Received 11 January 2024; accepted 12 February 2024. Date of publication 15 February 2024; date of current version 14 March 2024.
The review of this article was arranged by Editor P. Pavan.

Digital Object Identifier 10.1109/JEDS.2024.3366199

Unsupervised Learning in a Ternary SNN Using STDP

ABHINAV GUPTA AND SNEH SAURABH (Senior Member, IEEE)

Department of Electronics and Communication Engineering, Indraprastha Institute of Information Technology, Delhi 110020, India

CORRESPONDING AUTHOR: A. GUPTA (e-mail: abhinavg@iiitd.ac.in)

This work was supported in part by the University Grants Commission through the Senior Research Fellowship Scheme, and in part by the Science and Engineering Research Board, Department of Science and Technology, India, under Grant SRG/2021/001777.

ABSTRACT This paper proposes a novel implementation of a ternary Spiking Neural Network (SNN) and investigates it using a hierarchical simulation framework. The proposed ternary SNN is trained in an unsupervised manner using the Spike Timing Dependent Plasticity (STDP) learning rule. A ternary neuron is implemented using a Dual-Pocket Tunnel Field effect transistor (DP-TFET). The synapse consists of a Magnetic Tunnel Junction (MTJ) with a Heavy Metal (HM) underlayer, allowing for the adjustment of its conductance by directing a current through the HM layer. Further, we show that a pair of dual-pocket Fully-Depleted Silicon-on-Insulator (FD-SOI) MOSFETs can be utilized to generate a current, which reduces exponentially with increasing duration of firing events between pre- and post-synaptic neurons. This current modulates the synapse's conductance according to STDP. Furthermore, it is demonstrated that the proposed ternary SNN can be trained to classify digits in the MNIST dataset with an accuracy of 82%, which is better (75%) than that obtained using a binary SNN. Moreover, the runtime required to train the proposed ternary SNN is 8× less than that required for a binary SNN.

INDEX TERMS Ternary neuron, STDP, ternary SNN, neuromorphic computing, BTBT.

I. INTRODUCTION

Spiking Neural Networks (SNN) aim to model the behavior of the biological nervous system in an energy-efficient manner. While SNNs have proven to be a suitable contender to Artificial Neural Networks (ANN) due to their high energy efficiency, their use is still not prevalent. This is due to the lack of efficient training algorithms that efficiently utilize the temporal information embedded in discrete spikes. Unsupervised training algorithms like Spike Timing Dependent Plasticity (STDP) are among the most popular algorithms to train an SNN wherein the weight of the synapse connecting the two neurons is modulated in accordance with the time duration of firing events between the pre-synaptic and post-synaptic neurons. The weight of the synapse is potentiated (or depressed) when the firing event at the post-synaptic neuron is observed after (or before) the firing event at the pre-synaptic neuron.

The classification accuracy obtained by training an SNN using STDP is still not at par with its ANN counterparts, which are trained in a supervised manner using the gradient-descent backpropagation algorithm. Moreover, the training time for SNN is significantly longer in comparison to ANNs. This is because no learning occurs in the network until some spiking activity exists in the neurons. This is particularly problematic in deep SNNs comprising multiple layers of neurons. This is due to the decreased spiking probability of neurons deep in the network, referred to as vanishing forward-spike propagation. Thus, learning in deeper network layers is time-consuming and often requires multiple training epochs. A ternary SNN, comprising a ternary neuron that generates a $V_{DD}/2$ spike when its membrane potential crosses a threshold, say $v_{thresh1}$ and a V_{DD} spike when it crosses a higher threshold $v_{thresh2}$, can lead to a substantial speedup in training the SNN. This is due to the larger spiking probability of a ternary neuron compared to a conventional spiking neuron. Moreover, the ternary encoding of the rate-based spike train is a more accurate representation of the input dataset than the binary-encoded rate-based spike train. Fig. 1 compares

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

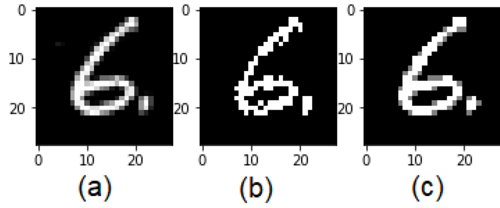

FIGURE 1. Comparison of the reconstructed input image in the MNIST dataset (a) Original image (b) Reconstructed image with binary spikes (c) Reconstructed image with ternary spikes.

the reconstructed image from the MNIST dataset [1] using a binary and a ternary spike. It can be observed that the reconstructed image with ternary spikes is a more accurate representation of the input image compared to its binary counterpart.

Neurons and synapses constitute the two basic components of an SNN. Numerous literary works comprise biologically plausible [2] or physiologically-inspired [3] models of the biological neuron. A Leaky-Integrate-and-Fire (LIF) neuron is frequently used due to its easy implementation and great biological plausibility [4], [5], [6], [7]. A synapse interconnects the two neurons and preserves the weight of the connection. Static Random Access Memories (SRAMs) can store this weight with finite accuracy. For instance, 32 transistors are required per synaptic element for storing the weight with 4-bit precision using an 8-T SRAM [8]. Non-volatile memories (NVM) like Phase Change Memories (PCM) [9], Ferroelectric RAM (FeRAM) [10], memristors [11], [12], [13], spintronic devices [14], [15], [16], floating gate transistors [17], and others can also be used to implement synapses.

In this work, a Ge-based Dual-Pocket Tunnel Field Effect Transistor (DP-TFET) is employed to implement a ternary spiking neuron. The ternary spiking neuron outputs a $V_{DD}/2$ spike when the membrane potential of the neuron surpasses a threshold, say $v_{thresh1}$ and a V_{DD} spike when it crosses a higher threshold $v_{thresh2}$. The weight of the interconnection between neurons is stored as the conductance of a synapse using a Magnetic Tunnel Junction (MTJ) with a Heavy Metal (HM) underlayer. Further, a pair of dual-pocket FD-SOI MOSFETs are employed to produce a current that tunes the synapse's conductance according to STDP. The paper is structured as follows. Section II introduces a hierarchical simulation framework to illustrate unsupervised learning in the proposed ternary SNN using STDP. Section III presents the implementation of STDP for training the ternary SNN. In Section IV, the ternary SNN is trained to classify digits in the MNIST dataset. Section V analyzes the impact of process-induced variations on the ternary neuron characteristics. Finally, Section VI concludes the work.

II. SIMULATION FRAMEWORK

In this section, a hierarchical simulation framework, proposed in [18], is employed to illustrate unsupervised learning in the ternary SNN using STDP. First, device-level simulations of a Ge-based DP-TFET are performed in the device simulator Synopsys Sentaurus [19]. Next, the synapse is simulated in mumax3 [20] to determine how the synapse's conductance can be tuned by the application of a current through the HM layer. Further, device-circuit co-simulation involving a pair of dual-pocket FD-SOI MOSFETs is performed in Synopsys Sentaurus to obtain a current, which tunes the synapse's conductance according to STDP. Finally, the simulation results are collated to train a ternary SNN with STDP to perform digit classification on the MNIST dataset.

A. TERNARY SPIKING NEURON

This section describes how the Ge-based DP-TFET can implement a ternary spiking neuron. Fig. 2 shows the cross-sectional view of the DP-TFET.

It comprises two pockets- one adjacent to the source, called the source pocket, and the other at a controlled distance from the first (L_I), called the channel pocket. The source pocket is a thin n^+-doped fully depleted pocket with a concentration of N_{NP} and length L_{NP}. The channel pocket is doped with p^+ carriers with a concentration of N_{PP} and has a length L_{PP}. The channel pocket can be fabricated lithographically, while the source pocket can be fabricated by a tilted implant after gate definition, followed by spike annealing, as proposed in [22]. There might be an uncertainty associated with the doping profile of the n^+ pocket formed after annealing in the direction underneath the gate. More advanced fabrication techniques like Molecular Beam Epitaxy (MBE) can also be employed for the formation of highly doped delta layers in vertical device configuration. Due to the stringent fabrication process employed, the impact of process-induced variations on the device characteristics is analyzed in Section V. The gate oxide used is HfO_2. Table 1 contains other critical device simulation parameters.

In this work, Germanium is preferred over Silicon. This is attributed to its smaller bandgap and the prevalence of a dominant direct tunneling mechanism [23]. This leads to a higher Band-to-Band Tunneling (BTBT) generating rate. The non-local BTBT model is employed with fitting parameters taken from [23]. We simulated a Ge-based TFET by simultaneously employing the direct and indirect tunneling parameters, as suggested in [23]. The characteristics of the Ge-based TFET (illustrated in the inset of Fig. 3) is compared against the results reported in [23] in Fig. 3.

Their reasonable similarity confirms the applicability of the calibrated BTBT model. The simulations incorporated the Shockley-Read-Hall (SRH) recombination model and a concentration-dependent Philips unified mobility model. Additionally, the Slotboom Band-Gap Narrowing (BGN) model was enabled in the simulations.

A ternary inverter has been implemented using a DP-TFET in [21]. Two tunneling regions exist in the DP-TFET - one within the channel and another at the source-channel junction. The tunneling region within the channel comprises a larger tunneling width (see Fig. 4(a)) than at the source-channel junction (see Fig. 4(d)). Thus, the within-channel

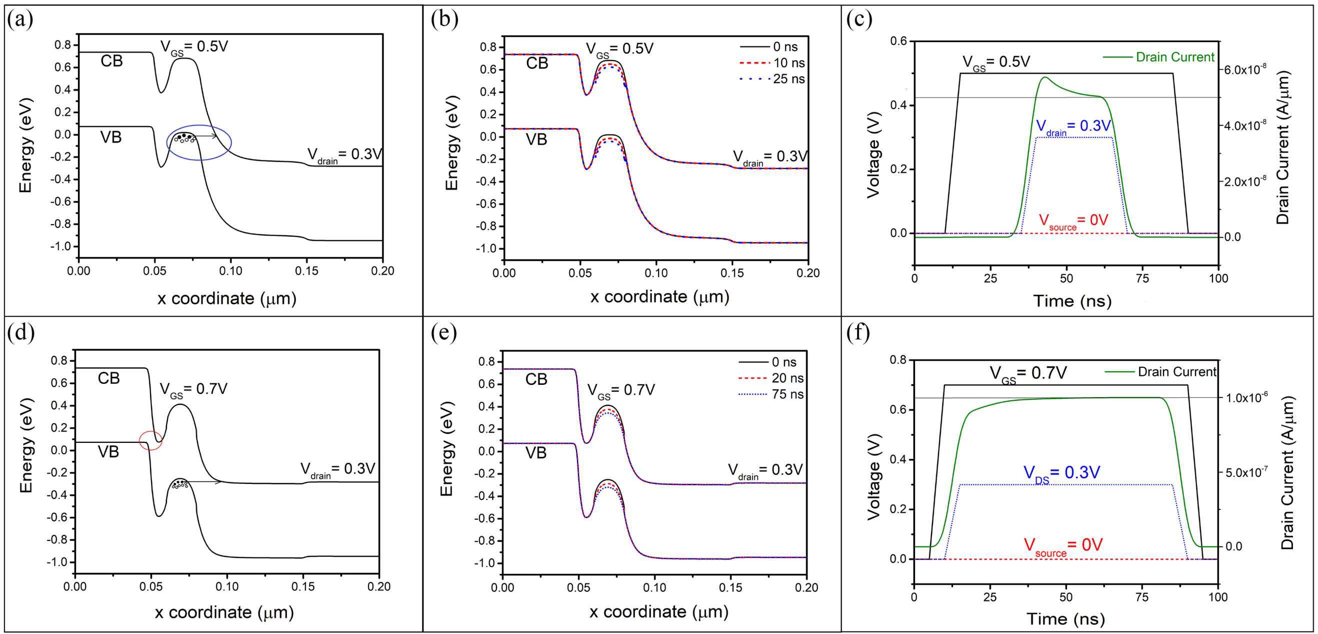

FIGURE 4. **Principle of operation of a ternary spiking neuron (a)-(c) Generation of a $V_{DD}/2$ voltage spike, and (d)-(f) Generation of a V_{DD} voltage spike.**

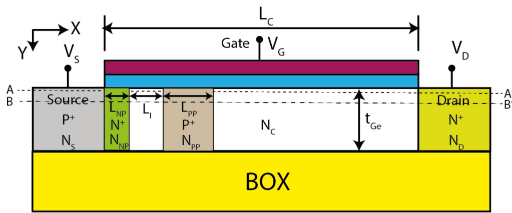

FIGURE 2. **The DP-FET used to implement a ternary spiking neuron [21].**

TABLE 1. **DP-TFET ternary neuron parameters.**

Parameter	Symbol	Value
Channel thickness (nm)	T_{ch}	20
Channel Length (nm)	L_C	100
Gate Oxide thickness (nm)	t_{ox}	5
Gate workfunction (eV)	ϕm	4.1
Channel Doping (p-type) $(atoms/cm^3)$	N_C	1×10^{17}
Source Doping (p-type) $(atoms/cm^3)$	N_S	1×10^{20}
Drain Doping (n-type) $(atoms/cm^3)$	N_D	5×10^{18}
Source Pocket Doping (n-type) $(atoms/cm^3)$	N_{NP}	1.5×10^{19}
Source Pocket Length (nm)	L_{NP}	4
Distance between pockets (nm)	L_I	6
Channel Pocket Doping (p-type) $(atoms/cm^3)$	N_{PP}	3×10^{19}
Channel Pocket Length (nm)	L_{PP}	20

tunneling current is much smaller in magnitude compared source-channel tunneling current.

A summed voltage from the pre-synaptic layer of neurons is applied as input to the gate terminal of the device. Fig. 5 shows the summer circuitry used to sum the pre-synaptic stimuli and generate an input potential for the ternary neuron,

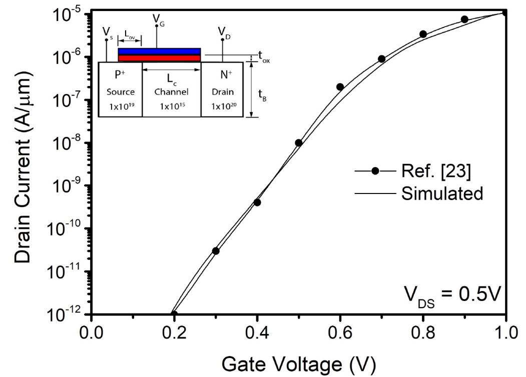

FIGURE 3. **The transfer characteristics of a Ge-based TFET (depicted in the inset) are contrasted between the simulation model and the findings documented in [23]. The same simulation model is used in [18].**

similar to the one used in [24], [25], [26]. The integration of charge, however, is happening inside the DP-TFET ternary neuron.

During the integration phase, the reset circuitry generates a voltage of 0.3V, which is applied to the drain while the source terminal is grounded. We now describe how the charge is integrated inside the device and how the $V_{DD}/2$ and V_{DD} spikes are generated. Fig. 6 shows the band diagram along cutline BB' showing a decrease in within-channel tunneling width and an increase in band overlap with an increase in the gate voltage (V_{GS}). Consequently, an increase in the V_{GS} leads to a rise in the BTBT generation rate, resulting in an increase in the within-channel tunneling current. In Fig. 4(a), the band diagram along cutline BB' at a gate voltage ($V_{GS} = 0.5V$) is displayed, illustrating within-channel tunneling of electrons. As a consequence, an accumulation of holes occurs

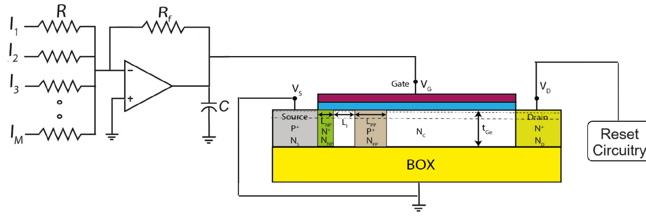

FIGURE 5. Ternary neuron architecture showing how the pre-synaptic stimuli are summed and the reset circuitry controlling the potential applied onto the drain terminal.

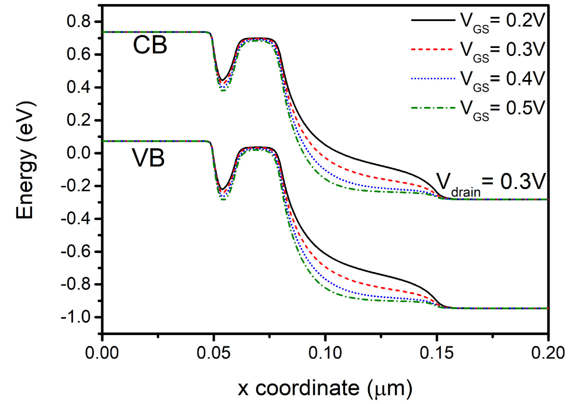

FIGURE 6. Band diagram along cutline BB' showing a decrease in tunneling width and an increase in band overlap with an increase in the gate voltage.

in the channel pocket region (hump) within the floating body of the device, leading to a gradual reduction in the height of the potential barrier over time, as depicted in Fig. 4(b). This causes a decline in the BTBT generation rate, and a consequent decrease in the within-channel tunneling current is observed. Simultaneously, due to thermionic emission, the accumulated holes in the channel pocket region undergo leakage into the source, causing an increase in the potential barrier at the same region. At equilibrium, the rate of holes leaking from the channel pocket region becomes equal to the rate of holes accumulating in the same region. When the current due to within-channel tunneling reaches a threshold value ($I_{th1} = 5 \times 10^{-8} A/\mu m$), the drain voltage is removed with the help of a reset circuitry, causing a rapid decrease in current. Note that the drain current has an initial overshoot that crosses I_{th1} when V_{DS} transitions to 0.3V from 0V. The reset circuitry removes the V_{DS} at this stage and triggers an external circuitry to generate a $V_{DD}/2$ voltage spike. For smaller gate voltage (for example, V_{GS}=0.4V), I_{th1} is never reached, and meanwhile, there is an integration of holes in the hump region in the device. Fig. 4(c) is shown only to illustrate the integration of holes (charge) happening in the hump region of the device with the evolution of the drain current with time.

Once the neuron fires a $V_{DD}/2$ spike, its drain voltage is removed using the reset circuitry, and the neuron enters into a refractory state. During the refractory period, its summed potential is allowed to climb further due to incoming pre-synaptic stimuli. V_{DS} is re-applied by the reset circuitry after a refractory period has elapsed. As long as the neuron's summed potential stays above the threshold potential (V_{th1}), it does not fire another $V_{DD}/2$ spike. However, it may fire a V_{DD} spike if it crosses a higher threshold potential, V_{th2}. In the absence of the pre-synaptic stimuli, the summed potential decreases with time. Now, after the refractory period has elapsed, if it goes below V_{th1}, the neuron can fire a $V_{DD}/2$ spike again when its potential crosses V_{th1}.

Once the accumulated potential resulting from the spiking activity of pre-synaptic neurons (V_{GS}) reaches 0.7V, the onset of source-channel tunneling current occurs. The presence of a hump in the band diagram in the channel causes the current flow through the device to involve two mechanisms. First, BTBT of electrons from the source results in an accumulation of electrons in the region between the two pockets. Subsequently, due to thermionic emission, the accumulated electrons surmount the barrier and reach the drain.

Fig. 4(d) displays the band diagram along cutline AA' at a gate voltage ($V_{GS} = 0.7V$) showing within-channel tunneling of electrons. As a consequence, holes accumulate in the channel pocket region, gradually reducing the potential barrier within the channel pocket region over time, as depicted in Fig. 4(e). As the potential barrier reduces, a greater number of electrons that had previously tunneled due to source-channel tunneling can now reach the drain. This leads to an increase in the current flowing through the device. Additionally, the accumulated holes in the channel pocket region leak away into the source, causing an increase in the height of the potential barrier. At equilibrium, the rate of holes leaking from the channel pocket region becomes equal to the rate of holes accumulating in the same region. At this stage, the current reaches a threshold value ($I_{th2} = 1 \times 10^{-6} A/\mu m$), and the drain voltage is removed, causing a rapid decrease in current, as shown in Fig. 4(f). At this stage, an external circuitry is triggered to generate a V_{DD} voltage spike. After the neuron has fired a V_{DD} voltage spike, the accumulated potential due to pre-synaptic stimuli is reset to $0V$. Such a reset circuitry has been employed in prior literature [6], [24], [25], [26] as well (for a binary neuron) and can be tailored for a ternary neuron as well. The implementation of such a reset circuitry is beyond the scope of this work. An external circuitry will be required to generate the $V_{DD}/2$ and V_{DD} spikes. A control circuitry will sense when the neuron has reached the threshold currents I_{th1} and I_{th2} and trigger the external circuitry to generate the $V_{DD}/2$ and V_{DD} spikes respectively. The design of such a control circuitry is beyond the scope of this work.

B. SYNAPSE

The principle of operation of the synapse is described in this section. Fig. 7 shows the cross-sectional view of the synapse. It comprises a Magnetic Tunnel Junction (MTJ) with an HM

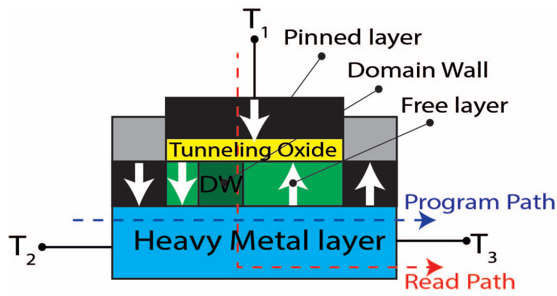

FIGURE 7. The cross-sectional view of the synapse [18].

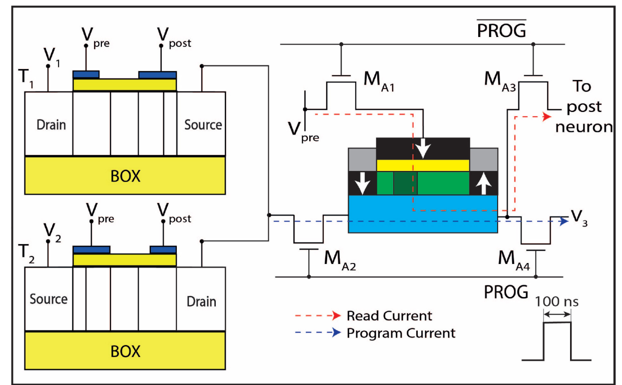

FIGURE 8. The pair of Ge-based dual-pocket FD-SOI MOSFETs that are utilized to produce a current, which tunes the synapse's conductance as per the STDP learning rule [18].

underlayer. The MTJ comprises a pinned ferromagnetic layer and a free ferromagnetic layer. The two ferromagnetic layers sandwich a tunneling oxide barrier (MgO). At the ends of the free ferromagnetic layer, two pinned ferromagnetic layers with opposing magnetization axes are present.

A Domain Wall (DW) is created by spin-orbit coupling at the free ferromagnetic layer and the HM interface, which induces Dzyaloshinskii-Moriya Interaction (DMI) [27], [28], [29], [30]. When an in-plane current flows through the HM underlayer, it deflects opposite spin-polarized electrons to the top and bottom surfaces of the HM layer, generating a transverse spin current [29], [30]. Consequently, an in-plane current generates a Spin Orbit Torque (SOT), which moves the DW in the free ferromagnetic layer. The Landau-Lifshitz-Gilbert (LLG) equation expresses the magnetization dynamics of the free ferromagnetic layer [27]. A shift in the DW's position leads to a modulation in the conductance of the synapse. A detailed simulation model used to simulate the movement of the DW as a result of the current passing through the HM using Mumax3 [20] is explained in [18]. The Tunnelling Magnetoresistance Ratio (TMR) represents the ratio of the maximum and minimum conductance of the synapse. A TMR value of 604% is reported for the MTJ in [31].

III. IMPLEMENTATION OF STDP

This section shows how a pair of Ge-based dual-pocket FD-SOI MOSFETs can implement unsupervised learning in a ternary SNN using STDP. Fig. 8 shows the pair of dual-pocket FD-SOI MOSFETs, which produce a current based on the correlation between spiking events in the pre- and post-synaptic layer of neurons. This current exponentially reduces in magnitude as the duration of spiking events between the pre-synaptic and the post-synaptic neurons increases. This current feeds into the HM layer in the FM-DW synapse.

A detailed description of the device-circuit co-simulation framework employed to produce a current, which exponentially reduces in magnitude as the duration of spiking events between the pre-synaptic and the post-synaptic neurons increases, is explained in [18]. The pair of dual-pocket FD-SOI MOSFETs takes pre- and post-synaptic voltage spikes as inputs to generate a current, which tunes the synapses' conductance as per the STDP learning rule. A pre-synaptic $V_{DD}/2$ voltage spike can result in a $V_{DD}/2$ or a V_{DD} post-synaptic voltage spike. Similarly, a pre-synaptic V_{DD} voltage spike can result in a $V_{DD}/2$ or a V_{DD} post-synaptic voltage spike. We choose the $V_{DD}/2$ voltage spike of magnitude -0.6V and the V_{DD} voltage spike of magnitude -0.7V. This is because the BTBT generation rate reduces exponentially with the applied voltage. Fig. 9 shows the current generated by the pair of dual-pocket FD-SOI MOSFETs for different pre- and post-synaptic firing events, which exponentially reduces as the duration of spiking events between the pre-synaptic and the post-synaptic neurons increases. A current density ($J = 10^{11} A/m^2$) is necessary to displace the domain wall in a CoFe strip with cross-section $160nm \times 0.6nm$ by $1\mu m$ in $30ns$ [32]. This corresponds to a current of $9.6\mu A$. The peak current generated by the pair of dual-pocket FD-SOI MOSFETs is around $8\mu A/\mu m$. Thus, a gate width of $1-1.2\mu m$ for the MOSFETs would be sufficient to generate this current.

If it is desired that the synapse be initialized to a fixed value and the stored information be erased, the pair of dual pocket FD-SOI MOSFETs need to be programmed via an alternative path using pre- and post-synaptic spikes to set the synapse to its maximum or minimum conductance states, as desired.

IV. APPLICATION OF TERNARY SNN

In this section, the proposed ternary SNN is trained to perform digit classification in the MNIST dataset using STDP. Results produced from the device- and circuit-level simulations are utilized to tune the synapse's weight based on the interval of firing events between the pre-synaptic and post-synaptic neurons. The ternary SNN consists of three layers. The first layer contains 784 neurons, while the second (excitatory) and third (inhibitory) layers comprise 800 neurons each. The first layer neurons are completely interconnected with the 800 excitatory neurons in the second layer through excitatory synapses. Each neuron in the third layer is connected one-to-one with the neuron in the excitatory layer such that when an excitatory neuron fires, an inhibitory

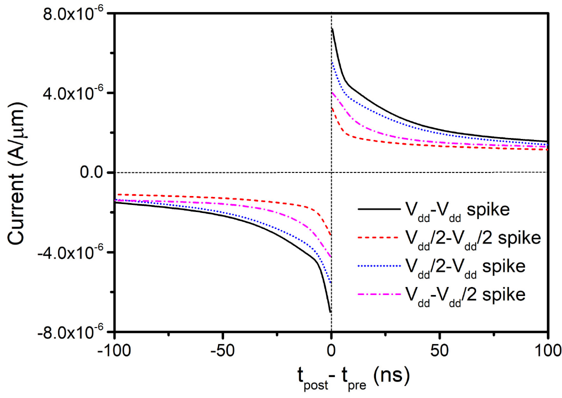

FIGURE 9. The current generated by the pair of dual-pocket FD-SOI MOSFETs for different pre and post-synaptic spiking events plotted as a function of the duration of firing events between the pre- and post-synaptic neurons.

neuron fires in response. Lateral inhibition is implemented wherein an inhibitory neuron firing event suppresses all other excitatory neurons except the one it obtains a connection from. The ternary SNN is trained by Diehl and Cook's algorithm [34]. LIF dynamics is utilized to model the neuron. The LIF neuron's membrane potential ($v(t)$) is governed by the following equation:

$$\tau_e \frac{dv}{dt} = -(v(t) - v_{rest}) + I(t) \tag{1}$$

where v_{rest} signifies neuron's resting potential, τ_e denotes the excitatory neuron's membrane time constant, and $I(t)$ signifies neuron's input voltage at time t. A pre-synaptic neuron spiking event, after being weighted by the excitatory synapse, results in an increase in the post-synaptic neuron's membrane potential. When the membrane potential of the neuron reaches the lower threshold value (v_{thres1}), it emits a $V_{DD}/2$ spike and enters into its refractory state. At this stage, its membrane potential is allowed to increase further due to incoming pre-synaptic stimuli. However, in the absence of incoming voltage spikes, the membrane potential decreases with time due to the leaky nature of the neuron. After the refractory period has elapsed, if the membrane potential goes below v_{thres1}, the neuron can fire a $V_{DD}/2$ spike again. However, if it remains above v_{thres1}, the neuron can emit a V_{DD} spike upon crossing the higher threshold value (v_{thres2}). After firing a V_{DD} spike, the neuron's membrane potential is reset to v_{reset}. Following the firing of a V_{DD} spike, the neuron enters into its refractory state, where its membrane potential is clamped to v_{rest}. After the refractory period, another LIF cycle begins. Tab. 2 lists a few key variables that were employed in the simulation. The time constants' units are defined in terms of the time step (dt) utilized in the simulation.

The network is trained using 80 images, selected at random, from each class of digits in the MNIST dataset. A binary spike train of length 350 × dt for each pixel in the image is created using rate encoding of pixels in the image. The frequency of firing activity at a particular pixel is proportional to that pixel's intensity in the image. This binary spike train is further converted to a ternary spike train. A sample window is defined comprising 35 time instances each, and the spike count is summed across all 35 time instances for each pixel in the image. This procedure results in the generation of a ternary spike train of 10 time instances for every pixel in the image based on the summed spike count as follows:

TABLE 2. System-level simulation parameters.

Parameter	Symbol	Value
Membrane time constant	τ_e	20
Resting potential	v_{rest}	-65 mV
Reset potential	v_{reset}	-65 mV
Lower threshold potential	v_{thres1}	-58 mV
Higher threshold potential	v_{thres2}	-52 mV
Refractory period	t_{ref}	5

$$Spike = \begin{cases} 0 & \text{if Spike Count} \le 2 \\ 1 & \text{if } 2 < \text{Spike Count} \le 4 \\ 2 & \text{otherwise} \end{cases} \tag{2}$$

The ternary spike train is now fed to the ternary SNN. At the beginning of the training process, the synapse's weights are initialized with random values. When the network receives the ternary spike train, the synaptic weights undergo modulation through STDP. The synaptic weights slowly settle to the desired values, and the training is stopped at that point. The classification accuracy of 75% was obtained using the binary SNN on the same benchmark dataset. However, the proposed ternary SNN resulted in a higher classification accuracy of 82%. This is because the ternary encoding of the dataset is a more accurate representation of the dataset than its binary counterparts since encoding involves some loss of information. The classification accuracy obtained in this work is compared against existing literature in Tab. 3. It can be observed that the classification accuracy obtained by training the ternary SNN on the MNIST dataset is lesser in comparison to [18] and [34], despite employing a larger number of neurons in the network. This can be attributed to the fact that only a subset of the MNIST dataset (80 randomly selected images for each digit) is presented to the network during training. On the other hand, in [18], [33], [34], the entire dataset (60,000 images) was used to train the network. This technique was adopted due to the limited computation resources available. Moreover, our aim was to compare the classification achievable with a ternary SNN and compare it with a binary SNN and not to demonstrate the maximum achievable classification accuracy. Thus, the accuracy drop was due to only a subset of the dataset provided to train the network.

Further, due to the smaller ternary spike train (10 time instances) compared to the much larger binary spike train (350 time instances), the inference time per image is

TABLE 3. Comparison of classification accuracy by training different SNN architectures on MNIST dataset.

Reference	Architecture	Learning Method	No. of excitatory neurons	Accuracy
[18]	SNN [34]	STDP (2 layer)	400	84%
[33]	SNN [34]	STDP (2 layer)	100 400	57% 73%
[34]	SNN [34]	STDP (2 layer)	100 400 1600 6400	82.9% 87% 91.9% 95%
This work (subset of dataset)	SNN [34] Ternary SNN	STDP (2 layer)	800	75% 82%
[35]	Spiking DBN	Offline learning, Conversion		95%
[36]	Spiking CNN	Offline learning, Conversion		99.1%

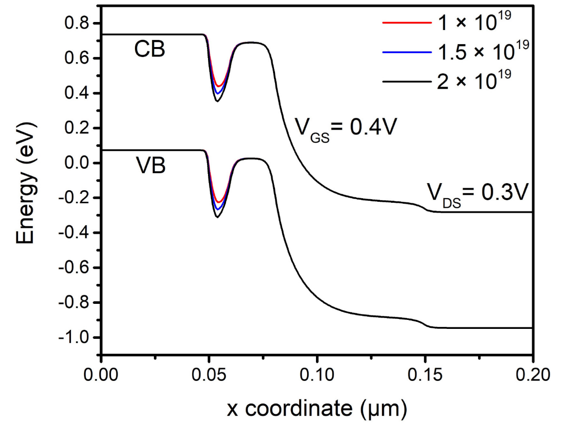

FIGURE 10. Band diagram along cutline BB' for different N_{NP} ($N_{PP} = 3 \times 10^{19}$ cm^{-3}, $L_I = 6nm$, $t_{ox} = 5nm$).

observed to reduce 8 × for the ternary SNN compared to the binary spiking neural network. Hence, the system-level simulations demonstrate that the proposed ternary spiking neural network can be more accurate and easier to train than the traditional binary spiking neural network.

V. VARIABILITY ANALYSIS

The results presented in this work have been obtained while considering ideal DP-TFET device characteristics with no variability. We will now consider the device-to-device variability as there might be process-induced variations during the fabrication process. These might impact the behavior of the ternary spiking neuron and the classification accuracy achievable with the implemented ternary SNN. Some of the parameters of the DP-TFET that are more susceptible to process-induced variations are the length of the intrinsic region between the two pockets (L_i), the doping of the n^+ pocket (N_{NP}), the doping of the p^+ pocket (N_{PP}), the thickness of the gate dielectric (t_{ox}) and the positional deviation of the gate with respect to the source/drain regions. We will analyze the impact of varying these parameters one at a time while keeping the others fixed on the DP-TFET ternary neuron characteristics.

A. IMPACT OF N$^+$ POCKET DOPING (N_{NP})

We vary N_{NP} from $1 \times 10^{19} - 2 \times 10^{19}$ cm^{-3} around its nominal value of 1.5×10^{19} cm^{-3} while keeping the other parameters at their nominal values ($N_{PP} = 3 \times 10^{19}$ cm^{-3}, $L_I = 6nm$, $t_{ox} = 5nm$). Fig. 10 shows the band diagrams along cutline BB' for different N_{NP}. It can be observed from the band diagram that with an increase in the N_{NP}, the sharpness of the band profile at the source-channel junction increases. This causes an alignment of the Valence Band (VB) in the source and the Conduction Band (CB) in the channel at a smaller V_{GS} compared to the case with a smaller N_{NP}. Thus, the neuron can fire a V_{DD} spike at a smaller accumulated potential (V_{GS}) compared to the case with a lower N_{NP}.

B. IMPACT OF CHANGE IN LENGTH OF INTRINSIC REGION BETWEEN POCKETS (L_I)

We vary L_I from 4-8 nm around the nominal value of 6 nm while keeping the other parameters at their nominal values ($N_{NP} = 1.5 \times 10^{19}$ cm^{-3}, $N_{PP} = 3 \times 10^{19}$ cm^{-3}, $t_{ox} = 5$ nm). Fig. 11 shows the band diagrams along cutline BB' for different L_i. It can be observed from the band diagram that as L_I decreases, the abruptness in the change of doping in the channel from n-type to p-type increases, while it is more gradual for a larger L_I. Thus, a lower L_I results in an early reversal of the band profile while going from the n^+ pocket to the p^+ pocket. Consequently, a neuron with a lower L_I will exhibit a delayed V_{DD} spiking event (at a higher V_{GS}) compared to the one with a higher L_I. It should be ensured that a minimum distance is maintained between the two pockets; otherwise, a very high V_{GS} will be required to cause source-channel tunneling. Such a high V_{GS} might not be achievable, and thus the affected neuron might never fire a V_{DD} spike.

C. IMPACT OF P$^+$ POCKET DOPING (N_{PP})

We vary the doping concentration of the p+ pocket from $2.5 \times 10^{19} - 3.5 \times 10^{19} cm^{-3}$ around the nominal value of $2.5 \times 10^{19} cm^{-3}$ while keeping the other parameters at their nominal values ($N_{NP} = 1.5 \times 10^{19} cm^{-3}$, $L_I = 6nm$, $t_{ox} = 5nm$). Fig. 12 shows the band diagrams along cutline BB' for different N_{PP}. It can be observed from the band diagram that as N_{PP} increases, the height of the barrier increases. This causes an increase in band overlap and results in an increase in the within-channel tunneling current. Due to this, the neuron fires a $V_{DD}/2$ spike for a smaller accumulated potential (V_{GS}) than the neuron with a lower N_{PP}.

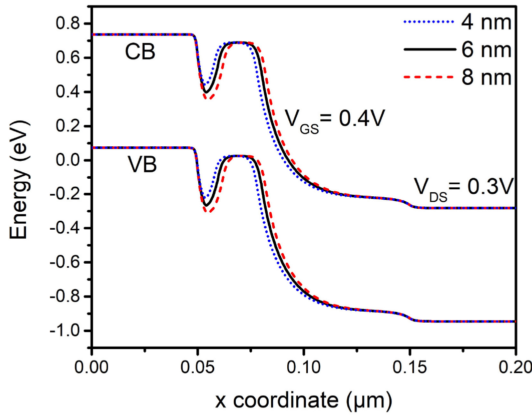

FIGURE 11. **Band diagram along cutline BB' for different L_I ($N_{NP} = 1.5 \times 10^{19}$ cm^{-3}, $N_{PP} = 3 \times 10^{19}$ cm^{-3}, $t_{ox} = 5nm$).**

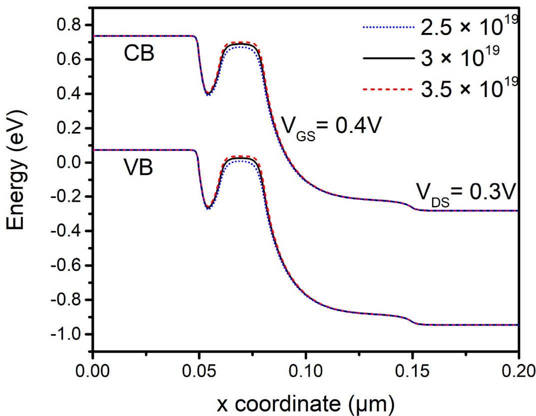

FIGURE 12. **Band diagram along cutline BB' for different N_{PP} ($N_{NP} = 1.5 \times 10^{19}$ cm^{-3}, $L_I = 6$ nm, $t_{ox} = 5nm$).**

D. IMPACT OF CHANGE IN THICKNESS OF GATE DIELECTRIC (T_{OX})

We vary t_{ox} from 4-6 nm around the nominal value of 5 nm while keeping the other parameters at their nominal values ($N_{NP} = 1.5 \times 10^{19}$ cm^{-3}, $L_I = 6$ nm. $N_{PP} = 3 \times 10^{19}$ cm^{-3}). Fig. 13 shows the band diagrams along cutline BB' for different t_{ox}. It can be observed from the band diagram below that with a decrease in t_{ox}, the tunneling width for within-channel tunneling decreases, resulting in an increase in the within-channel tunneling current. Consequently, the neuron with a thinner t_{ox} fires a $V_{DD}/2$ spike at a smaller V_{GS} compared to the one with a thicker t_{ox}. Also, it can be observed that a neuron with a thinner t_{ox} can fire a V_{DD} spike at a smaller V_{GS} compared to that with a thicker t_{ox}.

E. IMPACT OF CHANGE IN GATE ALIGNMENT

The results shown so far have considered an ideal alignment of the gate electrode with respect to the source/drain regions. However, due to process-induced variations, there may be a misalignment of the gate, resulting in an overlap/underlap of the gate with respect to the source/drain. An overlap/underlap of up to $5nm$ is considered on the source and drain sides

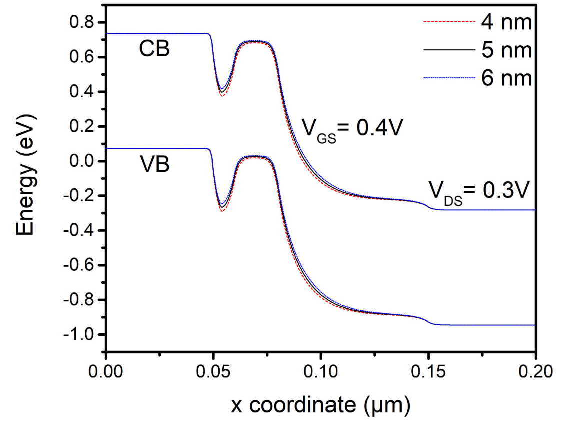

FIGURE 13. **Band diagram along cutline BB' for different N_{PP} ($N_{NP} = 1.5 \times 10^{19}$ cm^{-3}, $L_I = 6$ nm. $N_{PP} = 3 \times 10^{19}$ cm^{-3}).**

around the ideal case while keeping the pocket parameters at their nominal values ($N_{NP} = 1.5 \times 10^{19}$ cm^{-3}, $L_I = 6$ nm. $N_{PP} = 3 \times 10^{19}$ cm^{-3}, $t_{ox} = 5$ nm). Fig 14 shows the band diagrams along cutline BB' for different gate underlap/overlap with respect to the source/drain regions. It can be observed that with a $5nm$ underlap of the gate with respect to the source side, the gate no longer influences the n^+ pocket region, and there is no band bending with an increase in gate voltage. The ternary neuron can never fire a V_{DD} spike in such a scenario. Hence, the gate underlap can be detrimental to the functionality of the device, and this situation should be avoided by allowing sufficient margins for process-induced variations. As the gate underlap decreases from $5nm$, the gate regains control over the n^+ pocket region, and the neuron can fire a delayed V_{DD} spike. A band profile similar to the ideal gate alignment is obtained for a gate overlap with the source region. Hence, the gate overlap is not expected to impact the device functionality significantly. However, the increased overlap capacitance can impact the dynamic response of the device.

From a system-level standpoint, it can be inferred that due to device-to-device variability, there can be an earlier or delayed firing event between two neurons in adjacent layers. For instance, suppose that a neuron in the pre-synaptic layer was skewed such that it fired a V_{DD} spike earlier than it was supposed to (at a lower V_{GS}) and a post-synaptic neuron was skewed such that it is skewed to fire a V_{DD} spike later than it was supposed to (at a higher V_{GS}), or vice-versa, then change in the weight of the synapse connecting them would be small as the two spikes would have been further apart in time. This can lead to a slower training of the network compared to the case when both neurons are nominal.

VI. CONCLUSION

This paper utilizes a well-calibrated device-to-system level simulation framework to illustrate unsupervised learning in a ternary SNN using STDP. We demonstrate the implementation of a ternary spiking neuron using a Ge-based DP-TFET. Further, a device-circuit co-simulation framework shows that

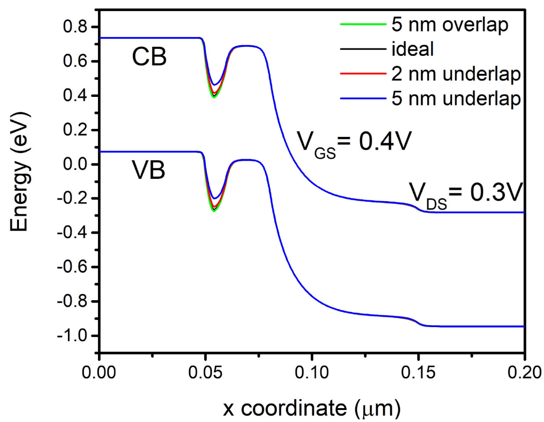

FIGURE 14. **Band diagram along cutline BB′ for different gate electrode underlap/overlap with respect to source/drain regions (N_{NP} = 1.5 × 10^{19} cm^{-3}, L_I = 6nm. N_{PP} = 3 × 10^{19} cm^{-3}, t_{ox} = 5nm).**

a pair of dual-pocket FD-SOI MOSFETs can be utilized to produce a current, which tunes the synapse's conductance according to STDP. The proposed ternary Spiking Neural Network (SNN) is trained to perform digit classification on the MNIST dataset. An accuracy of 82% was achieved in the classification, which is superior to the accuracy obtained with a binary SNN (75%). Moreover, the inference time is reduced by about 8× compared to a binary SNN. It must be ensured that the process-induced variations do not result in a large device-to-device variability to avoid training to slow down. In particular, the two pockets should be fabricated at a minimum controlled distance from one another, and the gate underlap should be controlled to allow firing activity for those neurons. Hence, the device-, circuit-, and system-level results demonstrate that ternary spiking neural networks can be a promising framework for brain-inspired computing.

REFERENCES

[1] 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, doi: 10.1109/5.726791.

[2] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *J. Physiol.*, vol. 117, no. 4, p. 500, 1952, doi: 10.1113/jphysiol.1952.sp004764.

[3] E. M. Izhikevich, "Simple model of spiking neurons," *IEEE Trans. Neural Netw.*, vol. 14, no. 6, pp. 1569–1572, Nov. 2003, doi: 10.1109/TNN.2003.820440.

[4] N. Kamal and J. Singh, "A highly scalable junctionless FET leaky integrate-and-fire neuron for spiking neural networks," *IEEE Trans. Electron Devices*, vol. 68, no. 4, pp. 1633–1638, Apr. 2021, doi: 10.1109/TED.2021.3061036.

[5] A. Gupta and S. Saurabh, "An energy-efficient Ge-based leaky integrate and fire neuron: Proposal and analysis," *IEEE Trans. Nanotechnol.*, vol. 21, pp. 555–563, 2022, doi: 10.1109/TNANO.2022.3209078.

[6] B. Das, J. Schulze, and U. Ganguly, "Ultra-low energy LIF neuron using Si NIPIN diode for spiking neural networks," *IEEE Electron Device Lett.*, vol. 39, no. 12, pp. 1832–1835, Dec. 2018, doi: 10.1109/LED.2018.2876684.

[7] W. H. Brigner et al., "Shape-based magnetic domain wall drift for an artificial spintronic leaky integrate-and-fire neuron," *IEEE Trans. Electron Devices*, vol. 66, no. 11, pp. 4970–4975, Nov. 2019, doi: 10.1109/TED.2019.2938952.

[8] B. Rajendran et al., "Specifications of nanoscale devices and circuits for neuromorphic computational systems," *IEEE Trans. Electron Devices*, vol. 60, no. 1, pp. 246–253, Jan. 2013, doi: 10.1109/TED.2012.2227969.

[9] S. Oh, Y. Shi, X. Liu, J. Song, and D. Kuzum, "Drift-enhanced unsupervised learning of handwritten digits in spiking neural network with PCM synapses," *IEEE Electron Device Lett.*, vol. 39, no. 11, pp. 1768–1771, Nov. 2018, doi: 10.1109/LED.2018.2872434.

[10] K.-Y. Hsiang et al., "Ferroelectric $HfZrO_2$ with electrode engineering and stimulation schemes as symmetric analog synaptic weight element for deep neural network training," *IEEE Trans. Electron Devices*, vol. 67, no. 10, pp. 4201–4207, Oct. 2020, doi: 10.1109/TED.2020.3017463.

[11] 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, 2010, doi: 10.1021/nl904092h.

[12] G. Indiveri, R. Legenstein, G. Deligeorgis, and T. Prodromakis, "Integration of nanoscale memristor synapses in neuromorphic computing architectures," *Nanotechnology*, vol. 24, no. 38, 2013, Art. no. 384010, doi: 10.1088/0957-4484/24/38/384010.

[13] X. Yang, B. Taylor, A. Wu, Y. Chen, and L. O. Chua, "Research progress on memristor: From synapses to computing systems," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 69, no. 5, pp. 1845–1857, May 2022, doi: 10.1109/TCSI.2022.3159153.

[14] D. Zhang et al., "All spin artificial neural networks based on compound spintronic synapse and neuron," *IEEE Trans. Biomed. Circuits Syst.*, vol. 10, no. 4, pp. 828–836, Aug. 2016, doi: 10.1109/TBCAS.2016.2533798.

[15] A. Sengupta, A. Banerjee, and K. Roy, "Hybrid spintronic-CMOS spiking neural network with on-chip learning: Devices, circuits, and systems," *Phys. Rev. Appl.*, vol. 6, no. 6, 2016, Art. no. 064003, doi: 10.1103/PhysRevApplied.6.064003.

[16] A. Amirany, M. H. Moaiyeri, and K. Jafari, "Nonvolatile associative memory design based on spintronic synapses and CNTFET neurons," *IEEE Trans. Emerg. Topics Comput.*, vol. 10, no. 1, pp. 428–437, Jan.–Mar. 2022, doi: 10.1109/TETC.2020.3026179.

[17] S. Ramakrishnan, P. E. Hasler, and C. Gordon, "Floating gate synapses with spike-time-dependent plasticity," *IEEE Trans. Biomed. Circuits Syst.*, vol. 5, no. 3, pp. 244–252, Jun. 2011, doi: 10.1109/TBCAS.2011.2109000.

[18] A. Gupta and S. Saurabh, "On-chip unsupervised learning using STDP in a spiking neural network," *IEEE Trans. Nanotechnol.*, vol. 22, pp. 365–376, Jul. 2023, doi: 10.1109/TNANO.2023.3293011.

[19] *Synopsys Sentaurus Device User Guide, T-2022.03*, Synopsys, Inc., Mountain View, CA, USA, 2022.

[20] A. Vansteenkiste, J. Leliaert, M. Dvornik, M. Helsen, F. Garcia-Sanchez, and B. Van Waeyenberge, "The design and verification of MuMax3," *AIP Adv.*, vol. 4, no. 10, 2014, Art. no. 107133, doi: 10.1063/1.4899186.

[21] A. Gupta and S. Saurabh, "Implementing a ternary inverter using dual-pocket tunnel field-effect transistors," *IEEE Trans. Electron Devices*, vol. 68, no. 10, pp. 5305–5310, Oct. 2021, doi: 10.1109/TED.2021.3106618.

[22] V. Nagavarapu, R. Jhaveri, and J. C. S. Woo, "The tunnel source (PNPN) n-MOSFET: A novel high performance transistor," *IEEE Trans. Electron Devices*, vol. 55, no. 4, pp. 1013–1019, Apr. 2008, doi: 10.1109/TED.2008.916711.

[23] K. H. Kao, A. S. Verhulst, W. G. Vandenberghe, B. Soree, G. Groeseneken, and K. De Meyer, "Direct and indirect band-to-band tunneling in germanium-based TFETs," *IEEE Trans. Electron Devices*, vol. 59, no. 2, pp. 292–301, Feb. 2012, doi: 10.1109/TED.2011.2175228.

[24] S. Dutta, V. Kumar, A. Shukla, N. R. Mohapatra, and U. Ganguly, "Leaky integrate and fire neuron by charge-discharge dynamics in floating-body MOSFET," *Sci. Rep.*, vol. 7, no. 1, pp. 1–7, Dec. 2017, doi: 10.1038/s41598-017-07418-y.

[25] D. Chatterjee and A. Kottantharayil, "A CMOS compatible bulk FinFET-based ultra low energy leaky integrate and fire neuron for spiking neural networks," *IEEE Electron Device Lett.*, vol. 40, no. 8, pp. 1301–1304, Aug. 2019, doi: 10.1109/LED.2019.2924259.

[26] T. Chavan, S. Dutta, N. R. Mohapatra, and U. Ganguly, "Band-to-band tunneling based ultra-energy-efficient silicon neuron," *IEEE Trans. Electron Devices*, vol. 67, no. 6, pp. 2614–2620, Jun. 2020, doi: 10.1109/TED.2020.2985167.

[27] J. C. Slonczewski, "Conductance and exchange coupling of two ferromagnets separated by a tunneling barrier," *Phys. Rev. B*, vol. 39, no. 10, pp. 6995–7002, 1989, doi: 10.1103/PhysRevB.39.6995.

[28] K. S. Ryu, L. Thomas, S. H. Yang, and S. Parkin, "Chiral spin torque at magnetic domain walls," *Nat. Nanotechnol.*, vol. 8, no. 7, pp. 527–533, 2013, doi: 10.1038/nnano.2013.102.

[29] S. Emori, U. Bauer, S. M. Ahn, E. Martinez, and G. S. Beach, "Current driven dynamics of chiral ferromagnetic domain walls," *Nature Mater.*, vol. 12, no. 7, pp. 611–616, 2013, doi: 10.1038/nmat3675.

[30] E. Martinez, S. Emori, N. Perez, L. Torres, and G. S. Beach, "Current driven dynamics of Dzyaloshinskii domain walls in the presence of in-plane fields: Full micromagnetic and one-dimensional analysis," *J. Appl. Phys.*, vol. 115, no. 21, 2014, Art. no. 213909, doi: 10.1063/1.4881778.

[31] S. Ikeda et al., "Tunnel magnetoresistance of 604% at 300 K by suppression of Ta diffusion in CoFeB/MgO/CoFeB pseudo-spin-valves annealed at high temperature," *Appl. Phys. Lett.*, vol. 93, no. 8, p. 2508, 2008, doi: 10.1063/1.2976435.

[32] A. Sengupta, Y. Shim, and K. Roy, "Proposal for an all-spin artificial neural network: Emulating neural and synaptic functionalities through domain wall motion in ferromagnets," *IEEE Trans. Biomed. Circuits Syst.*, vol. 10, no. 6, pp. 1152–1160, Dec. 2016, doi: 10.1109/TBCAS.2016.2525823.

[33] G. Srinivasan, A. Sengupta, and K. Roy, "Magnetic tunnel junction based long-term short-term stochastic synapse for a spiking neural network with on-chip STDP learning," *Sci. Rep.*, vol. 6, Jul. 2016, Art. no. 29545, doi: 10.1038/srep29545.

[34] P. U. Diehl and M. Cook, "Unsupervised learning of digit recognition using spike-timing-dependent plasticity," *Front. Comput. Neurosci.*, vol. 9, p. 99, Aug. 2015, doi: 10.3389/fncom.2015.00099.

[35] E. Stromatias, D. Neil, M. Pfeiffer, F. Galluppi, S. B. Furber, and S. C. Liu, "Robustness of spiking deep belief networks to noise and reduced bit precision of neuro-inspired hardware platforms" *Front. Neurosci.*, vol. 9, p. 222, Jul. 2015, doi: 10.3389/fnins.2015.00222.

[36] 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," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Killarney, Ireland, 2015, pp. 1–8, doi: 10.1109/IJCNN.2015.7280696.