a batch size of 400, and a dropout probability of 10%. The trained ANN weights were normalized and quantized based on the synaptic weight levels and then rescaled to fit within the conductance range of the synaptic weight. To prevent loss of accuracy, the positive and negative weights from ANN should be implemented identically in SNN, requiring the two synapse lines: excitatory and inhibitory synapse lines. Before the weight transfer, all synapses in the SNN system were initialized to the first weight level. During the weight transfer process, synapses in either inhibitory or excitatory lines were updated based on the sign of the weight. The weight is described as the following equation:

$$W = (G_E - G_I) \quad (1)$$

G_E and G_I are the conductance of the excitatory and inhibitory synapses, respectively. Fig. 6(b) shows the change in weight distribution before and after weight transfer to synapses. During the SNN simulation, the V_{mem} of a postsynaptic neuron is determined by the following equation:

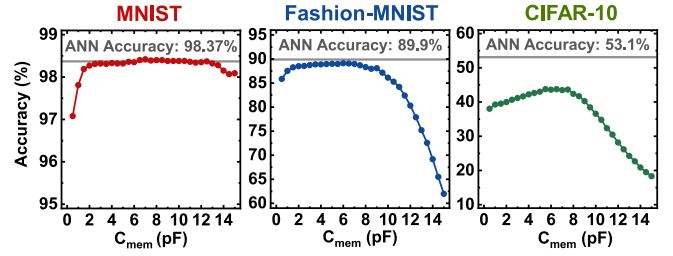


FIGURE 7. SNN accuracy results according to C_{mem} for three datasets (identical C_{mem}).

$$V_{mem}(t) = V_{mem}(t-1) + \frac{1}{C_{mem}} \sum_i^N W_i \times V_{spike,i} \times T_{spike,i} \quad (2)$$

$V_{mem}(t)$ is the membrane voltage at time step t . N is the number of presynaptic neurons. Since the extent of V_{mem} change is affected by C_{mem} , selecting the optimal value of C_{mem} is essential to enable proper information transfer across the layers. For example, if C_{mem} is excessively large, it can result in underflow, which suppresses spike generation in the neurons and hinders information transfer to the output neurons. On the other hand, if C_{mem} is too small, it can lead to overflow, causing V_{mem} to significantly exceed V_{MTH} . Since V_{mem} is reset to 0 V regardless of how much it exceeds V_{MTH} , any voltage exceeding V_{MTH} is not utilized for the next spike generation. Furthermore, due to overflow, spikes can be erroneously generated in neurons where spikes should not occur. For these reasons, optimization of C_{mem} is crucial to minimize the accuracy drop that occurs during the ANN-to-SNN conversion.

Therefore, we first examined the impact of C_{mem} on SNN accuracy and attempted to find optimal C_{mem} in our system, assuming that neurons in all layers have identical C_{mem} . For each dataset, SNN simulations were performed over a range of C_{mem} from 0.5 pF to 15 pF, as shown in Fig. 7. For the MNIST dataset, the accuracy drop is negligible for most C_{mem} values. SNN accuracy even exceeds ANN accuracy for C_{mem} of 7 pF, where the SNN and ANN accuracies are 98.42% and 98.37%, respectively. The similar accuracy between ANN and SNN might be due to the reason that it is a simple dataset. However, for the Fashion-MNIST dataset, SNN accuracy starts to decrease when C_{mem} exceeds 10 pF, indicating that the dataset is sensitive to information loss caused by underflow. The system shows the highest SNN accuracy of 89.12% at C_{mem} of 6 pF, which is a 0.78% drop compared to ANN accuracy. In the case of the CIFAR-10 dataset, SNN accuracy is more sensitively affected by C_{mem} , and the difference in accuracy between ANN and SNN is relatively larger compared to other datasets. Specifically, the SNN accuracy decreases to 38.03% when C_{mem} is 0.5 pF and further decreases to 18.31% when C_{mem} is 15 pF. Although the system can achieve its highest SNN accuracy of 43.76% at C_{mem} of 5.5 pF, there is still a large accuracy drop of 9.34% compared to ANN accuracy. The more

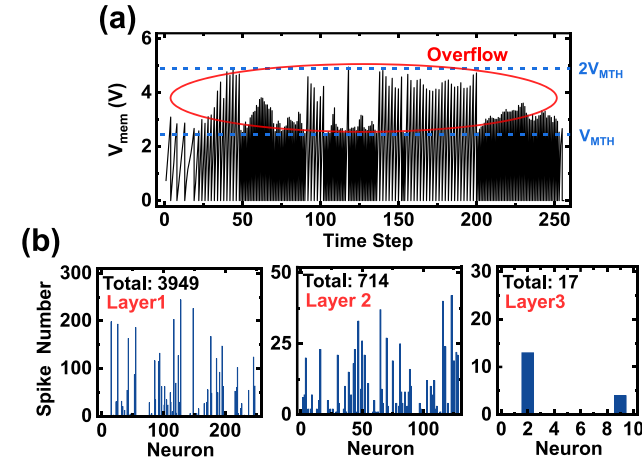


FIGURE 8. (a) V_{mem} according to the time step for the 55th neuron in layer 1 ($C_{mem} = 5.5$ pF). (b) The number of generated spikes during 255 time steps for all neurons when a CIFAR-10 input image is used.

complex dataset shows greater sensitivity to C_{mem} , accompanied by larger accuracy drop during the ANN-to-SNN conversion.

To investigate the reason for the large accuracy drop in the CIFAR-10 dataset, we analyzed the V_{mem} change in the neuron that actively generates the spikes during the image classification. The V_{mem} irregularly fluctuates according to the time step because the incoming synapse current varies depending on the spikes generated by presynaptic neurons. The neuron exhibits substantial overflow for most time steps, with V_{mem} even reaching about twice V_{MTH} at certain time steps, as shown in Fig. 8(a). In addition, it can be observed that neurons in layer 1 generate more spikes compared to other layers, as shown in Fig. 8(b). Thus, the first layer in SNN is more likely to experience overflow than other layers because it directly receives spikes from the input image, which exhibits the most active spike generation. However, the previous C_{mem} optimization method that uses identical C_{mem} for all layers has a limitation in that it does not consider the different degrees of overflow or underflow in each layer.

To overcome this limitation, we introduced a layer-by-layer C_{mem} optimization method that adjusts C_{mem} individually for each layer. Fig. 9 shows the SNN accuracy for three datasets as a function of C_{mem} for each layer. We narrowed the adjustment range of C_{mem} from the first layer to the third layer (16 pF-8 pF-6 pF), considering the decrease in the number of generated spikes as the layer becomes deeper. Regarding the MNIST dataset, SNN accuracy remains nearly identical regardless of the change in each layer's C_{mem} . When using C_{mem} of 16/6/3 pF for layers 1/2/3, the system can achieve the highest SNN accuracy of 98.47%. For the Fashion-MNIST dataset, using large C_{mem} for both layers 2 and 3 (>4 pF) leads to accuracy drop due to underflow, and the accuracy drop becomes more severe as C_{mem} of layer 1 increases. The accuracy decreases up to 85.38% when using C_{mem} of 16/8/6 pF for layers 1/2/3. Therefore, it is crucial to select appropriate C_{mem} for layers 2 and

3 to prevent underflow because these layers receive fewer spikes compared to layer 1. When using C_{mem} of 8/2/2 pF for layers 1/2/3, the system can achieve the highest SNN accuracy of 89.76%, which is close to ANN accuracy. For the CIFAR-10 dataset, increasing C_{mem} of layer 1 leads to improved accuracy, possibly due to the suppression of overflow. It is noted that accuracy starts to decrease when the C_{mem} of both layers 2 and 3 exceeds 4 pF. Therefore, configuring smaller C_{mem} as the layers become deeper is beneficial in minimizing the accuracy drop. Notably, setting C_{mem} as 16/8/1 pF for layers 1/2/3 results in the highest SNN accuracy of 48.06%, which represents a 4.3% improvement compared to the identical C_{mem} configuration of 5.5 pF.

To analyze the effect of layer-by-layer optimization on the CIFAR-10 dataset, we quantitatively compared the extent of overflow occurring in all neurons within each layer. The metric for comparison was calculated using the following equations:

$$V_{overflow} = \begin{cases} V_{mem} - V_{MTH}, & \text{if } V_{mem} > V_{MTH} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

$$Overflow_{total} = \left(\sum_i^M \sum_j^N \sum_k^T V_{overflow} \right) / (N * M) \quad (4)$$

M is the number of input images used for inference. N is the total number of neurons in a layer. T is the total number of time steps. $Overflow_{total}$ represents the cumulative overflow per neuron through all time steps when using M input images. The values of $Overflow_{total}$ were calculated according to C_{mem} configuration using 100 input images in the CIFAR-10 dataset, as shown in Fig. 10. When using the identical C_{mem} (5.5 pF), layer 1 exhibits the highest $Overflow_{total}$ of 38.42, while layers 2 and 3 show much lower $Overflow_{total}$ of 1.06 and 0.12, respectively, indicating that substantial amount of overflow still occurs in layer 1. As expected from the neuron behavior in Fig. 8(a), using identical C_{mem} for all layers is not enough to mitigate the overflow of layer 1, which receives the highest number of spikes among the three layers. In contrast, with optimized C_{mem} (16/8/1 pF), the $Overflow_{total}$ of layers 1 and 2 decreased to 6.34 and 0.17, respectively, which is attributable to the increased C_{mem} of layers 1 and 2. For layer 3, employing C_{mem} of 1 pF is advantageous for minimizing information loss because increasing C_{mem} of both layers 1 and 2 diminishes spike propagation to layer 3, which could result in underflow. These results indicate that layer-by-layer C_{mem} optimization can reduce information loss by adjusting C_{mem} in a way that minimizes either overflow or underflow in each layer.

Fig. 11 represents the result of SNN accuracy according to the time step between identical and optimized C_{mem} for three datasets. It is noteworthy that the SNN system can achieve ANN close accuracy for the MNIST and Fashion-MNIST datasets using optimized C_{mem} . In addition, an improvement in accuracy can be observed for the CIFAR-10 dataset after time step 150. Table 1 summarizes the SNN accuracy

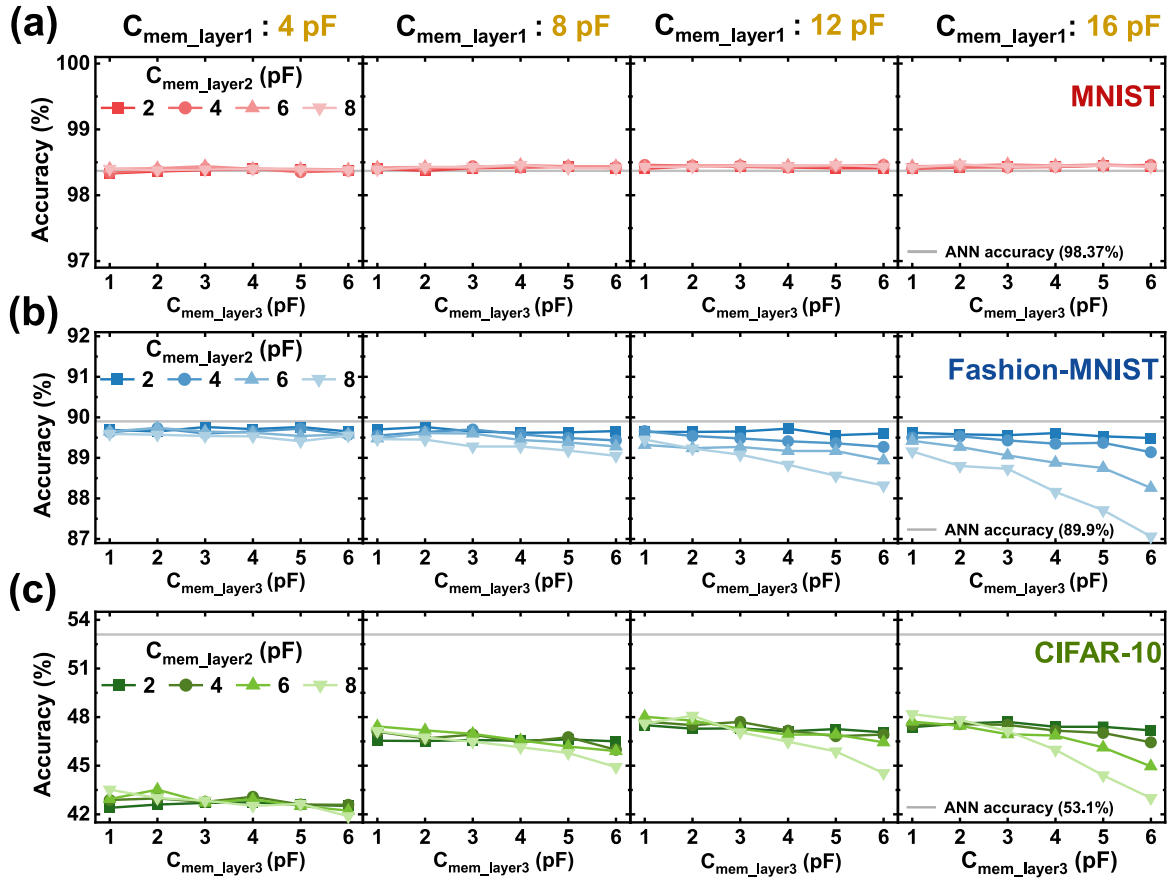


FIGURE 9. SNN accuracy result according to the C_{mem} for each layer using (a) MNIST, (b) Fashion-MNIST, and (c) CIFAR-10 datasets.

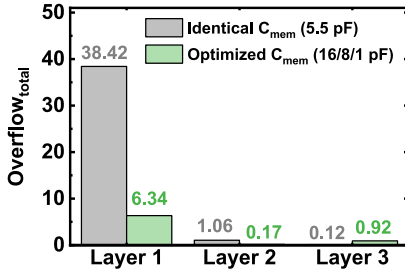


FIGURE 10. Calculated $Overflow_{total}$ for three layers with respect to C_{mem} configuration.

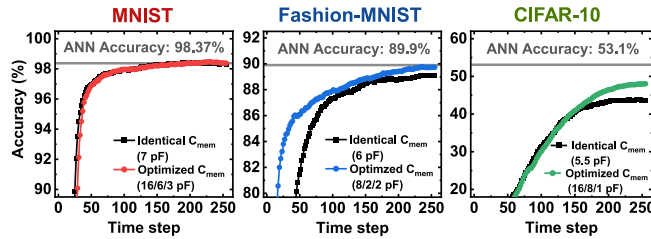


FIGURE 11. SNN accuracy results according to the time step for three datasets.

according to the C_{mem} configuration for three datasets. Using C_{mem} of 16/8/1 pF for layers 1/2/3 not only reduces the accuracy drop for the CIFAR-10 dataset but also maintains

TABLE 1. Accuracy of spiking neural network according to C_{mem} .

Dataset	Layer 1/2/3 [pF]	ANN accuracy	SNN accuracy	Accuracy drop
MNIST	7/7/7	98.37%	98.42%	-0.05%
	16/6/3		98.47%	-0.1%
	16/8/1		98.42%	-0.05%
Fashion-MNIST	6/6/6	89.9%	89.12%	0.78%
	8/2/2		89.76%	0.14%
	16/8/1		89.16%	0.74%
CIFAR-10	5.5/5.5/5.5	53.1%	43.76%	9.34%
	16/8/1		48.06%	5.04%

ANN close accuracy for other datasets. This result suggests that solely one C_{mem} configuration can achieve high SNN accuracy for three datasets without the need for different C_{mem} configurations for each dataset. We expect that C_{mem} optimization, aimed at reducing information loss, will be more critical in deeper networks to minimize accuracy drop.

Finally, we investigated the impact of the weight variation on SNN accuracy, where the variation can be attributed to the device-to-device variation in synaptic transistors. The standard variation at each weight level of the synaptic transistor was extracted, and random Gaussian variation was applied to all synapses in SNN. The simulations were conducted 20 times for each dataset. Compared to

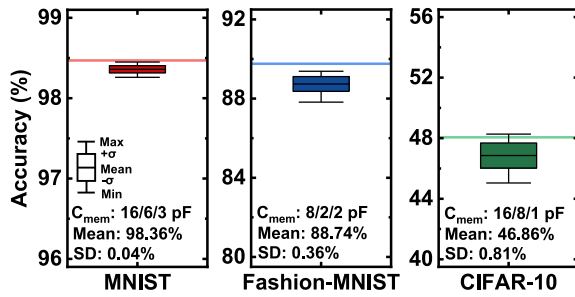


FIGURE 12. SNN accuracy for three datasets under device-to-device variation of IGZO synaptic transistors (Solid line indicates the SNN accuracy without the variation).

the accuracy without variation, accuracy with variation for MNIST, Fashion-MNIST, and CIFAR-10 decreased by 0.11%, 1.02%, and 1.2%, respectively, as shown in Fig. 12. Despite the weight variation, the SNN system can operate well with little performance degradation.

IV. CONCLUSION

In summary, we performed a simulation of the IGZO-based neuromorphic system for SNN with MNIST, Fashion-MNIST, and CIFAR-10 datasets. The simulation utilized an offline learning method and models based on the fabricated IGZO synaptic transistors and simulated I&F neuron circuit. Since information loss can occur due to overflow or underflow depending on C_{mem} , we optimized the C_{mem} of neurons to minimize the accuracy drop using identical C_{mem} for all layers. However, simulation results showed a significant accuracy drop (9.34%) for the CIFAR-10 dataset, primarily due to substantial overflow in layer 1. To address this issue, we introduced a layer-by-layer C_{mem} optimization method, which adopts different C_{mem} for each layer, considering the different degrees of overflow or underflow in each layer. This approach reduced the accuracy drop for the CIFAR-10 dataset from 9.54% to 5.04% and made ANN close accuracy for the MNIST and Fashion-MNIST datasets, suggesting that effective reduction of information loss is possible by layer-by-layer C_{mem} optimization. Further, we investigated the impact of synaptic weight variation, and our SNN system maintained high accuracy with little accuracy degradation (less than 2%). Based on the results, we expect that the IGZO-based SNN system could be a promising candidate for high-performance neuromorphic systems.

REFERENCES

- [1] J. Tang et al., "Bridging biological and artificial neural networks with emerging neuromorphic devices: Fundamentals, progress, and challenges," *Adv. Mater.*, vol. 31, no. 49, 2019, Art. no. 1902761, doi: [10.1002/adma.201902761](https://doi.org/10.1002/adma.201902761).
- [2] M. Davies et al., "Advancing neuromorphic computing with Loihi: A survey of results and outlook," *Proc. IEEE*, vol. 109, no. 5, pp. 911–934, Apr. 2021, doi: [10.1109/JPROC.2021.3067593](https://doi.org/10.1109/JPROC.2021.3067593).
- [3] A. Shrestha, H. Fang, Z. Mei, D. P. Rider, Q. Wu, and Q. Qiu, "A survey on neuromorphic computing: Models and hardware," *IEEE Circuits Syst. Mag.*, vol. 22, no. 2, pp. 6–35, May 2022, doi: [10.1109/MCAS.2022.3166331](https://doi.org/10.1109/MCAS.2022.3166331).
- [4] C. D. Schuman, S. R. Kulkarni, M. Parsa, J. P. Mitchell, and B. Kay, "Opportunities for neuromorphic computing algorithms and applications," *Nat. Comput. Sci.*, vol. 2, no. 1, pp. 10–19, 2022, doi: [10.1038/s43588-021-00184-y](https://doi.org/10.1038/s43588-021-00184-y).
- [5] S. Choi, J. Yang, and G. Wang, "Emerging memristive artificial synapses and neurons for energy-efficient neuromorphic computing," *Adv. Mater.*, vol. 32, no. 51, 2020, Art. no. 2004659, doi: [10.1002/adma.202004659](https://doi.org/10.1002/adma.202004659).
- [6] J.-Q. Yang et al., "Neuromorphic engineering: From biological to spike-based hardware nervous systems," *Adv. Mater.*, vol. 32, no. 52, 2020, Art. no. 2003610, doi: [10.1002/adma.202003610](https://doi.org/10.1002/adma.202003610).
- [7] Y. Jang, J. Park, J. Kang, and S.-Y. Lee, "Amorphous InGaZnO (a-IGZO) synaptic transistor for neuromorphic computing," *ACS Appl. Electron. Mater.*, vol. 4, no. 4, pp. 1427–1448, 2022, doi: [10.1021/acsaem.1c01088](https://doi.org/10.1021/acsaem.1c01088).
- [8] R. A. Martins et al., "Emergent solution based IGZO memristor towards neuromorphic applications," *J. Mater. Chem. C*, vol. 10, no. 6, pp. 1991–1998, 2022, doi: [10.1039/D1TC05465A](https://doi.org/10.1039/D1TC05465A).
- [9] Y. Zeng et al., "Solution-processed InGaZnO-based artificial neuron for neuromorphic system," *IEEE Trans. Electron Devices*, vol. 70, no. 4, pp. 2170–2174, Feb. 2023, doi: [10.1109/TED.2023.3247363](https://doi.org/10.1109/TED.2023.3247363).
- [10] J. C. Costa et al., "Flexible IGZO TFTs and their suitability for space applications," *IEEE J. Electron Devices Soc.*, vol. 7, pp. 1182–1190, Jul. 2019, doi: [10.1109/JEDS.2019.2931614](https://doi.org/10.1109/JEDS.2019.2931614).
- [11] W. Lu et al., "Monolithically stacked two layers of a-IGZO-based transistors upon a-IGZO-based analog/logic circuits," *IEEE Trans. Electron Devices*, vol. 70, no. 4, pp. 1697–1701, Feb. 2023, doi: [10.1109/TED.2023.3247364](https://doi.org/10.1109/TED.2023.3247364).
- [12] D. V. Christensen et al., "2022 roadmap on neuromorphic computing and engineering," *Neuromorphic Comput. Eng.*, vol. 2, no. 2, 2022, Art. no. 022501, doi: [10.1088/2634-4386/ac4a83](https://doi.org/10.1088/2634-4386/ac4a83).
- [13] N. Rathi et al., "Exploring neuromorphic computing based on spiking neural networks: Algorithms to hardware," *ACM Comput. Surv.*, vol. 55, no. 12, p. 243, 2023, doi: [10.1145/3571155](https://doi.org/10.1145/3571155).
- [14] M. Bouvier et al., "Spiking neural networks hardware implementations and challenges: A survey," *J. Emerg. Technol. Comput. Syst.*, vol. 15, no. 2, p. 22, 2019, doi: [10.1145/3304103](https://doi.org/10.1145/3304103).
- [15] J. K. Eshraghian et al., "Training spiking neural networks using lessons from deep learning," *Proc. IEEE*, vol. 111, no. 9, pp. 1016–1054, Sep. 2023, doi: [10.1109/JPROC.2023.3308088](https://doi.org/10.1109/JPROC.2023.3308088).
- [16] N. Rathi, P. Panda, and K. Roy, "STDP-based pruning of connections and weight quantization in spiking neural networks for energy-efficient recognition," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 38, no. 4, pp. 668–677, Apr. 2019, doi: [10.1109/TCAD.2018.2819366](https://doi.org/10.1109/TCAD.2018.2819366).
- [17] M. Kumar, S. S. Bezugam, S. Khan, and M. Suri, "Fully unsupervised spike-rate-dependent plasticity learning with oxide-based memory devices," *IEEE Trans. Electron Devices*, vol. 68, no. 7, pp. 3346–3352, Jul. 2021, doi: [10.1109/TED.2021.3077346](https://doi.org/10.1109/TED.2021.3077346).
- [18] N. Zheng and P. Mazumder, "A low-power hardware architecture for on-line supervised learning in multi-layer spiking neural networks," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, 2018, pp. 1–5, doi: [10.1109/ISCAS.2018.8351516](https://doi.org/10.1109/ISCAS.2018.8351516).
- [19] Y. Guo, H. Wu, B. Gao, and H. Qian, "Unsupervised learning on resistive memory array based spiking neural networks," *Front. Neurosci.*, vol. 13, Aug. 2019, Art. no. 457670, doi: [10.3389/fnins.2019.00812](https://doi.org/10.3389/fnins.2019.00812).
- [20] Y. Hao, X. Huang, M. Dong, and B. Xu, "A biologically plausible supervised learning method for spiking neural networks using the symmetric STDP rule," *Neural Netw.*, vol. 121, pp. 387–395, Jan. 2020, doi: [10.1016/j.neunet.2019.09.007](https://doi.org/10.1016/j.neunet.2019.09.007).
- [21] S. G. Hu, G. C. Qiao, T. P. Chen, Q. Yu, Y. Liu, and L. M. Rong, "Quantized STDP-based online-learning spiking neural networks," *Neural Comput. Appl.*, vol. 33, no. 19, pp. 12317–12332, 2021, doi: [10.1007/s00521-021-05832-y](https://doi.org/10.1007/s00521-021-05832-y).
- [22] L. Ma, G. Wang, S. Wang, and D. Chen, "Simulation of in-situ training in spike neural network based on non-ideal memristors," *IEEE J. Electron Devices Soc.*, vol. 11, pp. 497–502, Sep. 2023, doi: [10.1109/JEDS.2023.3311763](https://doi.org/10.1109/JEDS.2023.3311763).
- [23] T. Kim, H. Kim, J. Kim, and J.-J. Kim, "Input voltage mapping optimized for resistive memory-based deep neural network hardware," *IEEE Electron Device Lett.*, vol. 38, no. 9, pp. 1228–1231, Sep. 2017, doi: [10.1109/LED.2017.2730959](https://doi.org/10.1109/LED.2017.2730959).

- [24] S. Hwang et al., "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, doi: [10.1109/LED.2018.2853635](https://doi.org/10.1109/LED.2018.2853635).
- [25] G. Srinivasan, C. Lee, A. Sengupta, P. Panda, S. S. Sarwar, and K. Roy, "Training deep spiking neural networks for energy-efficient neuromorphic computing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2020, pp. 8549–8553, doi: [10.1109/ICASSP40776.2020.9053914](https://doi.org/10.1109/ICASSP40776.2020.9053914).
- [26] A. Sengupta, Y. Ye, R. Wang, C. Liu, and K. Roy, "Going deeper in spiking neural networks: VGG and residual architectures," *Front. Neurosci.*, vol. 13, p. 95, Mar. 2019, doi: [10.3389/fnins.2019.00095](https://doi.org/10.3389/fnins.2019.00095).
- [27] L. Deng et al., "Rethinking the performance comparison between SNNs and ANNs," *Neural Netw.*, vol. 121, pp. 294–307, Jan. 2020, doi: [10.1016/j.neunet.2019.09.005](https://doi.org/10.1016/j.neunet.2019.09.005).
- [28] E. Chicca et al., "A VLSI recurrent network of integrate-and-fire neurons connected by plastic synapses with long-term memory," *IEEE Trans. Neural Netw.*, vol. 14, no. 5, pp. 1297–1307, Sep. 2003, doi: [10.1109/TNN.2003.816367](https://doi.org/10.1109/TNN.2003.816367).
- [29] B. Han, G. Srinivasan, and K. Roy, "RMP-SNN: Residual membrane potential neuron for enabling deeper high-accuracy and low-latency spiking neural network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13558–13567.
- [30] 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," *Front. Neurosci.*, vol. 11, Dec. 2017, Art. no. 294078, doi: [10.3389/fnins.2017.00682](https://doi.org/10.3389/fnins.2017.00682).
- [31] J. Park, Y. Jang, J. Lee, S. An, J. Mok, and S.-Y. Lee, "Synaptic transistor based on In-Ga-Zn-O channel and trap layers with highly linear conductance modulation for neuromorphic computing," *Adv. Electron. Mater.*, vol. 9, no. 6, 2023, Art. no. 2201306, doi: [10.1002/aelm.202201306](https://doi.org/10.1002/aelm.202201306).
- [32] E. M. Izhikevich, "Which model to use for cortical spiking neurons?" *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1063–1070, Sep. 2004, doi: [10.1109/TNN.2004.832719](https://doi.org/10.1109/TNN.2004.832719).
- [33] S.-T. Lee and J.-H. Bae, "Investigation of deep spiking neural networks utilizing gated Schottky diode as synaptic devices," *Micromachines*, vol. 13, no. 11, p. 1800, 2022, doi: [10.3390/mi13111800](https://doi.org/10.3390/mi13111800).