Implementation of Neuromorphic Computing Framework using Tunneling-based Devices

By

Abhinav Gupta

PhD-18106

A Thesis

submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Under the guidance of

Dr. Sneh Saurabh

Associate Professor, IIIT-Delhi

Indraprastha Institute of Information Technology, Delhi

April, 2024

Implementation of Neuromorphic Computing Framework using Tunneling-based Devices

By

Abhinav Gupta

A Thesis

submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Electronics and Communication Engineering,

Indraprastha Institute of Information Technology Delhi

New Delhi– 110020

April, 2024

Certificate

This is to certify that the thesis titled "***Implementation of Neuromorphic Computing Framework using Tunneling-based Devices***" being submitted by ***Abhinav Gupta*** to the Indraprastha Institute of Information Technology Delhi, for the award of the degree of Doctor of Philosophy, is an original research work carried out by her under my supervision. In my opinion, the thesis has reached the standard fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree or diploma.

April, 2024

Dr. Sneh Saurabh

Associate Professor,

Indraprastha Institute of Information Technology Delhi-110020, India.

Declaration

This is certified that the thesis entitled "***Implementation of Neuromorphic Computing Framework using Tunneling-based Devices***" being submitted by me to the Indraprastha Institute of Information Technology Delhi, for the award of degree of **Doctor of Philosophy**, is a bonafide work carried out by me. This research work has been carried out under the supervision of **Dr. Sneh Saurabh**. The study pertaining to this thesis has not been submitted in part or in full, to any other University or Institution for the award of any other degree.

April, 2024

Abhinav Gupta

Abhinav Gupta

Ph.D. Student,

Indraprastha Institute of Information Technology Delhi-110020, India.

Abstract

In recent years, Machine Learning (ML) and Artificial Intelligence (AI) have become one of the hot topics for research and have found their use in various applications across different sectors like healthcare, automotive, marketing, finance, agriculture, Natural Language Processing (NLP), etc. However, training the current state-of-the-art AI-based algorithms are highly energy intensive. For instance, an energy of 932 MWh is required to train OpenAI's GPT-3 NLP model. The large power consumption stems from training these algorithms on conventional computing systems based on the von-Neumann architecture. In the von-Neumann architecture, memory and computation are decoupled from one another, making it energy intensive.

The human brain, comprising about 10^{11} neurons and 10^{15} synapses, operates at a power budget of just 20W. Taking inspiration from the highly dense and energy-efficient architecture of the biological brain, Spiking Neural Networks (SNN) aim to model the behavior of the biological neural network in an energy-efficient manner. The neurons in an SNN communicate via discrete action potentials or "spikes," which are sparse in time.

In this work, an energy-efficient SNN is proposed, which can be trained on-chip in an unsupervised manner using Spike Timing Dependent Plasticity (STDP). Firstly, to implement an energy-efficient SNN, a Leaky Integrate and Fire (LIF) neuron has been proposed. The proposed neuron, comprising a Ge-based PD-SOI MOSFET, can directly receive the incoming voltage spikes and avoid energy dissipation in generating a summed potential. The smaller bandgap with dominant direct tunneling of Ge allows the device to operate at a lower voltage level. The energy consumption per spike of the proposed neuron is 0.07fJ, which is lower than LIF neuron implementations (experimental or simulated) reported in the literature. A Ferromagnetic Domain Wall (FM-DW) based device has been employed to function as a synapse. It comprises a Magnetic Tunnel Junction (MTJ) with a Heavy Metal (HM) underlayer. The MTJ consists of a free

FM (CoFe) layer (whose magnetization can be varied) and a pinned FM layer (whose magnetization is fixed) separated by a tunneling oxide barrier (MgO). A DW separates two oppositely polarised magnetic regions in the free FM layer. A programming current flowing through the HM layer results in the movement of the DW in the free FM layer. A displacement in the position of the DW results in a change in the conductance of the FM-DW synapse. Secondly, a Ge-based dual-pocket Fully-Depleted Silicon-on-Insulator (FD-SOI) MOSFETs with dual asymmetric gates has been proposed that implements on-chip unsupervised learning using STDP in the SNN. Using a comprehensive device-to-system level simulation framework, it is demonstrated that a pair of proposed dual-pocket FD-SOI MOSFETs with dual asymmetric gates can generate a current, whose magnitude depends exponentially on the temporal correlation of spiking events between the pre- and post-synaptic neuronal layers. This current drives the HM layer in the FM-DW synapse and programs the conductance of the synapse in accordance with the STDP learning rule. The proposed implementation requires 2-3 $\times$ fewer transistors and offers a lower latency to implement STDP than existing literature.

While SNNs have emerged as a suitable contender to Artificial Neural Networks (ANN) due to their high energy efficiency, their use is still not prevalent. One of the major reasons preventing the widespread applicability of SNNs is the lack of efficient training algorithms that efficiently utilize the temporal information embedded in discrete spikes. Moreover, the time required to train the SNN can be substantially longer than ANNs. This is because no learning occurs in the network until some spiking activity exists in the neurons. Thus, learning in deeper network layers is time-consuming and often requires multiple training epochs. A ternary SNN, comprising a ternary neuron, outputs a $V_{DD}/2$ spike when the membrane potential of the neuron crosses a lower threshold, say $v_{thresh1}$ and a V_{DD} spike when it crosses a higher threshold $v_{thresh2}$, can result in a substantial speedup in the time required to train the SNN. This is due to the larger spiking probability of a ternary neuron compared to a conventional spiking neuron. Moreover, the ternary encoding is a more accurate representation than the binary encoding and can result in a higher classification accuracy compared to a conventional SNN. A Dual-Pocket Tunnel Field effect transistor (DP-TFET) has been proposed to implement a ternary spiking neuron. Two distinct tunneling mechanisms exist in the device - within-channel tunneling and source-channel tunneling, which are responsible for the generation of $V_{DD}/2$ and V_{DD} voltage spikes, respectively. An FM-DW based device is employed as the synapse, and

the network is trained on-chip in an unsupervised manner using STDP. Using a device-to-system level simulation framework, it is demonstrated that the 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 $8\times$ less than that required for a binary SNN.

To summarize, the goal of this work is to develop an energy-efficient framework for Neuromorphic Computing using an SNN. It involves developing an insight into the state-of-the-art hardware required to implement an SNN and proposing novel devices that aid in implementing and training the SNN in an energy-efficient manner.

To my teachers and parents, who have been a great source of inspiration and motivation, for pushing me every day to be the best version of myself, and for always being there for me . . .

Acknowledgements

"*Success, like happiness, can not be pursued. It comes only as a result of dedication.*"

– Viktor Frankl

To begin, I would like to praise and thank God for the countless blessings bestowed upon me. In addition to my efforts, my thesis's success relies heavily on others' help and direction. I want to take this opportunity to thank everyone who helped me finish this thesis.

I express my sincere gratitude to my advisor, Dr. Sneh Saurabh, for his invaluable guidance and support. He has been a constant source of inspiration to me and this dissertation would have been impossible if not for his patience, knowledge, and expertise. I owe him a debt of gratitude for his encouragement during my research work. To this day, I am grateful to him for all of his help and advice in nurturing my career.

I would like to thank my supervisory committee members - Dr. Anuj Grover and Dr. Sujay Deb for providing me with helpful suggestions and constructive criticism throughout my research work. Their feedback led me to explore new directions of research and helped me improve the quality of my research work.

I would also like to thank the Nanoscale Devices and Circuits research group, in particular Dr. Shelly Garg, Ms. Amina Haroon, Ms. Pooja Beniwal, and Mrs. Jasmine Kaur for providing helpful suggestions and moral support. I would also like to thank Mrs. Neelam Singh and Dr. Alok Nikhil Jha for their support and encouragement. I will cherish the friendships I have built with each of them for the rest of my life. I would like to take this opportunity to thank all the academic staff, the IT support staff, and the FMS staff at IIIT Delhi for always providing a speedy resolution to any problems I faced during the course of my research

work.

The support and encouragement of my parents have been crucial to my success. I would like to take this opportunity to thank them for all the sacrifices they have made throughout these years. Their belief in me has kept my spirits and motivation high during this arduous journey. I have been blessed to have them as my parents. I would like to extend my gratitude to The University Grants Commission, India for providing me with NET JRF scholarship to undertake research. I would also like to thank SERB, DST, India, for funding the research tools.

Contents

List of Figures

List of Tables

List of Abbreviations

AI Artificial Intelligence

ANN Artificial Neural Network

BTBT Band-to-band tunneling

CMOS Complementary metal oxide semiconductor

CNN Convolutional Neural Network

DBN Deep Belief Network

DNN Deep Neural Network

DP-TFET Dual Pocket Tunnel Field Effect Transistor

EHP Electron Hole Pairs

FD-SOI Fully-Depleted Silicon-on-Insulator

FeRAM Ferroelectric Random Access Memory

FM-DW Ferromagnetic-Domain wall

FN Fowler-Nordheim tunneling

Ge Germanium

Gnd Ground potential or $0\ V$

HCI Hot Carrier Injection

HM Heavy Metal

IF Integrate and Fire

II Impact Ionization

IoT Internet of Things

ITC Interface trap charges

LIF Leaky-Integrate and Fire

ML Machine Learning

MNIST Modified National Institute of Standards and Technology

MOSFET Metal oxide semiconductor field-effect transistor

MTJ Magnetic Tunnel Junction

NLP Natural Language Processing

NVM Non-Volatile Memory

PCM Phase Change Memory

PD-SOI Partially-Depleted Silicon on Insulator

SP Source pocket

ReLU Rectified Linear Unit

RMS Root Mean Square

RRAM Resistive Random access Memory

SNM Static Noise Margin

SNN Spiking Neural Network

SRAM Static Random Access Memory

STDP Spike Timing Dependent Plasticity

STI Standard Ternary Inverter

TFET Tunnel field-effect transistor

TMR Tunneling Magnetoresistance Ratio

VLSI Very Large Scale Integration

WTA Winner Take All

Chapter 1

Introduction

1.1 Motivation

In recent years, Machine Learning (ML) and Artificial Intelligence (AI) have become increasingly prominent in various sectors, including healthcare, automotive, marketing, finance, agriculture, and Natural Language Processing (NLP). One of the primary reasons for the widespread success of ML algorithms is the rapid advancement in the semiconductor industry. Gordon Moore, in 1965, gave the famous Moore's law of transistor scaling, which stated that the number of transistors that could be integrated onto a single chip at minimal cost would double every year [1]. By down-scaling the dimensions of the transistors, not only did we increase density, but we also improved the performance and reduced the cost of the chip. This led to faster data processing and cheaper data storage.

ML/AI algorithms utilize Artificial Neural Networks (ANNs) to classify the dataset into labels by intelligently extracting features from the dataset. Google AlphaGo is the first AI-based algorithm to defeat world champions in the board

game Go [2], demonstrating the potential of current AI algorithms, but requires significant CPU and GPU resources to train. OpenAI's GPT-3, an auto-regressive NLP model comprising 175 billion tunable parameters, is estimated to consume an energy of 932 MWh to train [3]. The large power consumption stems from training these ML algorithms on conventional computing systems based on the von-Neumann architecture. In the von-Neumann architecture, memory and computation are decoupled from one another. The processor needs to fetch the data from off-chip memory, process it, and store it back in the memory, making the process energy-intensive. The carbon footprint to train the current state-of-the-art ML algorithms on conventional computing systems has been increasing significantly over time [4]. Neuromorphic computing, on the other hand, takes inspiration from the functioning of the biological brain and aims to perform computation in an energy-efficient manner. With the approaching end of Moore's scaling law, neuromorphic computing presents an opportunity towards the realization of ultra-low power chips in the future. This requires a departure from traditional methods of computation and presents the need to develop novel energy-efficient hardware.

The human brain comprises about 10^{11} neurons and 10^{15} synapses while operating at a power budget of just 20 W. Thus, it is an extremely dense, inherently parallel, and energy-efficient architecture. The neurons in the biological brain communicate with each other via discrete action potentials or "spikes", which are sparse in time. Taking inspiration from the highly dense and energy-efficient architecture of the biological brain, newer forms of computation can

emerge that benefit from the advancements in ANNs and, at the same time, are energy-efficient. Spiking Neural Network (SNN), the successor to the widely popular ANN, aims to model the behavior of the biological neural network in an energy-efficient manner. In this work, various energy-efficient implementations of SNN are proposed and investigated.

1.2 Objectives

The goal of this work is to develop an energy-efficient framework for Neuromorphic computing using an SNN. It involves developing an insight into the state-of-the-art hardware required to implement an SNN and proposing novel devices that aid in implementing and training the SNN in an energy-efficient manner. Specifically, the objectives of this work are as follows.

- To propose an energy-efficient implementation of a Leaky-Integrate and Fire (LIF) neuron and investigate its behavior.
- To propose an implementation of an on-chip unsupervised learning framework in an SNN using Spike Timing Dependent Plasticity (STDP) and assess its performance.
- To propose a ternary SNN and develop a framework to implement unsupervised learning using STDP and compare its performance with a binary SNN on standard benchmarks.

1.3 Contributions

The major contributions of this work are summarized below.

- An energy-efficient LIF neuron is proposed to be employed in an SNN. The proposed neuron, comprising a Ge-based PD-SOI MOSFET, can directly receive the incoming voltage spikes and avoid energy dissipation in generating a summed potential. The smaller bandgap with dominant direct tunneling of Ge [5] allows the device to operate at a lower voltage level. The energy consumption per spike of the proposed neuron is 0.07fJ, which is lower than LIF neuron implementations (experimental or simulated) reported in the literature.
- A Ferromagnetic Domain Wall (FM-DW) based device is employed to function as a synapse. It comprises a Magnetic Tunnel Junction (MTJ) with a Heavy Metal (HM) underlayer. The MTJ consists of a free FM (CoFe) layer (whose magnetization can be varied) and a pinned FM layer (whose magnetization is fixed) separated by a tunneling oxide barrier (MgO). A DW separates two oppositely polarised magnetic regions in the free FM layer. A programming current flowing through the HM layer results in the movement of the DW in the free FM layer in the direction of the current flow. A displacement in the position of the DW results in a change in the conductance of the FM-DW synapse. The weight of the interconnection between two neurons is stored as the conductance of the synapse.

- An energy-efficient Ge-based device is proposed that implements on-chip unsupervised learning in an SNN using STDP. The proposed device configuration comprises a dual pocket FD-SOI MOSFET with dual asymmetric gates. Using a comprehensive device-to-system level simulation framework, it is demonstrated that a pair of such devices can generate a current, whose magnitude depends exponentially on the temporal correlation of spiking events between the pre- and post-synaptic neuronal layers. This current drives the HM layer in the FM-DW synapse and programs the conductance of the synapse in accordance with the STDP learning rule.

- A ternary SNN is proposed, which comprises a ternary spiking neuron and a spintronic synapse. A Dual-Pocket Tunnel Field effect transistor (DP-TFET) is employed to implement a ternary neuron. Two distinct tunneling mechanisms exist in the device - within-channel tunneling and source-channel tunneling, which are responsible for the generation of $V_{DD}/2$ and V_{DD} voltage spikes, respectively. The network is trained on-chip in an unsupervised manner using STDP. Using a device-to-system level simulation framework, it is demonstrated that the ternary SNN can offer a better classification accuracy and requires a smaller runtime to classify handwritten digits in the MNIST dataset than a conventional SNN.

1.4 List of Publications

1. **A. Gupta** and S. Saurabh, “An Energy-Efficient Ge-Based Leaky Integrate and Fire Neuron: Proposal and Analysis,” in *IEEE Transactions on Nanotechnology*, vol. 21, pp. 555-563, 2022, doi: 10.1109/TNANO.2022.3209078.

2. **A. Gupta** and S. Saurabh, “On-chip Unsupervised Learning using STDP in a Spiking Neural Network,” in *IEEE Transactions on Nanotechnology*, vol. 22, pp. 365-376, 2023, doi: 10.1109/TNANO.2023.3293011.

3. **A. Gupta** and S. Saurabh, “Implementing a Ternary Inverter Using Dual-Pocket Tunnel Field-Effect Transistors,” in *IEEE Transactions on Electron Devices*, vol. 68, no. 10, pp. 5305-5310, Oct. 2021, doi: 10.1109/TED.2021.3106618.

4. **A. Gupta** and S. Saurabh, “Novel attributes of a dual pocket tunnel field-effect transistor,” in *Japanese Journal of Applied Physics*, vol. 61, no. 3, p. 035001, 2022. doi: 10.35848/1347-4065/ac3722.

5. **A. Gupta** and S. Saurabh, “Unsupervised Learning in a Ternary SNN Using STDP," in *IEEE Journal of the Electron Devices Society*, vol. 12, pp. 211-220, 2024, doi: 10.1109/JEDS.2024.3366199

1.5 Organization

The rest of the thesis is organized as follows.

- Chapter 2 presents the background and related work of this dissertation.

Initially, a brief overview of neural networks is presented. Later, the current state-of-the-art is reviewed to implement an SNN.

- Chapter 3 presents an energy-efficient LIF neuron. The proposed neuron can directly accept voltage spikes as input and prevents energy dissipation in generating a summed potential.

- Chapter 4 presents an energy-efficient Ge-based device that implements on-chip unsupervised learning in an SNN using STDP. The proposed device configuration consists of a dual pocket FD-SOI MOSFET with dual asymmetric gates. A pair of these devices can be employed to generate a current that is exponentially dependent on the temporal correlation of spiking events between the pre-synaptic and the post-synaptic neuronal layers. This generated current modulates the conductance of the synapse in accordance with the STDP learning rule.

- Chapter 5 presents a ternary SNN. In the proposed implementation, a ternary spiking neuron has been implemented using a Dual-Pocket Tunnel Field Effect Transistor (DP-TFET). The network is trained in an unsupervised manner using STDP. The ternary SNN was trained to classify handwritten digits in the MNIST dataset.

- Chapter 6 concludes the dissertation and discusses the possible future research directions.

Chapter 2

Background and Related Work

In this chapter, the first part provides a brief overview of neural networks. The second part of this chapter reviews the current state-of-the-art in building an SNN.

2.1 Introduction to Neural Networks

A neural network is essentially a network of neurons interconnected via synapses designed for some information-processing task. The synapse stores the weight of the interconnection between two neurons. Fig. 2.1 shows a simple feed-forward neural network. It consists of an input layer, one or more hidden layers, and an output layer. The input layer of neurons receives stimuli from the external environment. The hidden layers of neurons process that information received from the input layer in a meaningful manner, and finally, the output layer generates suitable output stimuli. There can be different configurations of neural networks other than the feed-forward connection, as shown in Fig.

2.1. These are recurrent (connections between neurons in the same layer) or backward propagation (feedback connection) [6, 7].

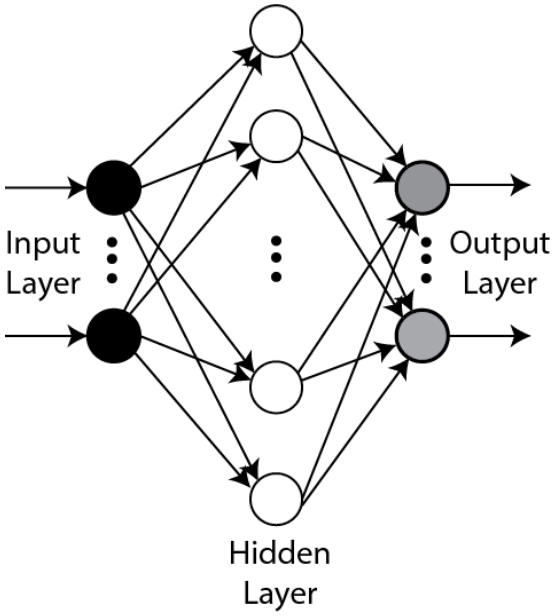

Figure 2.1: A simple feed-forward neural network

Training a neural network involves modulating the weights stored in the synapses such that the network gives the correct output for a given input. The neural network can be trained in a supervised or unsupervised manner. In supervised training, for each input pattern, the desired output is available. A loss function compares the output generated by the network with the desired output. This loss function is then utilized to modulate the weights of the synapses with the goal of minimizing this loss function. In unsupervised training, the desired output for a given input is not available with the network. The network must adjust its weights accordingly to generate the correct output for a given input pattern. Now that we have a brief understanding of what a neural network is, it will be beneficial to discuss the types of neural networks.

2.1.1 Biological Neural Network

A biological neural network consists of a vast network of biological neurons, which form the fundamental unit of our brain and nervous system. Sensory neurons receive information from the external environment, which is then processed in our brain, and finally, the motor neurons direct our muscles to take the appropriate action. Fig. 2.2 shows the structure of the neuron [8].

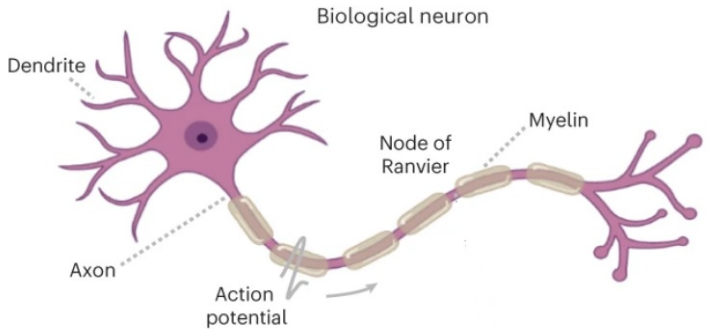

Figure 2.2: The biological neuron [8]

The biological neuron comprises the soma (cell body), dendrites, and axon. A gap between two neurons is called a synapse. The dendrites receive the inputs due to spiking activity in the pre-synaptic neurons. The membrane potential of the neuron is stored in the soma. Spiking activity in the neighborhood of a particular neuron results in an increment in the membrane potential. When it exceeds a threshold, the neuron outputs an electrical spike, which transfers along the length of the axon. The nerve endings present at the end of the axon convert the electrical impulse into a chemical signal in the form of neurotransmitters. This neurotransmitter is sensed by the dendrite of a post-synaptic neuron, which

results in a change in the membrane potential of the post-synaptic neuron.

2.1.2 Artificial Neural Network

Artificial Neural Networks (ANNs) aim to model the functionality of the biological neural network with the help of artificial neurons and synapses. Maass classified the ANN models into three generations on the basis of their computational units, i.e., neurons and synapses [9].

The first generation of neural networks consists of a network of artificial neurons called perceptrons. The perceptron comprises two sections: sum and threshold, as shown in Fig. 2.3. The sum section is connected with the preceding layer of perceptrons via synapses. Boolean output from the perceptrons is weighted by the weights ($w_1, ...w_N$) stored in the synapses, and a thresholding operation is performed on the result of the sum. If the sum is greater than the threshold, the state of the perceptron is active (output = 1). Otherwise, the perceptron is inactive (output = 0). The first generation of neural networks can also be trained by adjusting the weights of the synapses. The weights can be updated in accordance with some learning rule such that for a given input the network generates the desired output. Training the first generation of neural networks having one or more hidden layers in a supervised manner was infeasible because it required the output generated by the perceptron to be continuous in nature. However, the outputs generated by the perceptron are discontinuous. It was not feasible to compute the derivative of the loss function and back-propagate the error to modulate the weights of the synapses.

Unsupervised training employing some form of Hebbian learning rule, which stated that “neurons that fire together, wire together," was much more suited to train the perceptron.

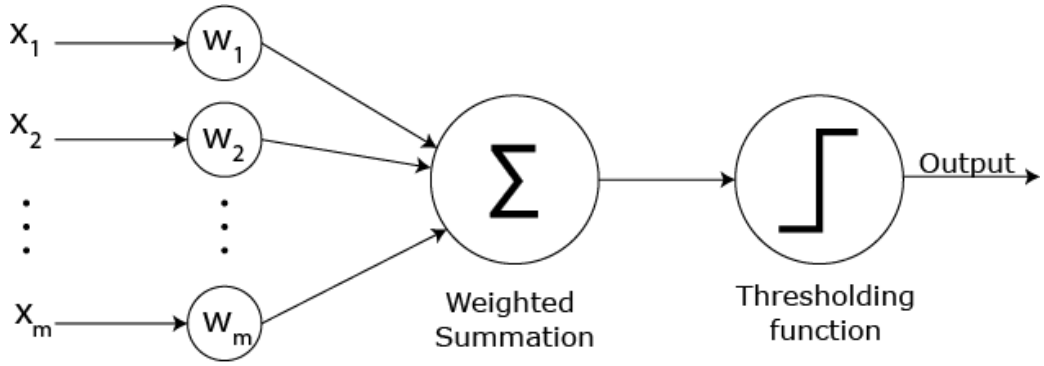

Figure 2.3: The first generation of ANNs, as categorized by Maass, comprising a network of artificial neurons, called perceptrons [9].

The second generation of neural networks comprises the perceptron, as in the first generation of neural networks. However, the thresholding function used in the first generation of neural networks was replaced with an activation function, which mapped the input to a continuous output $y \in [0, 1]$. Thus, the state of the neuron no longer represented whether the neuron fired or not, but rather represented a probability of firing. The second generation of ANNs, as categorized by Maass, is shown in Fig. 2.4. Since, the neuron now had a continuous nature of the output, training the network in a supervised manner using algorithms such as gradient descent was feasible. It garnered a lot of success in recent times.

The third generation of neural networks, as categorized by Maass, comprising spiking neurons as the computational units are shown in Fig. 2.5. The transition from the discontinuous spike outputs in the first generation of neural networks to

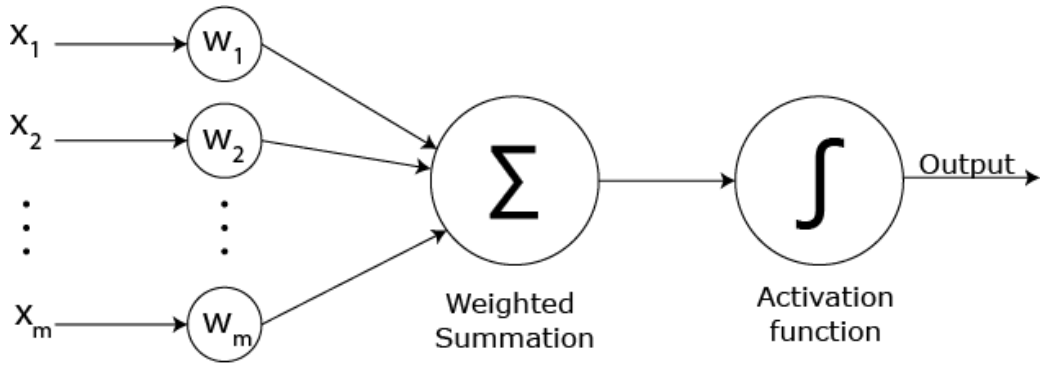

Figure 2.4: The second generation of ANNs, as categorized by Maass, comprised a network of artificial neurons with an activation function, which mapped the input to a continuous output.

continuous spike outputs in the second generation of neural networks resulted in the development of supervised training algorithms like gradient descent, which led to a lot of success in the fields of ML, AI, and deep learning. However, the power consumed in training these networks became prohibitively large and presented a need to develop energy-efficient models of neurons, which could be used to train the network in an energy-efficient manner. The spiking neuron had a discontinuous output, similar to that in the first generation of artificial neurons, but had a notion of time built into it, in the form of the membrane potential of the neuron changing over time. When the membrane potential crosses a particular threshold, the neuron outputs a discrete spike, and subsequently, its membrane potential is reset. Thus, the output of the spiking neuron will be in the form of a spike train. The information can be encoded by the temporal firing event of a neuron. The third generation of artificial neural networks, comprising spiking neurons, is popularly known as the Spiking Neural Network (SNN).

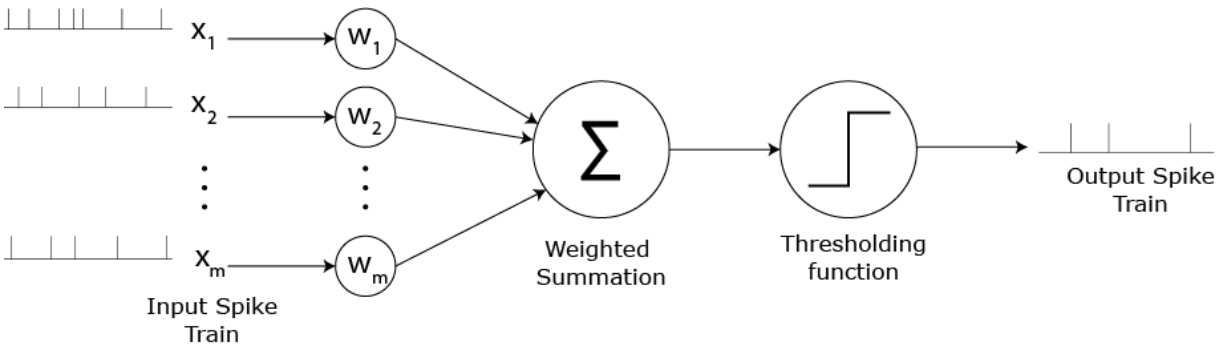

Figure 2.5: The third generation of ANNs, as categorized by Maass, comprising spiking neurons, is popularly known as the Spiking Neural Network (SNN).

2.2 Spiking Neural Network

The two fundamental building blocks in an SNN are neurons and synapses. In this section, we study how spiking neurons and synapses are modeled in literature. Next, we will study how these spiking neurons can be interconnected using synapses in an energy-efficient manner. Further, we will look at some techniques that can be used to train the SNN by modulating the weights stored in the synapses for a particular application.

2.2.1 Modeling Spiking Neurons

The biological neuron is widely modeled in literature, ranging from biologically plausible to those inspired by biology having simpler computational models. The latter models are inspired by ANN models rather than by biological brains. The spiking neuron models can be divided into the following three broad categories:

- Biologically plausible: Models the behavior observed in the biological neuron.

- Biologically inspired: Aims to mimic the behavior of the biological neuron, but not necessarily in a way that is consistent with biology.
- Integrate-and-fire: A simpler implementation of biologically-inspired neurons.

1. *Biologically plausible*: The Hodgkin-Huxley neuron model is the most popular biologically plausible model [10]. It is a highly complex neuron model comprising four-dimensional non-linear differential equations used to describe the ion transfer dynamics of the neuron. The Hodgkin-Huxley neurons have been widely used in neuromorphic implementations that attempt to model biological neural systems due to their biological plausibility accurately.

2. *Biologically inspired*: The biologically inspired neuron models aim to model the behavior of the biological neuron rather than emulating the physiological activity in the biological nervous systems. The computational complexity of these models is much less than that of biologically plausible models. The Izhikevich neuron model is popular in neuromorphic literature due to its simpler implementation and biologically realistic behavior [11].

3. *Integrate-and-fire*: The integrate-and-fire (IF) neuron models have lesser computational complexity than their biologically inspired counterparts. These are less biologically realistic but are easier to implement in hardware [12]. The simplest IF neuron model integrates incoming spikes from the pre-synaptic layer onto its membrane potential. In the absence of in-

coming spikes, the membrane potential does not decay with time. When the membrane potential exceeds a certain threshold the neuron fires and then resets its membrane potential. The Leaky Integrate and Fire (LIF) neuron model introduces a leaky term where the membrane potential decays with time. The LIF model is one of the most popular models in neuromorphic systems [13].

The appropriate model to choose from the three broad categories of spiking neurons discussed above can vary from application to application. For example, suppose it is desired to simulate the biological brain to study neuroscience. In that case, a biologically plausible neuron model can be chosen since it would emulate the behavior of the biological brain realistically. However, if it is required to train an ML model for image classification, then a biologically inspired model or an IF spiking neuron model with a simpler hardware implementation may be used. In this work, our main focus is to implement an energy-efficient SNN, which can be used to train an ML model in hardware. Hence, the neuron models used in this work belong to the IF family. In particular, we will focus on the energy-efficient hardware implementation of a LIF model due to its computational simplicity and biological realism.

Several hardware implementations of the LIF neuron exist in the literature [14–21]. In [15], a CMOS-based implementation of the LIF neuron is proposed, which comprises about 20 transistors to realize a single neuron. Considering the highly dense and power-efficiency requirements of the neuromorphic system,

such an implementation is infeasible in terms of area and power consumption. In [16–18], a Partially-Depleted Silicon on Insulator (PD-SOI) based MOSFET was used to implement the LIF neuron. The device used to implement the LIF neuron in [16] is as shown in Fig. 2.6(a) with the biasing scheme shown in Fig. 2.6(b).

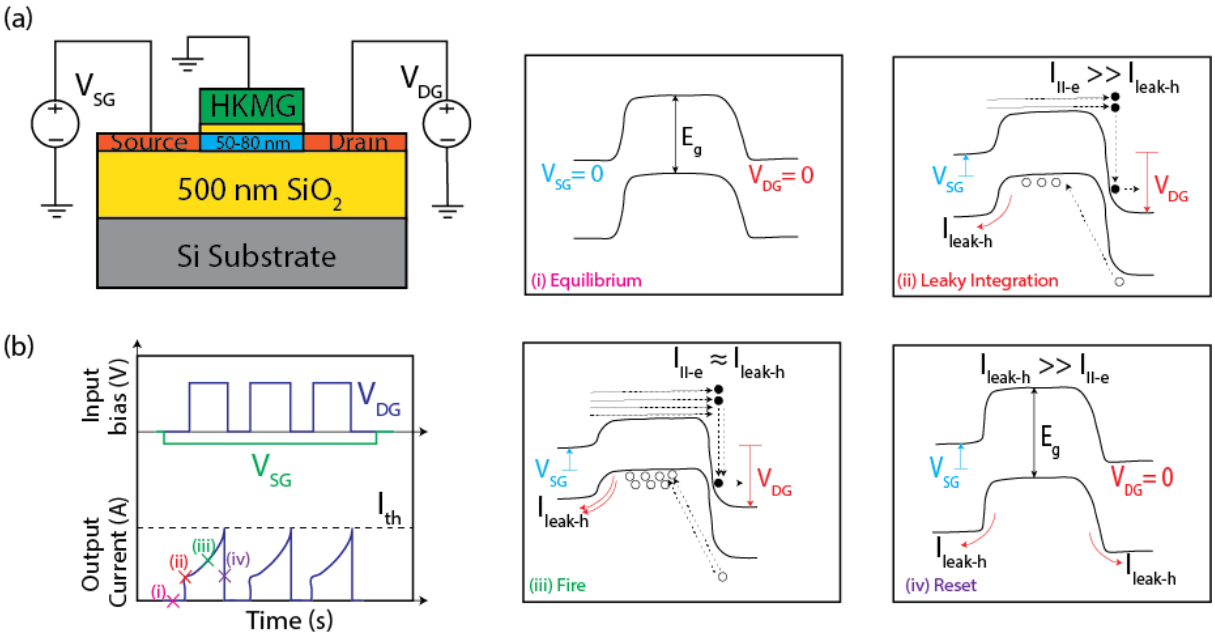

Figure 2.6: Impact Ionization based neuron [16]

Device operation can be explained using the following four phases [16]. During the equilibrium phase, the voltage applied on the drain (V_{DG}) and source (V_{SG}) terminals are $0V$. The electrons present at the source do not have sufficient energy to surmount the energy barrier and reach the drain. Hence, no current flows through the device. During the leaky integration phase, a large voltage ($V_{DG} = 2.8V$) is applied to cause Impact Ionization (II) in the device. Now, at $V_{SG} = 0V$, some current due to the thermionic emission of carriers flows through the device, but it is not sufficient to generate enough excess carriers required for the neuron to spike. Due to the spiking activity of neurons in the

pre-synaptic layer, a negative voltage appears at the source, i.e., $V_{SG} < 0V$. A negative V_{SG} results in decreasing the height of the potential barrier and results in an exponential increase in the number of electrons reaching the drain. This causes an exponential increase in the generation of excess carriers in the device due to II and results in the storage of holes in the potential well formed in the channel. At the same time, some of the accumulated holes leak through the source barrier. Due to the accumulation of charge in the channel region, the potential barrier seen by the electrons present in the source reduces, causing more II in the device. This results in positive feedback in the device, and there is a sharp increase in current through the device. At steady state, the rate of accumulation of holes becomes equal to the rate of leakage of holes through the source barrier. The current flowing at steady state is called the threshold current. When this state is reached, the neuron is said to have fired. The neuron is then reset by triggering a reset circuitry, which removes the voltage applied to the drain. This causes the accumulated holes in the channel to leak away through the source and drain barriers, and eventually, the equilibrium condition is reached. After a fixed period, called the refractory period, the entire process can be repeated.

The PD-SOI-based implementation of the LIF neuron used Impact Ionization (II) to generate excess carriers and required a large voltage to be applied at the drain to cause II in the device, making it inefficient in terms of energy consumption. Moreover, II is not an efficient method to generate excess carriers, i.e., electron-hole pairs (EHP) in the device. It requires imparting high kinetic

energy to the carriers using a high electric field, and only 0.1% of the drain current is responsible for the generation of EHPs [20]. Hence, voltage spikes, which go away after a short duration of time, cannot lead to an efficient generation of EHP. Thus, a summing circuitry, as shown in Fig. 2.7, comprising an op-amp and a capacitor, is required to sum up the currents received in the form of spikes from the pre-synaptic neuronal layer to generate a fixed voltage, which is then applied to a terminal (source terminal in ref. [16]) of the device. The capacitor keeps charging continuously with incoming spikes and discharges if there is no spiking activity, thereby dissipating significant power. Considering the highly dense architecture of an SNN, this implementation has a considerable overhead in terms of area and energy dissipation. The assumption of this summing circuitry is not accounted for in the energy consumption per spike reported in these studies [16–18]. Using SPICE simulations, we show that charging and discharging of the capacitor leads to about 10-50% overhead in terms of energy consumption. Hence, we cannot ignore this component of power dissipation. Additionally, the reset circuitry also consumes significant power due to the requirement of using a large drain voltage to cause II in the device. Das et al. reported an II-based energy-efficient NIPIN diode-based LIF neuron, which did not require a large voltage for its operation [19]. However, it required the fabrication of a thin P^+ pocket in the channel, making it difficult to fabricate.

LIF neurons which operate on the principle of Band to Band Tunneling (BTBT), as proposed in [20, 21], can offer a higher energy efficiency when compared to II-based neurons because BTBT-based LIF neurons can operate

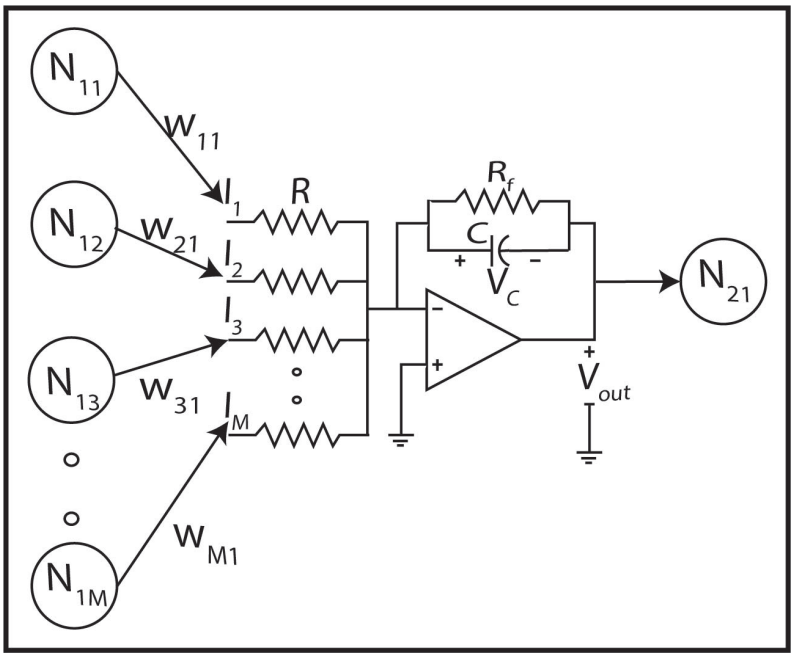

Figure 2.7: Summing circuitry used to generate a fixed potential, which is then applied to the subsequent neuron [16–18].

at smaller supply voltages and can spike at much smaller currents compared to II-based neurons. A Si-based PD-SOI MOSFET was used to implement the LIF neuron in [20]. The gate terminal is grounded and is used as the reference terminal. The source is kept at a small positive value ($V_{SG} = 0.4V$) to bias the device in deep OFF-state and reduce leakage. The input, in the form of accumulated potential from pre-synaptic neurons, is applied to the drain terminal ($V_{DG} = V_{input}$), and the voltage output is taken from the body terminal (V_{BG}). When the accumulated potential is of the order ($V_{DG} \sim 1V$), the minority electrons in the body tunnel into the drain, leaving behind holes in the floating body, resulting in an increase in body potential. A fraction of these accumulated holes leak through the source barrier. Due to the storage of holes in the body, the potential barrier seen by the electrons in the source decreases, resulting in an increase in the leakage of holes through the source barrier. At steady state,

the rate of accumulation of holes equals the rate of leakage of holes. The body potential, at steady state, is called the threshold potential. When the steady state is reached, the neuron fires and subsequently resets by connecting the device's body terminal to the ground. This results in the removal of the stored holes. The larger the accumulated potential at the drain, the larger the BTBT generation rate will be, resulting in a higher spiking frequency.

In this work, a Ge-based PD-SOI MOSFET employing BTBT has been proposed to implement the LIF neuron. BTBT current in a Ge-based device that has a smaller bandgap and dominant direct tunneling mechanism [5] can easily generate EHP in contrast to Si-based devices that utilize II or BTBT. Thus, even if voltage spikes are used as input, EHP are generated efficiently in the proposed neuron. Hence, we can directly apply the spikes to the neuron and at lower voltages, thus saving power dissipated in both the neuron and the summing circuitry.

2.2.2 Modeling the Synapse

Two neurons are interconnected via a synapse. We can classify the synapse models into two broad categories: biologically plausible and those inspired by biology. The biologically plausible synapse models aim to replicate the chemical interactions of biological synapses, such as the ion pumps or neurotransmitter interactions [22–24]. The synapse used in an SNN stores the weights of the interconnection between two neurons. Since the synapse is the most repeated element in an SNN, it is paramount that its implementation in hardware is area

and energy-efficient.

Synaptic plasticity refers to the ability of the synapse to modulate its weight over time and thus effectuate learning. Algorithms such as gradient descent cannot be used to train an SNN because taking a derivative of discrete action potentials is not feasible. We can train an SNN using Spike Timing Dependent Plasticity (STDP) by following the Hebbian form of learning [25], which states that “neurons that fire together wire together." This essentially means that if a pre (post) synaptic neuron spiking event is followed by the spiking event on the post (pre) synaptic neuron, the relationship is considered to be causal (anti-causal), and the strength of the synapse is increased (decreased). This increase (decrease) in the synapse’s strength depends exponentially on the temporal correlation between the two layers of spiking neurons, in accordance with the STDP learning rule.

There are several implementations of the artificial synapse in hardware. In [26], the synaptic weight is stored in an 8-T Static Random Access Memory (SRAM) with a precision of 4 bits. This implementation required 32 transistors per synaptic element and was inefficient in terms of area and energy consumption. Moreover, the stored weight is lost when the SRAM is powered off. Artificial synapses based on emerging memory technologies are promising from the perspective of the area and energy-efficient hardware implementation. A synapse has been modeled in the literature using Non-Volatile Memories (NVM) like memristors [27–30], floating gate transistors [31], Phase Change Memories (PCM) [32], Ferroelectric RAM (FeRAM) [33], spintronic devices [34–36], etc.

A memory resistor or memristor is perhaps the most ubiquitous device-level component in neuromorphic systems. The conductance of the memristor (G_{RRAM}) can be incrementally modified by controlling the current through it. In [27], a one-transistor/one-resistor (1T1R) Resistive Random access Memory (RRAM) is proposed to implement a binary synapse. A binary synapse has only two states - a low resistance state (LRS) and a high resistance state (HRS). The RRAM device consists of a Si-doped HfO_2 layer sandwiched between a TiN bottom electrode (BE) and a Ti top electrode (TE). A positive voltage ($V_{set} \sim 1.5V$) at the TE with respect to the BE results in the set transition where the device switches from the HRS to the LRS. Application of a negative voltage ($V_{reset} \sim -1V$) at the TE with respect to the BE results in the reset transition where the device switches from the LRS to the HRS. Both the set and reset mechanisms are abrupt, resulting in a binary synapse. A pre-synaptic neuron firing event (V_{spike}) results in the generation of a current ($I = G_{RRAM} \times V_{spike}$), which is added at the post-synaptic neuron and results in an increment in its internal potential. The 1T1R structure, as proposed in [27] is shown in Fig. 2.8.

The PRE spike voltage is applied to the gate of the transistor (V_G). The POST spike controls the TE voltage and is set to a low constant voltage ($V_{TE} = 20mV$) called the communication voltage by default. Every PRE spike activates the transistor and results in a current flow, which varies inversely with the resistance of the RRAM device. This current is collected by the POST neuron, which also collects currents from other synapses. Every POST neuron spike not only results in a voltage spike but also generates a V_{TE} signal, as shown in Fig. 2.8. The

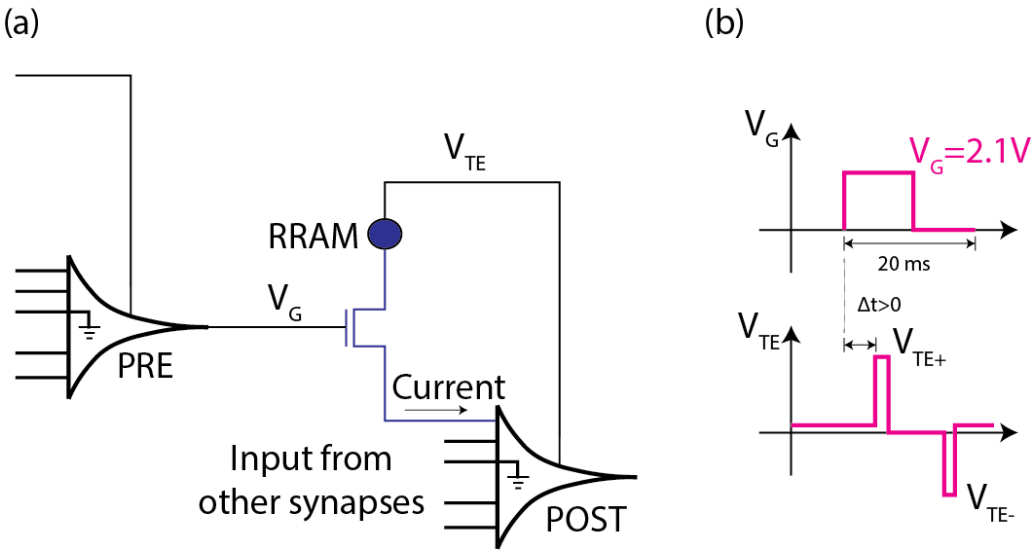

Figure 2.8: The 1T1R synapse structure proposed in [27]

signal V_{TE} has two phases after a POST spike, the first phase with $V_{TE+} = 2.5V$ for a duration of 1 ms followed by $0V$ for 9 ms and the second phase with $V_{TE-} = -1.6V$ for a duration of 1 ms followed by $0V$ for 9 ms. After the two phases have elapsed, V_{TE} is set to the communication voltage of 20 mV. The PRE spike corresponds to the first phase of the voltage of $2.1V$ for a time duration of 10 ms and a second phase of voltage $0V$ for 10 ms. When the first phase of the PRE spike overlaps with the first phase of V_{TE} (i.e., V_{TE+}), the RRAM device undergoes the set transition (or potentiation). However, when the first phase of the PRE spike overlaps with the second phase of V_{TE} (i.e., V_{TE-}), the RRAM device undergoes the reset transition (or depression). Thus, the abrupt set/reset transitions in the RRAM result in bistable STDP, which contrasts with gradual weight tuning in biological STDP.

In [28], a memristor-based synapse was proposed and STDP was demonstrated using CMOS circuits. The temporal correlation between spiking events is

translated into pulse width with the help of CMOS transistors, which is applied to the memristive synapse. A larger pulse width results in a larger increase or decrease in conductance depending on the polarity of the applied voltage to the memristor, resulting in synaptic plasticity.

A floating gate transistor is used to implement a synaptic element that stores the weight in a non-volatile manner [31]. The synapse's strength can be modulated based on the temporal correlation of spiking events between the pre- and post-synaptic neurons in accordance with the STDP learning rule. If the post-synaptic neuron firing event follows the pre-synaptic neuron firing event, electrons are added to the floating gate via Hot Carrier Injection (HCI) mechanism, increasing the weight of the synapse. However, when the pre-synaptic firing event follows the post-synaptic firing event, electrons are removed from the floating gate via the Fowler-Nordheim (FN) tunneling mechanism, decreasing the weight of the synapse. A pre-synaptic computation block is required per pre-neuron, which generates a triangular waveform after every pre-synaptic neuron spike. All the post-neurons share this computation block. Large voltages ($\sim 15V$)are required to cause FN tunneling in the device, which is energy-inefficient. Also, the circuitry operates at a timescale of a few ms, resulting in a higher latency to modulate the weight of the synapse.

A phase-change memory (PCM) is utilized to emulate biological synapses [32]. A phase change material exhibits switching between its amorphous (high resistance) and crystalline (low resistance) states by application of voltage pulses, which generate the heat required for the material to transform its phase. GST

$(Ge_2Sb_2Te_5)$ is used as the phase-change material, which is sandwiched between the bottom electrode (W) and top electrode (TiN). Switching from the High Resistance State (HRS) to the Low Resistance State (LRS) is called set, while switching from the LRS to HRS is called reset. The GST layer is polycrystalline with a resistance of about 500Ω in the fully set state and amorphous with a resistance of about $2M\Omega$ in the fully reset state. Intermediate resistance states can be programmed between the fully set and reset states, in line with the analog nature of biological synapses. A pre-spike, in the form of a pulse train, is applied to the top electrode of the PCM synapse, and a post-spike, also in the form of a spike train, is applied to the bottom electrode. STDP is used as the learning mechanism to modulate the resistance of the PCM synapse. An energy of about $50pJ$ is required per synaptic update during the reset phase, and the circuitry operates at a timescale of a few ms, resulting in a higher latency to modulate the weight of the synapse.

A Ferroelectric-gate Field effect transistor (FeFET) is used to demonstrate a synaptic element [33]. The FeFET comprises a $ZnO/Pb(Zr,Ti)O_3$ (PZT) structure where the conductance of the device can be switched by changing the polarization of the ferroelectric layer by applying the appropriate voltage on the gate. The time difference of spiking activity between the pre- and post-synaptic neuron is translated into a pulse height of a fixed width. A positive pulse is generated if the post-synaptic neuron firing event follows the pre-synaptic neuron firing event. However, if the pre-synaptic neuron firing event follows the post-synaptic neuron firing event, a negative pulse is generated. The larger

the temporal correlation of spiking events between the pre- and post-synaptic neurons, the larger the height of the voltage pulse generated. This voltage pulse is applied to the gate of FeFET and results in a change in polarization of the ferroelectric layer. Thus, the device's conductance is modulated per the STDP learning rule. The FeFET-based synapse is a three-terminal device that simultaneously supports signal processing and learning, unlike resistive switching devices (2 terminal devices), where learning can only be performed when signal transmission between neurons is aborted. The STDP circuitry operates at a timescale of a few μs, offering a lower latency than floating-gate or PCM-based synapse implementations.

In [34], a Magnetic Tunnel Junction (MTJ) Heavy Metal (HM) based binary stochastic synapse is proposed. In a binary synapse, potentiation (depression) would result in the maximum (minimum) conductance states of the synapse. However, in a binary stochastic synapse, synaptic plasticity is achieved by stochastically switching the conductance of the MTJ-HM synapse between their high and low conductance states. The conductance state of the MTJ stochastically switches in the presence of the thermal noise with a finite probability, which depends on the temporal correlation of spiking events between the pre- and post-synaptic spiking activity in accordance with the STDP learning rule. Thus, the higher the temporal correlation in spiking events, the higher the switching probability. A stochastic binary synapse can be considered equivalent to its analog counterparts and results in gradual weight tuning as in biological STDP. The MTJ-based binary stochastic synapse and CMOS-based circuitry used to

implement synaptic plasticity using STDP as proposed in [34] is shown in Fig. 2.9.

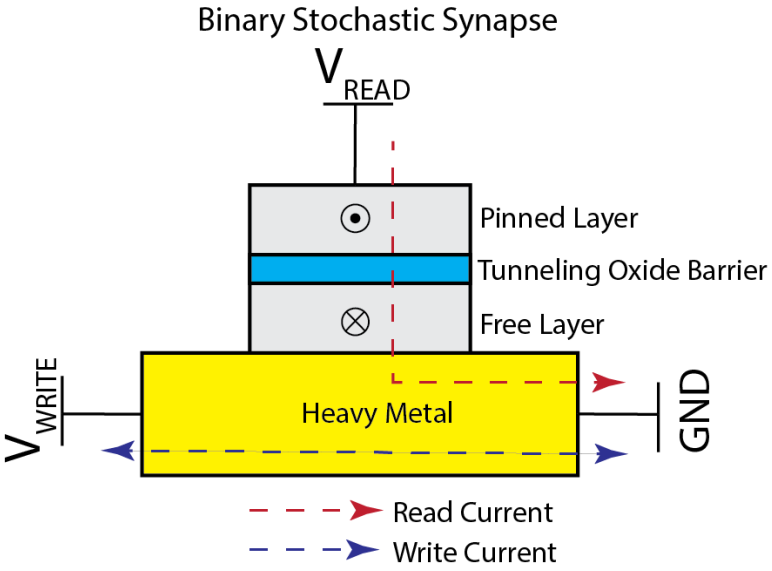

Figure 2.9: The MTJ-based binary stochastic synapse

The MTJ shown in Fig. 2.9 consists of a free Ferromagnetic (FM) layer (whose magnetization can be varied) and a pinned FM layer (whose magnetization is fixed) separated by a tunneling oxide barrier (MgO). The MTJ exhibits two stable conductance states depending on the orientation of the free FM layer with respect to the pinned FM layer. It is said to be in the high (low) conductance state if the magnetization of the free FM layer is parallel (anti-parallel) with respect to the pinned FM layer. The magnetization of the free FM layer can be varied by passing a current through the Heavy metal (HM) layer beneath the free FM layer. A CMOS-based circuitry is used to generate the current, whose magnitude depends on the temporal correlation of spiking events in the pre- and post-synaptic neurons.

In [35], a Ferromagnetic-Domain Wall (FM-DW) synapse is proposed wherein

the conductance of the synapse varies continuously in an analog manner between the minimum and maximum values. The synaptic element, comprising an MTJ with an HM underlayer, is shown in Fig. 2.10. The MTJ consists of a free FM (CoFe) layer (whose magnetization can be varied) and a pinned FM layer (whose magnetization is fixed) separated by a tunneling oxide barrier (MgO). A DW separates two oppositely polarised magnetic regions in the free FM layer. A programming current flowing through the HM layer (between terminal T2 and T3) results in the movement of the DW in the free FM layer in the direction of the current flow. At the extreme ends of the free FM layer, two pinned FM layers with opposite directions of magnetization exist to stabilize the DW for sufficiently large magnitudes of current flowing through the HM layer. A displacement in the position of the DW results in a change in the conductance of the FM-DW synapse. The FM-DW synapse has decoupled read and program paths wherein learning and spike transmission between neurons can occur simultaneously.

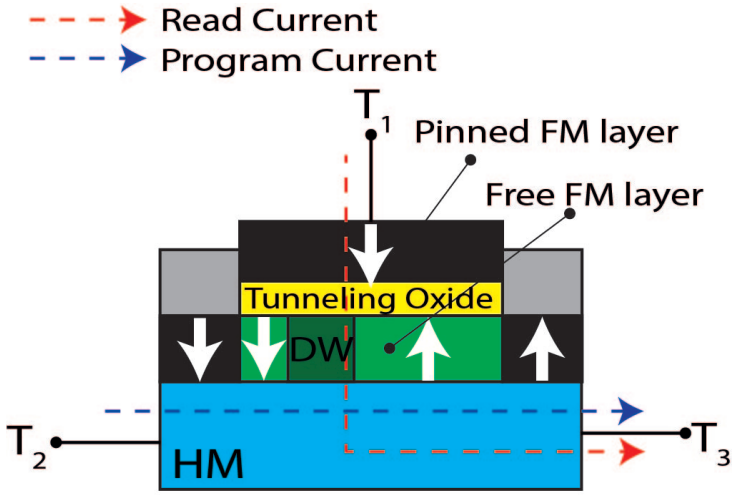

Figure 2.10: FM-DW synapse with decoupled read and program paths. The read current flows between terminals T_1 and T_3 while the programming current flows between terminals T_2 and T_3.

Let us denote the conductance of the FM-DW synapse in the parallel (anti-

parallel) configuration of magnetization of the free layer with respect to the fixed FM layer to be G_P (G_{AP}). The conductance of the synapse (G_S), when the DW is displaced by x along the length of the FM free layer, is given as follows [35]:

$$G_s(x) = G_P(\frac{x}{L}) + G_{AP}(1 - \frac{x}{L}) \tag{2.1}$$

where L denotes the length of the FM free layer in the MTJ. The parameters G_P, L, and G_{AP} are constants; hence, G_S varies linearly with the position of the DW in the FM free layer. It is to be noted that the maximum (minimum) value of conductance is G_P (G_{AP}), and their ratio is called Tunneling Magnetoresistance Ratio (TMR).

The FM-DW synapse described above requires a programming current (in the range of a few μA) to flow through the HM layer of the synapse for a short duration of time (a few ns). Thus, the FM-DW synapse offers a better energy efficiency (in fJ) and lower latency than synapse implementations with other NVMs like memristive switching devices [27–30], floating gate transistors [31], PCM [32], and FeFET [33]. Therefore, in this work, we have employed an FM-DW-based device as a synaptic element.

While the FM-DW synapse can be programmed in the timescale of a few ns, the CMOS-based circuitry necessary to generate a suitable programming current operates at a timescale of a few μs. Thus, in the existing literature [34–36], STDP is implemented with a higher latency (μs) than is supported by the FM-DW synapse (ns). The CMOS-based circuitry used to generate the

programming current is shown in Fig. 2.11. The following paragraph explains the working of the same.

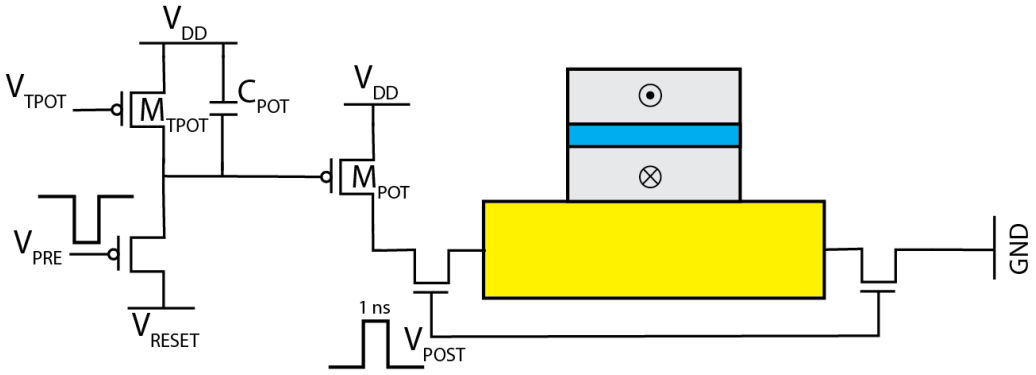

Figure 2.11: The CMOS-based circuitry used to generate a programming current in accordance with the STDP learning rule as proposed in [34]

When a pre-synaptic neuron firing event is observed (V_{PRE}), an external capacitor (C_{POT}), initially at a potential V_{RESET}, starts to linearly charge towards V_{DD} via a PMOS transistor (M_{TPOT}) operating in the subthreshold regime. This capacitor is connected to a PMOS transistor (M_{POT}). Now, if a post-synaptic neuron firing event is observed shortly after the pre-synaptic neuron firing event, a programming current (order of few μA) flows through the synapse for a short duration of time (1 ns), which modulates the conductance of the synapse. The smaller the time difference between the spiking events in the pre- and post-neuron, the smaller will be the potential at the gate of M_{POT}, and consequently, a larger current will flow through the synapse. This will result in a larger probability of switching the conductance of the synapse. If there is no post-neuronal spiking activity, C_{POT} slowly charges to V_{DD} (in a few μs) via a subthreshold current (a few nA) and returns to its resting potential V_{RESET}. Since the spiking activity in the pre-synaptic neuronal layer is much higher than

in the post-synaptic neuronal layer, every spiking activity at the pre-synaptic neuron will not result in spiking activity in the post-synaptic neuron. Thus, energy is expended in needlessly charging and discharging C_{POT} when there is no need to modulate the conductance of the synapse (no post-neuron firing event was observed). A similar problem exists in the CMOS-based implementation of the STDP learning circuitry proposed in [35, 36] as well.

In this work, the STDP learning rule is implemented wherein such a CMOS-based circuitry is not required, and there is no charging/discharging of external capacitors. A programming current is generated only when it is required to program the conductance of the synapse. Also, the proposed implementation operates at a time frame of the order of a few nanoseconds (because no external capacitor needs to be charged). It offers much lower latency when compared with existing literature [34–36], which operates at the order of a few μs. Moreover, the proposed implementation requires 2-3$\times$ fewer transistors to implement the STDP learning rule compared to CMOS-based implementations [28, 29, 34–36].

2.2.3 Network Models

Network models describe the topology with which neurons are interconnected via synapses and how spikes are transmitted along the network. The selection of an appropriate network model depends on several factors. One of these factors can be whether the target application involves biological inspiration or the complexity of the chosen neuron and synapse models. Another factor can be the applicability of the existing training algorithms to train the chosen network

model. The network models range from Spiking feed-forward networks [37–39], Spiking Deep Neural Networks (DNN) [40–42], Spiking Recurrent networks [43–45], Spiking Convolutional Neural Network (CNN) [46], Spiking Deep Belief Networks (DBN) [47], Spiking Winner Take All (WTA) networks [48,49], etc.

The Spiking feed-forward neural network is by far the most popular implementation of an SNN. It comprises multiple layers of neurons, where each neuron in a layer is connected to neurons in the subsequent layer. Neurons in this type of network do not form a cycle, i.e., no back connections exist. The information in these types of networks always traverses forward, hence the name. As the training progresses in a spiking feed-forward network, the weights stored in the synapses get modulated with each forward pass of the input in accordance with the training algorithm employed. A feed-forward SNN comprises an input layer of neurons, an output layer of neurons, and one or more hidden layers of neurons. When the number of hidden layers in an SNN becomes substantially large, these are referred to as spiking DNNs. Recurrent Neural Networks (RNN), unlike feed-forward networks, allow feedback between internal nodes in the network, thereby forming cycles. Thus, it exhibits a dynamic temporal behavior, causing the output from a previous time step to affect the input to the current time step. RNNs are commonly employed for temporal problems, like Natural Language Processing (NLP), speech recognition, etc. Fig. 2.12 shows a spiking feed-forward and recurrent neural network.

Convolutional Neural Networks (CNNs) are widely used for image classifica-

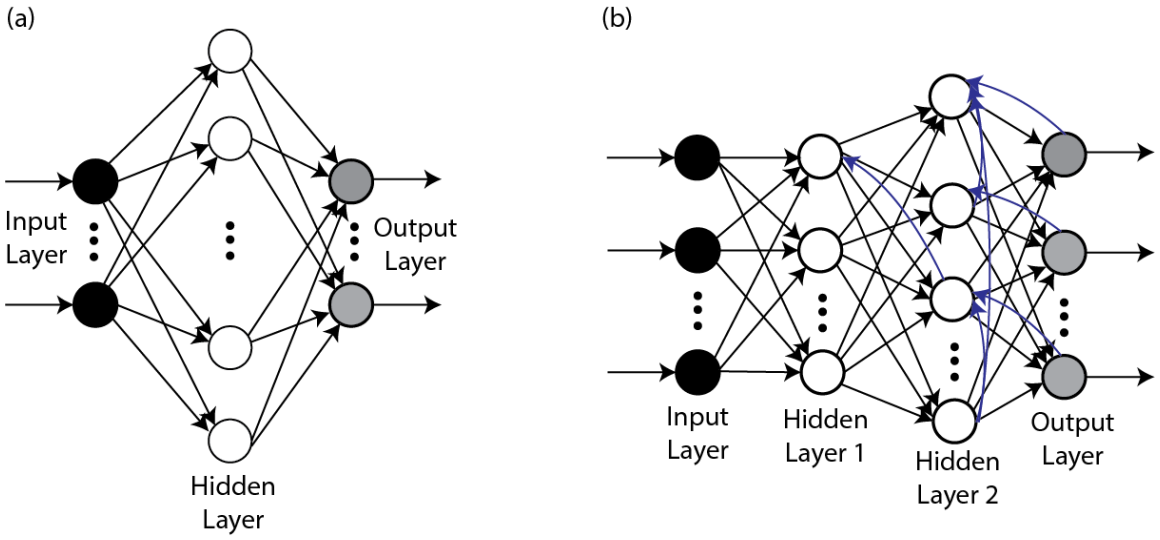

Figure 2.12: Spiking Feed-forward and recurrent neural network models

tion and computer vision applications. It leverages principles of linear algebra, specifically matrix multiplication, to identify patterns within images. CNNs comprise three main layers of neurons, viz., a convolutional layer, a max pooling layer, and a fully connected layer of neurons. The convolutional layer is the main building block of a CNN. In this layer, a dot product operation is performed between the convolution kernel (of size $n \times n$) and the input matrix. The kernel is shifted, and the process is repeated until the kernel has swept across the entire image. A feature map is obtained after convolving the entire image with the kernel. This feature map is fed to one or more convolutional layers or to a max pooling layer. In the max pooling layer, a filter sweeps across the entire feature map and, during each stride, selects the maximum value as output, thereby reducing the dimensions of the data. Finally, the fully-connected layer performs the task of classification based on the features extracted through the previous layers. A spiking CNN is shown in Fig. 2.13.

Winner-take-all (WTA) networks utilize recurrent inhibitory connections such

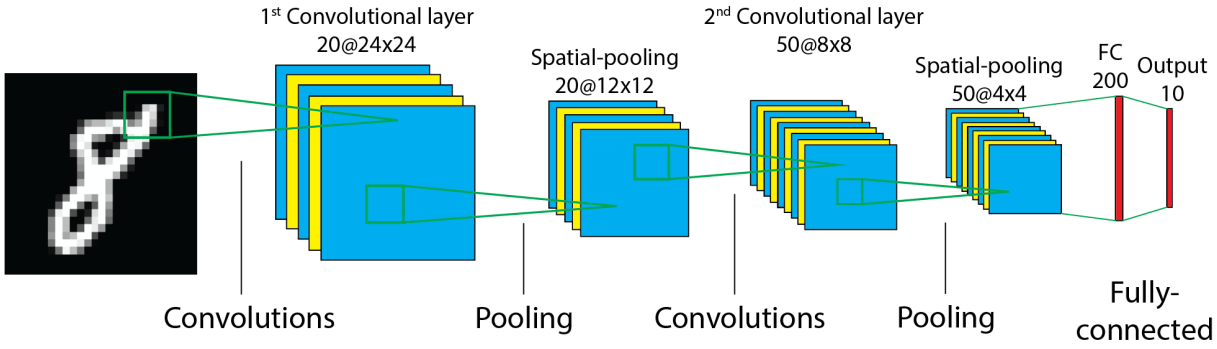

Figure 2.13: A spiking Convolutional Neural network

that the spiking activity of all output neurons except one is inhibited, i.e., only one output neuron can fire at a time. For example, in [49], a WTA SNN is proposed to classify handwritten images in the MNIST dataset. The network comprised 784 input neurons, one corresponding to each pixel in the input image. Each neuron in the input layer is fully connected to each of the 100 neurons in the excitatory layer. The third layer comprises 100 inhibitory neurons, each connected to the corresponding neuron in the excitatory layer. A spiking event at a neuron in the excitatory layer results in a spiking event at the corresponding neuron in the inhibitory layer. Each neuron in the inhibitory layer is connected to all the neurons in the excitatory layer via inhibitory synapses, except the one it receives a connection from. The weights of the inhibitory synapses are programmed such that a spiking event at an inhibitory neuron inhibits all the excitatory neurons except the excitatory neuron from which it receives a connection.

In hardware, a layer of pre-synaptic neurons can be efficiently connected to a layer of post-synaptic neurons by following a crossbar architecture with synapses present between neurons, as shown in Fig. 2.14. The crossbar architecture has the advantage that currents due to spiking activity in the pre-neuronal layer can

be weighted by the conductance stored in the synapse and then added at each post-neuron. This essentially means that memory (weight stored in synapse) and compute (multiply and accumulate) are collocated. In a traditional von-Neumann-based architecture, during each training cycle, weights would have to be fetched from memory, then updated by the processor, and finally stored back in memory, resulting not only in performance bottlenecks but also in additional power consumption. Thus, the crossbar architecture allows us to implement an SNN in an energy-efficient manner.

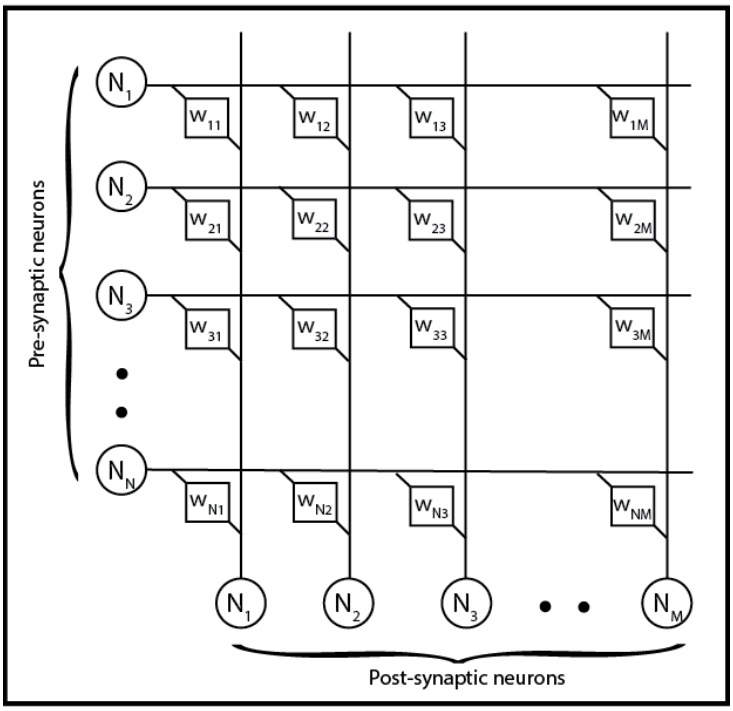

Figure 2.14: A crossbar architecture that can be employed to efficiently connect a layer of pre-synaptic neurons to a layer of post-synaptic neurons [35].

2.2.4 Learning Algorithms

The algorithms employed to train the SNN are broadly classified into two categories - supervised and unsupervised. The choice of a suitable algorithm to

train an SNN depends on the chosen models for neurons and synapses and the network topology employed. Another issue that dictates the choice of an algorithm is whether the training is performed on-line in an unsupervised manner (training necessarily on-chip), whether off-chip supervised training is necessary, or whether a combination of the two is necessary.

2.2.4.1 Supervised Learning

Supervised learning is the most popular learning methodology in Deep Artificial Neural Networks (ANNs). The most common supervised learning algorithm is that of the gradient descent where the Root Mean Square (RMS) error (calculated from the desired response and the observed response for a particular input pattern) is backpropagated all the way from the output layer to the input layer [50]. The weights of the connections are adjusted such that the RMS error is minimized. Thus, with each forward pass of the network, a backward pass is performed, which adjusts the weights of the connections such that the difference between the desired response and the observed response for a given input is minimized. This requires the computation of the gradient of the RMS error with respect to the weight of the connection. The backpropagation algorithm can be implemented in ANNs since the activation functions (Sigmoid, ReLU, etc.) employed are continuous and thus differentiable in nature. However, in an SNN, the computation of the gradient is not feasible since the output of a spiking neuron is a discrete action potential and not a continuous function, as in the case of an ANN. A summary of the supervised training algorithm used to train an

ANN is shown in Fig. 2.15.

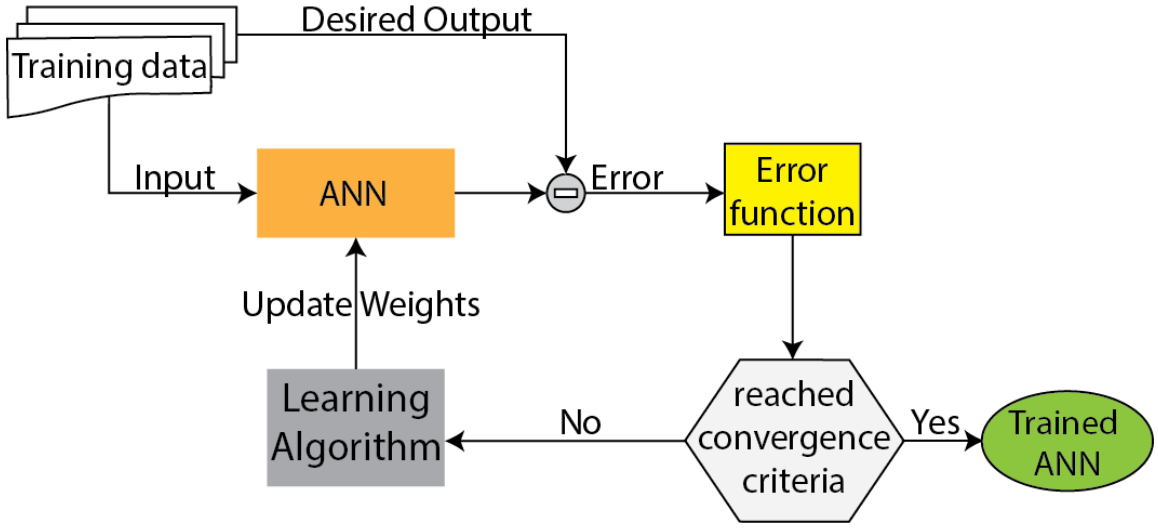

Figure 2.15: A summary of the supervised training algorithm used to train an ANN

There have been several approaches in the literature that involve training a traditional ANN and devising conversion algorithms that translate the ANN weights to equivalent spiking counterparts [13, 51, 52]. For example, in [13], a softened rate model is proposed where the hard threshold response of an LIF neuron is substituted with a continuous differentiable function to make it suitable for use in backpropagation. Subsequently, they trained the ANN with the rate model and transformed it into an SNN made up of LIF neurons. In all these prior conversion-based algorithms, the training is performed using continuous signals, which do not imbibe the temporal information embedded in spikes.

2.2.4.2 Unsupervised Learning

Due to the non-differentiable nature of the discrete action potentials in spiking neurons, the gradient-descent-based supervised learning techniques involving backpropagation cannot be employed to train SNNs. Hence, unsupervised

learning is the most popular learning rule in SNNs. Some early neuromorphic implementations of unsupervised learning involved some form of Hebbian learning rule and its derivatives. The Hebbian learning rule states that “neurons that fire together, wire together." This rule means that if a pre-synaptic neuron firing event results in a firing event at the post-synaptic neuron, then the strength of the connection between them should be increased because the relationship between firing events was causal. However, if the firing event at the pre-synaptic neuron follows the post-synaptic neuron firing event, the strength of the connection between them should be decreased because the relationship between firing events was anti-causal. Spike Timing Dependent Plasticity (STDP), the most widely used on-line unsupervised learning mechanism in neuromorphic systems, has recently become prominent [53, 55]. The STDP learning rule states that the change in the synapse’s strength depends exponentially on the temporal correlation of spiking events between the pre- and post-synaptic neurons, as shown in Fig. 2.16.

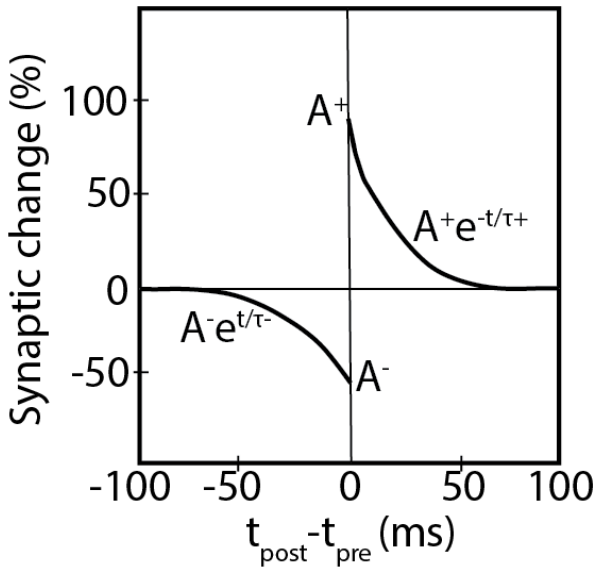

Figure 2.16: STDP [53]

If the pre (post)-synaptic neuron firing event is followed by the firing event at the post (pre)-synaptic neuron, the strength of the synapse connecting the two is increased (decreased). The smaller the time difference between spiking events, the larger the change in the synapse's strength.

In this work, the SNN is trained on-line in an unsupervised manner using the STDP learning rule. Training the network using the STDP learning rule is more biologically plausible than conversion-based supervised learning rules, which do not imbibe the temporal information embedded in discrete spikes.

Chapter 3

An Energy-efficient Leaky Integrate and Fire Neuron

An energy-efficient Ge-based Leaky Integrate and Fire (LIF) neuron is proposed in this chapter, and its analysis is conducted using a well-calibrated 2D simulation model. Direct reception of incoming voltage spikes by the proposed neuron prevents energy dissipation in generating a summed potential. This process involves the accumulation of holes in the channel due to the incoming voltage spikes, resulting in the reduction of the potential barrier and a subsequent increase in current. When the current reaches a predefined threshold, a firing, and subsequent reset circuitry are activated. The device's operation at a lower voltage level is facilitated by the smaller bandgap and the dominant direct tunneling of Ge. The energy consumed per spike in the proposed implementation amounts to 0.07fJ, a value lower than that of LIF neuron implementations (experimental or simulated) documented in the existing literature. Moreover, the power consumed by the reset circuitry can be decreased due to the lower drain voltage requirement

of the proposed device. The work done in this chapter is published in [54].

3.1 Device Structure and Simulation Model

Fig. 3.1 depicts the schematic cross-sectional view of the device employed to realize a LIF neuron. The gate oxide utilized in this study is SiO_2. Additional crucial parameters of the device are listed in Tab. 3.1.

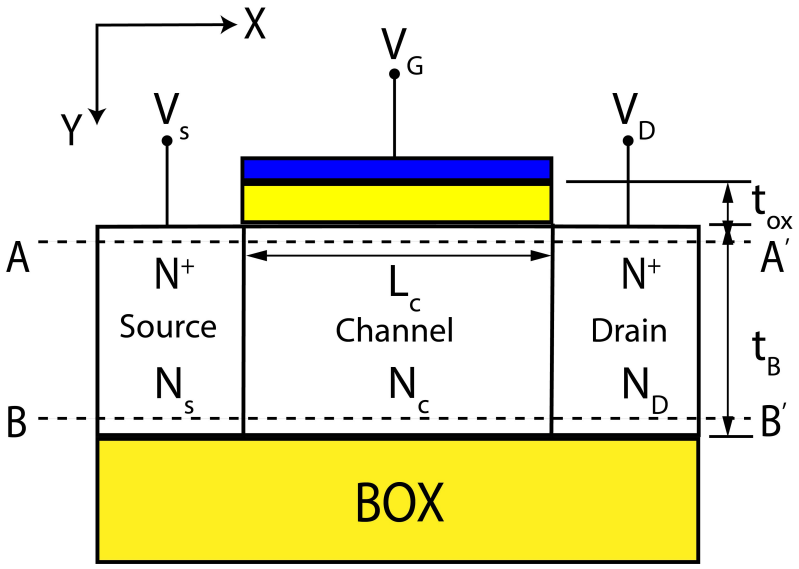

Figure 3.1: Structure of the proposed LIF neuron.

Table 3.1: Device Parameters of the proposed LIF neuron

Device Parameter	Symbol	Value
Drain Voltage (V)	V_D	0.35(Ge)/0.85(Si)
Source Voltage (V)	V_S	0.3
Substrate Voltage (V)	V_{sub}	0
Channel Length (nm)	L_C	50
Gate Oxide thickness (nm)	t_{ox}	3
Buried Oxide thickness (nm)	t_{BOX}	20
Channel thickness (nm)	t_B	20
Gate workfunction (eV)	ϕ_m	4.6(Ge)/4.8(Si)
Channel Doping (p-type) (cm^{-3})	N_C	1×10^{17}
Source Doping (n-type) (cm^{-3})	N_S	1×10^{20}
Drain Doping (n-type) (cm^{-3})	N_D	1×10^{20}

The simulations were carried out using Synopsys Sentaurus, version N-

2020.09-SP2 [56]. Germanium, due to its smaller bandgap and a dominant direct tunneling mechanism compared to Silicon, is used as a base material in this work [5]. This results in a larger BTBT generation rate. Thus, the LIF neuron can be operated at a smaller supply voltage and has the potential to deliver higher energy efficiency compared to Si-based devices. However, there are certain drawbacks associated with using Ge over Si. It has a higher leakage current, and the absence of a high-quality native oxide in Ge makes Si the go-to choice for designers [57]. However, due to the advent of high-k dielectrics, the interface between Ge and the dielectric is of high quality and free from defects. To suppress high leakage currents due to a lower bandgap of Ge, SiGe homojunction [58] or a Ge source heterojunction [59] may be employed to curb leakage, but they have a lower BTBT generation rate than pure Ge. Furthermore, Ge has a lower thermal stability compared to Si and, thus, cannot be operated at high temperatures. Since the device shall be employed in an energy-efficient SNN, the heat dissipated in the network is expected to be small. However, the application of such a device in an environment with high ambient temperature should be avoided, and in such conditions, SiGe can be the material of choice. The non-local BTBT model, employing fitting parameters from [5], was employed for simulations. The proposed device, a Ge-based PD-SOI MOSFET, utilizes BTBT as the mechanism for charge (hole) storage within the floating body of the device. Consequently, calibration of the BTBT parameters utilized in the non-local BTBT model was necessary. This calibration process involved simulating a Ge-based TFET (illustrated in the inset of Fig. 3.2) while concurrently

considering direct and indirect tunneling parameters, as suggested in ref. [5]. The comparison between simulation outcomes and the published results for the Ge-based device in ref. [5] is presented in Fig. 3.2.

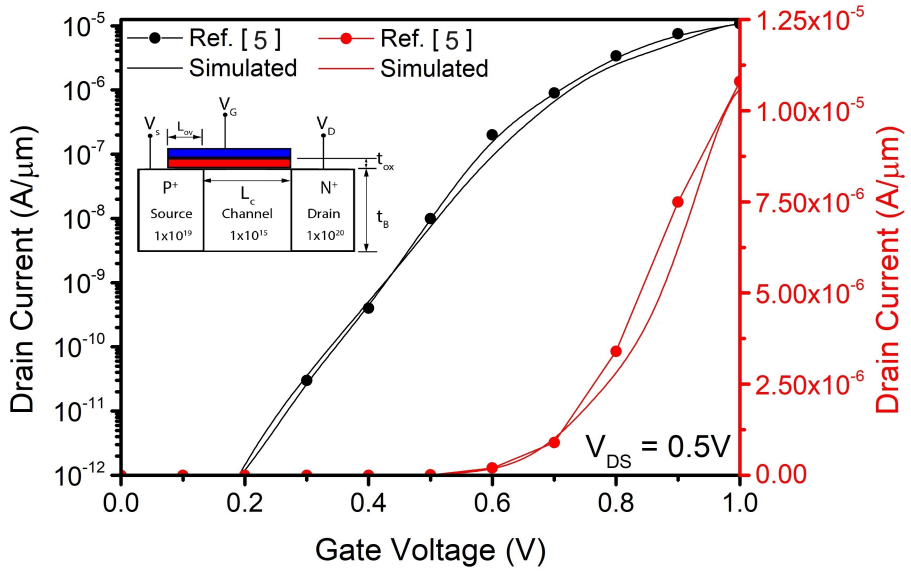

Figure 3.2: Comparison of the results produced by the simulation model and published results for the Ge device in ref. [5]. The device structure used in our simulation and ref. [5] is shown in the inset.

A close agreement between these results serves as a confirmation of the appropriateness of the calibrated BTBT model. Shockley-Read-Hall (SRH) recombination model has been accounted for. Additionally, the Slotboom Band-Gap Narrowing (BGN) model has been incorporated to accommodate the impact of highly doped source and pocket regions. Moreover, a concentration-dependent Philips unified mobility model has been activated. Tunneling through the gate oxide has been neglected [60, 61].

3.2 LIF neuron characteristics

The operational principle of the LIF neuron is explained in this section. Fig. 3.3 provides comprehensive band diagrams along the cutline AA', presenting the carrier movement within the device during distinct operational phases: leaky integration, firing, and reset. At equilibrium, when no voltages are applied, the band diagram of the device is depicted in Fig. 3.3(a). A constant voltage of 0.3 V is applied at the source to create a potential barrier for electrons at the source and prevent them from reaching the drain. Moreover, a voltage of 0.35 V is applied at the drain to facilitate leaky integration (Fig. 3.3(b)). This arrangement ensures minimal BTBT at the drain-channel interface in the absence of any incoming spike. Throughout the integration phase, constant voltages are applied at the drain and source terminals. The occurrence of an incoming spike from the pre-neuronal layer (-0.7V and 1ns duration) at the gate triggers BTBT at both the drain-channel and source-channel interfaces (illustrated in Fig. 3.3(c)). This BTBT process leads to the creation of vacancies (holes) in the channel as electrons undergo BTBT, consequently inducing the generation of electron-hole pairs (EHP) [56].

For demonstration purposes, the spikes are presented in a deterministic pattern with arrivals every few nanoseconds at the gate. In practical scenarios, however, the spikes are expected to arrive randomly. Fig. 3.3(d) illustrates the evolution of the band diagram with time in response to incoming spikes. The 2D contour depicting hole density (cm^{-3}) within the device illustrates the accumulation

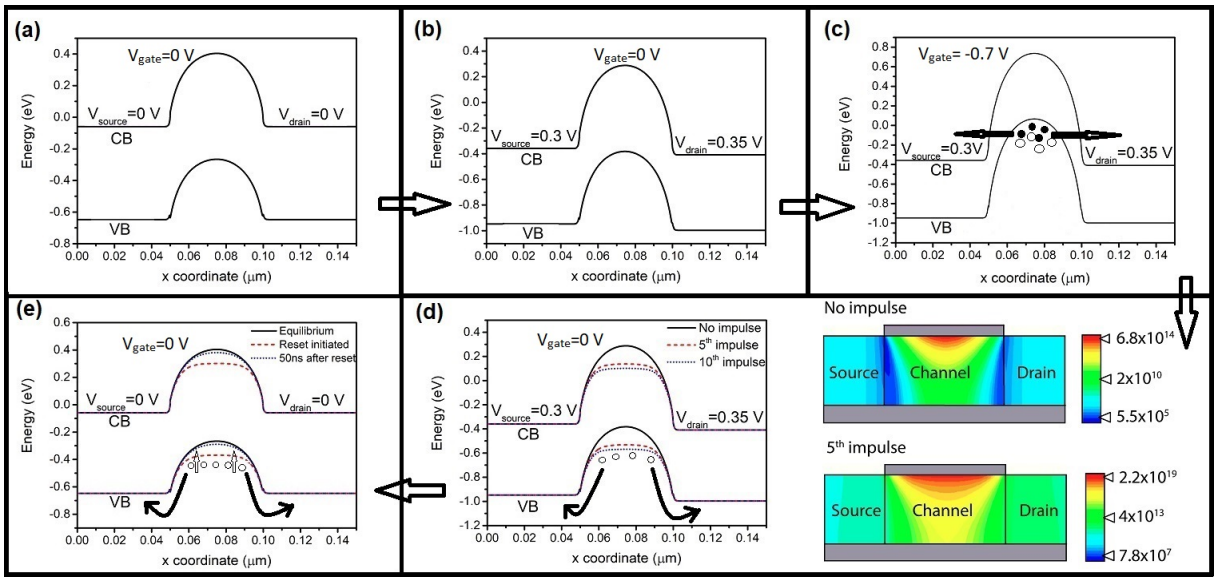

Figure 3.3: (a) LIF neuron band diagram at equilibrium, (b) LIF neuron band diagram with biasing applied to enable the integration phase, (c) LIF neuron band diagram during application of a voltage spike causing BTBT of electrons at the drain-channel and source-channel junctions leaving behind vacancies (holes) in the channel, (d) Evolution of the LIF neuron band diagram with incoming spikes and 2D contour of hole concentration (cm^{-3}) demonstrating accumulation of holes in the channel with the incoming spikes, and (e) LIF neuron band diagram after the reset event demonstrating leakage of holes into the drain and source regions, thereby causing the neuron to return to its equilibrium state.

of holes within the channel due to incoming spikes. This accumulation of holes leads to a reduction in the potential barrier encountered by electrons at the source, consequently yielding an increase in current. Simultaneously, the accumulated holes in the channel leak into the source and drain regions due to thermionic emission. In the absence of incoming spikes, leakage of accumulated holes results in an increase in the potential barrier seen by the electrons in the source. This correlates to a reduction in current as shown in Fig. 3.4, demonstrating leakage. Also, in the absence of incoming spikes, the neuron retains the stored charge for a notable duration (on the order of tens of μs). This attribute establishes the proposed LIF neuron as a reliable candidate for neuromorphic applications.

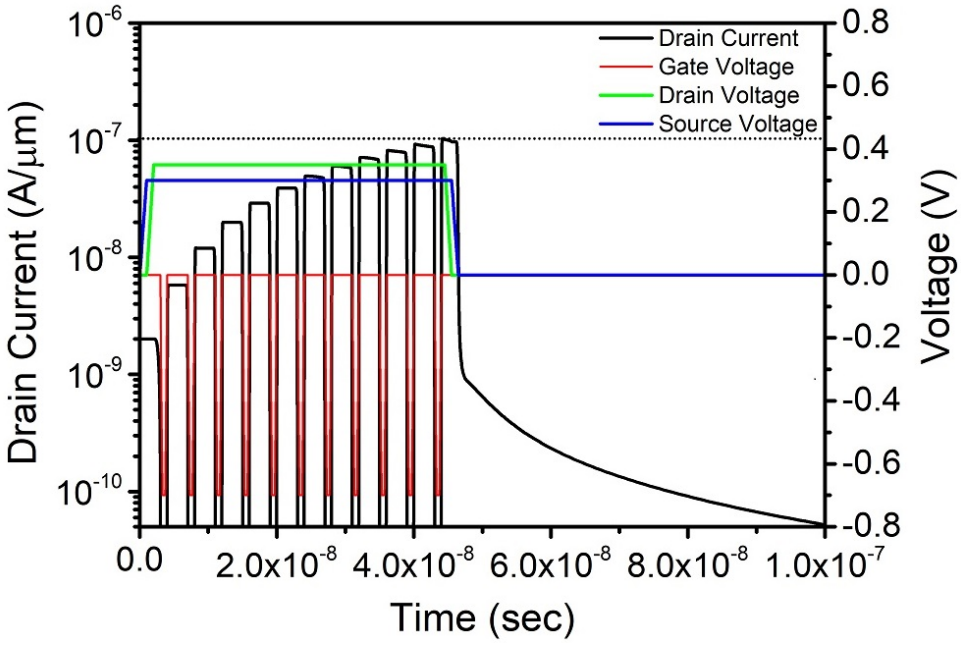

Figure 3.4: LIF neuron characteristics demonstrating rise in current with incoming spikes

At equilibrium, the rate of accumulation of holes equals the rate of leakage of accumulated holes. At this stage, the current through the device reaches a threshold ($I_{th} = 10^{-7} A/\mu m$), and the neuron fires a spike. Thereafter, the neuron is reset by removing the voltages applied on the source/drain terminals. This reduces the potential barrier for the accumulated holes, and they leak away quickly into the source and drain regions (as can be seen from Fig. 3.3(e)). Thus, the neuron is rapidly restored to its equilibrium state in a short period (about 50ns). This phenomenon is associated with a steep drop in current, as depicted in Fig. 3.4.

It should be noted that Fig. 3.4 considers deterministic occurrence of spikes. In reality, however, the occurrence of spikes shall be random. Fig. 3.5 shows the evolution of the drain current through the device with random occurance of spikes. The maximum number of spikes that the LIF neuron can handle

before reaching threshold will depend on the spiking activity in the pre-synaptic layer of neurons. Since the rate of spiking activity in Fig. 3.4 was large the neuron reached the the threshold in fewer number of spikes. However, if the frequency of spiking activity in the pre-synaptic layer of neurons is low, the LIF neuron will require a larger number of spikes to reach the threshold than the case when the frequency of spiking activity in the pre-synaptic layer of neurons is high. This is because, in the absence of spiking activity in the pre-synaptic layer of neurons, the accumulated holes in the channel will leak away with time and greater number of spikes shall be required for the LIF neuron to reach its threshold.

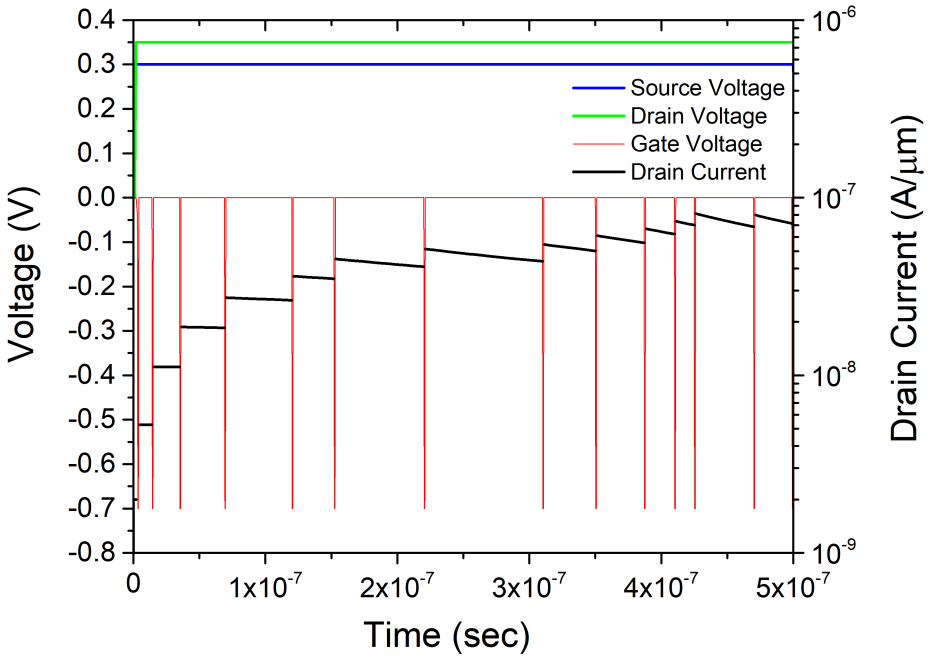

Figure 3.5: LIF neuron characteristics demonstrating rise in current with random spikes

The architecture of the proposed LIF neuron is depicted in Fig. 3.6. It consists of a control circuitry, which is responsible for identifying when the drain current

attains the threshold, I_{th}. The input to the neuron is voltage spike, which in turn results in a flow of current through the device. This current is sensed by the control circuitry and does not propagate to the downstream neurons. Once the threshold current is reached for a specific neuron (let's say N_{11}), the control circuitry generates a voltage signal V_{11} lasting for a duration of t_{spike} (spike duration). This voltage signal is then applied to a high threshold voltage pass transistor (M_{11}), resulting in the generation of a current spike. When the current remains below I_{th}, the gate of such a pass transistor is held at V_{SS}. Several such current spikes (if they are generated simultaneously) are weighted and summed for the neurons in the pre-neuronal layer at a particular time step for a particular post-neuron. The resulting current spike is transformed into a voltage spike by utilizing a transistor (L_1 in Fig. 3.6) operating in the triode region. Following the firing of neuron N_{11}, the control circuitry generates a reset signal. This signal serves to eliminate the voltages applied to the source ($V_{S,11}$) and drain ($V_{D,11}$) terminals of the device, thereby eliminating the accumulated holes within the channel. This action enables the neuron to return to the equilibrium, as depicted in Fig. 3.3(e). The inclusion of a reset circuitry is essential for other neurons proposed in the existing literature as well [16–20]. The operation of the reset circuitry depends on the voltages applied to the drain/source terminals. In contrast to other LIF neurons, such as II-based neurons that operate at around 3 V, the proposed neuron is capable of operating at a lower voltage of 0.7 V. Consequently, a reduction in power dissipation within the reset circuitry is anticipated in the proposed implementation.

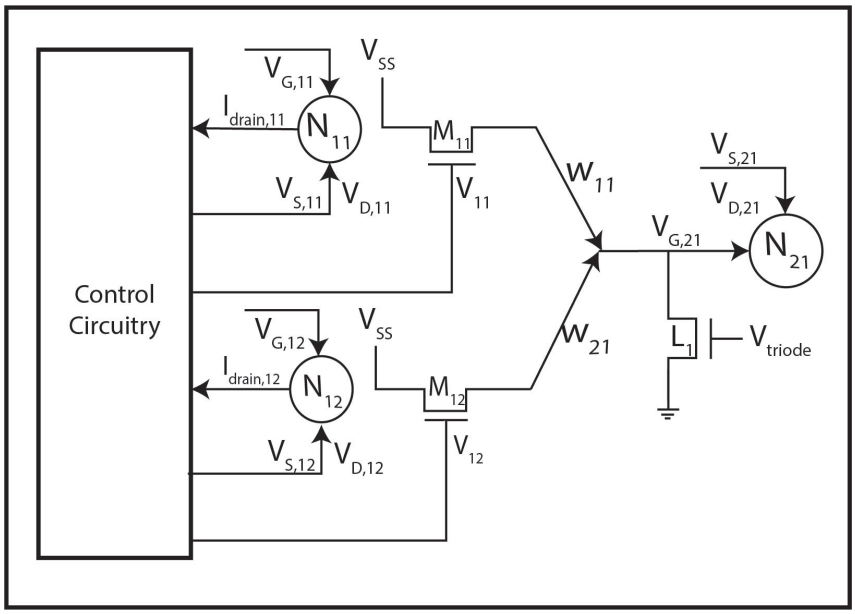

Figure 3.6: Architecture of the proposed LIF neuron

The architecture diagram displayed in Fig. 3.6 shows a control circuitry which is shared by a group of neurons. If we employ a control circuit for every neuron, then its area and energy consumption would blow up out of proportion and such a configuration would not be optimal for a dense and energy-efficient SNN architecture. Thus, sharing of the control circuitry among a group of neurons seems to be the best way forward. Since, the application does not demand to be operated at very high frequencies, we can share the control circuitry for a group of neurons in a particular layer and sense whether a particular neuron in that group has reached its threshold or not at a given frequency. The fact that the frequency of spiking activity reduces in deeper layers in the SNN can be leveraged to save power consumed by the control circuitry for a group of neurons present deeper in the network. For instance, in the input layer if the spiking activity is 50 Hz, then a group of 20 neurons can be formed and control circuitry be shared among them, which senses the threshold current for each neuron at,

say 10KHz. If there are a 1000 neurons in the input layer, 50 such groups shall be formed (saving power at the cost of area). Further, suppose a layer deep inside the network has a spiking activity of 5 Hz, then a group of 200 neurons may be formed and control circuitry be shared among them, which senses the threshold current for each neuron at 10KHz. If there are a 1000 neurons in this layer, only 5 such groups shall be formed (saving area at the cost of power). Thus, there is a direct trade-off between area and power consumed by the control circuitry and such an architectural choice needs to be made on the basis of actual overhead in terms of power, performance and area consumed, and which parameter is more critical for the desired application.

The aforementioned process is repeated after a specific duration known as the refractory period, during which the neuron remains in its equilibrium state. In this refractory period, if the neuron receives a spike, there is minimal band overlap at the drain-channel and source-channel junctions, as depicted in Fig. 3.7(a). Consequently, the BTBT generation rate is low, resulting in negligible hole accumulation within the channel, as illustrated in Fig. 3.7(b). Any accumulated holes disperse rapidly due to the comparatively shorter potential barrier than when drain/source voltages were applied.

The minimal refractory period corresponds to the time needed for the neuron to restore its equilibrium state after firing, a value dictated by the device's characteristics. For the proposed neuron, the minimum refractory period was determined to be approximately 50 ns. The control circuitry can generate suitable biases (V_S, V_D) with a refractory period surpassing the defined minimum

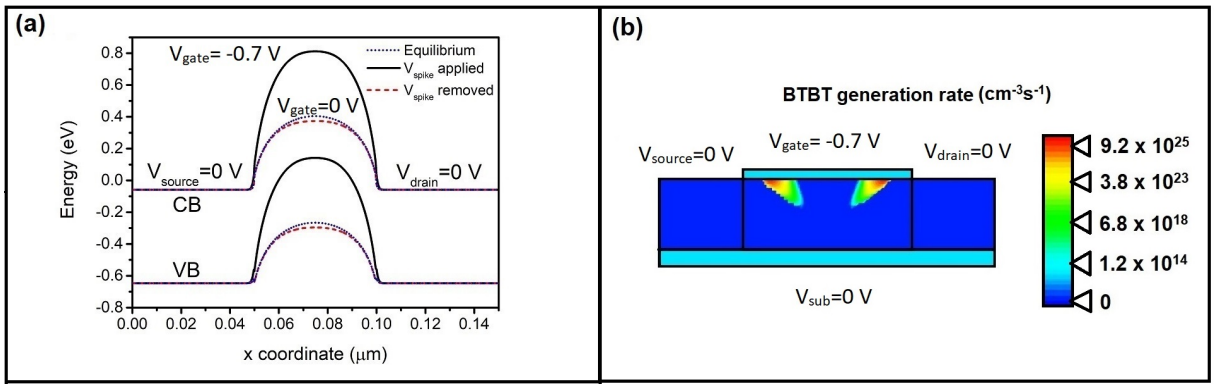

Figure 3.7: (a) Band diagram along cutline AA' and (b) 2D BTBT generation contour ($cm^{-3}s^{-1}$) demonstrating negligible BTBT when a spike is incident during the refractory period of the neuron

refractory period for a given neuron. Moreover, an optimal refractory period for neurons will depend on the specific application and will be determined by design objectives such as accuracy, functionality, energy efficiency, and other system-level considerations. Following the refractory period, when the neuron is prepared to process additional spikes, the control circuitry applies appropriate voltages to the source and drain terminals, initiating a new LIF cycle.

In this study, the transient simulations have been performed using the device simulator Synopsys Sentaurus. For small-sized circuits, mixed-mode simulations can be performed in Sentaurus. Additionally, circuit simulations can be conducted in SPICE by constructing a model for the proposed LIF neuron. Given that the device utilizes BTBT to generate holes, it is imperative to incorporate a model for the BTBT generation rate. Prior research has introduced compact SPICE models for Tunnel FETs [62–64]. The proposed LIF neuron can also be appropriately modeled in SPICE using a Verilog-A model. This process involves extracting the device's Current-Voltage (I-V) characteristics (both transfer and output) and parasitics from the device simulator. Subsequently, these

extracted data can be compiled into a look-up table format for incorporation into the Verilog-A model. This approach allows the SPICE simulation platform to emulate the behavior of the proposed LIF neuron.

3.3 Device Optimization

This section analyzes various parameters pertaining to the Ge-based LIF neuron and their influence on the characteristics of the neuron. Specifically, the consequences of altering the channel thickness and the gate oxide thickness on the attributes of the LIF neuron are explored.

3.3.1 Channel Thickness

LIF neuron characteristics strongly depend on the choice of channel thickness. As the channel's thickness increases, the channel's capacity to store charge increases. With incoming spikes, holes get accumulated in the channel. As a result, there is an increase in current. At the steady state, the rate of accumulation of holes with incoming spikes equals the rate of leakage of holes. Consequently, with incoming spikes, a negligible increase in current is observed at the steady state. This trend is evident in Fig. 3.8, where it can be observed that more spikes are required to attain the steady state for a channel thickness of 20 nm in comparison to thicknesses of 15 nm and 10 nm, respectively. This observation signifies that the ability to store holes is greater for a channel thickness of 20 nm compared to thicknesses of 15 nm and 10 nm.

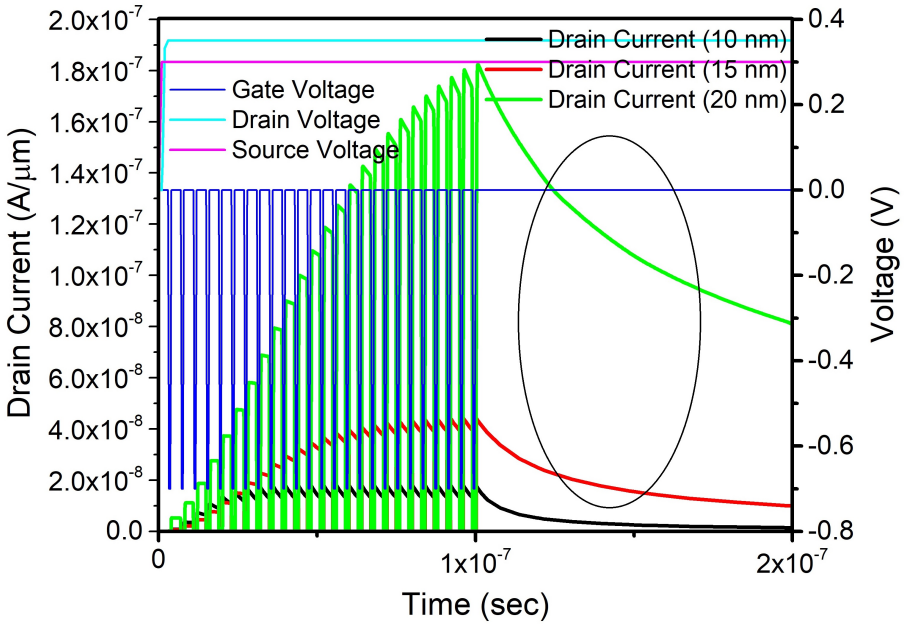

Figure 3.8: Plot of Drain current for different channel thicknesses

A channel with a thickness of 10 nm is fully depleted, making it challenging to store charge effectively in such a thin structure. As a result, a thicker channel that is only partially depleted is chosen for the LIF neuron, enabling it to better retain stored charge for an extended duration.

In the absence of incoming spikes, the accumulated holes gradually dissipate into the source and drain regions due to thermionic emission over the potential barrier. This process is indicated by a decrease in current, as demonstrated in Fig. 3.8. As the holes continue to leak over time, the potential barrier gradually increases, leading to a reduction in the rate of hole leakage. Consequently, the rate at which the current declines also decreases with time. Fig. 3.9 presents band diagrams for a channel thickness of 20 nm at various time intervals (after achieving steady-state) in the absence of incoming spikes. Notably, the potential barrier exhibits a rapid initial increase, followed by a deceleration in the rate

of increase over time. Furthermore, for a thicker channel, the stored holes will remain within the channel for a longer duration (i.e., a longer retention time). For instance, with a channel thickness of 10 nm, the retention time is approximately 200 ns, while for a thickness of 20 nm, it extends to around $17.4\mu s$. The retention time is measured as the time taken by the current to reduce to its value on reset from its value at threshold.

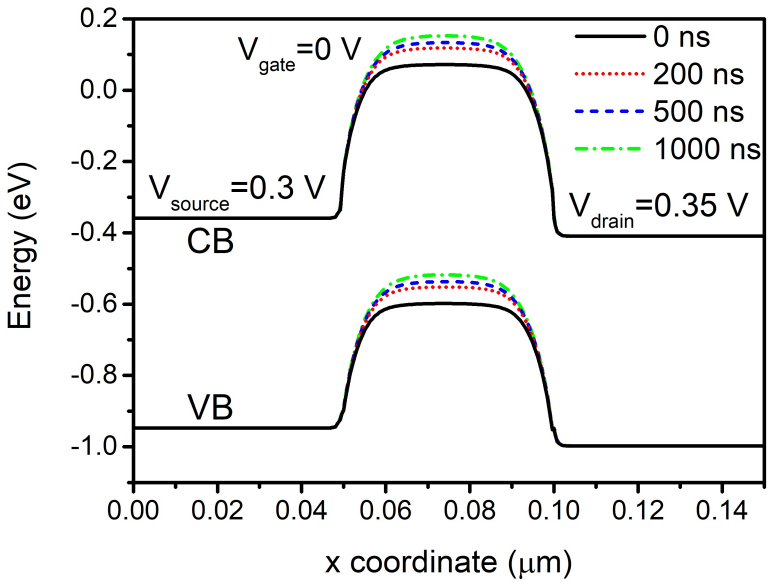

Figure 3.9: Band diagrams for 20 nm channel thickness along cutline AA' at different time instances after reaching the steady state in the absence of incoming spikes.

Nevertheless, it is important to note that a thicker device will result in a higher average current flow. Consequently, the threshold current will be significantly higher for a thicker channel, reaching approximately $1 \times 10^{-6} A/\mu m$ for a 30 nm channel thickness. Consequently, the energy consumption per spike (E_{spike}) is likewise elevated compared to a narrower channel device. To mitigate this, opting for a narrower channel becomes crucial in order to reduce E_{spike}. Selecting a 20 nm channel thickness balances between a high retention time for stored holes

and maintaining E_{spike} at acceptable levels.

It is important to consider that the presence of a floating body in the PD-SOI MOSFET can potentially lead to phenomena like the kink effect or history effect, particularly when Impact Ionization (II) phenomena occurs in the device [65]. This condition might also contribute to increased subthreshold slope, heightened parasitic effects, and elevated power dissipation, generally surpassing what is observed in FD-SOI MOSFETs. However, since the proposed device relies on BTBT for its charge storage mechanism rather than II, it is not anticipated to exhibit the kink effect. It is worth noting that while a PD-SOI MOSFET might introduce some performance loss, this trade-off can be acceptable within the context of the proposed application. This is due to the fact that the PD-SOI structure allows for the storage of charge in the floating body for a longer period of time. This extended storage duration prevents rapid leakage of accumulated charge, resulting in significant memory retention for the neuron. Furthermore, the potential loss in performance can be tolerated within this application, as it is not a stringent requirement to operate the neural network at extremely high frequencies. This aligns well with the intended functionality of the proposed neuron and its application.

3.3.2 Gate dielectric thickness

A gate dielectric layer (SiO_2) with a thickness of 3 nm is used in this work. Similar to previous research work [60, 61, 66–68], tunneling through the gate dielectric has been neglected. To further curtail gate leakage, we investigate the

impact of employing a thicker gate dielectric on the characteristics of the LIF neuron.

As the gate dielectric thickness increases, the degree of gate control over the channel diminishes. Fig. 3.10(a) illustrates the band diagram along cutline AA' when a spike is applied to the gate of the LIF neuron, considering various gate dielectric thicknesses. Noticeably, as the gate dielectric thickness increases, the overlap of the valence band in the channel with the conduction bands of the drain and source decreases. This translates to an exponential reduction in the BTBT generation rate, as depicted in the 2D BTBT generation rate contours in Fig. 3.10(b). This leads to an increased requirement on the number of spikes required to reach the threshold current, thereby resulting in a reduction in the output firing frequency for a given input firing frequency.

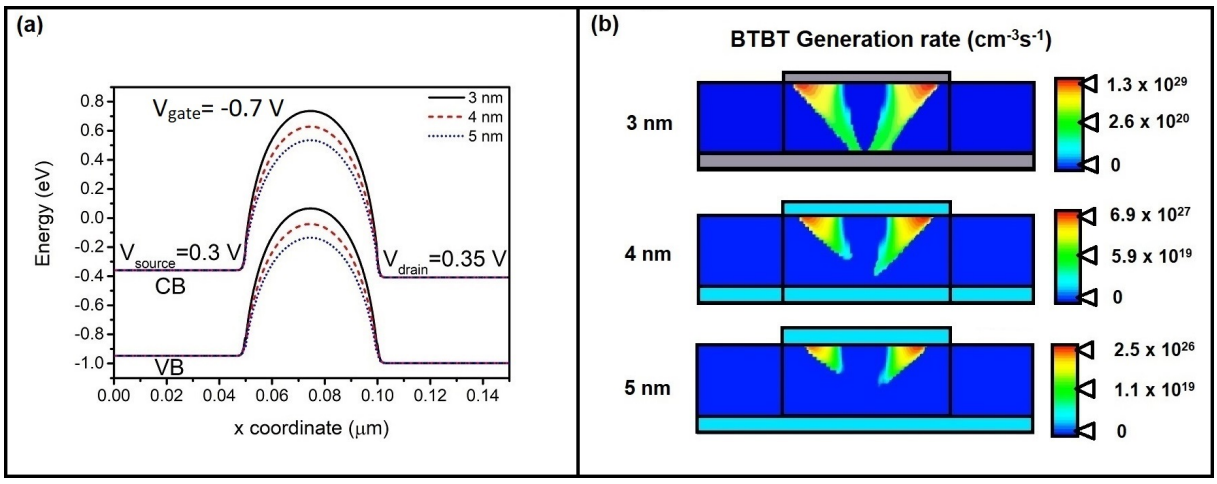

Figure 3.10: (a) Band diagram along cutline AA' and (b) 2D BTBT generation ($cm^{-3}s^{-1}$) contours when a spike is incident on the neuron for different gate dielectric thickness

To sustain an equivalent output firing frequency with a thicker dielectric, a larger V_{spike} must be applied. This, however, leads to a an increased energy consumption per spike. Hence, for the implementation of a low-power LIF

neuron, a smaller gate dielectric thickness is desirable. However, it is crucial to acknowledge that a reduction in gate dielectric thickness can potentially result in an elevation of gate leakage current. Consequently, a slight increase in the static power dissipation of the LIF neuron can be expected. Nonetheless, this gate leakage can be effectively mitigated by employing a thicker high-κ dielectric like HfO_2, which simultaneously suppresses gate leakage and maintains performance.

3.4 Energy efficiency

In this section, we explain the superior energy efficiency exhibited by a Ge-based LIF neuron in comparison to II-based or BTBT-based Si LIF neurons. The energy consumption per spike (E_{spike}) in an LIF neuron can be defined as follows:

$$E_{spike} = V_{spike} \times I_{th} \times t_{spike} \tag{3.1}$$

where V_{spike} is the magnitude of the incoming spike, I_{th} is the threshold current at which the firing circuitry is triggered, and t_{spike} is the time duration of the spike. Tab. 3.2 compares the energy consumption per spike for the proposed implementation with other implementations in the literature. In the proposed implementation, the energy consumption per spike is measured as 0.07fJ, a value notably lower than that of all LIF neuron implementations (whether experimental or simulated) documented in the existing literature.

It should be noted that the energy consumption of 0.07 fJ per spiking event corresponds to 20 nm channel thickness. Since the ability of a thinner channel to store charge is lesser than that of a thicker channel, a neuron with a thinner channel has a smaller I_{th} than a neuron with a thicker channel. Consequently, the energy consumption per spike for 10 nm channel (with I_{th}=$1.5 \times 10^{-8} A/\mu m$) thickness is 0.01 fJ per spiking event, for a 15 nm channel (with I_{th}=$4 \times 10^{-8} A/\mu m$) is $\sim$0.03 fJ per spiking event and that for 20 nm channel thickness (with I_{th}=$1 \times 10^{-7}\ A/\mu m$) is $\sim$0.07 fJ per spiking event. A thicker channel can store the charge for a longer period of time (retention time $\sim$20 μs). A suitable choice of channel thickness could be application dependent and a channel thicker than 20 nm can be chosen if a larger retention time is desired at the cost of high energy consumption per spike.

The duration of spiking activity (t_{spike}) plays a big role in the high energy-efficiency of the proposed LIF neuron. The proposed LIF neuron is capable of operating with t_{spike} in the order of $\sim$ns due to the dominant direct tunneling mechanism in Ge, which allows the device to be operated at low voltages. The frequency of spiking activity considered in this work is of the order of $\sim$MHz and operating the neuron at such a high frequency can result in a higher power dissipation in the peripheral circuitry of the neuron. The same circuit, with a few modifications, can be used to implement an LIF neuron where the operating frequency is more bio-realistic. Currently, the LIF neuron is implemented with a Ge-based PD-SOI MOSFET with channel thickness of 20 nm. The neuron can accept spikes in a timescale of a few ns and has a retention time of $\sim$20μs. If

the neuron is to be operated at bio-realistic frequencies, it must have a larger retention time in accordance with the frequency it is operated at. A thicker channel shall be required to increase the retention time of the neuron and bring it closer to bio-realistic numbers.

Analyzing Tab. 3.2, it becomes evident that II-based LIF neuron implementations [16–19] consume more energy when compared to BTBT-based implementations. This observation aligns with expectations, as II-based implementations necessitate a higher voltage to induce impact ionization within the device. Additionally, these implementations exhibit significantly larger threshold currents (I_{th}) than their BTBT-based counterparts.

Table 3.2: Comparison of energy consumption per spike for the proposed implementation with the state-of-the-art

Reference	Device Type	V_{spike} (V)	I_{th} (A)	t_{spike} (ns)	E_{spike} (fJ)	E_{add} (fJ)	E_{total} (fJ)
[14]	PCMO RRAM	–	–	–	4.8×10^3	–	4.8×10^3
[15]	CMOS	–	–	–	9×10^5	–	9×10^5
[16]	PD-SOI MOSFET (II)	2.8	0.5×10^{-3}	25	3.5×10^4	2.23×10^3	3.72×10^4
[17]	Bulk FinFET (II)	3	0.35×10^{-6}	6	6.3	0.96	7.26
[18]	JL-FET neuron (II)	0.4	0.5×10^{-3}	5.7	1.14×10^3	0.58×10^3	1.72×10^3
[19]	Si NIPIN diode (II)	0.8	2.4×10^{-3}	62.5	1.2×10^5	0.3×10^5	1.5×10^5
[20]	Si PD-SOI MOSFET (BTBT)	1.5	8×10^{-6}	0.26	3.2	1.4	4.6
Proposed neuron	Ge PD-SOI MOSFET (BTBT) Si PD-SOI MOSFET (BTBT)	0.7 2.0	1×10^{-7}	1	0.07 0.2	–	0.07 0.2

First, we analyze why a BTBT-based Ge LIF neuron demonstrates superior energy efficiency compared to a BTBT-based Si LIF neuron [20]. For the sake of facilitating an effective comparison, I_{th} and t_{spike} have identical values across different materials. The BTBT generation rate in a Ge-based device is significantly higher due to its dominant direct tunneling mechanism and a smaller

band gap. Consequently, with incoming spikes, there is a rapid accumulation of holes within the channel. This leads to the attainment of the threshold current using a smaller V_{spike}. In a BTBT-based Si LIF neuron, a larger band overlap is necessary (manifested as a larger V_{spike}) to balance the BTBT generation rate in a BTBT-based Ge LIF neuron. This requirement for a larger V_{spike} results in sub-optimal energy efficiency in a BTBT-based Si LIF neuron.

Fig. 3.11 illustrates the band diagram of the LIF neuron both with and without spike voltage at the gate, alongside the corresponding 2D band-to-band generation rate contours for both Si and Ge-based devices. Notably, in the Si case, a substantially higher V_{spike} of -2V is employed as compared to -0.7V for Ge. This adjustment in voltage is necessary to balance the elevated BTBT generation rate in Ge-based devices. Additionally, the band profile in pure Ge device appears more symmetrical. This symmetry promotes BTBT at both the drain-channel and source-channel interfaces, thereby facilitating a swifter accumulation of holes in the channel when compared to a Si-based device.

Furthermore, the examination of Fig. 3.11 reveals that a higher drain voltage (0.85V) must be applied in the case of Si, in contrast to 0.35V for Ge, to facilitate leaky integration. This difference in drain voltage requirements leads to a reduction in the power consumed by the reset circuitry in Ge-based LIF neuron implementations compared to their Si-based counterparts. This serves as an additional contributing factor to the improved energy efficiency exhibited by the Ge-based BTBT LIF neuron.

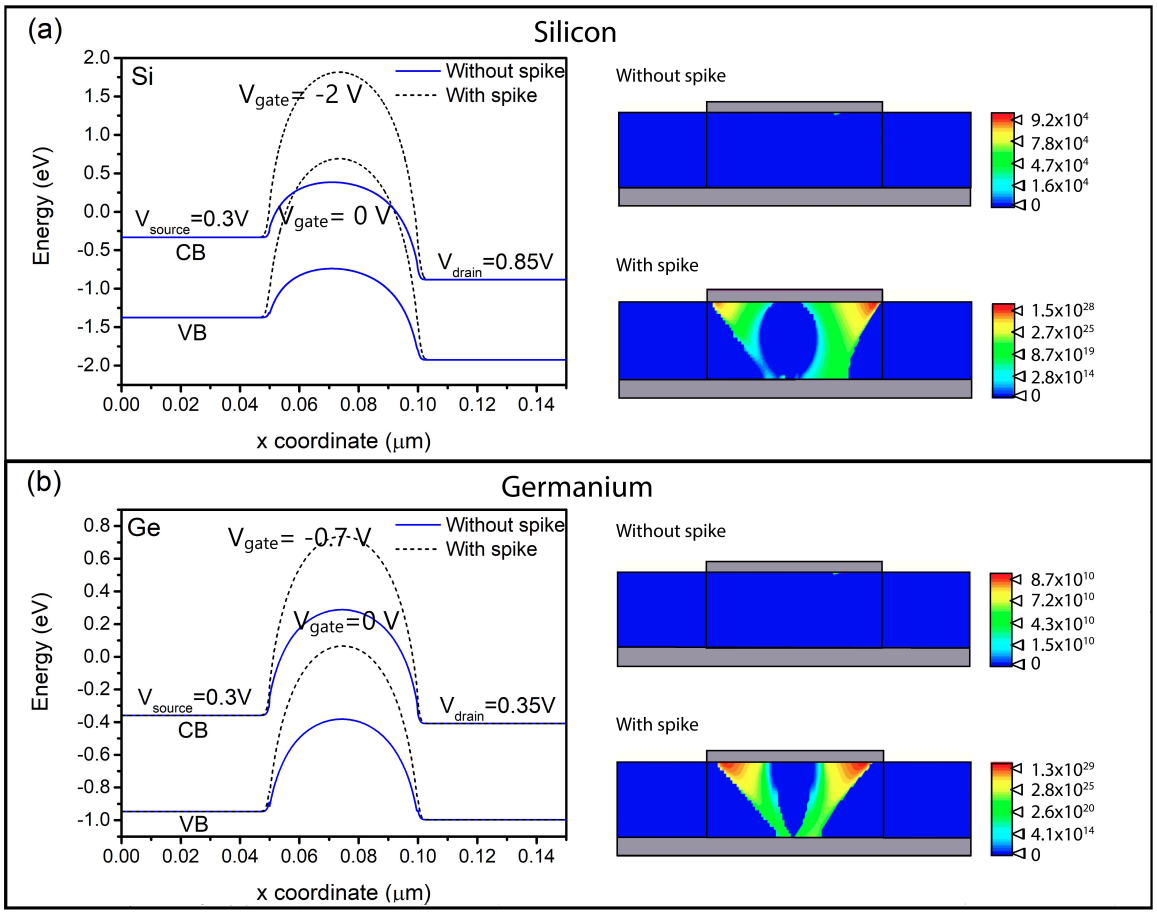

Figure 3.11: LIF neuron band diagram with and without the presence of spike voltage at the gate and the corresponding 2D band-to-band generation rate ($cm^{-3}s^{-1}$) contours for (a) Si and (b) Ge-based device

Through SPICE simulations, an assessment of the energy dissipated per spike (E_{add}) was conducted in the charging and discharging process of the capacitor within the fixed voltage generating summer circuit, illustrated in Fig. 2.7. The key attributes of the spike (V_{spike}, I_{th}, and t_{spike}) have been obtained from the relevant studies [16–20]. By selecting appropriate resistance (R) and capacitance (C) values, we have ensured that an increment in potential on the order of millivolts ($\sim$mV) is achieved across the capacitor in response to an input current spike. The fixed number of spikes needed to generate a specific voltage, as indicated in the respective studies, is represented as N_{spike}. Using SPICE simulations, the

energy dissipated in the resistor-capacitor over N_{spike} spikes has been computed, and the resulting average energy dissipated per spike (E_{add}) is shown in Tab. 3.2. A noteworthy observation is that E_{add} contributes to roughly 10-50% of the energy consumption reported in the literature. It is essential to recognize and incorporate this component of power dissipation when evaluating overall energy consumption in the implementation of LIF neurons.

3.5 Dynamic response

The impulse parameters (V_{spike}, I_{th}, t_{spike}, and frequency of pre-neuronal spiking activity (f_{in})) play a significant role in influencing the output firing frequency (f_{out}) of a given LIF neuron. However, predicting the exact dynamic response of the proposed LIF neuron can be challenging. In a real-world system-level implementation, where multiple LIF neurons are interconnected via synapses, the voltage spike incident on a specific LIF neuron would typically be scaled down by the synapse's weight. Furthermore, the precise timing of spike arrivals is often uncertain. Initial layers within the network might experience higher spiking activity than deeper layers. Consequently, the achievable spiking frequency can greatly vary based on the specific application and the architecture of the SNN, particularly the number of layers involved. For the purpose of illustration, we have selected the following values for the impulse parameters ($V_{spike} = -0.7V$, $I_{th} = 1 \times 10^{-7} A/\mu m$, $t_{spike} = 1ns$, and $f_{in} = 250MHz$). Using these values, we have explored the behavior of f_{out} by individually varying each parameter while keeping the others constant. It is important to note that these values are

chosen for illustrative purposes and may not correspond to real-world scenarios due to the aforementioned complexities in actual network implementations and varying application demands.

3.5.1 Magnitude of incoming spike

If the magnitude of the incoming spike (V_{spike}) is increased, meaning a more negative value, it would result in a greater band overlap of the channel with the source/drain. This occurrence subsequently results in a larger BTBT generation rate and an accelerated accumulation of holes within the channel. Consequently, the output firing frequency (f_{out}) would exhibit an exponential increase as V_{spike} grows. This behavior is driven by the exponential relationship between the BTBT generation rate and the band overlap. Fig. 3.12 illustrates how f_{out} changes for various V_{spike} values. The resultant curve appears linear on this logarithmic plot due to the exponential nature of the relationship between the parameters. These results demonstrate the strong impact of V_{spike} on the LIF neuron's output firing frequency.

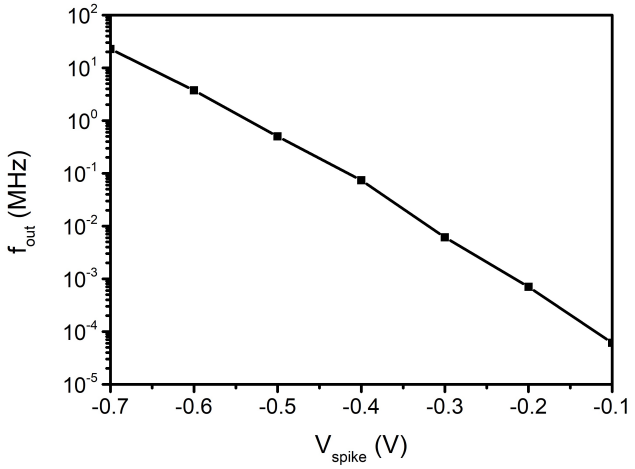

Figure 3.12: Output firing frequency for different magnitudes of incoming spikes

3.5.2 Time duration of incoming spike

Fig. 3.13, illustrates how the output firing frequency (f_{out}) varies for different values of t_{spike}. When the duration of the incoming spike increases, the time available for electrons to tunnel also increases, ultimately resulting in a more substantial accumulation of holes within the channel. This phenomenon results in a linear increase in the output spiking frequency as the spike's width increases.

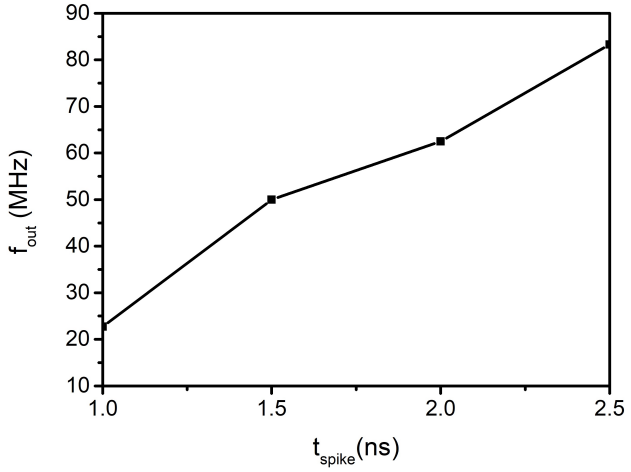

Figure 3.13: Output firing frequency for different t_{spike}

3.5.3 Threshold current

As I_{th} increases, it is expected that the output firing frequency (f_{out}) will decrease. This expectation arises from the fact that a larger number of spikes would be necessary to significantly reduce the potential barrier, and generate a larger current. This increment in spike count leads to a decrease in f_{out}. When a larger I_{th} is chosen, the rate of leakage will be considerably higher as the output current approaches the vicinity of I_{th}. This increased leakage is attributed to the presence of a smaller barrier that facilitates the rapid leakage of holes into the source and

drain regions. Fig. 3.14 demonstrates a decline in f_{out} as I_{th} rises.

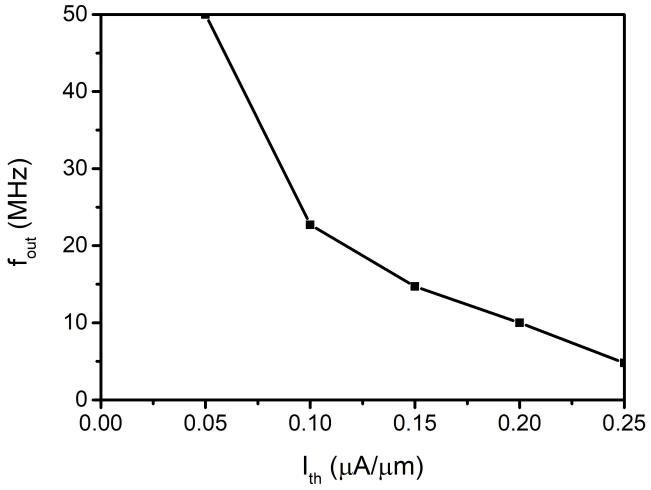

Figure 3.14: Output firing frequency for different I_{th}

3.5.4 Input firing frequency

An increase in the frequency of spiking activity in the pre-neuronal layer (f_{in}) is expected to result in an increase in the output firing frequency (f_{out}). The relationship between the two variables is likely to be linear for higher values of f_{in}. This is because when incoming spikes are closer together temporally, the potential barrier has less time to recover between spikes, leading to reduced leakage. However, the scenario changes for lower f_{in} values. In this case, significant leakage of holes occurs, necessitating more spikes to reach the threshold current (I_{th}). Consequently, for decreasing f_{in}, there is an exponential decrease in the output spiking frequency (f_{out}). The curve depicted in Fig. 3.15 demonstrates the relationship between f_{out} and f_{in}. Notably, the curve's slope is higher for smaller f_{in} values, gradually decreasing as f_{in} becomes larger. This trend can be attributed to the more significant reduction in f_{out} when f_{in} decreases, driven by

the higher component of leakage.

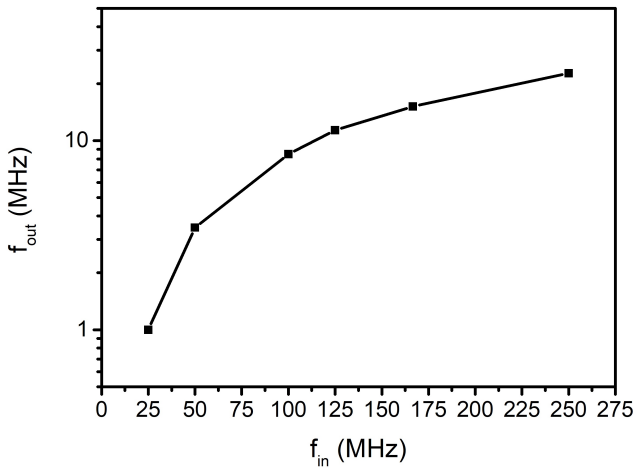

Figure 3.15: Output firing frequency for different input firing frequency

3.5.5 Interface Trap Charges

The presence of trap charges at the semiconductor-oxide interface can have detrimental effects on the device performance due to Trap Assisted Tunneling (TAT). This phenomenon not only degrades device performance but also reduces device reliability and overall longevity. Thus, investigating the impact of Interface Trap Charges (ITCs) on device performance becomes crucial.

The ITCs can be positively charged, negatively charged or neutral depending on their energy level with respect to the Fermi energy level and the midband energy level. It behaves like an acceptor trap if its energy level is above the midband energy level and a donor trap if its energy level is below the midband level [69]. An acceptor trap is negatively charged if its energy level is below the fermi level by accepting an electron and neutral if its energy level is above the fermi level. Similarly, a donor trap is positively charged if its energy level

is above the fermi level and neutral if its energy level is below the fermi level. For a p-type channel, having its fermi level near to the valence band, the donor trap states are partially filled, while the acceptor trap states are empty, making the interface charge positive in nature. On the other hand, for an n-type channel, the acceptor states are partially filled and the donor states are empty, making the interface charge negative in nature.

In this particular use case of the LIF neuron, the PD-SOI MOSFET is operated in the accumulation mode, due to the application of a negative voltage on the gate. Thus, the channel is predominantly p-type in nature. This results in positively charged donor traps at the interface. These repel the accumulated holes and an electron-hole recombination occurs due to the electron emitted by the donor state. Now, more number of voltage spikes shall be required to counter this positive charge at the interface. Thus, there is no trapping of holes after repeated LIF operations.

To incorporate ITCs into our analysis, we introduced a fixed charge of $\pm(1 \times 10^{12})cm^{-2}$ at the semiconductor-oxide interface. This choice of value is based on prior simulation studies [70, 71], which explored ITC densities in the range of $\pm(1 \times 10^{11} - 1 \times 10^{13})cm^{-2}$. Fig. 3.16 illustrates f_{out} for varying magnitudes of ITCs. The figure shows that negative (acceptor) ITCs tend to enhance the accumulation of holes in the channel by boosting the BTBT generation rate. Conversely, positive (donor) ITCs reduce the BTBT generation rate, decreasing the output spiking frequency.

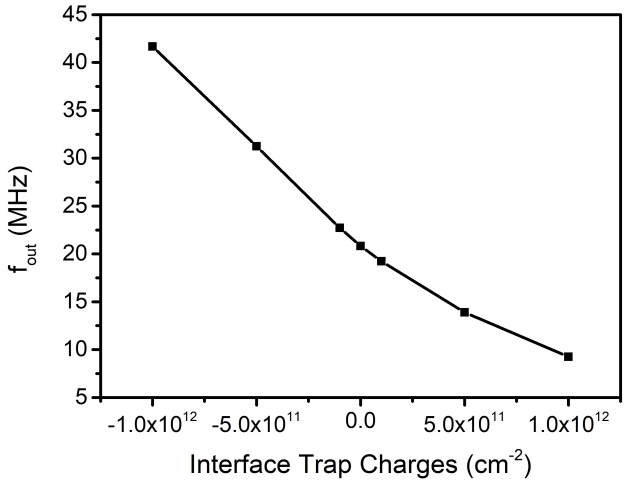

Figure 3.16: Output firing frequency for different interface charge concentrations

Notably, positive ITCs at the semiconductor-oxide interface can lead to performance degradation and a decrease in energy efficiency (assuming a constant spiking frequency). However, this reduction in spiking frequency due to positive ITCs remains acceptable in neuromorphic chips, as these chips do not require extremely high operating frequencies. The trade-off between performance and reliability is essential in designing such devices.

3.6 Conclusions

In this chapter, using a well-calibrated simulation model, an energy-efficient BTBT-based Ge LIF neuron has been demonstrated. The energy consumption per spike in the proposed LIF neuron is 0.07fJ, which is lower than the LIF neurons reported in the literature. The proposed implementation is also more biologically plausible than those presented in the literature, wherein summer circuits are required. From the system-level standpoint, such additional circuitry can incur a much larger area and power penalty. Further, a reduction in power

consumed by the reset circuitry is also achieved by lowering the drain bias. The channel thickness should be carefully optimized such that the stored charge can be retained for a longer duration of time while keeping energy consumption at acceptable levels.

Chapter 4

On-chip Unsupervised Learning using STDP in a Spiking Neural Network

This chapter proposes an energy-efficient Ge-based device that implements on-chip unsupervised learning in an SNN using STDP. The approach involves utilizing a Ferromagnetic Domain Wall (FM-DW) based device as a synapse. The proposed device configuration consists of a dual pocket Fully-Depleted Silicon-on-Insulator (FD-SOI) MOSFET equipped with dual asymmetric gates. Using a carefully calibrated 2D device simulation model, we demonstrate that a pair of these devices can produce a current output that is exponentially dependent on the temporal correlation between spiking events in the pre- and the post-synaptic neuronal layers. This current drives the FM-DW synapse, leading to the adjustment of the synapse's conductance in line with the STDP learning rule. The proposed implementation requires 2-3$\times$ fewer transistors and offers a lower latency than the existing literature. Further, we demonstrate the real-world applicability of the proposed device at the system level. Specifically,

we showcase its performance by employing it to train an SNN for recognizing handwritten digits in the MNIST dataset. Remarkably, this implementation achieves a classification accuracy of 84%, thereby underlining the device's potential for practical applications and its utility in advancing neural network capabilities. The work done in this chapter is published in [72].

4.1 Simulation framework

In this section, a hierarchical simulation framework is developed to demonstrate on-chip unsupervised learning using STDP in an SNN. Fig. 4.1 shows an overview of the hierarchical simulation framework employed in this work. The framework is comprised of several stages, each addressing a specific aspect of the overall process. Initially, a device-level simulation is performed in mumax3 [73]. This simulation focuses on the FM-DW spintronic synapse, aiming to capture the change in conductance resulting from a current of a specific magnitude and duration traversing through the HM layer of the FM-DW synapse. Following this, device-level simulation is performed in the device simulator Synopsys Sentaurus [81] to show the functionality of the proposed dual-pocket FD-SOI MOSFET. Subsequently, mixed-mode circuit simulations are performed in the device simulator Synopsys Sentaurus, employing a pair of dual-pocket FD-SOI MOSFETs, which generate a programming current. The magnitude of this current is expected to show an exponential dependence on the temporal correlation between spiking events in the pre- and post-synaptic neuronal layers. This phenomenon essentially executes the STDP learning algorithm.

The programming current serves to modulate the conductance of the FM-DW synapse. The final step in the framework involves integrating the results and characteristics from both the device-level and circuit-level simulations to derive the system-level behavior. With the modeled synapse in place, an SNN is trained using STDP to recognize handwritten digits in the MNIST dataset. This training and benchmarking is achieved via the BindsNET platform [74].

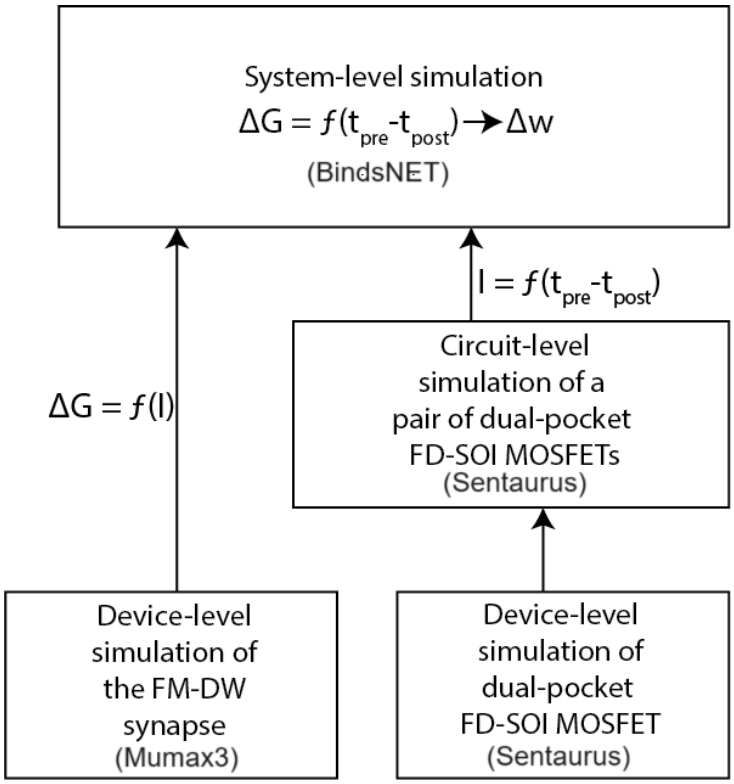

Figure 4.1: The hierarchical simulation framework to demonstrate on-chip unsupervised learning in an SNN.

4.1.1 Spintronic Synapse

This section describes the device physics that governs the functioning of the FM-DW synapse. The synaptic element, which comprises a Magnetic Tunnel Junction (MTJ) coupled with an HM underlayer, is depicted in Fig. 4.2. The

MTJ, at its core, is constructed by combining two key components: a free Ferromagnetic (FM) layer comprised of CoFe, whose magnetization can be varied, and a pinned FM layer, whose magnetization remains fixed. These two FM layers are separated by a tunneling oxide barrier made of MgO. An intrinsic component of this setup is the presence of a Domain Wall (DW), a boundary that demarcates two distinct magnetic regions within the free FM layer, each characterized by an opposite polarization.

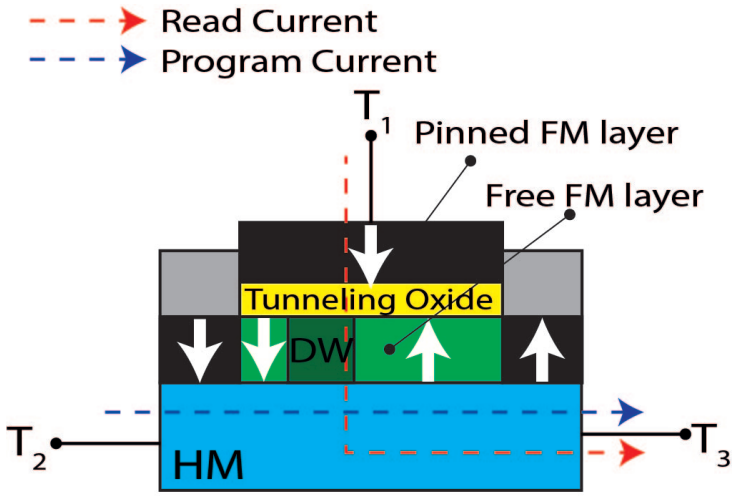

Figure 4.2: FM-DW synapse with decoupled read and program paths. The read current flows between terminals T_1 and T_3 while the programming current flows between terminals T_2 and T_3.

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) [75–78]. 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 [77, 78]. Consequently, the flow of current through the HM layer induces a Spin-Orbit Torque (SOT) that propels the DW in the free FM layer to shift along its length. This DW movement is reinforced by opposing magnetizations in the two pinned FM

layers located at the ends of the free FM layer, resulting in the stabilization of the DW under the influence of sufficiently high currents within the HM layer. The magnetization dynamics of the free FM layer's chiral DW are described by the Landau-Lifshitz-Gilbert (LLG) equation [75]. To simulate the DW's motion due to the current flow through the HM layer, comprehensive simulations were conducted using mumax3 [73]. Some essential simulation parameters used in these simulations are enumerated in Tab. 4.1, with their values adopted from a previous study [78]. The simulation aims to model the synapse behavior, quantifying how its conductance changes due to to the flow of a current through the HM layer.

Table 4.1: FM-DW synapse simulation parameters

Device Parameter	Symbol	Value
Saturation Magnetization	M_S	700 KA/m
Exchange Correlation constant	A_{ex}	$1 \times 10^{-11} J/m$
Perpendicular Magnetic Anisotropy	K_u	$4.8 \times 10^5 J/m^3$
Effective DMI constant	D_{ind}	$-1.2 \times 10^{-3} J/m^2$
Gilbert damping factor	α	0.3
Spin Hall angle	θ	0.07
Heavy metal thickness	t_{HM}	10 nm
Free FM layer dimensions		$500 \times 20 \times 0.6\ nm^3$
Grid size		$4 \times 1 \times 0.6\ nm^3$

The FM-DW synapse described above has decoupled read and program paths. Specifically, the program path pertains to the programming current that flows through the HM layer, traversing between terminals T_2 and T_3. The consequence of this programming current is the motion of the DW within the free FM layer. This behavior is depicted in Fig. 4.3(a), where the displacement of the DW in response to a current density ($J = 1 \times 10^{11} A/m^2$) is illustrated. The corresponding DW velocity for varying current densities is shown in Fig.

4.3(b). The Domain Wall's (DW) velocity does not depend on the time duration of current spike through the Heavy Metal (HM) layer. A longer duration of current spike results in a larger displacement of the DW in the free ferromagnetic layer, it does not result in a larger velocity of the DW. It is worth noting that these micromagnetic simulation results, which detail the DW's position modulation due to the flow of a current density ($J = 1 \times 10^{11} A/m^2$) through the HM layer, correlate well with previously published results [78].

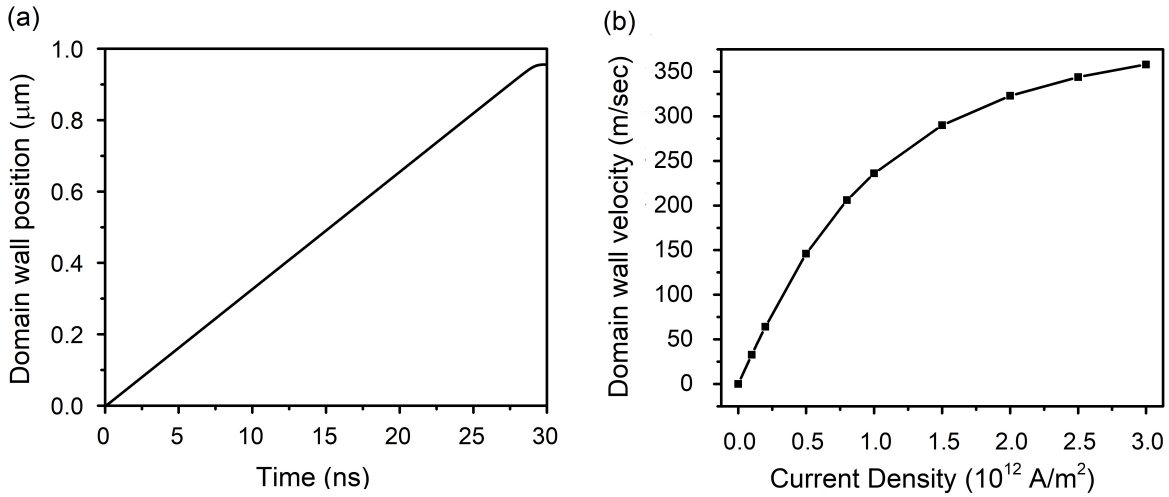

Figure 4.3: (a) DW displacement due to the application of a current density ($J = 1 \times 10^{11} A/m^2$) along the HM layer and (b) DW velocity plotted as a function of current density through the HM layer.

A spatial shift in the DW position results in a corresponding modulation in the conductance of the FM-DW synapse. When the magnetization of the free layer is aligned parallel to the fixed FM layer, the synapse exhibits a conductance of G_P. Conversely, when the magnetization of the free layer is oriented anti-parallel to the fixed FM layer, the synapse's conductance is G_{AP}. The conductance of the synapse (G_S), accounting for the DW's displacement by x along the FM free layer's length, is described as follows [35]:

$$G_s(x) = G_P\left(\frac{x}{L}\right) + G_{AP}\left(1 - \frac{x}{L}\right) \tag{4.1}$$

where L denotes the length of the FM free layer in the MTJ. The conductance G_S follows a linear relationship with the DW's position in the FM free layer due to the constant values of G_P, L, and G_{AP}. In this context, the maximum and minimum values of the conductance are G_P and G_{AP}, respectively. The ratio between these values constitutes the Tunneling Magnetoresistance Ratio (TMR). The Resistance-Area (RA) product of the MTJ, set at 10 $\Omega\mu m^2$, coupled with a TMR value of 600%, as reported in [79], further characterizes the properties of the synapse.

The linearity parameters represent the number of pulses (of a certain amplitude and duration) required to change the conductance of the synapse from minimum to maximum (α_p) and vice-versa (α_d). Linearity parameters are essential in synaptic devices where the conductance change is non-linear due to the abrupt set and reset transitions. For STDP, the change of conductance depends on the temporal correlation of spiking events, so linearity does not play a big role here.

On the other hand, read current flows through the MTJ between terminals T_1 and T_3. The voltage V_S applied at terminal T_1 drives a read current $I_S = G_S \cdot V_S$ through the synapse. The conductance G_S of the synapse depends on the DW's position in the free FM layer, and this dependence arises from the linear relationship between G_S and DW position. It is essential to ensure that the magnitude of the read current does not surpass the DW depinning current. The

DW depinning current is the current threshold required to initiate the movement of the DW along the length of the free FM layer. Keeping the read current below this threshold is crucial to prevent unintentional modulation of the DW's position, which would otherwise disturb the synaptic weight and could result in undesirable changes to the synapse behavior. This control ensures the stability and reliability of the synaptic operation.

The main reason behind opting for a spin-based synapse was the decoupled nature of spike transmission and learning in the synapse. The 3-terminal FM-DW synapse allowed learning during network operation, which is a crucial requirement for STDP based learning to take place. Moreover, the FM-DW synapse allows continuous conductance states, i.e., the conductance (weight) of the synapse can be modulated by the position of the domain wall in the free ferromagnetic layer, which in turn can be tuned by passing a current through the HM layer. Generating such a current whose magnitude was a function of the temporal correlation of spiking events between the pre- and post-synaptic layers of neurons with a few CMOS transistors or a pair of dual-pocket FD-SOI MOSFET's with dual asymmetric gates (proposed in this work) was indeed possible. Furthermore, the conductance of the FM-DW synapse varies linearly with the position of the domain wall, while for the case of memristive synapses, the conductance may switch abruptly between high and low resistance states.

4.1.2 Device-Level simulation

The schematic cross-sectional view of the proposed device, which includes a dual pocket Ge-based FD-SOI MOSFET with dual asymmetric gates, is depicted in Fig. 4.4. In the next subsection, it will be demonstrated that a pair of such devices can generate a current based on the temporal correlation of spiking events between the pre-synaptic and post-synaptic neurons. This current is fed to the HM layer of the FM-DW synapse, and it governs the modulation of the synapse's conductance in accordance with the STDP learning rule.

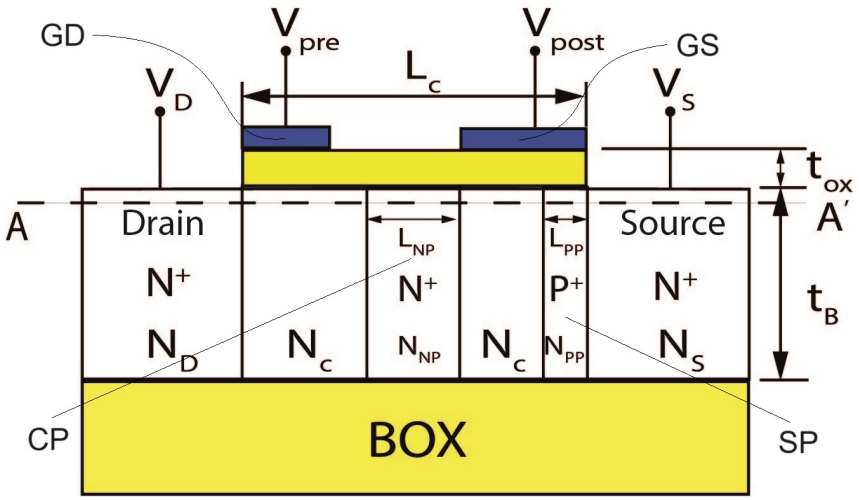

Figure 4.4: Structure of the proposed device used to train the SNN using STDP.

The proposed device comprises two pockets. The first pocket, known as the Source Pocket (SP), is doped with p^+ dopants, having a length of L_{PP} and a doping concentration of N_{PP}. This narrow, fully depleted pocket is placed at the source-channel interface, enhances the sharpness of the band profile at this interface and thereby enhances the BTBT generation rate. To achieve a sharp band profile, the optimization of pocket parameters N_{PP} and L_{PP} has been undertaken. The second pocket, termed as Channel Pocket (CP), is doped

with n^+ dopants, possessing a length of L_{NP} and a doping concentration of N_{NP}. Located at the center of the channel, its purpose is to establish electrical isolation between the two channel regions. Fabrication of these pockets can be accomplished by initially defining the region for the larger n^+ pocket through a suitable mask and subsequent doping. The second p^+ pocket can then be produced by implementing a tilt implant followed by spike annealing, as outlined in [80]. The proposed device is equipped with dual-asymmetric gates, with one allocated for each channel region. The gate situated near the source (GS) has a length of $L_{G,S}$, whereas the gate adjacent to the drain (GD) possesses a length of $L_{G,D}$. An HfO_2 gate oxide with a thickness of 5 nm is employed to enhance gate control over the channel while concurrently minimizing leakage attributed to the gate tunneling current. Additional device parameters are presented in Tab. 4.2. The rationale for utilizing a thin, fully-depleted body in conjunction with dual-asymmetric gates will be elucidated during the discussion of the operating principle of the proposed device.

Table 4.2: Device Parameters of the proposed device used to train the SNN using STDP

Device Parameter	Symbol	Value
Drain Voltage (V)	V_D	0.4
Source Voltage (V)	V_S	0.3
Channel Length (nm)	L_C	100
Gate Oxide thickness (nm)	t_{ox}	5
Body thickness (nm)	t_B	10
Gate workfunction (eV)	ϕ_m	4.4
Channel Doping (p-type) (cm^{-3})	N_C	1×10^{17}
Source Doping (n-type) (cm^{-3})	N_S	1×10^{20}
Drain Doping (n-type) (cm^{-3})	N_D	1×10^{20}
p^+ Pocket Doping (p-type) ($atoms/cm^3$)	N_{PP}	1×10^{19}
p^+ Pocket Length (nm)	L_{PP}	4
n^+ Pocket Doping (n-type) ($atoms/cm^3$)	N_{NP}	4×10^{19}
n^+ Pocket Length (nm)	L_{NP}	20
Length of gate near source (nm)	$L_{G,S}$	40
Length of gate near drain (nm)	$L_{G,D}$	25

Germanium, due to its smaller bandgap and a dominant direct tunneling mechanism compared to Silicon, is used as a base material in this work [5]. This results in a larger BTBT generation rate compared to a Si-based device. A non-local BTBT model has been used for simulations with fitting parameters adopted from [5]. The detailed simulation model employed in this study has been presented in section 3.1.

The principle of operation for the proposed device will now be explained. Fig. 4.5 shows detailed band diagrams along the cut-line AA' to illustrate the movement of carriers across the device based on the temporal correlation of spiking events between pre- and post-synaptic neurons for the case when a post-synaptic neuron spiking event follows a pre-synaptic neuron spiking event.

The initial state, without any applied voltages to the device, is represented by the band diagram in Fig. 4.5(a). Subsequently, by applying a constant voltage of 0.4V to the drain and 0.3V to the source, the resulting band diagram is depicted in Fig. 4.5(b). In the presence of a pre-neuronal spike (V_{pre}= -0.7 V and 1ns duration) at GD, as illustrated in the band diagram of Fig. 4.5(c), BTBT of electrons occurs at the drain-channel interface. This results in the generation of electron-hole pairs due to the formation of vacancies (holes) in the channel [81]. Due to the gate underlap of GD with respect to CP, BTBT is suppressed at the CP-channel interface. This ensures equivalent BTBT generation rates within the device during both pre- and post-neuronal spiking occurrences. Subsequent to the accumulation of holes in the channel, the removal of V_{pre} results in a reduction of the potential barrier seen by electrons at the drain terminal, as

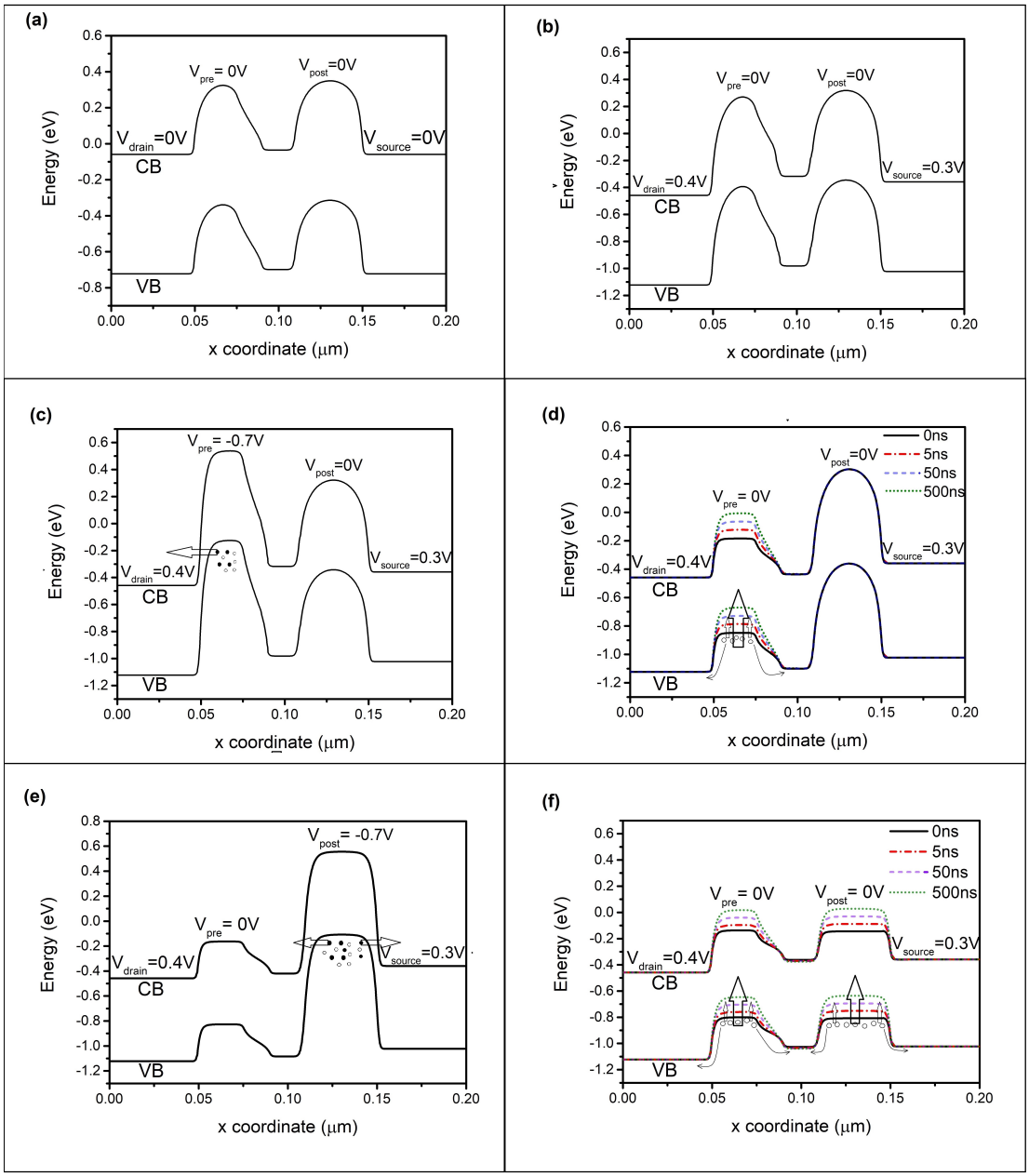

Figure 4.5: Detailed band diagrams for the proposed device along cutline AA' to illustrate the movement of carriers across the device based on temporal correlation between pre- and post-neuronal spiking activity a post-synaptic neuron spiking event follows a pre-synaptic neuron spiking event. (a) Band diagram at equilibrium, (b) Band diagram with applied bias, (c) Band diagram during application of a pre-neuronal voltage spike (V_{pre}) at *GD* causing BTBT of electrons at the drain-channel interface leaving behind vacancies (holes) in the channel, (d) Evolution of the band diagram after removal of V_{pre} spike demonstrating leakage of accumulated holes in the channel with time, (e) Band diagram during application of the post-neuronal voltage spike V_{post} (immediately after V_{pre}) at *GS* causing BTBT of electrons at the source-channel and the CP-channel interfaces leaving behind vacancies (holes) in the channel, and (f) Evolution of the band diagram after removal of V_{post} demonstrating leakage of accumulated holes in the channel with time.

presented in Fig. 4.5(d). Despite this reduction in the potential barrier, the current due to thermionic emission remains minimal due to the large potential barrier at the source end. Due to the device's thin body, the accumulated holes quickly dissipate into the drain, consequently elevating the potential barrier at the drain end. As depicted in Fig. 4.5(d), the temporal evolution of band diagram illustrates an increase in the potential barrier, which is attributed to the leakage of holes from the channel.

When a post-neuronal spike (V_{post}= -0.7 V and 1ns duration) is applied at GS, as illustrated in the band diagram of Fig. 4.5(e), BTBT occurs both at the source-channel and the CP-channel interface. The addition of the SP increases the sharpness of the band profile at the source-channel interface, increasing the BTBT generation rate. Despite a smaller band overlap during the post-neuronal firing event due to a reduced source bias compared to the pre-neuronal firing event (drain side), it is ensured that the BTBT generation rate remains uniform throughout both pre- and post-neuronal events. The band diagram in Fig. 4.5(e) corresponds to the scenario where V_{post} follows immediately after V_{pre}. Upon the removal of V_{post}, the potential barrier at the source end reduces, as depicted in the band diagram of Fig. 4.5(f), resulting in the flow of current on the order of a few μA through the device. The accumulated holes gradually dissipate into the source, drain, and CP regions over time, causing an increase in the potential barrier at both the source and drain ends. This phenomenon leads to an exponential decline in current. The temporal evolution of the band diagram, as shown in Fig. 4.5(f), illustrates the increasing potential barrier at the source and

drain ends due to the leakage of accumulated holes from the channel.

The magnitude of current flow is dependent on the temporal correlation between the spiking events of the pre- and post-neurons. The proximity of these two spikes directly influences the amplitude of the current. This relationship arises from the rapid dissipation of accumulated holes triggered by the pre-neuronal spiking event, causing an increase in the potential barrier at the drain end. Prior to the occurrence of the post-neuronal spiking event, the potential barrier at the source end is sufficiently high, resulting in an almost negligible current flow. Simultaneously, the channel region near the drain end continues to experience hole leakage, thereby increasing the potential barrier at that end. If the temporal interval between the two spikes surpasses 100ns, the potential barrier at the drain end becomes prohibitively high. In such a scenario, even following the post-neuronal spiking event, the device exhibits minimal current flow. The thin body of the device facilitates rapid leakage of holes. Consequently, the current due to the thermionic emission diminishes in an exponential manner as the temporal gap between the pre- and post-neuronal spiking events increases. Fig. 4.6 shows the current flow across the device and demonstrates how the current changes as the temporal interval of spiking events between the pre- and the post-synaptic neuronal spiking events varies, for the case when the post-neuron firing event follows a pre-neuron firing event and vice-versa.

Subsequently, we consider the scenario in which the pre-neuronal firing event takes place after a post-neuronal firing event. In this situation, the initiation of a post-neuronal spike (V_{post} = -0.7 V and 1ns duration) triggers BTBT of

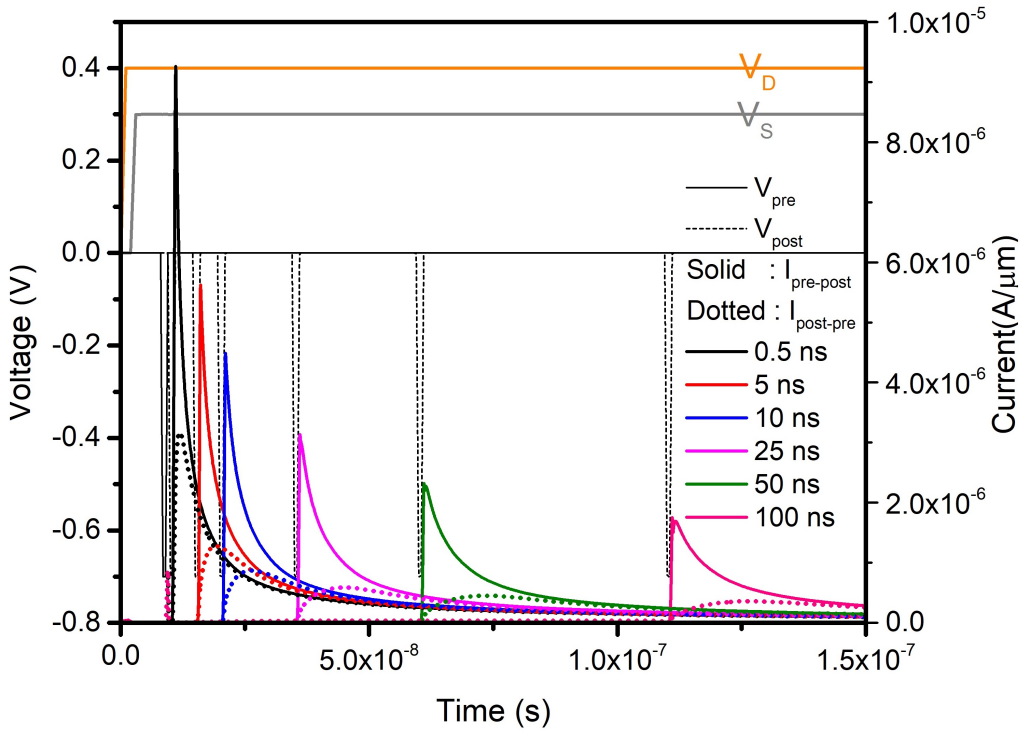

Figure 4.6: Current flow through the proposed device for the case when the post-neuron firing event follows a pre-neuron firing event and vice-versa for different time intervals between the two spiking events.

electrons at both the source-channel and CP-channel interfaces. This leads to the accumulation of holes within the channel. The potential barrier at the source end decreases upon removal of the post-neuronal spike. Due to the device's thin body, these accumulated holes swiftly leak into the source and CP regions. As a consequence, a rapid increase in the potential barrier at the source end ensues. By the time the pre-neuronal spiking event occurs, the potential barrier at the source end becomes sufficiently high to impede current flow across the device. Fig. 4.6 displays the current flow through the device when a pre-neuron firing event follows a post-neuron firing event. Notably, the current amplitude in this sequence is notably lower compared to the case where a post-neuron firing event follows a pre-neuron firing event. Furthermore, it is evident that

the current declines exponentially as the temporal gap between the two spiking events increases.

4.1.3 Circuit-level simulation

This section elaborates on the methodology employed to derive the STDP characteristics using the proposed circuit. To achieve the desired STDP characteristics, certain prerequisites need to be fulfilled. Firstly, the synapse's strength should be either potentiated or depressed depending upon the temporal correlation of spiking activity between the pre- and the post-synaptic neurons. Specifically, if a firing event in the pre-neuron is succeeded by a firing event in the post-neuron, the connecting synapse between the two neurons should be potentiated. Conversely, if a post-neuron firing event is followed by a pre-neuron firing event, the synapse should undergo depression. Additionally, the modulation in the synapse's weight should vary exponentially based on the temporal correlation between the spiking events of the pre- and post-synaptic neurons. A smaller temporal gap between these events should correspond to a more substantial increase or decrease in the synapse's weight. The proposed circuit configuration, which employs a pair of dual-pocket FD-SOI MOSFETs to drive the HM layer within the FM-DW synapse, is depicted in Fig. 4.7.

The proposed circuit's functionality and ability to achieve the desired STDP characteristics are explained as follows. A constant voltage ($V_1 = 0.4V$) is applied to the drain terminal of transistor T_1. The device T_2 is connected in series with T_1, with the source terminal of T_2 connected to a constant voltage source

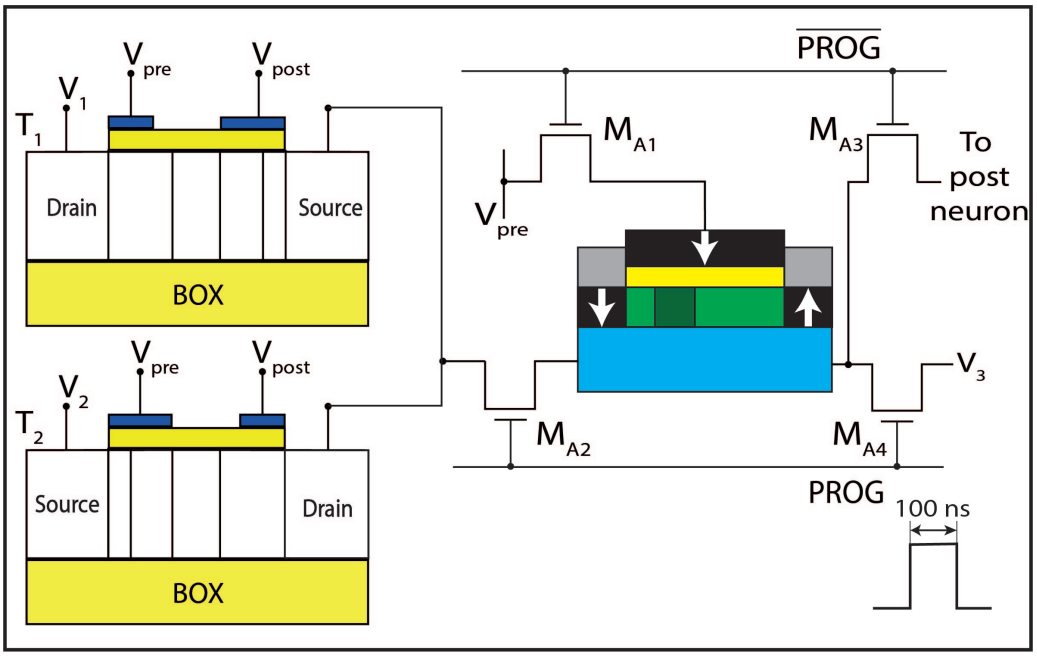

Figure 4.7: The proposed circuit used to tune the conductance of the FM-DW synapse using STDP.

($V_2 = 0.2V$). The drain terminal of T_2 is connected with the source terminal of T_1, and it is also connected to the HM layer of the FM-DW synapse via an access transistor M_{A2}. The HM layer's opposite end is connected to a constant voltage source ($V_3 = 0.3V$) through another access transistor M_{A4}. Upon detection of a post-neuronal spiking event, the $PROG$ signal is activated, which facilitates the programming of the synapse by enabling the access transistors M_{A2} and M_{A4}. Importantly, the $PROG$ signal's magnitude is set to a level that circumvents any significant threshold voltage (V_t) reduction across the access transistors. Subsequently, when the $PROG$ signal is not active, the synapse is configured in the read mode.

The dissimilarity in BTBT generation rates between T_1 and T_2 is due to the difference in the drain and source voltages applied to these transistors. To equalize the BTBT generation rates in both devices and achieve symmetrical

STDP characteristics, a 20% increase in the channel width for T_2 is made. Mixed-mode simulations are performed in the Sentaurus device simulator. The HM layer within the FM-DW synapse is simulated as a resistor, with a resistance of 200Ω, attributed to Pt's relatively low resistivity. When the post-neuronal firing event is observed after the pre-neuronal firing event, T_1 witnesses a higher current conduction compared to the capacity of T_2 to sink. The disparity in these currents manifests as a net current flowing through the HM layer within the FM-DW synapse. Notably, the time gap between these two spiking events inversely influences the magnitude of the current flowing through the HM layer. Following the post-neuronal firing event, the accumulated holes escape, leading to a rapid exponential decrease in current through T_1, ultimately equalizing it with the current through T_2 within a brief time span of the order of ns. As a result, the net current through the HM layer diminishes to zero. This transient surge of current in the HM layer causes the DW to shift in the direction of the current flow. The DW's velocity is proportional to the current's magnitude through the HM layer, increasing the synapse conductance. Importantly, this increase in conductance is exponentially dependent on the temporal correlation between the pre- and post-neuronal spiking events. The current inflow into the HM layer is plotted against the interval between the pre- and post-neuronal spiking events in Figure 4.8. The figure indicates the exponential reduction in current through the HM layer in proportion to the temporal interval ($t_{post} - t_{pre}$) between spiking activities of pre- and post-neuronal elements.

In the scenario where the pre-neuronal spiking event follows the post-neuronal

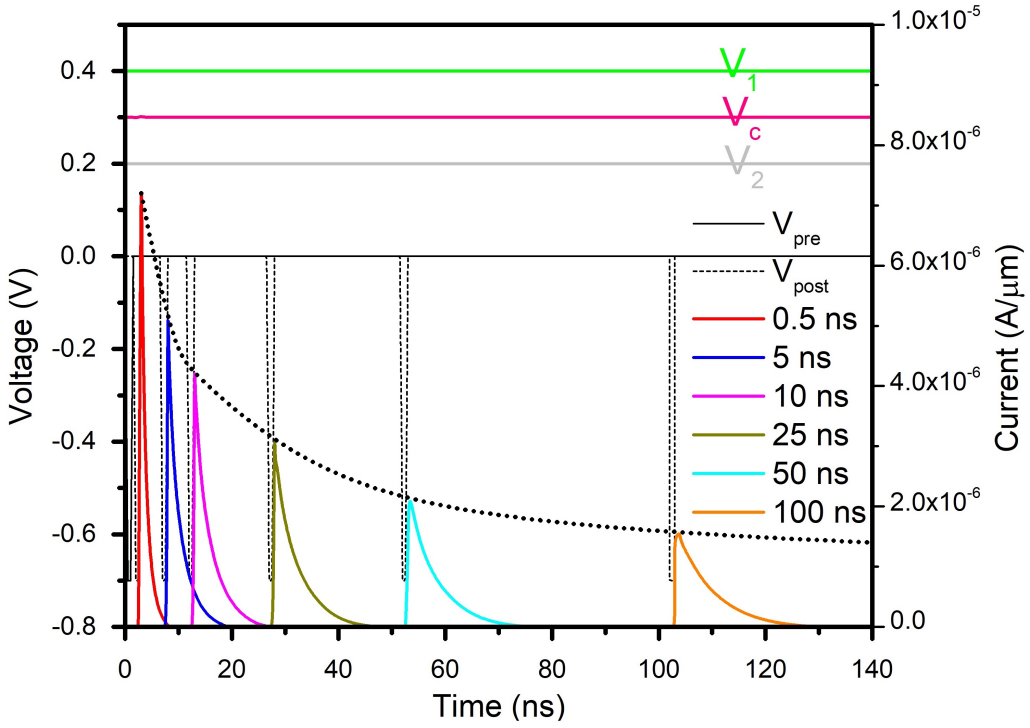

Figure 4.8: The current flowing through the HM layer plotted as a function of the time interval ($t_{post} - t_{pre}$) between the pre- and the post-spiking neuron.

spiking event, the current dynamics differ. In this case, transistor T_2 sinks more current compared to the current sourced by transistor T_1. This current is supplied by voltage source V_3. Consequently, the current through the HM layer flows in the opposite direction, contrary to the previous scenario. The domain wall (DW) movement is also altered; it now moves toward the current flow's direction, shifting toward the left side of the device. As a result, the synapse's conductance reduces, an effect that is once again exponentially related to the temporal correlation between the pre- and post-synaptic spiking neuronal events. The time interval between the pre- and post-spiking neuron activities influences the decrement in conductance. The representation of current inflow into the HM layer against the temporal interval between the pre- and post-spiking neurons is shown in Fig. 4.9. The negative current spikes displayed in the graph signify

current flow in the opposite direction, i.e., the current supplied by voltage source V_3. The trend depicted in Figure 4.9 demonstrates the exponential decline in the magnitude of current flow through the HM layer as a function of the temporal gap $(t_{post} - t_{pre})$ between pre- and post-spiking neuronal activities. The requirement for bidirectional current flow through the HM layer in the FM-DW synapse necessitates the use of a pair of dual-pocket FD-SOI MOSFETs in this work.

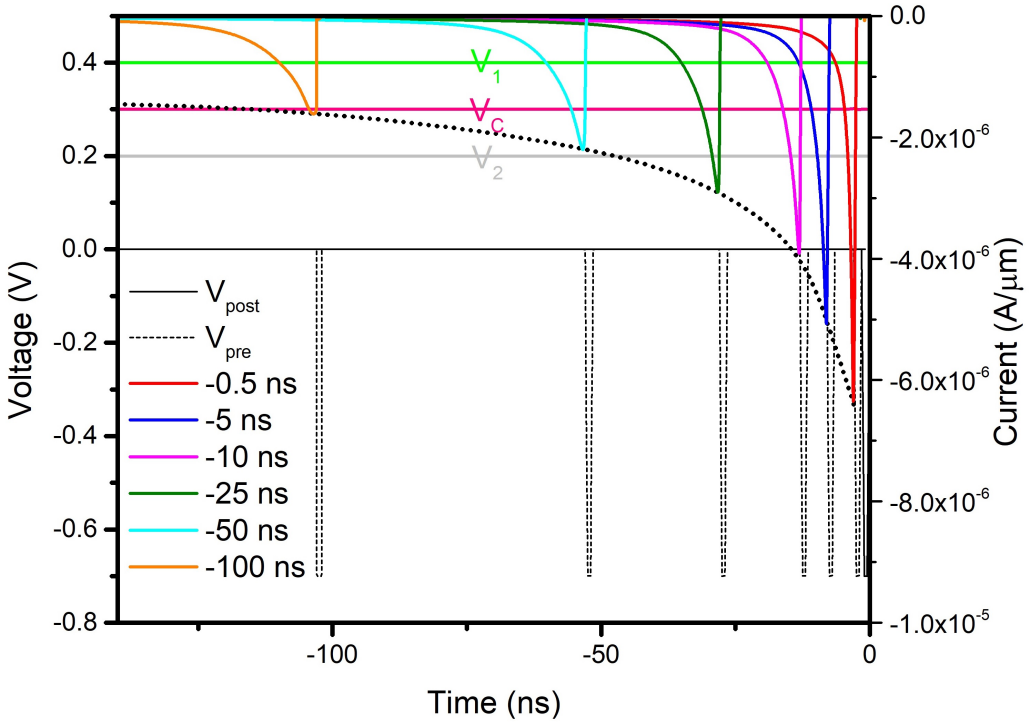

Figure 4.9: The current flowing into the HM layer plotted as a function of the time interval $(t_{post} - t_{pre})$ between the pre and the post-spiking neuron.

If the synapse is equally potentiated and depressed for the absolute value to the time difference ($\delta w \; \alpha \; |t_{pre} - t_{post}|$) between spiking activities in the pre-synaptic and post-synaptic layer of neurons, then the characteristics are symmetric. In the STDP based learning circuitry proposed in this work, the characteristics can be made symmetric by increasing the width of transistor T2, such that the magnitude of current generated by the learning circuitry is equal for both events

(pre-synaptic event followed by post and vice-versa).

Fig. 4.10 illustrates a crossbar architecture suitable for establishing connections between the pre-synaptic neurons (N_{pre}) and the post-synaptic neurons (N_{post}) through synapses. The architecture of the neuron is not shown in the figure. However, the same may be referred from the Fig. 3.6. The input to the neuron is a voltage spike, which also goes as input to the learning circuitry. Once a threshold current is reached, the control circuitry generates a voltage spike ($V_{pre,1}$) in Fig. 4.10. This voltage spike generates a current spike using the architecture proposed in Fig. 3.6. This current spike will transmit through the synapse and its magnitude is modulated in the process in accordance with the conductance of the synapse. A weighted sum of these current spikes from different neurons in the pre-synaptic neuronal layer at a particular time instant for a particular post-neuron is then converted into a post-synaptic voltage spike using an interface circuitry. This post-synaptic voltage spike is applied to the learning circuitry. Thus, on the basis of temporal correlation of spiking events between the pre- and post-synaptic neurons, an appropriate current is generated, which modulates the conductance of the synapse in accordance with the STDP learning rule. This architecture can efficiently add the current from the pre-neuronal spiking activity at the post-neurons. Essentially, this design collocates memory (where weights are stored within synapses) and computation (performing multiplication and accumulation operations).

The proposed circuitry operates by receiving voltage spikes from both the pre-synaptic (V_{pre}) and the post-synaptic (V_{post}) neurons. It then generates a current

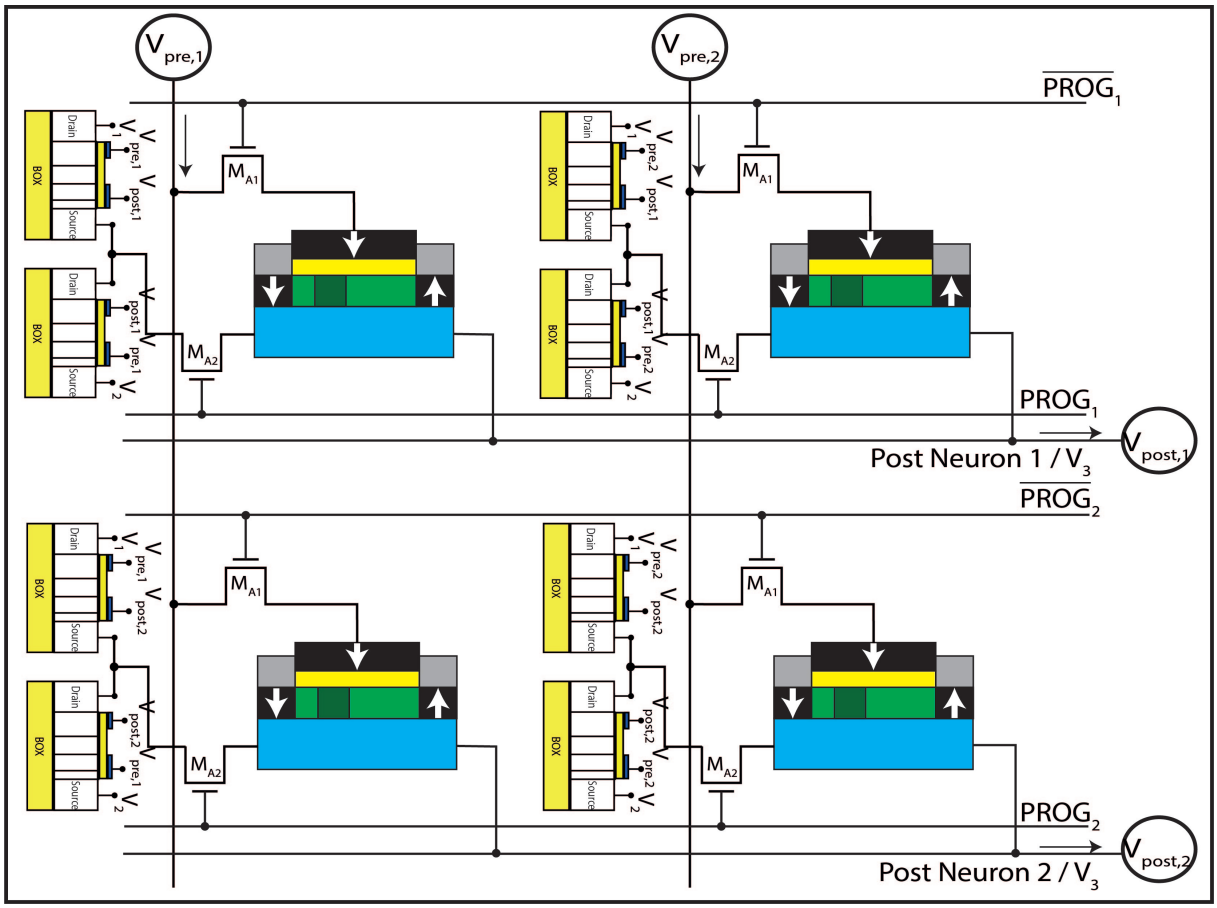

Figure 4.10: A crossbar architecture that can be employed to interconnect neurons via synapses.

that adjusts the conductance (weight) of the synapse when the corresponding $PROG$ signal is activated. Subsequently, when the $PROG$ signal is deactivated, the synapse enters read mode. When a voltage spike occurs in the pre-neuronal layer, it triggers a current spike whose magnitude depends on the weight of the synapse. This current is subsequently transmitted to the post-neuron. For each synaptic element, two pass transistors (M_{A1} and M_{A2}) are necessary. Additionally, per post-neuron (not depicted in Fig. 4.10), two pass transistors (M_{A3} and M_{A4} in Fig. 4.7) are required. These pass transistors (M_{A3} and M_{A4}) are shared among all pre-neurons and serve to multiplex between the voltage source V_3 during writing and the corresponding post-neuron during reading, depending

on the state of the $PROG$ signal. Such pass transistors are required in other reported implementations also [34–36]. These transistors enable the connection between post-neurons (during reading) and ground (during writing) to be multiplexed. Furthermore, access transistors like M_{A1} and M_{A2} ensure isolation between the read and write paths during the respective operations. Despite the incorporation of these pass transistors, the proposed approach requires fewer transistors within the crossbar array due to the minimized number of transistors necessary to generate current in compliance with the STDP learning rule.

4.1.4 System-level simulations

Using data obtained from device-level and circuit-level simulations, an SNN is trained to recognize handwritten digits from the MNIST dataset. The MNIST dataset comprises grayscale images of handwritten digits with dimensions of 28×28 pixels [82]. The constructed neural network is composed of three layers: the input layer containing 784 neurons (one per pixel), the second layer comprising 100 excitatory neurons fully interconnected with the input layer through excitatory synapses, and the third layer consisting of 100 inhibitory neurons, each linked to its corresponding neuron in the excitatory layer.

The proposed circuitry, which contains a pair of dual pocket FD-SOI MOSFETs, plays a crucial role in updating the weights of the FM-DW synapses between the input and excitatory neurons. When an excitatory neuron fires in the network, it triggers the corresponding inhibitory neuron to spike, thereby leading to lateral inhibition of all excitatory neurons except for the one connected to the

firing inhibitory neuron.

The training of the SNN utilizes the algorithm devised by Diehl and Cook [49], as illustrated in the crossbar array representation depicted in Fig. 4.11. The Diehl and Cook algorithm is implemented using BindsNET [74], which converts the differential equations governing neuronal behavior into discrete difference equations, which are solved at defined time intervals (dt). Furthermore, this code was extended to map the weight updates obtained from device- and circuit-level simulations, which reflect changes in the synapse's conductance, to corresponding weight updates in the excitatory synapses during system-level simulations.

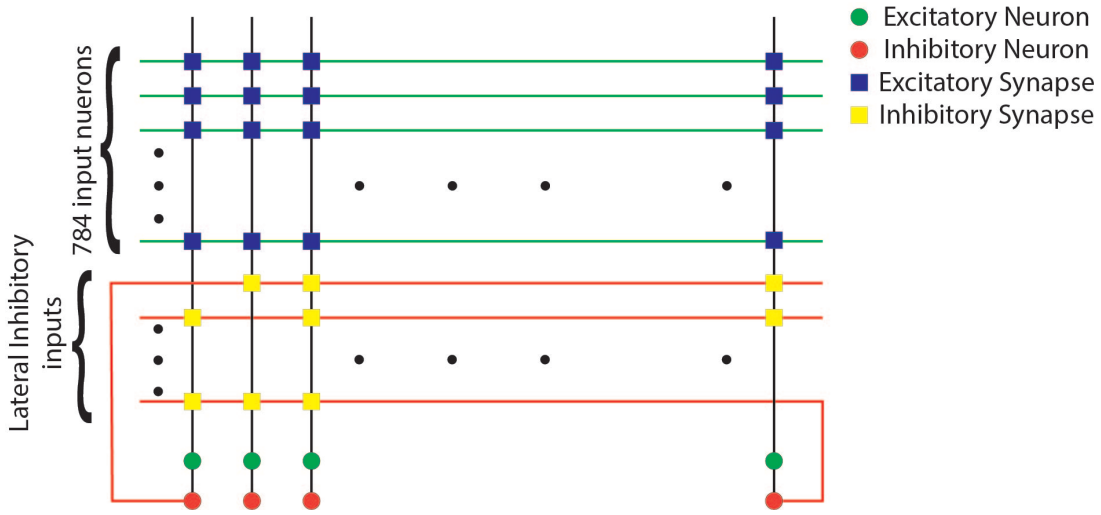

Figure 4.11: The SNN topology used to interconnect neurons via synapses in the form of a crossbar array.

The behavior of the neuron is characterized using Leaky Integrate-and-Fire (LIF) dynamics. The membrane potential ($v(t)$) of the LIF neuron evolves according to the following equation:

$$\tau_e \frac{dv}{dt} = -(v(t) - v_{rest}) + I(t) \tag{4.2}$$

where τ_e represents the membrane time constant of the excitatory neuron, v_{rest} denotes the resting potential of the neuron, and $I(t)$ is the input potential to the neuron at time instant t. The incoming spikes from pre-synaptic neurons are weighted by the synapse weight and summed together to form the total input potential $I(t)$ at a post-neuron at time t. When the membrane potential reaches the threshold (v_{th}), the neuron fires a spike and resets its membrane potential to v_{reset}. After firing a spike, the neuron remains at the resting potential during the refractory period (t_{ref}), during which it does not integrate incoming spikes. After the refractory period, the neuron resumes its LIF cycle. To prevent a single neuron from dominating the firing activity in the output layer, an adaptive thresholding technique is used. The threshold voltage of the neuron is not fixed at θ_0 but varies with time according to the equation:

$$v_{th} = \theta_0 + \theta(t) \tag{4.3}$$

Whenever a neuron fires, $\theta(t)$ is incremented by θ_{plus} and decays exponentially to zero in the absence of firing activity, as described by the equation:

$$\tau_\theta \frac{d\theta}{dt} = -\theta(t) \tag{4.4}$$

where the decay of the threshold voltage is governed by the time constant τ_θ. Tab. 4.3 presents various essential parameters utilized in the simulation. The

time constants are expressed in multiples of the simulation's time step duration.

Table 4.3: System-level simulation Parameters

Parameter	Symbol	Value
No. of excitatory/inhibitory neurons		100
No. of time steps per image		350
Membrane time constant	τ_e	20
Resting potential	v_{rest}	-65 mV
Reset potential	v_{reset}	-65 mV
Threshold Voltage	θ_0	-52 mV
Adaptive threshold voltage increase	θ_{plus}	5 mV
Time constant of Adaptive threshold voltage	τ_θ	1000
Refractory period	t_{ref}	5

The images are rate encoded using the Poisson distribution and are subsequently presented to the SNN's input layer. The spiking rate at the input layer is directly proportional to the pixel intensity in the input image. The weights of the excitatory synapses within the network are adjusted following the STDP learning rule proposed in this study. Utilizing device-level simulations of the FM-DW synapse, it is determined how the synapse's conductance changes when a current of specific amplitude and duration passes through the HM layer of the FM-DW synapse. Additionally, circuit-level simulations of a pair of dual pocket FD-SOI MOSFETs generated a suitable current based on the temporal correlation of spiking between pre- and post-synaptic neurons, conforming to the STDP learning rule. By combining data from both device and circuit-level simulations, the variation in synapse conductance is determined in relation to the temporal correlation of spiking events between pre- and post-synaptic neurons. Subsequently, the change in the conductance of the synapse, derived from these simulations, is translated to a corresponding modification in the weight of the synapse within the system-level simulations.

The MNIST dataset is composed of 60,000 training images and 10,000 test images. Initially, the weights of the synapse in the network are initialized with random values. As the training progresses, these weights are adjusted based on the STDP learning rule. These synapses connected to each neuron in the excitatory layer gradually learn the spiking patterns corresponding to the classes of different representative digits present in the input images. Consequently, the neurons become more responsive to the class of images representing the digit associated with the stored synaptic weight. In Fig. 4.12(a), the normalized synaptic weights (ranging from 0 to 0.3) are plotted in a 28 $\times$ 28 array for each neuron in the excitatory layer before the commencement of training. Fig. 4.12(b) depicts the normalized synaptic weights post-training the network with 60,000 training images, illustrating the diverse representative digits encoded in the synaptic weights.

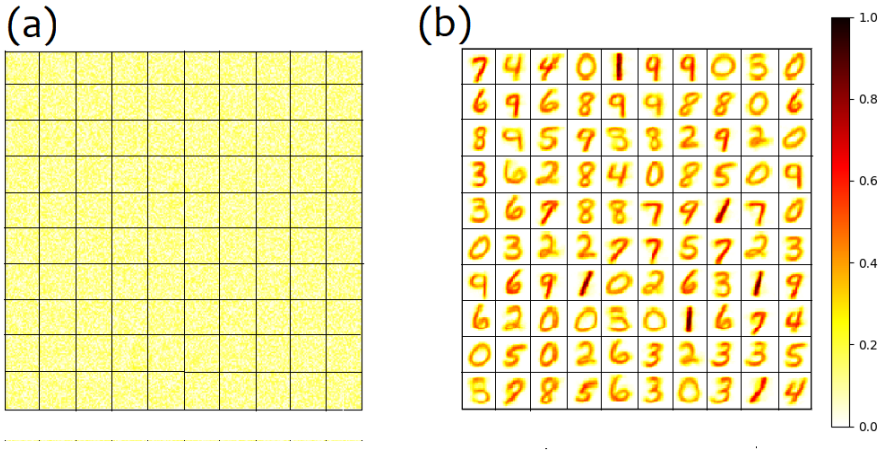

Figure 4.12: (a) Normalized synaptic weights plotted in a 28 $\times$ 28 array for each neuron in the excitatory layer before the beginning of the training process and (b) Normalized synaptic weights after training the network with 60,000 training images illustrating the various representative digits being stored in the synaptic weights.

The classification accuracy of 75% is obtained using 100 neurons in the exci-

tatory layer. The accuracy improves to 84% if the number of neurons is increased to 400. The primary reason for the relatively lower classification accuracy can be attributed to the architectural configuration employed in training the SNN. The training process is restricted to a single layer of synapses interconnecting the input and the excitatory neurons. The classification accuracy obtained in this work is compared with prior literature in Tab. 4.4. The results demonstrate that the accuracy obtained in this work is at par with existing literature that adopts a similar architecture and a comparable number of neurons in the excitatory layer [34, 49]. It is worth noting that the classification accuracy could be enhanced by introducing additional layers of neurons into the SNN; however, this would entail a considerable increase in the time required for training the network. Furthermore, for tasks requiring high accuracy in image classification, the adoption of a Spiking Deep Belief Network (DBN) or a Spiking Convolutional Neural Network (CNN) would be more suitable [40, 47].

Table 4.4: Comparison of classification accuracy on MNIST dataset among various SNN architectures

Reference	Architecture	Learning Method	No. of excitatory neurons	Accuracy
[34]	SNN [49]	STDP (2 layer)	100 400	57% 73%
This work	SNN [49]	STDP (2 layer)	100 400	75% 84%
[49]	SNN [49]	STDP (2 layer)	6400	95%
[47]	Spiking DBN	Offline learning, Conversion		95%
[40]	Spiking CNN	Offline learning, Conversion		99.1%

4.2 Impact of Variations

To obtain symmetrical STDP characteristics, matching the pair of devices responsible for synaptic plasticity is essential. This matching is necessary to achieve

equal potentiation and depression in the synaptic weight based on the relative occurrence of spiking activity in the pre- and the post-synaptic neuronal layers, thus facilitating efficient training of the SNN. The matching process accounts for the differences in the BTBT generation rate between the two devices—specifically, the BTBT rate in device T_2 is lower due to the application of smaller drain and source voltages. In order to equalize the current flowing into the heavy-metal (HM) layer of the synapse, the width of device T_2 is increased by 20%, as mentioned earlier.

However, it is important to recognize that the process-induced variations can potentially lead to asymmetrical STDP characteristics. Such asymmetry would favor either potentiation or depression in synaptic weight, thereby hampering efficient training by prolonging the time required to train the network. Process-induced variations can arise from various sources, but the impact of variations in parameters like body thickness (T_B) and gate oxide thickness (t_{ox}) can be particularly significant. Consequently, this section investigates the effect of process-induced variations — specifically variations in T_B and t_{ox} on the STDP characteristics in the pair of proposed devices.

4.2.1 Body Thickness

The symmetry of the STDP characteristics achieved using the pair of proposed devices will be influenced by the body thickness (T_B) of both devices. To achieve symmetrical STDP behavior, it is important that both devices have a body thickness of $10nm$, with device T_2 having a 20% larger channel width than

device T_1. However, real-world manufacturing processes introduce variations, resulting in potential deviations from the nominal values.

Considering these variations, each device can have a body thickness (T_B) with some standard deviation (σ) around a mean value (μ). Given the large number of synapses involved in constructing the SNN, some devices might have significantly skewed T_B values compared to the mean. To investigate the impact of such process-induced variations, we analyze the resulting STDP characteristics when the body thickness of each device is independently varied within a range of $\pm 2nm$ around the mean value of $10nm$. An effect of these variations is presented in Fig. 4.13, which illustrates the resulting STDP characteristics as the body thickness of both devices is varied within the specified range.

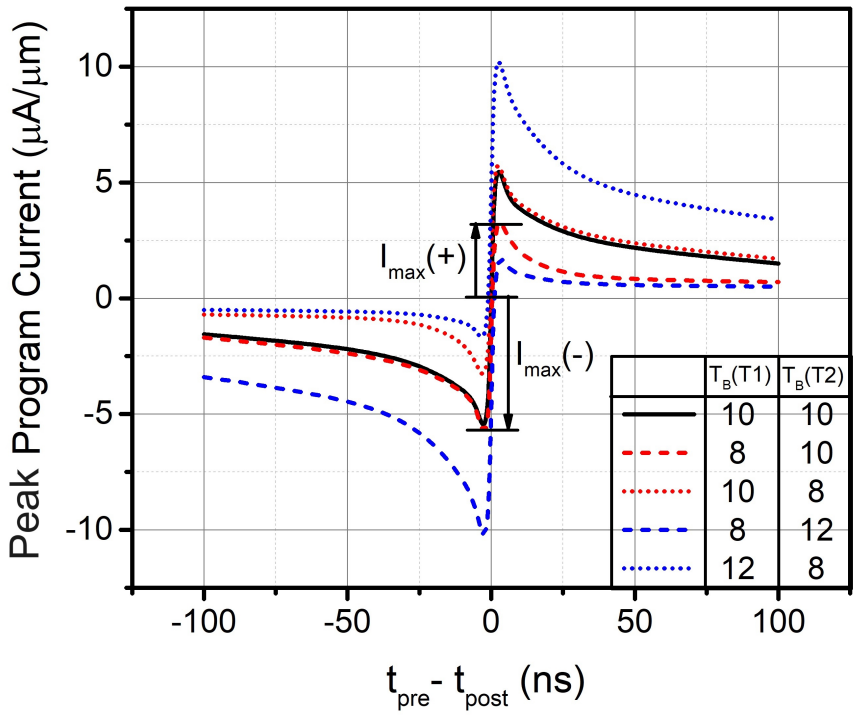

Figure 4.13: Impact of variation in body thickness (T_B) on the STDP characteristics.

An increase in T_B of the device leads to an increase in its cross-sectional area,

which subsequently lowers the device's resistance. Consequently, the current through the device increases. If a mismatch arises in the T_B of the two devices, such as device T_1 having a higher T_B compared to device T_2, it leads to different current flow behaviors during different spiking scenarios. For instance, when the post-synaptic neuron spiking event follows the pre-synaptic neuron spiking event, there is an increase in current flow into the HM layer. This is due to the increased cross-sectional area and reduced resistance of device T_1. On the other hand, when the pre-synaptic neuron firing event follows the post-synaptic neuron firing event, the current flow through the HM layer is relatively smaller. This is because the device T_2 is unable to effectively sink more current, resulting in a mismatched STDP response. As a result, such asymmetrical behaviors in the current flow can lead to asymmetrical STDP characteristics, impacting the efficacy of the learning process in the network. The implications of these variations will be explained in the next section.

4.2.2 Gate oxide thickness

The gate oxide thickness (t_{ox}) of the devices is another factor influencing the symmetry of the STDP characteristics. In the context of the proposed devices, when both devices have identical t_{ox} values, symmetrical STDP characteristics were obtained. This condition was met when both devices had $t_{ox} = 5nm$, and T_2 had a 20% larger channel width compared to T_1. However, due to the inherent variations in the manufacturing process, the gate oxide thickness can exhibit variations across different devices. This variation can lead to asymmetrical

STDP characteristics. To study this phenomenon, the gate oxide thickness of both devices was independently varied within a range of $\pm 1nm$ from the mean value of $5nm$. The resulting STDP characteristics are shown in Fig. 4.14.

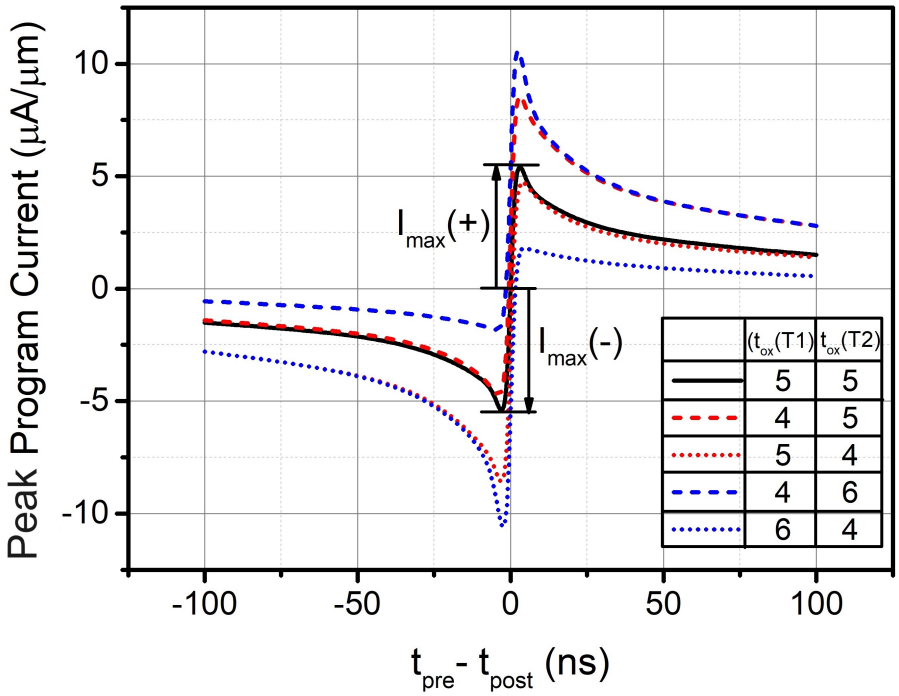

Figure 4.14: Impact of variation in oxide thickness (t_{ox}) on the STDP characteristics.

As t_{ox} is reduced, the coupling between the gate and the channel improves, leading to a higher rate of BTBT generation when a voltage spike occurs. This results in a larger accumulation of holes within the channel and a more pronounced reduction in the potential barrier. Consequently, the current flowing through the device increases. In cases where there is a mismatch in the gate oxide thickness between the two devices, such as when T_1 has a smaller t_{ox} compared to T_2, an increase in current flow into the HM layer is observed when the post-neuronal spiking event follows the pre-neuronal spiking event. Conversely, when the pre-neuronal firing event succeeds the post-neuronal firing event, the current

flow through the HM layer is hindered due to the limited capacity of T_2 to sink additional current.

Figs. 4.13 and 4.14 illustrate that when the pair of devices (T_1 and T_2) exhibit skewed device parameters (T_B and t_{ox}), significant asymmetry is observed in the resulting STDP characteristics. Such asymmetry can negatively impact the training of the SNN. Consequently, it becomes crucial to establish quantifiable boundaries for device variations that yield acceptable STDP characteristics even in the presence of process-induced variations. To accomplish this, a parameter called the Skew Difference (SD) has been introduced, which serves as a metric to gauge the degree of imbalance in the STDP characteristics. This parameter is defined as follows:

$$SD = ||I_{max}(+)| - |I_{max}(-)|| \tag{4.5}$$

$I_{max}(+)$ represents the maximum current flowing through the HM layer in the positive time window, while $I_{max}(-)$ represents the maximum current in the negative time window. The positive time window corresponds to cases where $(t_{post} - t_{pre}) > 0$, and the negative time window corresponds to cases where $(t_{post} - t_{pre}) < 0$. The behavior of SD with the percentage variation in T_B and t_{ox} of device T_1 with respect to T_2 is illustrated in Fig. 4.15.

When the STDP characteristics are symmetrical, $SD = 0$ is obtained, and in cases where the STDP characteristics are asymmetrical, SD will be greater than zero. The acceptable range for SD can vary depending on the specific

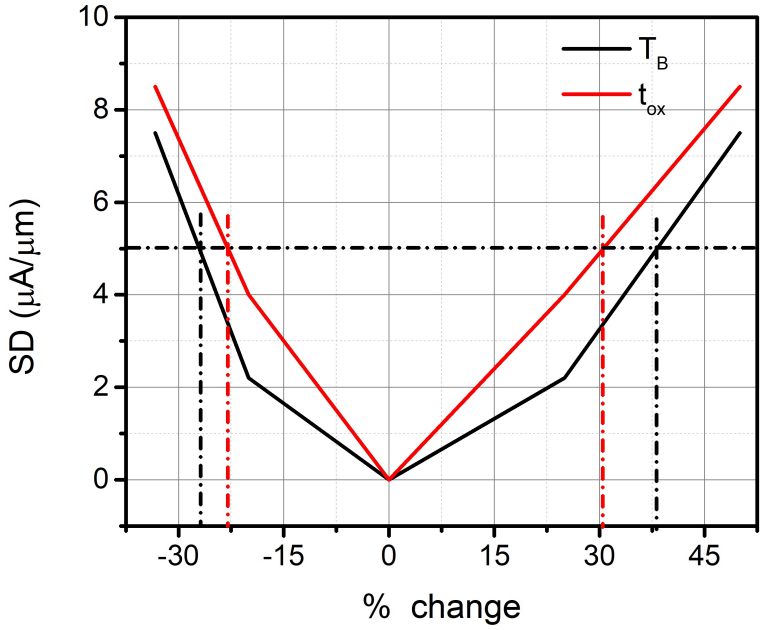

Figure 4.15: *SD* plotted as a function of the percentage variation in T_B and t_{ox} of device T_1 compared to T_2.

application's requirements. For example, assume that we define acceptable STDP characteristics for the pair of devices T_1 and T_2 as having $SD = 5\mu A/\mu m$. From Fig. 4.15, it is apparent that in order to achieve this SD value, the mismatch between device parameters (T_B and t_{ox}) should not exceed 20%.

4.3 Conclusions

In this chapter, using a well-calibrated hierarchical simulation framework, a Ge-based device has been demonstrated that enables on-chip unsupervised learning using STDP in an SNN. The proposed circuit generates a current, which depends exponentially on the temporal correlation of spiking events in the pre- and the post-synaptic neuronal layer. This current modulates the conductance of the synapse in accordance with the STDP learning rule. The proposed imple-

mentation requires 2-3$\times$ fewer transistors to implement the STDP learning rule compared to existing literature and is expected to result in an area- and energy-efficient implementation of the SNN. Further, the application of the proposed device to train an SNN to recognize handwritten digits in the MNIST dataset is demonstrated.

Chapter 5

A Ternary Spiking Neural Network

In this chapter, an implementation of ternary SNN is proposed. In the first part of this chapter, a novel device called Dual-Pocket Tunnel Field Effect Transistor (DP-TFET) is proposed and examined. Subsequently, a pair of DP-TFETs have been employed to implement a ternary inverter. Later, in the second part of this chapter, a ternary neuron is implemented using the proposed DP-TFET with appropriate biasing. The ternary neuron is employed in a ternary SNN, and the network is trained in an unsupervised manner using STDP. The behavior of the proposed ternary SNN implementation is investigated in detail.

5.1 Dual-Pocket Tunnel Field-effect Transistor

In this section, a novel device called a Dual-Pocket Tunnel Field Effect Transistor (DP-TFET) is proposed and examined. The work presented in this subsection is published in [83].

5.1.1 Tunnel Field-effect Transistor

This work proposes a special type of Tunnel Field-Effect Transistor (TFET) called a Dual Pocket Tunnel Field-Effect Transistor. Hence, for completeness, background information on TFETs is provided in this section. MOSFETs, which operate on the principle of thermionic emission of carriers over a potential barrier, are constrained by a fundamental limitation on achieving a minimum subthreshold swing, typically around $60\ mV/decade$ at room temperature [84]. This limitation poses a challenge when attempting to reduce supply voltage while maintaining operational speed and hampers further CMOS scaling [85–88]. As an alternative device, a TFET, which leverages the principle of band-to-band tunneling (BTBT), has emerged. TFETs offer distinctive advantages, including low OFF-state current (I_{OFF}) and a subthreshold swing (SS) below $60\ mV/decade$ at room temperature [89–92]. These properties enable a concurrent reduction in both static and dynamic power consumption within integrated circuits, making TFETs particularly promising for low-voltage, low-power applications. However, TFETs encounter a limitation in the ON-state current (I_{ON}), particularly evident at smaller supply voltages [61, 68, 93]. To surmount this challenge, innovative architectures, novel heterostructures, and advanced materials have been proposed [94–98, 111].

5.1.2 Device Structure and Simulation Model

Fig. 5.1 shows a schematic cross-sectional view of the proposed DP-TFET.

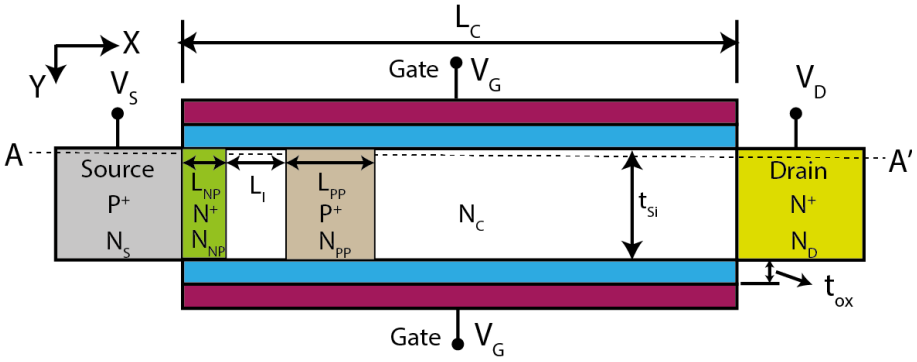

Figure 5.1: Structure of the proposed DP-TFET.

It consists of two pockets located on the source side of the device. The first pocket is n^+-doped with a concentration of N_{NP} and length of L_{NP}. The inclusion of this thin, fully depleted pocket adjacent to the source results in an increase in the built-in electric field at the source-channel junction [80]. Consequently, a sharp curvature of the conduction band at the source-channel junction is obtained, which leads to a reduced tunneling width. Hence, the probability of tunneling increases, and a smaller SS is obtained when compared to conventional TFET (C-TFET). This TFET with a single pocket adjacent to the source is referred to as a Single-Pocket TFET (SP-TFET). In this work, a second pocket, which is p^+-doped, with a concentration of N_{PP} and length L_{PP}, is employed at a controlled distance (L_I) from the first pocket. The existence of the p^+ pocket and its interaction with the n^+ pocket modify the energy bands at the source-channel junction such that a sharper band profile is obtained compared to the SP-TFET. Consequently, the proposed DP-TFET is expected to switch ON more abruptly and exhibit superior electrical characteristics compared to C-TFET and SP-TFET. In the later part of this work, the electrical characteristics of the proposed DP-TFET are compared with those of the SP-TFET and C-TFET.

The fabrication of the two pockets in the proposed DP-TFET can be done by first lithographically defining the region for the larger p^+ pocket using a suitable mask, followed by appropriate doping. Then, the second n^+ pocket can be fabricated by employing a tilted implant followed by spike annealing as proposed in [80]. The gate oxide used in this study is SiO_2. Other important device parameters are shown in Tab. 5.1.

Table 5.1: Device Parameters

Device Parameter	C-TFET	SP-TFET	DP-TFET
Supply Voltage (V_{DD}) (V)	1.0	0.5-1.0	0.5-1.0
Si film thickness (t_{Si}) (nm)	10	10	10
Channel Length (L_C) (nm)	25-100	25-100	25-100
Gate Oxide thickness (t_{ox}) (nm)	3	3	3
Gate workfunction (ϕ_m) (eV)	4.1	4.2	3.9-4.1
Channel Doping (N_C) (n-type) ($atoms/cm^3$)	1×10^{17}	1×10^{17}	1×10^{17}
Source Doping (N_S) (p-type) ($atoms/cm^3$)	1×10^{20}	1×10^{20}	1×10^{20}
Drain Doping (N_D) (n-type) ($atoms/cm^3$)	5×10^{18}	5×10^{18}	5×10^{18}
N^+ Pocket Doping (N_{NP}) (n-type) ($atoms/cm^3$)	-	3×10^{19}	1×10^{19}-4×10^{19}
N^+ Pocket Length (L_{NP}) (nm)	-	4	2-5
Length of intermediate region between pockets (L_I) (nm)	-	-	0-12
P^+ Pocket Doping (N_{PP}) (p-type) ($atoms/cm^3$)	-	-	5×10^{18}-2.5×10^{19}
P^+ Pocket Length (L_{PP}) (nm)	-	-	5-30

In this work, the simulations are done in Silvaco Atlas, version 5.22.1.R [99]. Non-local BTBT model has been used for simulations. Shockley-Read-Hall (SRH) recombination model has been taken into account. Moreover, the Band-Gap Narrowing (BGN) model has also been enabled to account for the presence of highly doped source and pocket regions. A concentration-dependent mobility model is also enabled. Tunneling through the gate oxide has been neglected, and the doping profiles are assumed to be abrupt [61, 66–68, 101, 102]. The narrow intrinsic region between the two pockets is prone to quantum confinement, and a bandgap widening model is implemented to include the effect of quantum confinement in the region. The model assumes a rectangular

infinite potential well and assumes zero density of states until the first sub-band in the conduction band and valence band is encountered, thus giving the effect of bandgap widening.

5.1.3 Electrical Characteristics

Next, the band diagrams of the proposed DP-TFET are compared with other devices. Subsequently, the root cause of its superior electrical characteristics is analyzed. For a fair comparison, in this work, the pocket doping and the pocket width of the SP-TFET have been optimized to obtain the best performance.

Fig. 5.2(a) compares the band diagrams at the onset of tunneling for the three TFET configurations, namely C-TFET, SP-TFET, and DP-TFET. For a fair comparison, the workfunction of the gate is adjusted in all three cases so that the current in the transfer characteristics takes off at $0\ V$. The gate voltage at which the current takes off in the transfer characteristics is referred as V_{OFF}. From Fig. 5.2(a), it can be inferred that the band profile at the source-channel junction is the sharpest for DP-TFET, followed by SP-TFET and C-TFET. The narrow tunneling width in the DP-TFET is expected to boost the I_{ON}.

Fig. 5.2(b) compares the transfer characteristics of the three devices. It is observed that DP-TFET switches ON more abruptly compared to C-TFET and SP-TFET. To quantify the improvement in the electrical characteristics, we extract the point subthreshold slope SS_{point} from the transfer characteristics as follows:

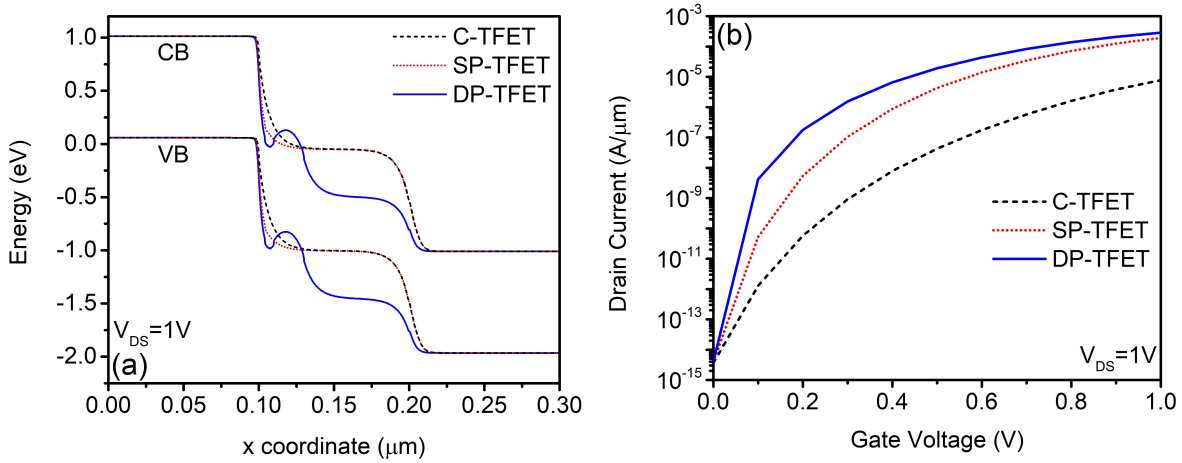

Figure 5.2: Comparison of C-TFET, SP-TFET ($N_{NP} = 3 \times 10^{19}/cm^3$, L_{NP}=4 nm) and DP-TFET (N_{NP}=$3 \times 10^{19}/cm^3$, L_{NP}=4 nm, L_I=6 nm N_{PP}=$1 \times 10^{19}/cm^3$, and L_{PP}=20 nm) at V_{DS}=1 V (a) band diagrams at the onset of tunneling (b) transfer characteristics

$$SS_{point} = \left[Max \left(\frac{d(log(I_D))}{d(V_{GS})} \right) \right]^{-1} \tag{5.1}$$

where I_D is the drain current and V_{GS} is the gate voltage. Furthermore, for future low-power applications average subthreshold slope (SS_{avg}) is an important measure for a device. Therefore, we have extracted SS_{avg} as follows:

$$SS_{avg} = \frac{V_{th} - V_{OFF}}{log(I_{Vt}) - log(I_{OFF})} \tag{5.2}$$

where V_{th} is the threshold voltage of the transistor, I_{Vt} is the drain current when the gate voltage is V_{th} and I_{OFF} is the drain current when the gate voltage is V_{OFF}. In this work, V_{th} is taken as the gate voltage when the drain current reaches $1 \times 10^{-7} A/\mu m$ with the drain terminal biased at the supply voltage. I_{ON} is the drain current when $V_{GS} = V_{DS} = V_{DD}$, where V_{DD} is the supply voltage. Another figure of merit useful in evaluating the performance of sub-60 mV/dec devices is I_{60}. It is defined as the drain current at which the point subthreshold

Table 5.2: Comparison of point SS, avg. SS, I_{ON}, and I_{60} for C-TFET, SP-TFET and DP-TFET

TFET	Point SS (mV/dec.)	Avg. SS (mV/dec.)	I_{ON} ($\mu A/\mu m$)	I_{60} ($\mu A/\mu m$)
C-TFET	39.3	70	7.7	1×10^{-5}
SP-TFET	24.2	41.8	90	4×10^{-3}
DP-TFET	12.2	25.4	290	6×10^{-2}

swing is 60 $mV/dec.$ The I_{60} current is essentially independent of the gate workfunction used, and those devices whose I_{60} is of the order of 1-10 $\mu A/\mu m$ can be considered as contenders to the state-of-the-art CMOS devices [103].

The extracted electrical parameters for the three devices are shown in Tab. 5.2. The point subthreshold swing reduces by 69% compared to C-TFET and by 49% compared to SP-TFET. The average subthreshold swing reduces by 64% compared to C-TFET and by 39% compared to SP-TFET. Moreover, the I_{ON} improves by $37\times$ compared to C-TFET and by $3\times$ compared to SP-TFET. Moreover, the extracted I_{60} for the DP-TFET is four orders greater than that of C-TFET and one order more compared to that of SP-TFET. Thus, the proposed technique of employing a double pocket is effective in improving the electrical characteristics of a TFET.

In Tab. 5.3, the I_{60} of the DP-TFET is compared with other TFETs published in the literature. The I_{60} of the DP-TFET is lower than many other TFETs proposed in the literature.

5.2 Ternary Inverter implementation

In this section, a Standard Ternary Inverter (STI) has been implemented by employing a pair of DP-TFETs. Through the utilization of a well-calibrated

Table 5.3: DP-TFET's I_{60} benchmarked against different TFETs published in literature

Published work	Material system	Geometry	I_{60} ($\mu A/\mu m$)
Gandhi [104]	Si	GAA NW	5×10^{-5}
Knoll [105]	s-Si	Planar	1×10^{-6}
Memisevic [106]	InAs/GaAsSb/GaSb	GAA NW	0.056
Wu [107]	Si-Ge GaAs-InAs	Planar	6×10^{-4} 1
Li [57]	Ge pocket	Planar	0.2
Cheng [108]	Si-Ge (T=77 K) Si-Ge (T=4.9 K)	Planar	0.08 0.1
DP-TFET (This work)	Si(EOT=3 nm) Si(EOT=0.75 nm)	Planar	0.06 0.45
Wang [43]	GeSn/SiGeSn	Planar	0.7
Lu [109]	AlGaSb/InAs	Planar	3
Zhang [62]	Bi_2Se_3	2D	10

two-dimensional device simulation framework, it is demonstrated that by carefully selecting suitable doping concentrations and lengths for the dual pocket, the device can exhibit Voltage Transfer Characteristics (VTC) resembling those of a ternary inverter. This unique VTC configuration encompasses three distinct stable output voltage levels. The emergence of these ternary inverter characteristics can be attributed to the combined influence of two tunneling mechanisms intrinsic to the device: (1) gate bias-independent within-channel tunneling and (2) gate bias-dependent source-channel tunneling. Furthermore, the ternary inverter can be operated at various supply voltage levels by controlling the pocket's doping concentration. The work presented in this section is published in [112].

5.2.1 Motivation

At sub-micron technologies, the power density has been increasing significantly with scaling, particularly due to the increasing leakage currents. To mitigate the increasing power density in the chip, the adoption of multi-valued logic, such as

ternary logic [113–115] is attractive. Ternary logic introduces three stable states, in contrast to the binary logic's two, offering enhanced information storage capacity per unit area. In the context of Very Large Scale Integration (VLSI) chip design, this directly translates to a reduced footprint and assists in minimizing the number of pins and interconnects required in the resultant chip. Numerous implementations of ternary inverters based on CMOS and Carbon Nanotube (CNT) Field-effect Transistors have been proposed in the literature [116, 117]. However, CMOS-based designs often necessitate multiple voltage sources and resistors, leading to increased area and power consumption. The resistive load CNTFET approach requires an off-chip resistor [118], while CNT diameter-controllable multi-threshold transistor-based ternary inverters necessitate at least six transistors [119]. More recently, TFET-based ternary inverters have emerged as a solution, exhibiting reduced area and static power dissipation compared to CMOS-based counterparts [120]. Nonetheless, these TFET-based inverters often face challenges in terms of their Static Noise Margins (SNM), which are comparatively low.

Ternary TFETs face intrinsic limitations due to their inherently low ON-state current (I_{ON}), especially when operating at smaller supply voltages, making them inferior to CMOS technology. However, in the context of neuromorphic applications that aim to replicate the functioning of the human brain, specific requirements emerge. These applications prioritize a high degree of parallelization and operation at low frequencies [121–123]. In scenarios where low power consumption, low-frequency operation, and a highly parallel architecture are

crucial, ternary TFETs can offer advantages over traditional CMOS technology.

5.2.2 Device Structure and Simulation Model

SiGe with 20% Ge content is used as a base material in the proposed DP-TFET to operate the ternary inverter at a lower supply voltage compared to a Si-based DP-TFET. An SiO_2 gate oxide is employed. Essential simulation parameters of the device can be found in Table 5.4.

Table 5.4: Device Parameters of the proposed DP-TFET

Device Parameter	Symbol	Value
Supply Voltage (V)	V_{DD}	0.8
Channel thickness (nm)	T_{ch}	10
Channel Length (nm)	L_C	100
Gate Oxide thickness (nm)	t_{ox}	3
Gate workfunction (eV)	ϕ_m	4.0
Channel Doping (n-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}
N^+ Pocket Doping (n-type) ($atoms/cm^3$)	N_{NP}	1.5×10^{19}
N^+ Pocket Length (nm)	L_{NP}	4
Length of intermediate region between pockets (nm)	L_I	6
P^+ Pocket Doping (p-type) ($atoms/cm^3$)	N_{PP}	2×10^{19}
P^+ Pocket Length (nm)	L_{PP}	20

For the present study, simulations have been conducted using Synopsys Sentaurus, version N-2017.09-SP2 [100]. The simulations have been performed by enabling the non-local band-to-band tunneling (BTBT) model, with fitting parameters $A_{BTBT} = 6.5 \times 10^{15} cm^{-3} s^{-1}$ and $B_{BTBT} = 8.1 \times 10^6 V/cm$ [110]. Shockley-Read-Hall (SRH) recombination model has been taken into account. Furthermore, the Slotboom Band-Gap Narrowing (BGN) model has been activated to account for the influence of highly doped source and pocket regions. A concentration-dependent Philips unified mobility model is also enabled. Tunneling through the gate oxide has been neglected. Additionally, for the sake of

simplicity, abrupt doping profiles have been assumed [61, 66–68, 101, 102]. To ensure the accuracy of the simulation model, calibration has been carried out based on the experimental results presented for SiGe TFET in [110]. In Fig. 5.3, a comparison is shown between the results produced by the simulation model and the results obtained from measurements for the SiGe device, as reported in [110].

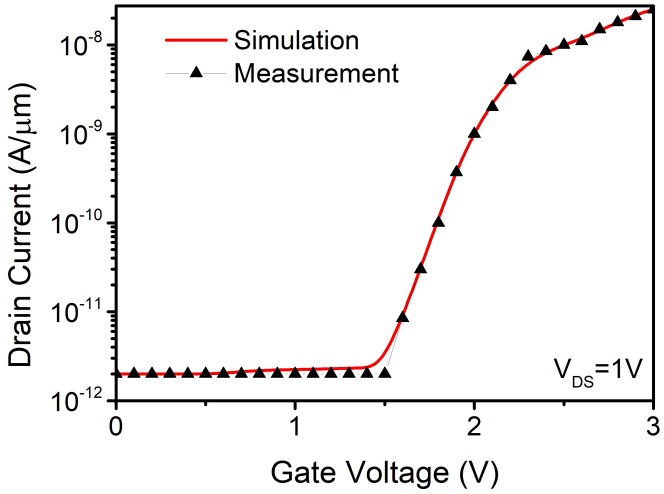

Figure 5.3: Calibration of the simulation model for SiGe TFET. The measurement data were taken from Fig. 2(a) in [110].

A good match between the simulation model's results and the experimental results confirms the suitability of the simulation model. The narrow intrinsic region between the two pockets is particularly susceptible to quantum confinement effects. In order to accurately account for this phenomenon, a bandgap widening model has been incorporated. This model operates under the premise of an infinite rectangular potential well, assuming a zero density of states until the first sub-band is encountered. This approach effectively captures the effect of bandgap widening resulting from quantum confinement [124].

5.2.3 Device operation

In this section, the operation of DP-TFET is discussed in detail. In Fig 5.4, a comparison is shown between the band diagrams at two distinct gate voltages: 0V and 0.4V. It becomes evident that tunneling occurs within the channel due to band overlaps [126]. This phenomenon is attributed to the presence of a highly doped p^+ pocket within the channel. The inset shown in Fig. 5.4 demonstrates the BTBT generation rate at V_{GS} = 0.4V, thereby confirming the existence of within-channel tunneling. However, a further increase in the gate voltage to 0.55V does not yield a substantial increase in current (as depicted in Fig. 5.6), owing to the misaligned valence band in the source and conduction band in the channel.

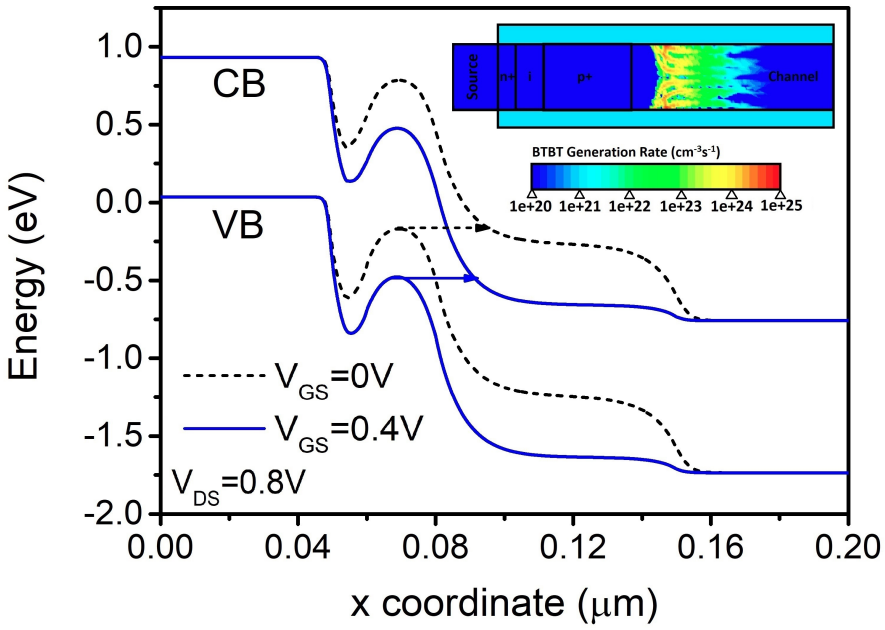

Figure 5.4: DP-TFET band diagrams along cut-line AA' for within-channel BTBT at different gate voltages (N_{NP}=1.5×10^{19}/cm^3, L_{NP}=4 nm, L_I=6 nm, N_{PP}=2×10^{19}/cm^3, L_{PP}=20 nm). The inset shows the BTBT generation rate at V_{GS} = 0.4V.

Once the gate voltage surpasses 0.55V (as shown in Fig. 5.5), an abrupt increase in current is observed due to the alignment in the valence band in the source and the conduction band in the channel. It can be observed that the drain current arising from within-channel tunneling remains unaffected by the gate voltage and exists even after the onset of source-channel tunneling, a fact validated by the band diagram presented in Fig. 5.5. The inset within Fig. 5.5 corroborates that the BTBT at the source-channel interface notably exceeds the within-channel tunneling as it showcases the BTBT generation rate at V_{GS} = 0.8V.

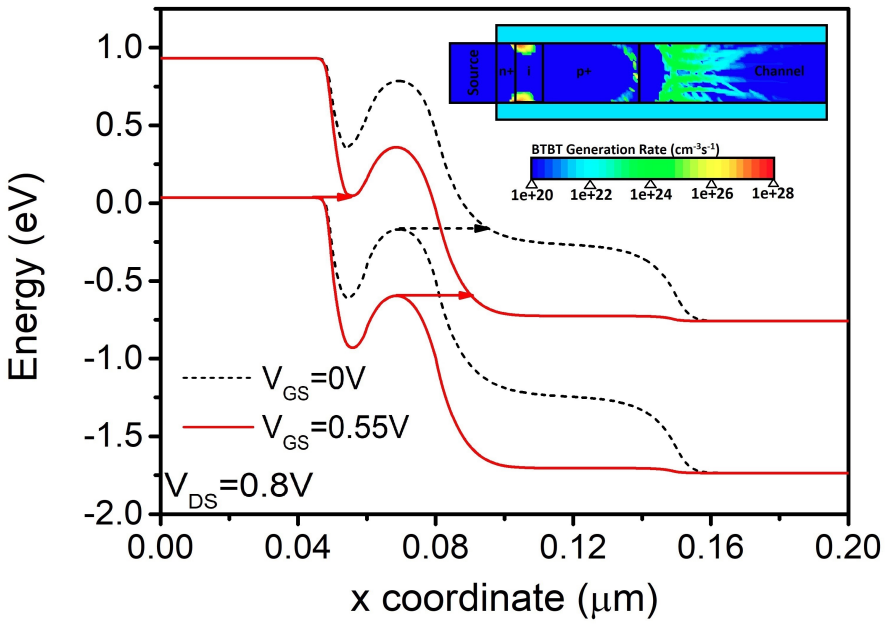

Figure 5.5: DP-TFET band diagrams along cut-line AA' for BTBT at the source–channel junction at different gate voltages (N_{NP}=1.5×10^{19}/cm^3, L_{NP}=4 nm, L_I=6 nm, N_{PP}=2×10^{19}/cm^3, L_{PP}=20 nm). The inset shows the BTBT generation rate at V_{GS} = 0.8V.

It should be noted that once the electrons are introduced into the channel through source-channel tunneling, they encounter a hump in the channel due to the presence of the highly doped p^+ pocket. These electrons must overcome this

potential barrier through thermionic emission to advance toward the drain and participate in conduction. This phenomenon explains why the increase in current does not exhibit the anticipated sharpness in TFETs.

To realize a ternary inverter, a p-type device was designed with similar characteristics as the n-type DP-TFET. The dimensions and doping concentrations of the pockets for the p-type device correspond to those of the n-type device. However, the doping type for all regions is the opposite of the n-type DP-TFET. The p-type device has a gate material with a workfunction of $5.35\ eV$. All other device parameters are similar to those specified in Tab. 5.1. As demonstrated in Fig. 5.6, the n-type and p-type devices exhibit nearly symmetrical transfer characteristics, a quality that is desirable for an inverter.

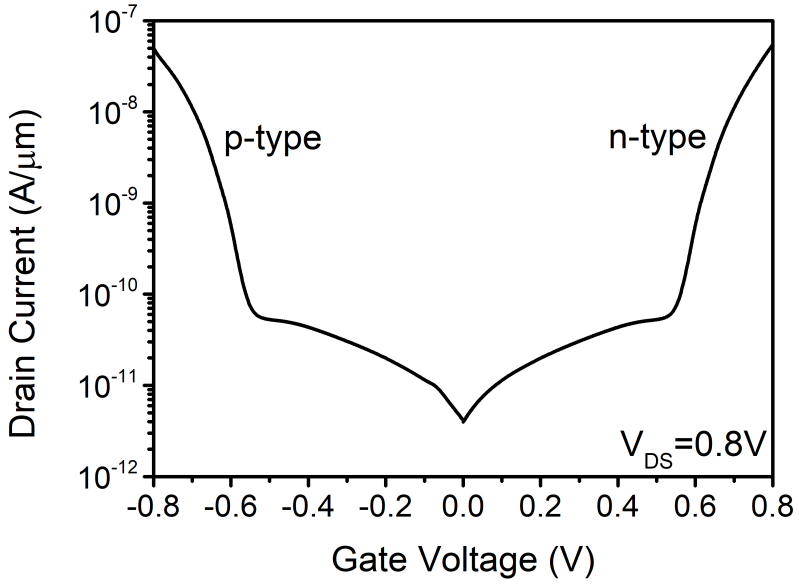

Figure 5.6: DP-TFET transfer characteristics (N_{NP}=1.5×10^{19}/cm^3, L_{NP}=4 nm, L_I=6 nm, N_{PP}=2×10^{19}/cm^3, L_{PP}=20 nm).

In conclusion, the current conduction within the device can be attributed to two predominant tunneling mechanisms: (1) within-channel tunneling and (2)

source–channel tunneling. The source–channel tunneling current is accountable for the binary inverter Voltage-Transfer Characteristic (VTC), while the within-channel tunneling current contributes to the establishment of a stable third voltage level around V_{DD}/2.

5.2.4 Ternary inverter

The ternary VTC derived from a pair of p-type and n-type DP-TFETs are shown in Fig. 5.7. Definitions for various voltage levels associated with the STI are introduced to facilitate easy explanation. The maximum input voltage that is regarded as logic '0' is denoted as V_{IL}, with the corresponding output voltage denoted as V_{OH} (logic '2'). Conversely, the minimum input voltage that is regarded as logic '2' is denoted as V_{IH}, with the corresponding output voltage being V_{OL} (logic '0'). Additionally, there exist two intermediate input voltages, V_{IML} and V_{IMH} (both representing logic '1'), accompanied by corresponding output voltages (both logic '1') V_{OMH} and V_{OML}, respectively. Achieving a distinct intermediate state in a ternary inverter necessitates the drain current to remain unaffected by variations in the gate voltage for $V_{IML} \leq V_{GS} \leq V_{OML}$. If the current displays significant fluctuations within the above mentioned gate voltage range, a stable third output voltage level might not be reliably established.

When the gate voltage falls below V_{IL}, the pull-up device operates in saturation while the pull-down device remains in the cut-off state, leading to an output voltage of V_{OH}. Conversely, when the gate voltage surpasses V_{IH}, the pull-down device enters saturation while the pull-up device is cut off, resulting

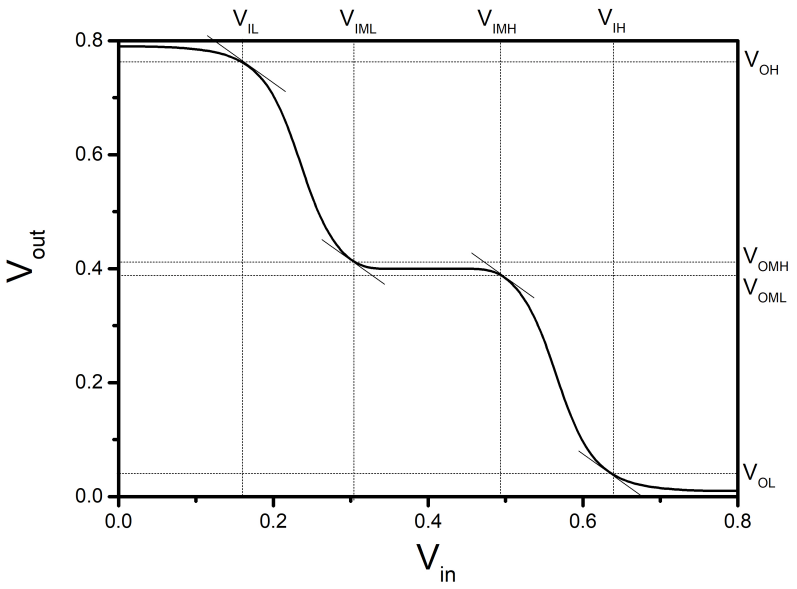

Figure 5.7: DP-TFET exhibiting ternary inverter VTC (N_{NP}=1.5×10^{19}/cm^3, L_{NP}=4 nm, L_I=6 nm, N_{PP}=2×10^{19}/cm^3, L_{PP}=20 nm).

in an output voltage of V_{OL}. For gate voltages near V_{DD}/2 (i.e., falling within the range of $V_{IML} < V_{in} < V_{IMH}$), both the pull-up and pull-down devices exhibit weak conduction, causing the output voltage level to rely on the resistive voltage divider network created between these devices. A stable third output voltage level mandates a balanced current between the pull-up and pull-down devices.

The butterfly curve, depicted in Fig. 5.8, is generated by reflecting the DP-TFET ternary VTC discussed earlier. Determining the noise margin entails measuring the diagonal of the largest square that can be inscribed within the butterfly curve. Compared to a binary CMOS inverter with only two noise margins, a ternary inverter has four noise margins. The Static Noise Margin (SNM) for the ternary inverter is derived from the smallest of these four diagonals, which is found to be 120 mV. To enhance the SNM, broadening the intermediate voltage level state within the ternary inverter would be beneficial.

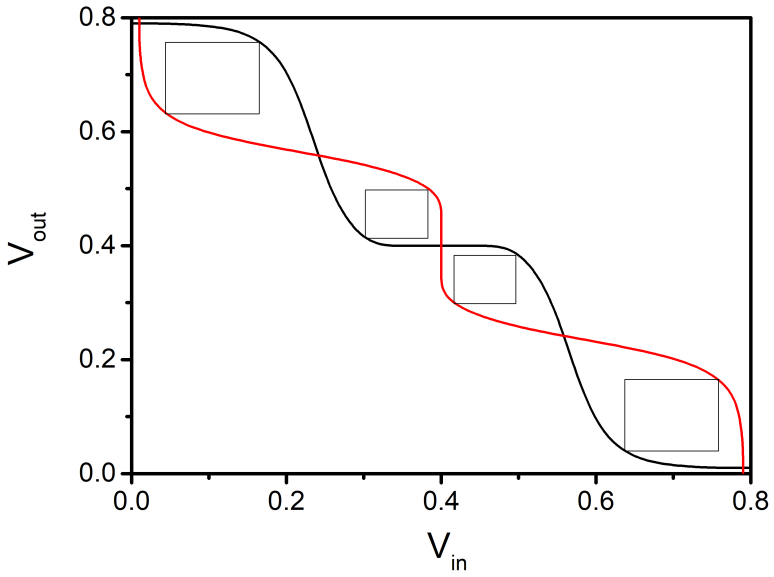

Figure 5.8: Butterfly curve for DP-TFET ternary VTC (N_{NP}=1.5×10^{19}/cm^3, L_{NP}=4 nm, L_I=6 nm, N_{PP}=2×10^{19}/cm^3, L_{PP}=20 nm).

It is important to highlight that the within-channel tunneling current is significantly lower than the source–channel tunneling current, primarily due to the larger tunneling width as shown in Fig. 5.4. This discrepancy in tunneling currents can result in time delays that are imbalanced across different transitions, owing to the substantial difference in average charging and discharging currents. For instance, suppose the input to the ternary inverter is initially at GND (0V) and then gradually rises toward V_{DD}. The output will drop from V_{DD} as the output capacitance discharges. Given the smaller within-channel tunneling current, the n-type DP-TFET will require more time to discharge the output capacitance to V_{DD}/2. On the other hand, the transition of the output from V_{DD}/2 to GND (0 level) will be driven by a considerably higher source–channel tunneling current. Consequently, the transition from V_{DD} to V_{DD}/2 will be slower than the transition from V_{DD}/2 to GND. In circuit design, these imbalances in delay must be taken

into account. Nevertheless, the slightly increased delay caused by the lower within-channel tunneling current can be tolerated in applications that involve low-frequency neuromorphic operations.

5.2.5 Device Optimization

This section examines the effect of various parameters of the DP-TFET on the ternary inverter VTC. The DP-TFET parameters are varied one at a time while keeping the other parameters at their nominal values. The within-channel tunneling current should be independent of the gate voltage to obtain a stable intermediate output voltage level. It is also desirable to increase the SNM of the resulting ternary inverter. We modify the DP-TFET's parameters to increase SNM and obtain a stable intermediate output voltage level.

5.2.5.1 Effect of change in N_{NP}

The n^+ pocket doping concentration (N_{NP}) is varied from $1 \times 10^{19} - 2 \times 10^{19}$ cm^{-3} around its nominal value of $1.5 \times 10^{19} cm^{-3}$ while keeping the other parameters at their nominal values (L_{NP}= 4nm, L_i= 6nm, N_{PP}= 2×10^{19} cm^{-3}, L_{PP}= 20nm). It can be observed from the band diagram in Fig. 5.9(a) that an increase in N_{NP} results in an increase in the sharpness of the band profile at the source-channel junction. Consequently, a higher N_{NP} results in an earlier onset of source-channel tunneling than a lower N_{NP}. Further, N_{NP} should be chosen to obtain symmetrical ternary inverter voltage transfer characteristics. If N_{NP} is increased to $2 \times 10^{19} cm^{-3}$, then the tunneling initiates at the source-channel

junction at a much smaller V_{GS} (0.5V), and ternary inverter VTC can be obtained at a smaller V_{DD} (0.65V). Thus, a higher N_{NP} may be used for low-power applications. However, the SNM of the resulting ternary inverter would be lower in this case.

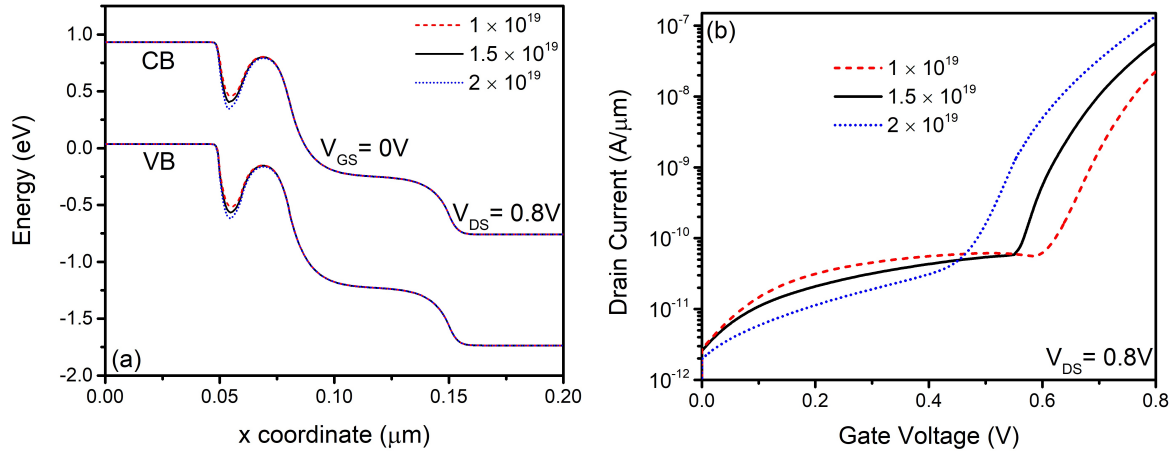

Figure 5.9: Band diagram and corresponding transfer characteristics for different N_{NP} (L_{NP}= 4nm, L_i= 6nm, N_{PP}= 2×10^{19} cm^{-3}, L_{PP}= 20nm) demonstrating earlier onset of source-channel tunneling at higher N_{NP}.

5.2.5.2 Effect of change in L_{NP}

The n^+ pocket length (L_{NP}) is varied from 3 - 5 nm around its nominal value of 4 nm while keeping the other parameters at their nominal values (N_{NP}= $1.5\times10^{19}cm^{-3}$, L_i= 6nm, N_{PP}= 2×10^{19} cm^{-3}, L_{PP}= 20nm). It can be observed from the band diagram in Fig. 5.10(a) that an increase in L_{NP} increases the sharpness of the band profile at the source-channel junction. Consequently, a larger L_{NP} results in an earlier onset of source-channel tunneling than a smaller L_{NP}. Thus, by opting for a higher L_{NP}, the ternary inverter could be operated at a smaller supply voltage but at the cost of a decrease in SNM. As a trade-off, the L_{NP} of 4 nm is selected, corresponding to the ternary inverter being operated at

V_{DD}=0.8V and an SNM of 120mV.

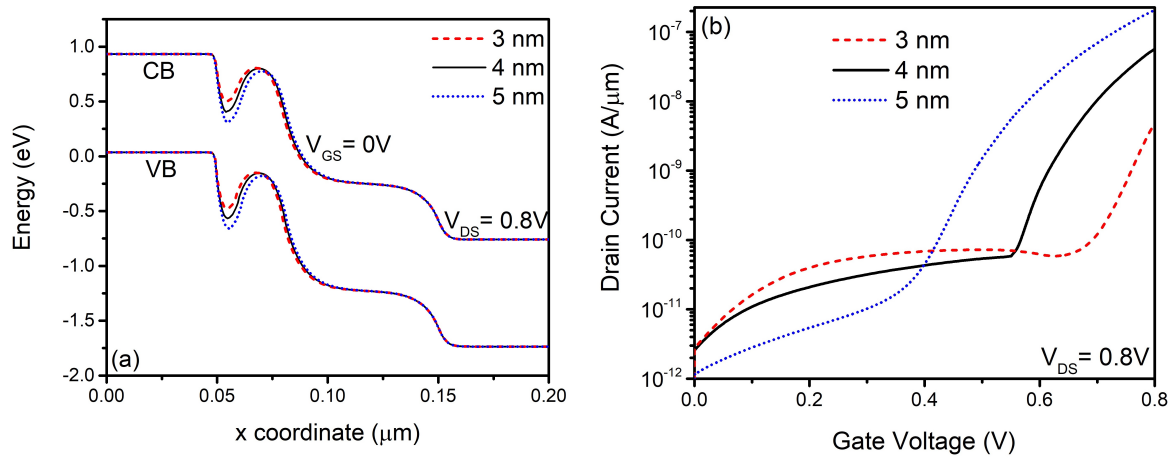

Figure 5.10: Band diagram and corresponding transfer characteristics for different L_{NP} (N_{NP}= $1.5 \times 10^{19} cm^{-3}$, L_i= 6nm, N_{PP}= 2×10^{19} cm^{-3}, L_{PP}= 20nm) demonstrating earlier onset of source-channel tunneling at higher L_{NP}.

5.2.5.3 Effect of change in L_i

The length of the intrinsic region between the two pockets (L_i) is varied from 4 - 8 nm around its nominal value of 6 nm while keeping the other parameters at their nominal values (N_{NP}= $1.5 \times 10^{19} cm^{-3}$, L_{NP}= 4nm, N_{PP}= 2×10^{19} cm^{-3}, L_{PP}= 20nm). It can be observed from the band diagram in Fig. 5.11(a) 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 and a delayed onset of source-channel tunneling.

5.2.5.4 Effect of change in N_{PP}

The p^+ pocket doping concentration (N_{PP}) is varied from 1.5×10^{19} - 2.5×10^{19} cm^{-3} around its nominal value of $2 \times 10^{19} cm^{-3}$ while keeping the other

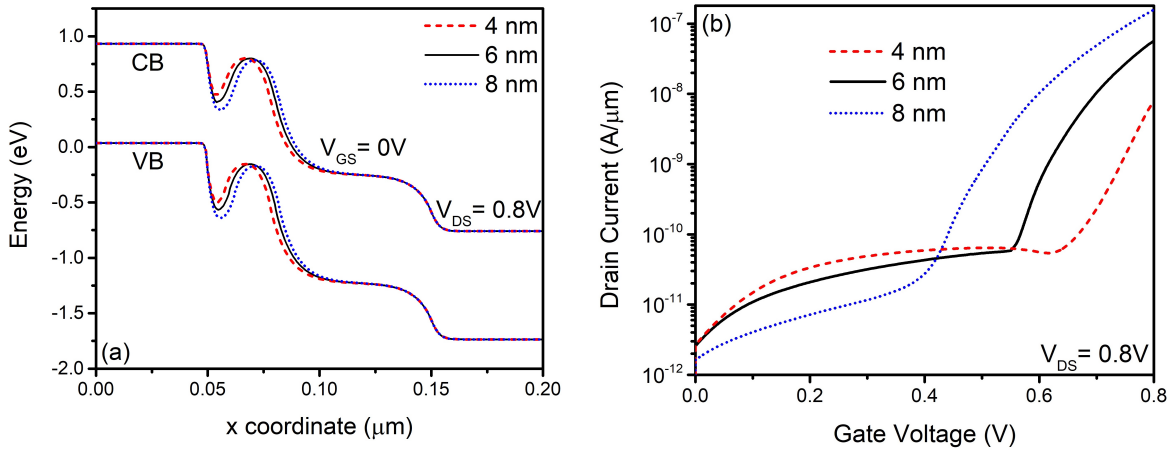

Figure 5.11: Band diagram and corresponding transfer characteristics for different L_i (N_{NP}= $1.5 \times 10^{19} cm^{-3}$, L_{NP}= 4nm, N_{PP}= 2×10^{19} cm^{-3}, L_{PP}= 20nm) demonstrating earlier onset of source-channel tunneling at higher L_i.

parameters at their nominal values (N_{NP}= $1.5 \times 10^{19} cm^{-3}$, L_{NP}= 4nm, L_i= 6nm, L_{PP}= 20nm). It can be observed from the band diagram in Fig. 5.12(a) that as N_{PP} is increased, the height of the barrier in the p+ pocket region (hump) increases, which results in an increase in within-channel tunneling current (due to increased band overlap). It can be observed that when N_{PP}= $1.5 \times 10^{19} cm^{-3}$, the height of the barrier is so low that there is no band overlap, and ternary inverter VTC is not obtained. It can also be observed that the step in the transfer characteristics is observable till N_{PP}=$2.5 \times 10^{19}/cm^3$. With a doping more than this, the ternary behavior is lost due to monotonically increasing current. Also, it is expected that as N_{PP} is increased, it would be difficult to obtain a stable third output voltage level since the gate voltage considerably changes within-channel tunneling current. A wider intermediate region is obtained for N_{PP}=$2.2 \times 10^{19} cm^{-3}$, resulting in a higher SNM of 140mV.

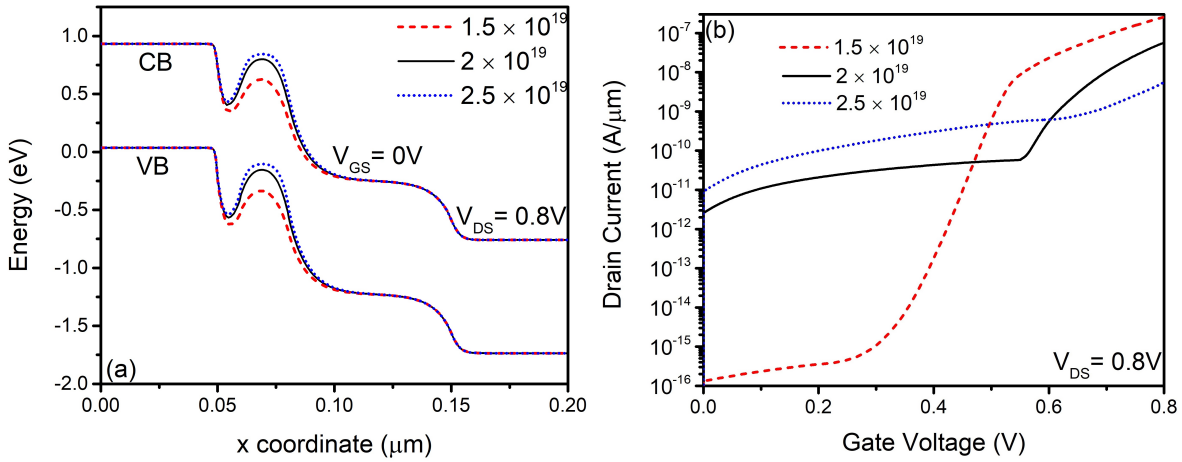

Figure 5.12: Band diagram and corresponding transfer characteristics for different N_{PP} (N_{NP}= $1.5 \times 10^{19} cm^{-3}$, L_{NP}= 4nm, N_{PP}= 2×10^{19} cm^{-3}, L_{PP}= 20nm).

5.2.5.5 Effect of change in L_{PP}

The p^+ pocket length (L_{PP}) is varied from 15 - 25 nm around its nominal value of 20 nm while keeping the other parameters at their nominal values (N_{NP}= $1.5 \times 10^{19} cm^{-3}$, L_{NP}= 4nm, L_i= 6nm, N_{PP}= $2 \times 10^{19} cm^{-3}$). It can be observed from the band diagram in Fig. 5.13(a) that as L_{PP} is decreased, the height of the barrier in the channel reduces. This results in a decrease in band overlap, and consequently, the within-channel tunneling current reduces, as seen in Fig. 5.13(b). Furthermore, with a smaller L_{PP}, the range of V_{GS} for which the current is relatively constant is small, which results in a thinner intermediate region and a smaller SNM.

5.2.6 Variability Analysis

In this section, the impact of variations of the device parameters is analyzed on the ternary inverter VTC. The p-type device parameters have also been

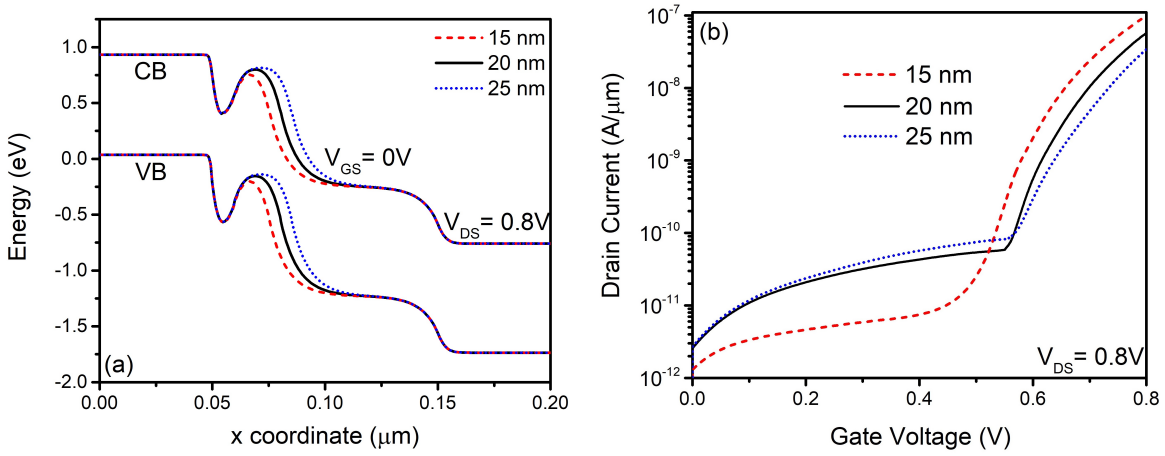

Figure 5.13: Band diagram and corresponding transfer characteristics for different L_{PP} (N_{NP}= $1.5 \times 10^{19} cm^{-3}$, L_{NP}= 4nm, N_{PP}= $2 \times 10^{19}\ cm^{-3}$, N_{PP}= $2 \times 10^{19} cm^{-3}$).

varied in accordance with the n-type device so as to obtain symmetrical transfer characteristics.

5.2.6.1 Gate workfunction

As the gate workfunction is varied, it is observed that the transfer characteristics shift along the X-axis (voltage axis), as demonstrated for the n-type DP-TFET in Fig. 5.14(a). When the workfunction of the n-type device is increased to 4.1eV, the source-channel tunneling initiates at a larger gate voltage when compared to a smaller gate workfunction. Consequently, the within-channel tunneling current flows for a larger range of gate voltage. The workfunction of the p-type device was decreased to 5.25eV to obtain symmetrical transfer characteristics. Due to the current being constant for a larger range of V_{GS} for a larger n-type gate workfunction, a wider intermediate state was obtained, as shown in Fig. 5.14(b). Thus, appropriate workfunctions need to be chosen for both p-type and n-type devices to obtain symmetrical ternary VTC and the required width of the

intermediate region.

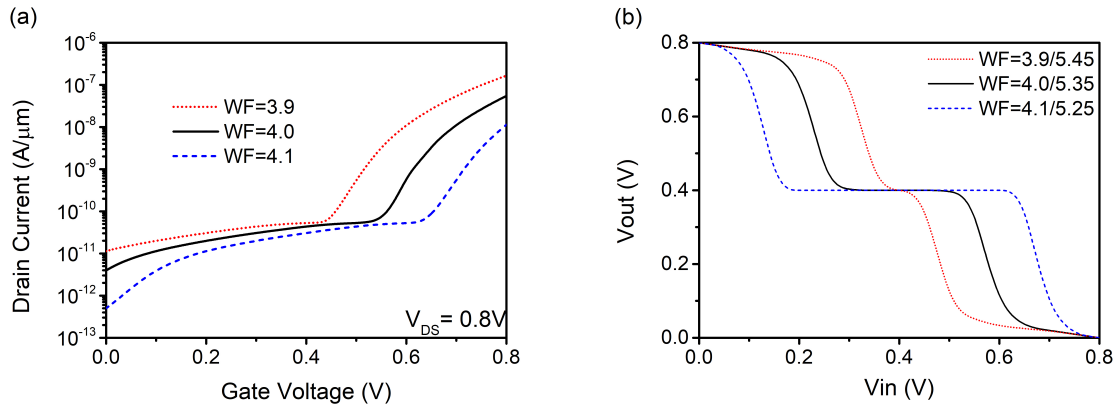

Figure 5.14: Effect of changing gate workfunction on (a) transfer characteristics of the TFET (b) ternary inverter VTC.

5.2.6.2 Gate dielectric thickness

As the gate dielectric thickness t_{ox} is reduced, the coupling of the gates with the channel becomes stronger. Consequently, the onset of source-channel tunneling occurs at a lower V_{GS}, as shown in the transfer characteristics in Fig. 5.15(a). It can be observed from Fig. 5.15(b) that for smaller gate dielectric thicknesses ($t_{ox} < 2.7nm$), the ternary behavior is lost. This is due to the current being constant for a relatively small range of V_{GS}. Thus, to observe two distinct mechanisms of tunneling and obtain a stable ternary behavior, t_{ox} must be maintained sufficiently high ($t_{ox} > 2.7nm$).

5.2.6.3 Channel thickness

The channel thickness is varied $\pm 10\%$ around its nominal value of 10 nm. Fig. 5.16(a) and 5.16(b) show the transfer characteristics and the corresponding VTC for different channel thicknesses. As the channel thickness decreases, the top and

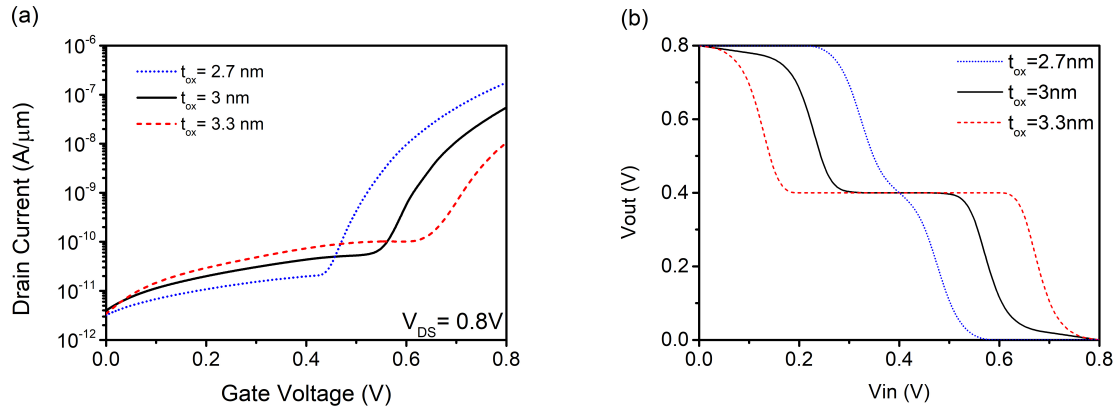

Figure 5.15: Effect of changing gate dielectric thickness on (a) transfer characteristics of the TFET (b) ternary inverter VTC.

the bottom gates come closer and control the channel potential more strongly. As a result, the valence band and the conduction band get misaligned, as shown in Fig. 5.16(c). Though a hump exists in the channel due to the highly doped p^+ pocket, the tunneling width inside the channel region increases significantly for a thinner channel (8 nm) compared to a thicker channel (10 nm). Hence, no within-channel tunneling is observed for a thin-channel (less than 8 nm) DP-TFET. In contrast, the source–channel tunneling enhances when the channel is made thinner due to a stronger coupling of the top and the bottom gates. Therefore, source–channel tunneling initiates at a lower gate voltage for a thinner channel. The combined result of the above effects is the suppression of the step-like behavior in the transfer characteristics for a thin-channel DP-TFET. Consequently, the ternary behavior disappears in the inverter implemented using thin DP-TFETs. Therefore, to obtain a stable ternary behavior, we need to keep the thickness of the channel sufficiently high (> 9 nm).

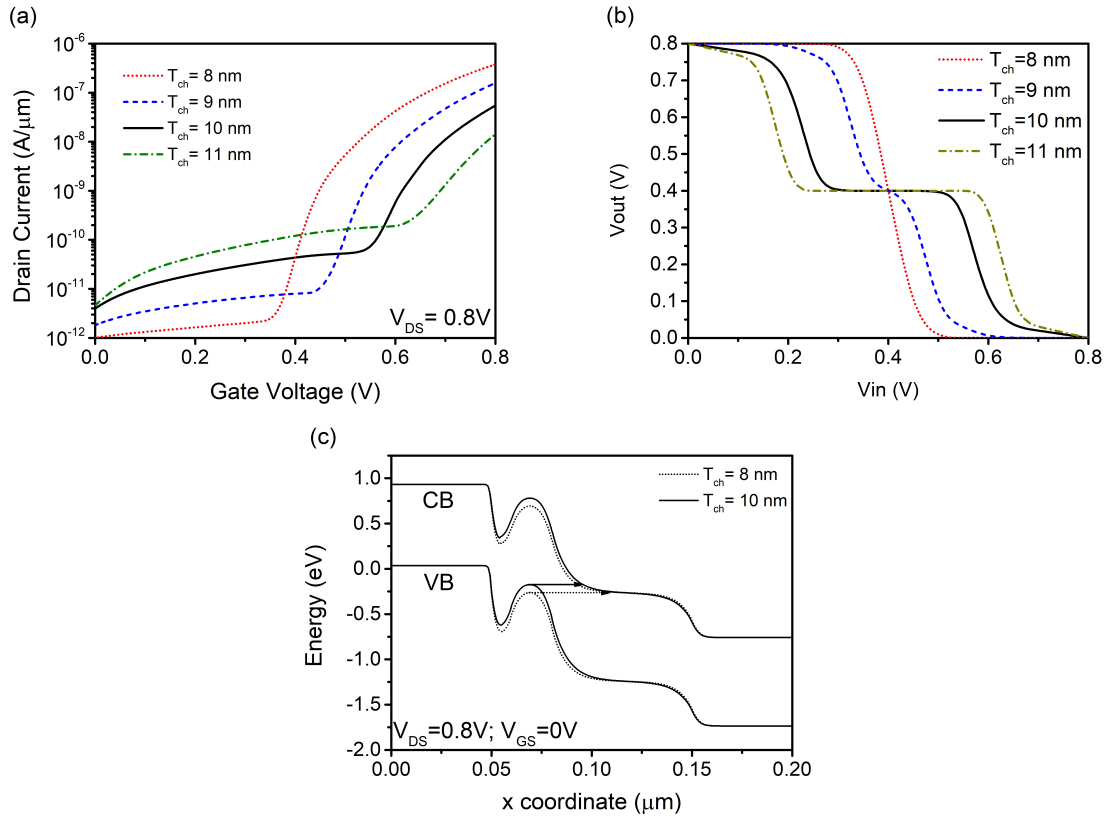

Figure 5.16: Effect of changing channel thickness on (a) transfer characteristics of the TFET (b) ternary inverter VTC (c) band diagram and tunneling width for within-channel tunneling.

5.2.6.4 Interface traps

Interface traps (D_{it}) at the semiconductor–oxide interface can deteriorate device performance. In TFETs, interface trap charges can also lead to trap-assisted tunneling (TAT). To account for the interface traps, a fixed charge of $\pm(5 \times 10^{12} - 1 \times 10^{13})cm^{-2}$ is considered at the semiconductor–oxide interface. Hurkx TAT model has been included to account for tunneling due to traps [100]. Fig. 5.17 shows the transfer characteristics and the corresponding VTC for different interface fixed trap charges. A positive fixed charge results in an early onset of source-channel tunneling, resulting in a narrow intermediate level in the corresponding VTC. Therefore, we should ensure that the interface trap charge

does not exceed $5 \times 10^{12} cm^{-2}$ to obtain a stable ternary operation.

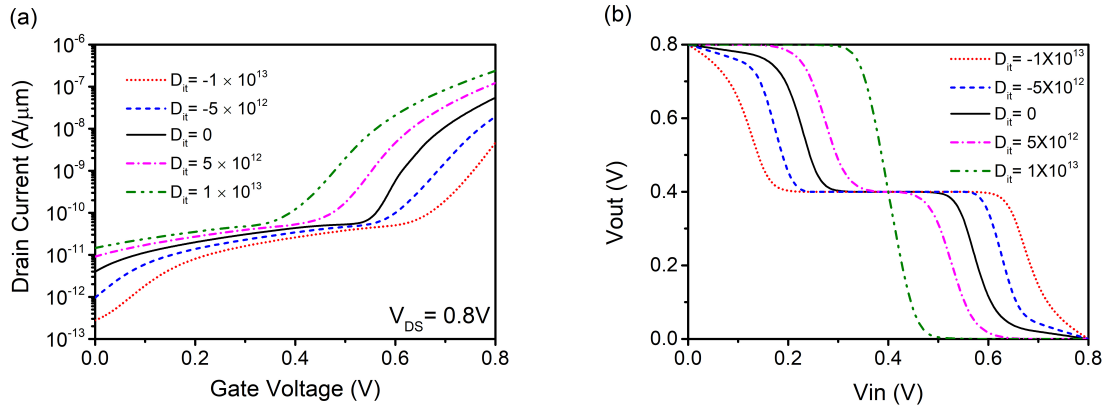

Figure 5.17: Effect of interface trap charge concentrations (D_{it} cm^{-2}) on (a) transfer characteristics of the TFET (b) ternary inverter VTC.

Thus, the design parameters, such as pocket dopings and width, gate workfunction, gate dielectric thickness, and channel thickness, must be chosen carefully to obtain a stable ternary operation with appreciable SNM. The optimum design parameters for the n-type DP-TFET are listed in Tab. 5.4. Further, the workfunction of the n-type (4.0 eV) and p-type (5.35 eV) devices must be matched to obtain symmetrical ternary VTC. A deviation from these optimized parameters can result in a deterioration in the SNM of the resulting ternary inverter.

5.3 Ternary Spiking Neural Network

This section proposes a novel implementation of a Ternary SNN and investigates it using a hierarchical simulation framework. The ternary neuron employed in the ternary SNN is implemented using the proposed DP-TFET, and the network is trained in an unsupervised manner using STDP. The behavior of the proposed ternary SNN implementation is investigated in detail. 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. The work presented in this section is published in [125].

5.3.1 Motivation

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. 5.18 compares

the reconstructed image from the MNIST dataset [82] 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.

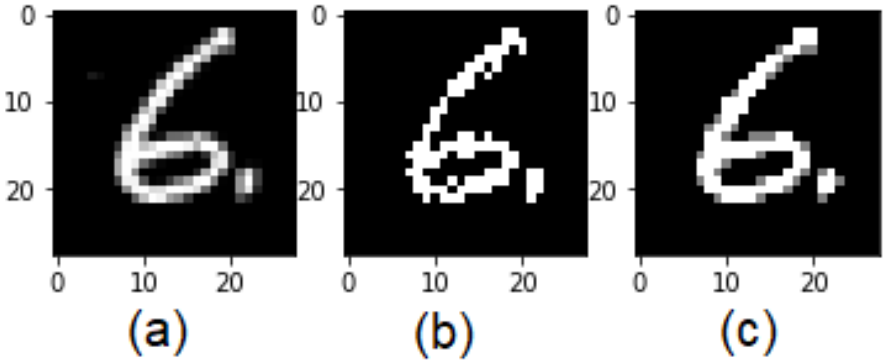

Figure 5.18: Comparison of the reconstructed input image in the MNIST dataset (a) Original image (b) Reconstructed image with binary spikes (c) Reconstructed image with ternary spikes [82]

A binary spike train of length 350 samples is created by rate encoding the original grayscale image in the MNIST dataset. The binary reconstructed image is produced by taking one such sample of the binary spike train. The binary reconstructed image shows that there is some loss of information as the probability of spiking activity for a pixel with low intensity is low, and that is represented as a black pixel in the binary reconstructed image. A ternary spike train is derived from the binary spike train itself by considering multiple samples from the binary spike train at a time. We defined 10 windows comprising of 35 samples each and the spike count was 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, in accordance with equation 5.4. The ternary reconstructed image presents one

such window. It can be observed that a ternary representation is much closer to the original image than its binary counterpart. Thus, it can be inferred that more information can be embedded in fewer number of ternary spike train (10 instances) than compared to a binary spike train (350 instances)

5.3.2 Ternary Spiking Neuron

This work employs a Ge-based DP-TFET, as shown in Fig. 5.19, to implement a ternary spiking neuron.

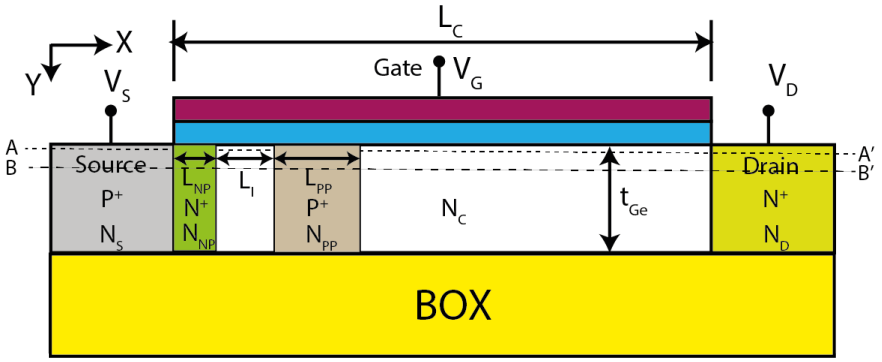

Figure 5.19: The DP-FET used to implement a ternary spiking neuron.

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 gate oxide used is HfO_2. Table 5.5 contains other device simulation parameters.

In this work, Germanium is preferred over Silicon. This is attributed to its

Table 5.5: 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)	ϕ_m	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

smaller bandgap and the prevalence of a dominant direct tunneling mechanism [5]. This leads to a higher BTBT generating rate. The non-local BTBT model is employed with fitting parameters taken from [5]. The detailed simulation model employed in this study has been presented in section 3.1.

A ternary inverter has been implemented using a DP-TFET, as described in the previous section. 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 (shown in Fig. 5.20(a)) than at the source-channel junction (shown in Fig. 5.20(d)). Thus, the within-channel 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. 5.21 shows the summer circuitry used to sum the pre-synaptic stimuli and generate an input potential for the ternary neuron, similar to the one used in [16, 17]. The integration of charge, however, is

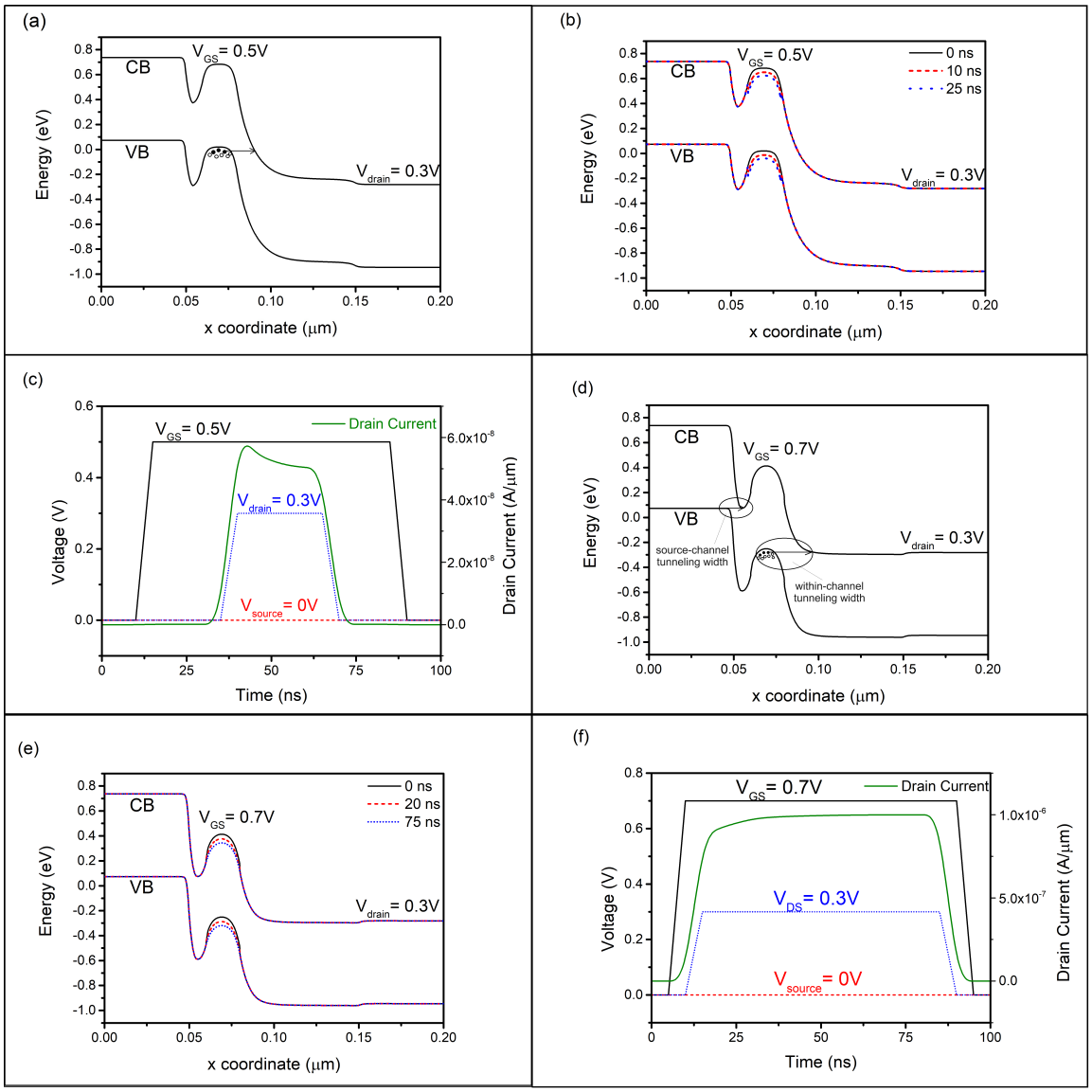

Figure 5.20: Principle of operation of a ternary spiking neuron (a)-(c) Generation of a $V_{DD}/2$ voltage spike, and (d)-(f) Generation of a V_{DD} voltage spike.

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.

The integration of charge inside the device and the generation of $V_{DD}/2$ and V_{DD} spikes is now discussed. Fig. 5.22 shows the band diagram along cutline BB' showing a decrease in the within-channel tunneling width and an increase

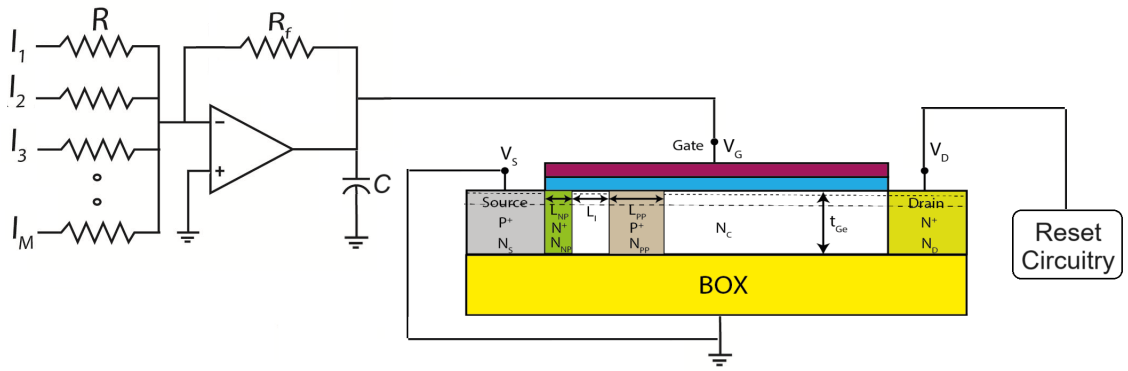

Figure 5.21: Ternary neuron architecture showing how the pre-synaptic stimuli are summed and the reset circuitry controlling the potential applied onto the drain terminal.

in the band overlap with an increase in the gate voltage, resulting in an increase in the within-channel tunneling current.

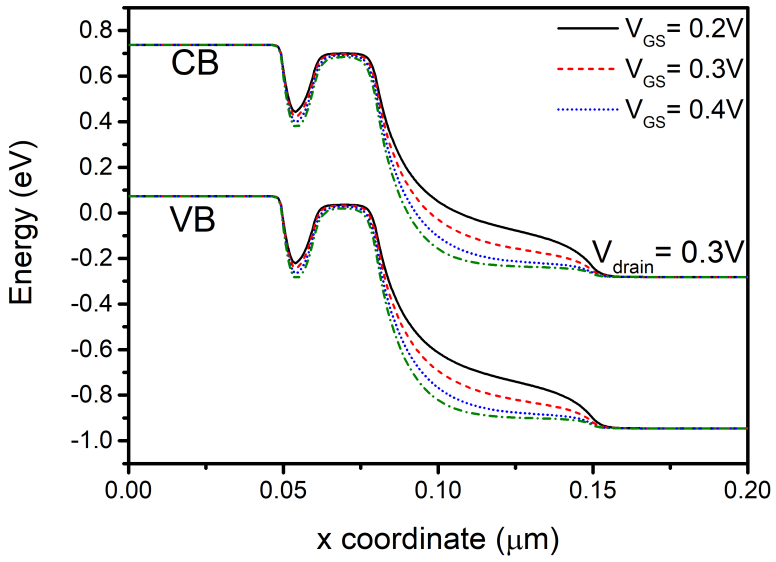

Figure 5.22: Band diagram along cutline BB' showing a decrease in tunneling width and an increase in band overlap with an increase in the gate voltage

In Fig. 5.20(a), the band diagram along cutline BB' at a gate voltage ($V_{GS} = 0.5V$) is shown, illustrating the within-channel tunneling of electrons. As a consequence, an accumulation of holes occurs in the channel pocket region 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. 5.20(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 the thermionic emission, the accumulated holes in the channel pocket region undergo leakage into the source, causing an increase in the potential barrier in 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. At this stage, the current due to within-channel tunneling reaches a threshold value ($I_{th1} = 5 \times 10^{-8} A/\mu m$), and the drain voltage is removed with the help of a reset circuitry, causing a rapid decrease in the current, as shown in Fig. 5.20(c). 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. 5.20(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 (gate voltage) reaches 0.7V, the onset of a source-channel tunneling current occurs. The presence of a hump in the band diagram in the channel causes the current flow through the device in a two-step process. First, BTBT of electrons from the source results in an accumulation of electrons in the region between the two pockets. Subsequently, due to the thermionic emission, the accumulated electrons surmount the barrier and reach the drain.

Fig. 5.20(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. 5.20(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. 5.20(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 [16, 17, 19, 20] as well (for a binary neuron) and can be tailored for a ternary neuron as well. 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 energy consumption per spiking event for the V_{DD}/2 spike is ∼0.45 fJ ($0.3V \times 50nA \times 30ns$) and for the V_{DD} spike is ∼22.5fJ ($0.3V \times 1\mu A \times 75ns$). These numbers do not contain the energy consumed by the external firing circuitry, reset circuitry, and the control circuitry. The actual energy consumption by the ternary neuron at the system-level is difficult to ascertain and shall depend on the architecture of the network.

The third state of the ternary neuron can lead to high stand-by power. After the firing of a $V_{DD}/2$ spike, the ternary neuron enters into a refractory state, where its drain voltage is removed, but its gate voltage is allowed to vary with time and incoming spikes from the pre-synaptic layer of neurons. During the refractory period, the leakage current is larger than "0" state, but since the drain voltage is not applied, it is still not very high. Thus, the intermediate stage can result in a larger stand-by power consumption in the device.

However, the proposed ternary SNN is expected to deliver a better classification accuracy at a smaller energy footprint due to faster convergence of the weights of synapses ($8\times$ lesser inference time, as shown later in section 5.3.4). This is attributed to the faster learning in the ternary SNN due to occurrence of V_{DD}/2 spikes and more information being embedded in ternary spikes in comparison to a binary SNN. Thus, even though the implementation of a ternary neuron can present some overhead in terms of a larger footprint and energy consumption compared to a binary neuron, it is expected be more energy-efficient from a system-level standpoint.

5.3.3 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. 4.7 shows the pair of dual-pocket FD-SOI MOSFETs, which produce a current that exponentially reduces in magnitude as the duration of the spiking events between the pre-synaptic and the post-synaptic neurons increases.

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 section 4.1. 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. 5.23 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 the 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 $30ns$ [35]. 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.

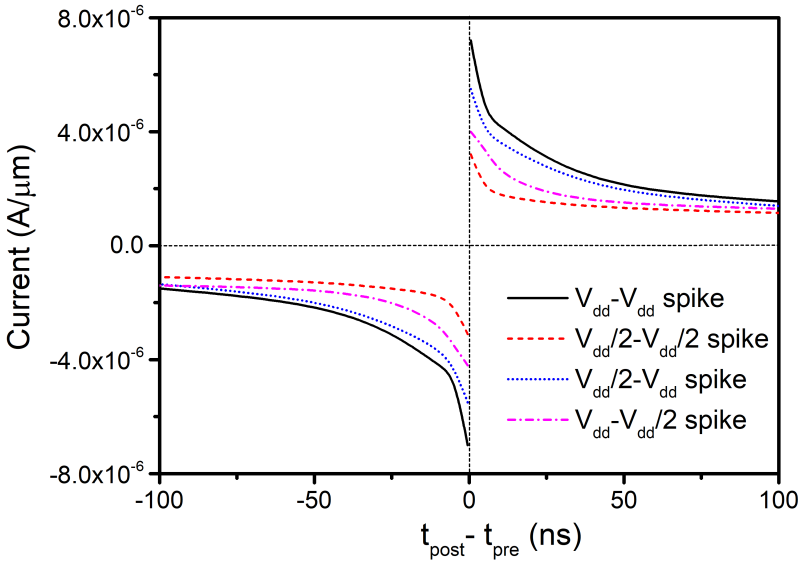

Figure 5.23: The current generated by the pair of dual-pocket FD-SOI MOSFETs for different pre and post-synaptic spiking events plotted as a function of the duration of firing events between the pre- and post-synaptic neurons.

5.3.4 Application of Ternary SNN

In this section, the proposed ternary SNN is trained to perform digit classification in the MNIST dataset using STDP. The 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, the second (excitatory) and third (inhibitory) layers comprise 800 neurons each. The first layer's 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 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 [49]. LIF dynamics is utilized to model the neuron. The LIF neuron's membrane potential ($v(t)$) is governed by the following equation:

$$\tau_e \frac{dv}{dt} = -(v(t) - v_{rest}) + I(t) \tag{5.3}$$

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. 5.6 lists the values of a few key parameters that were employed in the simulation. The time constants' units are defined in terms of the time step (dt) utilized in the simulation.

Table 5.6: 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

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 $\times$ 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:

$$Spike = \begin{cases} 0 & \text{if Spike Count} \leq 2 \\ 1 & \text{if } 2 < \text{Spike Count} \leq 4 \\ 2 & \text{otherwise} \end{cases} \tag{5.4}$$

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. 5.7. It can be observed that the classification accuracy obtained by training the ternary SNN on the MNIST dataset is lesser in comparison to [49] and [112], 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 [34, 49, 112], 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.

Table 5.7: Comparison of classification accuracy by training different SNN architectures on MNIST dataset

Reference	Architecture	Learning Method	No. of excitatory neurons	Accuracy
[112]	SNN [49]	STDP (2 layer)	400	84%
[34]	SNN [49]	STDP (2 layer)	100 400	57% 73%
[49]	SNN [49]	STDP (2 layer)	100 400 1600 6400	82.9% 87% 91.9% 95%
This work (subset of dataset)	SNN [49] Ternary SNN	STDP (2 layer)	800	75% 82%
[47]	Spiking DBN	Offline learning, Conversion		95%
[40]	Spiking CNN	Offline learning, Conversion		99.1%

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 observed to reduce 8 $\times$ 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.

5.3.5 Variability analysis

The results presented in this work have been obtained while considering ideal DP-TFET device characteristics with no variability. Device-to-device variability has now been considered, 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.

5.3.5.1 Impact of n^+ pocket doping (N_{NP})

N_{NP} was varied from $1 \times 10^{19} - 2 \times 10^{19}\ cm^{-3}$ around its nominal value of $1.5 \times 10^{19}\ cm^{-3}$ while keeping the other parameters at their nominal values ($N_{PP} = 3 \times 10^{19}\ cm^{-3}$, $L_I = 6nm$, $t_{ox} = 5nm$). Fig. 5.24 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}.

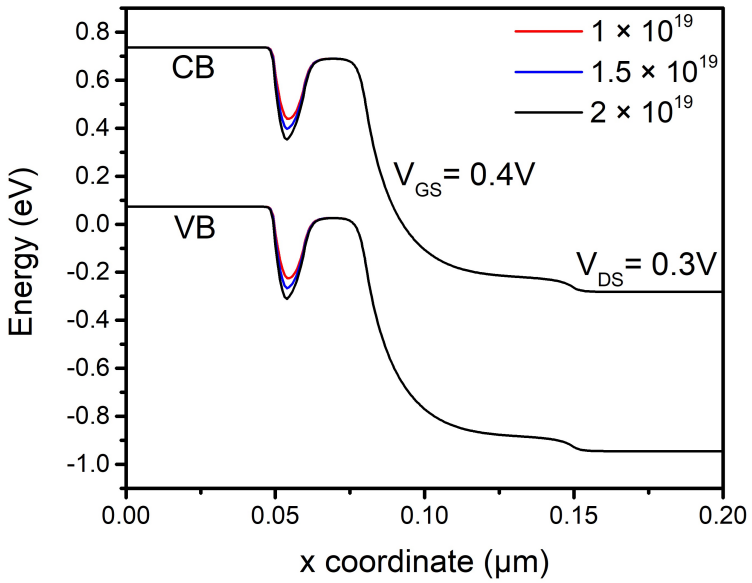

Figure 5.24: Band diagram along cutline BB' for different N_{NP} ($N_{PP} = 3 \times 10^{19}\ cm^{-3}$, $L_I = 6nm$, $t_{ox} = 5nm$)

5.3.5.2 Impact of change in length of intrinsic region between pockets (L_I)

L_I was varied from 4-8 nm around the nominal value of 6 nm while keeping the other parameters at their nominal values ($N_{NP} = 1.5 \times 10^{19}\ cm^{-3}$, $N_{PP} = 3\times 10^{19}\ cm^{-3}$, $t_{ox} = 5nm$). Fig. 5.25 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.

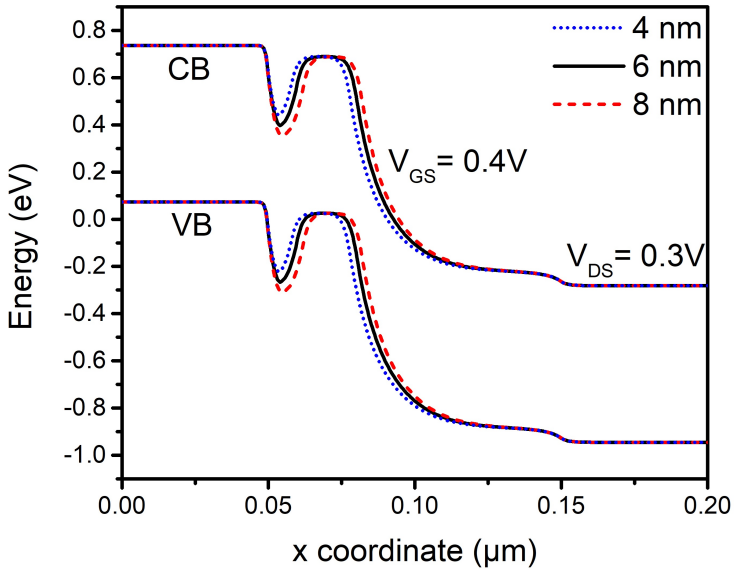

Figure 5.25: Band diagram along cutline BB' for different L_I ($N_{NP} = 1.5 \times 10^{19}\ cm^{-3}$, $N_{PP} = 3 \times 10^{19}\ cm^{-3}$, $t_{ox} = 5nm$)

5.3.5.3 Impact of p^+ pocket doping (N_{PP})

N_{PP} was varied 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. 5.26 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}.

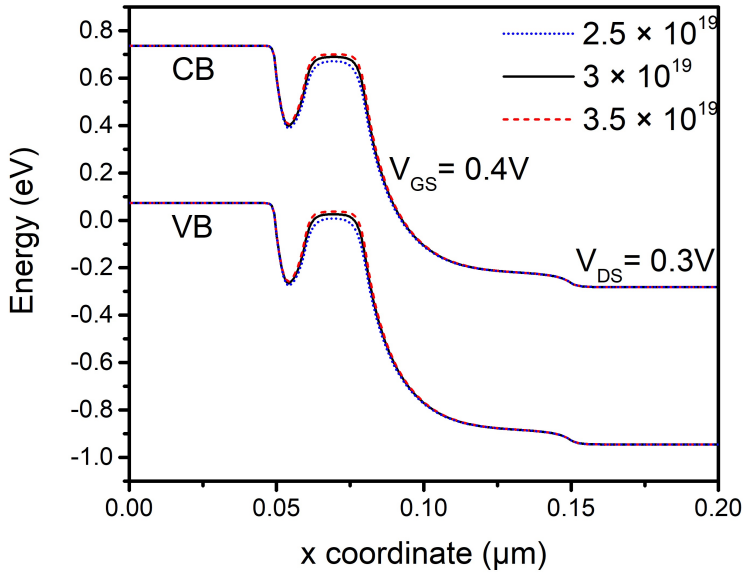

Figure 5.26: Band diagram along cutline BB' for different N_{PP} ($N_{NP} = 1.5 \times 10^{19}\ cm^{-3}$, $L_I = 6nm$, $t_{ox} = 5nm$)

5.3.5.4 Impact of change in thickness of gate dielectric (t_{ox})

t_{ox} was varied from 4-6 nm around the nominal value of 5 nm while keeping the other parameters at their nominal values ($N_{NP} = 1.5 \times 10^{19}\ cm^{-3}$, $L_I = 6nm$. $N_{PP} = 3 \times 10^{19}\ cm^{-3}$). Fig. 5.27 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}.

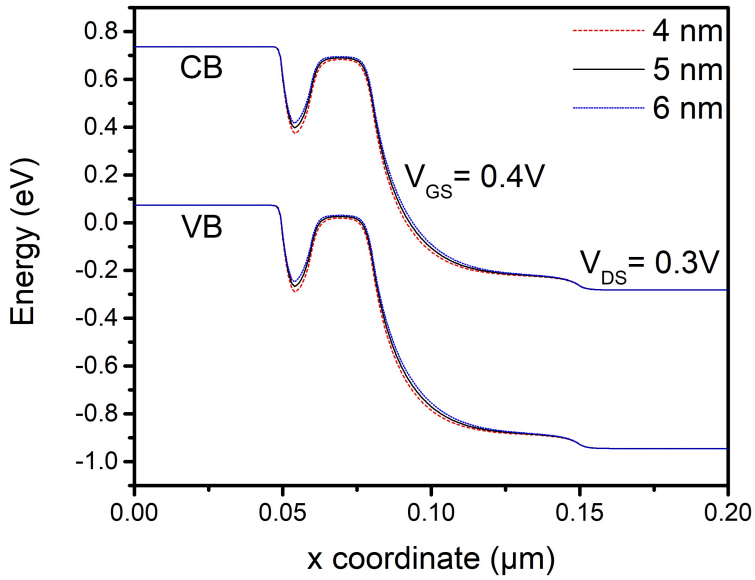

Figure 5.27: Band diagram along cutline BB' for different N_{PP} ($N_{NP} = 1.5 \times 10^{19}\ cm^{-3}$, $L_I = 6nm$. $N_{PP} = 3 \times 10^{19}\ cm^{-3}$)

5.3.5.5 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 around the ideal case while keeping the pocket parameters at their nominal values ($N_{NP} = 1.5 \times 10^{19}\ cm^{-3}$, $L_I = 6nm$. $N_{PP} = 3 \times 10^{19}\ cm^{-3}$, $t_{ox} = 5nm$). Fig. 5.28 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.

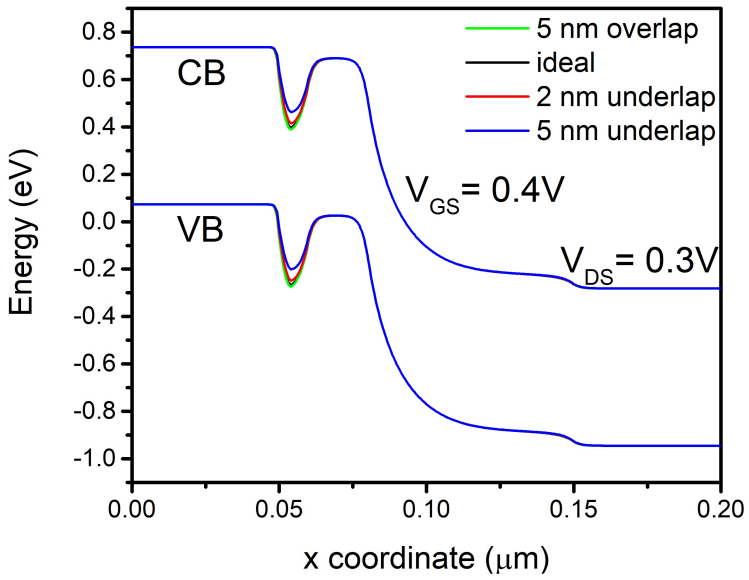

Figure 5.28: Band diagram along cutline BB' for different gate electrode underlap/overlap with respect to source/drain regions ($N_{NP} = 1.5 \times 10^{19}\ cm^{-3}$, $L_I = 6nm$. $N_{PP} = 3 \times 10^{19}\ cm^{-3}$, $t_{ox} = 5nm$)

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.

5.3.6 Conclusions

The classification accuracy of 75% was obtained using the binary SNN on the MNIST dataset, as shown in section 4.1.4. 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. Moreover, 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 observed to reduce $8\times$ for the ternary SNN compared to the binary spiking neural network. 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.

Chapter 6

Conclusion and Future Work

6.1 Summary

Machine Learning and Artificial Intelligence have been gaining a lot of traction in recent years and have found their use in various applications across different sectors like healthcare, automotive, finance, etc. Training the current state-of-the-art ML algorithms using Artificial Neural Networks is highly power intensive. Neuromorphic Computing, which draws inspiration from the functioning of the biological brain, presents an energy-efficient solution to train neural networks.

In this thesis, an energy-efficient SNN has been proposed, where training has been conducted on-chip in an unsupervised manner using STDP. In the proposed SNN, a LIF neuron has been implemented using a Ge-based PD-SOI MOSFET, which can directly receive the incoming voltage spikes as input and avoid energy dissipation in generating a summed potential. The smaller bandgap with dominant direct tunneling of Ge allows the device to operate at a lower voltage level. The energy consumption per spike in the proposed LIF neuron

is 0.07fJ, which is lower than LIF neuron implementations (experimental or simulated) reported in the literature. A Ferromagnetic Domain Wall (FM-DW) based device, which has decoupled read and write paths, is used as a synapse in this work. The conductance of the synapse can be modulated by passing a current through the Heavy Metal (HM) layer of the FM-DW synapse. Further, a pair of dual pocket Fully-Depleted Silicon-on-Insulator (FD-SOI) MOSFETs with dual asymmetric gates can be employed to generate a current, which depends exponentially on the temporal correlation of spiking events in the pre- and post-synaptic neuronal layers. This current is fed to the HM layer of the FM-DW synapse, which in turn modulates the conductance of the synapse in accordance with the STDP learning rule. The proposed implementation requires 2-3$\times$ fewer transistors and offers a lower latency to implement STDP than existing literature.

While SNNs have proven to be a suitable contender to ANNs due to their high energy efficiency, their use is still not prevalent. One of the major reasons preventing the widespread applicability of SNNs is the lack of efficient training algorithms that efficiently utilize the temporal information embedded in discrete spikes. The classification accuracy obtained by training an SNN using STDP is still not at par with its ANN counterparts. Moreover, the time required to train the SNN is substantially larger compared to ANNs. This is because no useful computation (learning) occurs in the network until and unless there is some spiking activity in the network. A ternary SNN, comprising a ternary neuron, which outputs a $V_{DD}/2$ spike when the membrane potential of the neuron crosses a threshold, say $v_{thresh1}$ and a V_{DD} spike when it crosses a higher threshold

$v_{thresh2}$, can result in a substantial speedup in the time required to train 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 binary encoding. This thesis explores an energy-efficient ternary SNN, where a ternary neuron has been implemented using a DP-TFET. Two distinct tunneling mechanisms exist in the device - within-channel tunneling and source-channel tunneling, which are responsible for the generation of $V_{DD}/2$ and V_{DD} voltage spikes, respectively. An FM-DW device is used as a synapse and a pair of dual-pocket FD-SOI MOSFETs with dual asymmetric gates are employed to implement on-chip learning using STDP in the ternary SNN. 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$\times$ less than that required for a binary SNN.

6.2 Future Work

There are numerous ways in which this work can be extended, improved, and employed for practical applications. The following are some potential future research directions:

- The behavior of the biological neurons and synapses in the brain is inherently stochastic in nature. Thus, an artificial Stochastic SNN (SSNN) comprising

a stochastic neuron and a stochastic synapse can result in a more biologically plausible implementation than a deterministic SNN due to the inherent stochasticity present in the biological nervous systems.

- A synapse is the most repetitive element in an SNN. Each neuron in the network often has an average fan-out of 10,000, i.e., each neuron is connected to 10,000 other neurons via synapses. Thus, the synaptic element should not only be highly energy-efficient but should present a very small load to the neuron driving it. The future work of this study involves research into more energy-efficient synaptic elements with a smaller load compared to the existing literature. This involves research from the materials perspective to the system-level implementation of the synaptic element, which can be deployed in future SNNs.

- Robotics, in particular, is a niche application area that can leverage the capabilities of SNNs and result in more human-like decision-making. Since the SNN is a highly energy-efficient architecture, a 3D interconnection of neurons can result in an area-efficient implementation of the SNN. Leveraging the extremely low-energy footprint of the network, a 3D stacking of neurons will not result in excessive generation of heat. Thus, the future work for this study involves developing an insight into different SNN topologies that are amenable to 3D integration.

- Orch theory, postulated by Sir Roger Penrose and Stuart Hameroff, infers that human cognition is based on quantum computation. The functioning

of the biological brain itself is not completely understood. Thus, implementing a Neuromorphic Computing framework using deterministic devices and circuits would be an incorrect approximation of the biological system. Exploring Neuromorphic Computing from the purview of Quantum Computation can not only lead to a more biologically plausible model of the biological nervous system but can also help in a better understanding of the biological nervous system.

References

[1] G. Moore, “Moore’s law." in *Electronics Magazine*, vol. 38, no. 8, p. 114, 1965.

[2] D. Silver, A. Huang, C. Maddison, et al. “Mastering the game of Go with deep neural networks and tree search” in *Nature*, vol 529, pp. 484–489, 2016, doi: 10.1038/nature16961.

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell and S. Agarwal, “Language models are few-shot learners” in *Advances in neural information processing systems*, vol. 33, pp.1877-1901, 2020. doi: 10.48550/arXiv.2005.14165.

[4] E. Strubell, A. Ganesh, and A. McCallum, “Energy and policy considerations for deep learning in NLP" in *arXiv preprint*, 2019, doi: 10.48550/arXiv.1906.02243.

[5] 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," in *IEEE Transactions on Electron Devices*, vol. 59, no. 2, pp. 292-301, Feb. 2012, doi: 10.1109/TED.2011.2175228.

[6] Jun Wang, "Analysis and design of a recurrent neural network for linear programming," in *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 40, no. 9, pp. 613-618, Sept. 1993, doi: 10.1109/81.244913.

[7] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," in *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673-2681, Nov. 1997, doi: 10.1109/78.650093.

[8] P. Gkoupidenis, “Artificial neurons emulate biological counterparts to enable synergetic operation," in *Nature Electronics*, vol. 5, pp. 721-722, 2022, doi: 10.1038/s41928-022-00862-3.

[9] Wolfgang Maass, “Networks of spiking neurons: The third generation of neural network models", in *Neural Networks*, vol. 10, pp. 1659-1671, 1997, doi: 10.1016/S0893-6080(97)00011-7.

[10] A. L. Hodgkin and A. F. Huxley, “A quantitative description of membrane current and its application to conduction and excitation in nerve,” in *The Journal of physiology*, vol. 117, no. 4, p. 500, 1952, doi: 10.1113/jphysiol.1952.sp004764.

[11] E. M. Izhikevich, “Simple model of spiking neurons," in *IEEE Transactions on Neural Networks*, vol. 14, no. 6, pp. 1569-1572, Nov. 2003, doi: 10.1109/TNN.2003.820440.

[12] A.N. Burkitt, “A Review of the Integrate-and-fire Neuron Model: I. Homogeneous Synaptic Input", in *Biol Cybern*, vol. 95, pp. 1–19, 2006, doi: 10.1007/s00422-006-0068-6.

[13] E. Hunsberger and C. Eliasmith, “Spiking deep networks with LIF neurons" in *arXiv preprint*, 2015, doi: 10.1038/323533a0.

[14] S. Lashkare, S. Chouhan, T. Chavan, A. Bhat, P. Kumbhare and U. Ganguly, “PCMO RRAM for Integrate-and-Fire Neuron in Spiking Neural Networks," in *IEEE Electron Device Letters*, vol. 39, no. 4, pp. 484-487, April 2018, doi: 10.1109/LED.2018.2805822.

[15] G. Indiveri, E. Chicca and R. Douglas, “A VLSI array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity." in *IEEE Transactions on Neural Networks*, vol. 17, no. 1, pp. 211-221, Jan. 2006, doi: 10.1109/TNN.2005.860850.

[16] 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." in *Sci. Rep.*, vol. 7, no. 1, pp. 1-7, Dec 2017, doi: 10.1038/s41598-017-07418-y.

[17] D. Chatterjee and A. Kottantharayil, “A CMOS Compatible Bulk FinFET-Based Ultra Low Energy Leaky Integrate and Fire Neuron for Spiking Neural Networks," in *IEEE Electron Device Letters*, vol. 40, no. 8, pp. 1301-1304, Aug. 2019, doi: 10.1109/LED.2019.2924259.

[18] N. Kamal and J. Singh, “A Highly Scalable Junctionless FET Leaky Integrate-and-Fire Neuron for Spiking Neural Networks," in *IEEE Transactions on Electron Devices*, vol. 68, no. 4, pp. 1633-1638, April 2021, doi: 10.1109/TED.2021.3061036.

[19] B. Das, J. Schulze and U. Ganguly, “Ultra-Low Energy LIF Neuron Using Si NIPIN Diode for Spiking Neural Networks," in *IEEE Electron Device Letters*, vol. 39, no. 12, pp. 1832-1835, Dec. 2018, doi: 10.1109/LED.2018.2876684.

[20] T. Chavan, S. Dutta, N. R. Mohapatra and U. Ganguly, “Band-to-Band Tunneling Based Ultra-Energy-Efficient Silicon Neuron," in *IEEE Transactions on Electron Devices*, vol. 67, no. 6, pp. 2614-2620, June 2020, doi: 10.1109/TED.2020.2985167.

[21] A. Beohar et al., “Compact Spiking Neural Network System with SiGe Based Cylindrical Tunneling Transistor for Low Power Applications" in *International Symposium on VLSI Design and Test*, Springer, Singapore, 2019, doi: 10.1007/978-981-32-9767-8$_5$4.

[22] E. Lazaridis, E. M. Drakakis and M. Barahona, “A biomimetic CMOS synapse," in *2006 IEEE International Symposium on Circuits and Systems*, Kos, Greece, 2006, pp. 4, doi: 10.1109/ISCAS.2006.1692694.

[23] M. Noack, C. Mayr, J. Partzsch and R. Schüffny, “Synapse dynamics in CMOS derived from a model of neurotransmitter release," in *2011 20th European Conference on Circuit Theory and Design (ECCTD)*, Linköping, Sweden, 2011, pp. 198-201, doi: 10.1109/ECCTD.2011.6043316.

[24] S. Thanapitak and C. Toumazou, “A Bionics Chemical Synapse," in *IEEE Transactions on Biomedical Circuits and Systems*, vol. 7, no. 3, pp. 296-306, June 2013, doi: 10.1109/TBCAS.2012.2202494.

[25] N. Caporale and Y. Dan, “Spike timing-dependent plasticity: A hebbian learning rule,” in *Annu. Rev. Neurosci.*, vol. 31, pp. 25–46, 2008, doi: 10.1146/annurev.neuro.31.060407.125639.

[26] B. Rajendran et al., “Specifications of Nanoscale Devices and Circuits for Neuromorphic Computational Systems," in *IEEE Transactions on Electron Devices*, vol. 60, no. 1, pp. 246-253, Jan. 2013, doi: 10.1109/TED.2012.2227969.

[27] S. Ambrogio et al., “Neuromorphic Learning and Recognition With One-Transistor-One-Resistor Synapses and Bistable Metal Oxide RRAM," in *IEEE Transactions on Electron Devices*, vol. 63, no. 4, pp. 1508-1515, April 2016, doi: 10.1109/TED.2016.2526647.

[28] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu, “Nanoscale memristor device as synapse in neuromorphic systems,” in *Nano letters*, vol. 10, no. 4, pp. 1297–1301, 2010, doi: 10.1021/nl904092h.

[29] G. Indiveri, R. Legenstein, G. Deligeorgis, T. Prodromakis et al., “Integration of nanoscale memristor synapses in neuromorphic computing architectures,” in *Nanotechnology*, vol. 24, no. 38, p. 384010, 2013, doi: 10.1088/0957-4484/24/38/384010.

[30] M. Hu, Y. Wang, W. Wen, Y. Wang and H. Li, “Leveraging Stochastic Memristor Devices in Neuromorphic Hardware Systems," in *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 6, no. 2, pp. 235-246, June 2016, doi: 10.1109/JETCAS.2016.2547780.

[31] S. Ramakrishnan, P. E. Hasler and C. Gordon, “Floating Gate Synapses With Spike-Time-Dependent Plasticity," in *IEEE Transactions on Biomedical Circuits and Systems*, vol. 5, no. 3, pp. 244-252, June 2011, doi: 10.1109/T-BCAS.2011.2109000.

[32] D. Kuzum, R. G. D. Jeyasingh, B. Lee, and H.-S. P. Wong, “Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing,” in *Nano Lett.*, vol. 12, no. 5, pp. 2179–2186, 2012, doi: 10.1021/nl201040y.

[33] Y. Nishitani, Y. Kaneko, M. Ueda, T. Morie, and E. Fujii , “Three-terminal ferroelectric synapse device with concurrent learning function for artificial neural networks", in *Journal of Applied Physics*, vol. 111, no. 12, p. 124108, 2012, doi: 10.1063/1.4729915.

[34] 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,” in *Scientific Reports*, vol. 6, 2016, doi: 10.1038/srep29545.

[35] A. Sengupta, A. Banerjee, and K. Roy, “Hybrid Spintronic-CMOS Spiking Neural Network with On-Chip Learning: Devices, Circuits, and Systems,”

in *Phys. Rev. Applied*, vol. 6, no. 6, p. 064003, 2016, doi: 10.1103/PhysRevApplied.6.064003.

[36] U. Sahu, A. Pandey, K. Goyal, and D. Bhowmik, “Spike time dependent plasticity (STDP) enabled learning in spiking neural networks using domain wall based synapses and neurons". in *AIP Advances*, vol. 9, no. 125339, 2019, doi: 10.1063/1.5129729.

[37] Q. Yu, H. Tang, K. C. Tan and H. Li, “Rapid Feedforward Computation by Temporal Encoding and Learning With Spiking Neurons," in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 10, pp. 1539-1552, Oct. 2013, doi: 10.1109/TNNLS.2013.2245677.

[38] B. Zhao, R. Ding, S. Chen, B. Linares-Barranco and H. Tang, “Feedforward Categorization on AER Motion Events Using Cortex-Like Features in a Spiking Neural Network," in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 9, pp. 1963-1978, Sept. 2015, doi: 10.1109/TNNLS.2014.2362542.

[39] H. Mostafa, “Supervised Learning Based on Temporal Coding in Spiking Neural Networks," in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 7, pp. 3227-3235, July 2018, doi: 10.1109/TNNLS.2017.2726060.

[40] 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 *2015 International Joint Conference on Neu-*

ral Networks (IJCNN), Killarney, Ireland, 2015, pp. 1-8, doi: 10.1109/IJCNN.2015.7280696.

[41] G. Indiveri, F. Corradi and N. Qiao, “Neuromorphic architectures for spiking deep neural networks," in *2015 IEEE International Electron Devices Meeting (IEDM)*, Washington, DC, USA, 2015, pp. 4.2.1-4.2.4, doi: 10.1109/IEDM.2015.7409623.

[42] B. Han, A. Sengupta and K. Roy, “On the energy benefits of spiking deep neural networks: A case study," in *2016 International Joint Conference on Neural Networks (IJCNN)*, Vancouver, BC, Canada, 2016, pp. 971-976, doi: 10.1109/IJCNN.2016.7727303.

[43] W. Wang, S. Hao, Y. Wei, S. Xiao, J. Feng and N. Sebe, “Temporal Spiking Recurrent Neural Network for Action Recognition," in *IEEE Access*, vol. 7, pp. 117165-117175, 2019, doi: 10.1109/ACCESS.2019.2936604.

[44] G. Pedretti et al., “A Spiking Recurrent Neural Network With Phase-Change Memory Neurons and Synapses for the Accelerated Solution of Constraint Satisfaction Problems," in *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, vol. 6, no. 1, pp. 89-97, June 2020, doi: 10.1109/JXCDC.2020.2992691.

[45] E. O. Neftci, H. Mostafa and F. Zenke, “Surrogate Gradient Learning in Spiking Neural Networks: Bringing the Power of Gradient-Based Optimization to Spiking Neural Networks," in *IEEE Signal Processing Magazine*, vol. 36, no. 6, pp. 51-63, Nov. 2019, doi: 10.1109/MSP.2019.2931595.

[46] C. Lee, G. Srinivasan, P. Panda and K. Roy, “Deep Spiking Convolutional Neural Network Trained With Unsupervised Spike-Timing-Dependent Plasticity," in *IEEE Transactions on Cognitive and Developmental Systems*, vol. 11, no. 3, pp. 384-394, Sept. 2019, doi: 10.1109/TCDS.2018.2833071.

[47] E. Stromatias, D. Neil, F. Galluppi, M. Pfeiffer, S. -C. Liu and S. Furber, “Scalable energy-efficient, low-latency implementations of trained spiking Deep Belief Networks on SpiNNaker," in *2015 International Joint Conference on Neural Networks (IJCNN)*, Killarney, Ireland, 2015, pp. 1-8, doi: 10.1109/IJCNN.2015.7280625.

[48] A. Gupta and L. N. Long, “Hebbian learning with winner take all for spiking neural networks," in *2009 International Joint Conference on Neural Networks*, Atlanta, GA, USA, 2009, pp. 1054-1060, doi: 10.1109/IJCNN.2009.5178751.

[49] P. U. Diehl and M. Cook, “Unsupervised learning of digit recognition using spike-timing-dependent plasticity," in *Frontiers in Computational Neuroscience*, vol. 9, no. 99, 2015, doi: 10.3389/fncom.2015.00099.

[50] D. Rumelhart, G. Hinton and R. Williams, “Learning representations by back-propagating errors" in *Nature*, vol. 323, pp. 533–536, 1986, doi: 10.1038/323533a0.

[51] P. O’Connor, D. Neil, S.-C. Liu, T. Delbruck, and M. Pfeiffer, “Realtime classification and sensor fusion with a spiking deep belief network," in *Frontiers in Neuroscience*, vol. 7, 2013, doi: 10.3389/fnins.2013.00178.

[52] J. A. Pérez-Carrasco et al., “Mapping from Frame-Driven to Frame-Free Event-Driven Vision Systems by Low-Rate Rate Coding and Coincidence Processing–Application to Feedforward ConvNets," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2706-2719, Nov. 2013, doi: 10.1109/TPAMI.2013.71.

[53] G. Q. Bi and M. M. Poo, “Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type" in *Journal of neuroscience*, vol. 18, no. 24, pp. 10464-10472, 1998, doi: 10.1523/JNEUROSCI.18-24-10464.1998.

[54] A. Gupta and S. Saurabh, “An Energy-Efficient Ge-Based Leaky Integrate and Fire Neuron: Proposal and Analysis,” in *IEEE Transactions on Nanotechnology*, vol. 21, pp. 555-563, 2022, doi: 10.1109/TNANO.2022.3209078.

[55] E. M. Izhikevich and N. S. Desai, “Relating STDP to BCM," in *Neural Computation*, vol. 15, no. 7, pp. 1511-1523, 2003, doi: 10.1162/089976603321891783.

[56] *Synopsys Sentaurus Device User Guide, N-2020.09*, Synopsys, Inc., Mountain View, CA, USA, 2020.

[57] W. Li and J. C. S. Woo, “Optimization and Scaling of Ge-Pocket TFET," in *IEEE Transactions on Electron Devices*, vol. 65, no. 12, pp. 5289-5294, Dec. 2018, doi: 10.1109/TED.2018.2874047.

[58] Q. -T. Zhao et al., “Strained Si and SiGe Nanowire Tunnel FETs for Logic and Analog Applications," in *IEEE Journal of the Electron Devices Society*, vol. 3, no. 3, pp. 103-114, May 2015, doi: 10.1109/JEDS.2015.2400371.

[59] K. Vanlalawpuia and B. Bhowmick, “Investigation of a Ge-Source Vertical TFET With Delta-Doped Layer," in *IEEE Transactions on Electron Devices*, vol. 66, no. 10, pp. 4439-4445, Oct. 2019, doi: 10.1109/TED.2019.2933313.

[60] K. Roy, S. Mukhopadhyay and H. Mahmoodi-Meimand, “Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits," in *Proceedings of the IEEE*, vol. 91, no. 2, pp. 305-327, Feb. 2003, doi: 10.1109/JPROC.2002.808156.

[61] K. Boucart and A. M. Ionescu, “Double-Gate Tunnel FET With High-κ Gate Dielectric," in *IEEE Transactions on Electron Devices*, vol. 54, no. 7, pp. 1725-1733, July 2007, doi: 10.1109/TED.2007.899389.

[62] L. Zhang and M. Chan, “SPICE Modeling of Double-Gate Tunnel-FETs Including Channel Transports," in *IEEE Transactions on Electron Devices*, vol. 61, no. 2, pp. 300-307, Feb. 2014, doi:10.1109/TED.2013.2295237

[63] Y. -Y. Chen et al., “A SPICE-Compatible Model of MOS-Type Graphene Nano-Ribbon Field-Effect Transistors Enabling Gate- and Circuit-Level Delay and Power Analysis Under Process Variation," in *IEEE Transactions on Nanotechnology*, vol. 14, no. 6, pp. 1068-1082, Nov. 2015, doi:10.1109/TNANO.2015.2469647.

[64] B. Lu et al., “Fully Analytical Carrier-Based Charge and Capacitance Model for Hetero-Gate-Dielectric Tunneling Field-Effect Transistors," in *IEEE Transactions on Electron Devices*, vol. 65, no. 8, pp. 3555-3561, Aug. 2018, doi:10.1109/TED.2018.2849742.

[65] A. Marshall and S. Natarajan, "PD-SOI and FD-SOI: a comparison of circuit performance," in *9th International Conference on Electronics, Circuits and Systems*, 2002, pp. 25-28 vol.1, doi: 10.1109/ICECS.2002.1045324.

[66] D. B. Abdi and M. Jagadesh Kumar, “Controlling Ambipolar Current in Tunneling FETs Using Overlapping Gate-on-Drain," in *IEEE Journal of the Electron Devices Society*, vol. 2, no. 6, pp. 187-190, Nov. 2014, doi: 10.1109/JEDS.2014.2327626.

[67] S. Sahay and M. J. Kumar, “Controlling the Drain Side Tunneling Width to Reduce Ambipolar Current in Tunnel FETs Using Heterodielectric BOX," in *IEEE Transactions on Electron Devices*, vol. 62, no. 11, pp. 3882-3886, Nov. 2015, doi: 10.1109/TED.2015.2478955.

[68] B. R. Raad, S. Tirkey, D. Sharma and P. Kondekar, “A New Design Approach of Dopingless Tunnel FET for Enhancement of Device Characteristics," in *IEEE Transactions on Electron Devices*, vol. 64, no. 4, pp. 1830-1836, April 2017, doi: 10.1109/TED.2017.2672640.

[69] A. Rassekh, F. Jazaeri, M. Fathipour and J. -M. Sallese, “Modeling Interface Charge Traps in Junctionless FETs, Including Temperature Effects," in *IEEE*

Transactions on Electron Devices, vol. 66, no. 11, pp. 4653-4659, Nov. 2019, doi: 10.1109/TED.2019.2944193.

[70] Y. Qiu, R. Wang, Q. Huang and R. Huang, "A Comparative Study on the Impacts of Interface Traps on Tunneling FET and MOSFET," in *IEEE Transactions on Electron Devices*, vol. 61, no. 5, pp. 1284-1291, May 2014, doi:10.1109/TED.2014.2312330.

[71] J. Madan and R. Chaujar, "Numerical Simulation of N+ Source Pocket PIN-GAA-Tunnel FET: Impact of Interface Trap Charges and Temperature," in *IEEE Transactions on Electron Devices*, vol. 64, no. 4, pp. 1482-1488, April 2017, doi:10.1109/TED.2017.2670603.

[72] A. Gupta and S. Saurabh, "On-chip Unsupervised Learning using STDP in a Spiking Neural Network," in *IEEE Transactions on Nanotechnology*, vol. 22, pp. 365-376, 2023, doi: 10.1109/TNANO.2023.3293011.

[73] A. Vansteenkiste, J. Leliaert, M. Dvornik, M. Helsen, F. Garcia-Sanchez, and B. Van Waeyenberge, "The design and verification of mumax3," in *AIP Advances*, vol. 4, no. 10, 2014, Art. no. 107133, doi: 10.1063/1.4899186.

[74] H. Hazan et al., "Bindsnet: A machine learning-oriented spiking neural networks library in python." in *Frontiers in neuroinformatics*, vol. 12, p. 89, 2018, doi: 10.3389/fninf.2018.00089.

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

[76] K. S. Ryu, L. Thomas, S. H. Yang, and S. Parkin, “Chiral spin torque at magnetic domain walls,” in *Nature Nanotechnology*, vol. 8, no. 7, pp. 527–533, 2013, doi: 10.1038/nnano.2013.102.

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

[78] 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,” in *Journal of Applied Physics*, vol. 115, no. 21, 2014, Art. no. 213909, doi: 10.1063/1.4881778.

[79] S. Ikeda, J. Hayakawa, Y. Ashizawa, Y. Lee, K. Miura, H. Hasegawa, M. Tsunoda, F. Matsukura, and H. Ohno, “Tunnel magnetoresistance of 604% at 300 K by suppression of Ta diffusion in CoFeB/MgO/CoFeB pseudo-spin-valves annealed at high temperature,” in *Applied Physics Letters*, vol. 93, no. 8, p. 2508, 2008, doi: 10.1063/1.2976435.

[80] V. Nagavarapu, R. Jhaveri and J. C. S. Woo, “The Tunnel Source (PNPN) n-MOSFET: A Novel High Performance Transistor," in *IEEE Transactions on Electron Devices*, vol. 55, no. 4, pp. 1013-1019, April 2008, doi: 10.1109/TED.2008.916711.

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

[82] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, “Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, Nov. 1998, doi: 10.1109/5.726791.

[83] A. Gupta and S. Saurabh, “Novel attributes of a dual pocket tunnel field-effect transistor,” in *Japanese Journal of Applied Physics*, vol. 61, no. 3, p. 035001, 2022. doi: 10.35848/1347-4065/ac3722.

[84] J. -. Colinge, “Subthreshold slope of thin-film SOI MOSFET’s," in *IEEE Electron Device Letters*, vol. 7, no. 4, pp. 244-246, April 1986, doi: 10.1109/EDL.1986.26359.

[85] D. J. Frank, R. H. Dennard, E. Nowak, P. M. Solomon, Y. Taur and Hon-Sum Philip Wong, “Device scaling limits of Si MOSFETs and their application dependencies," in *Proceedings of the IEEE*, vol. 89, no. 3, pp. 259-288, March 2001, doi: 10.1109/5.915374.

[86] W. Haensch et al., “Silicon CMOS devices beyond scaling," in *IBM Journal of Research and Development*, vol. 50, no. 4.5, pp. 339-361, July 2006, doi: 10.1147/rd.504.0339.

[87] M. Bohr, “A 30 Year Retrospective on Dennard’s MOSFET Scaling Paper," in *IEEE Solid-State Circuits Society Newsletter*, vol. 12, no. 1, pp. 11-13, Winter 2007, doi: 10.1109/N-SSC.2007.4785534.

[88] H. Esmaeilzadeh, E. Blem, R. St. Amant, K. Sankaralingam and D. Burger, “Dark Silicon and the End of Multicore Scaling," in *IEEE Micro*, vol. 32, no. 3, pp. 122-134, May-June 2012, doi: 10.1109/MM.2012.17.

[89] A. C. Seabaugh and Q. Zhang, “Low-Voltage Tunnel Transistors for Beyond CMOS Logic," in *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2095-2110, Dec. 2010, doi: 10.1109/JPROC.2010.2070470.

[90] W. Y. Choi, B. Park, J. D. Lee and T. K. Liu, “Tunneling Field-Effect Transistors (TFETs) With Subthreshold Swing (SS) Less Than 60 mV/dec," in *IEEE Electron Device Letters*, vol. 28, no. 8, pp. 743-745, Aug. 2007, doi: 10.1109/LED.2007.901273.

[91] H. Lu and A. Seabaugh, “Tunnel Field-Effect Transistors: State-of-the-Art," in *IEEE Journal of the Electron Devices Society*, vol. 2, no. 4, pp. 44-49, July 2014, doi: 10.1109/JEDS.2014.2326622.

[92] D. Sarkar, X. Xie, W. Liu, W. Cao, J. Kang, Y. Gong, S. Kraemer, P. M. Ajayan and K. Banerjee, “A subthermionic tunnel field-effect transistor with an atomically thin channel," in *Nature*, vol. 526, no. 7571, pp. 91-95, Oct. 2015. doi:10.1038/nature15387

[93] S. Saurabh and M. J. Kumar, “Fundamentals of tunnel field-effect transistors," in *CRC Press*, 2016. doi: 10.1201/9781315367354-3

[94] J. Knoch and J. Appenzeller, “Modeling of High-Performance p-Type III–V Heterojunction Tunnel FETs," in *IEEE Electron Device Letters*, vol. 31, no. 4, pp. 305-307, April 2010, doi: 10.1109/LED.2010.2041180.

[95] T. Krishnamohan, D. Kim, S. Raghunathan and K. Saraswat, “Double-Gate Strained-Ge Heterostructure Tunneling FET (TFET) With record high drive currents and $<<$60mV/dec subthreshold slope," in *2008 IEEE Inter-*

national Electron Devices Meeting, San Francisco, CA, 2008, pp. 1-3, doi: 10.1109/IEDM.2008.4796839.

[96] A. S. Verhulst, W. G. Vandenberghe, K. Maex, S. De Gendt, M. M. Heyns and G. Groeseneken, “Complementary Silicon-Based Heterostructure Tunnel-FETs With High Tunnel Rates," in *IEEE Electron Device Letters*, vol. 29, no. 12, pp. 1398-1401, Dec. 2008, doi: 10.1109/LED.2008.2007599.

[97] H. Ilatikhameneh, Y. Tan, B. Novakovic, G. Klimeck, R. Rahman and J. Appenzeller, “Tunnel Field-Effect Transistors in 2-D Transition Metal Dichalcogenide Materials," in *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, vol. 1, pp. 12-18, Dec. 2015, doi: 10.1109/JXCDC.2015.2423096.

[98] F. W. Chen, H. Ilatikhameneh, T. A. Ameen, G. Klimeck and R. Rahman, “Thickness Engineered Tunnel Field-Effect Transistors Based on Phosphorene," in *IEEE Electron Device Letters*, vol. 38, no. 1, pp. 130-133, Jan. 2017, doi: 10.1109/LED.2016.2627538.

[99] Silvaco Atlas user’s manual, 2015. [Online]. Available: http://www.silvaco.com.

[100] *Synopsys Sentaurus Process User Guide, N-2017.09*, Synopsys, Inc., Mountain View, CA, USA, 2017.

[101] S. Saurabh and M. J. Kumar, “Novel Attributes of a Dual Material Gate Nanoscale Tunnel Field-Effect Transistor," in *IEEE Transac-*

tions on Electron Devices, vol. 58, no. 2, pp. 404-410, Feb. 2011, doi: 10.1109/TED.2010.2093142.

[102] A. S. Verhulst, W. G. Vandenberghe, K. Maex, and G. Groeseneken, "Tunnel field-effect transistor without gate-drain overlap," in *Applied Physics Letters*, vol. 91, no. 5, p. 053102, 2007, doi: 10.1063/1.2757593.

[103] W. G. Vandenberghe, A. S. Verhulst, B. Sorée, W. Magnus, G. Groeseneken, Q. Smets, M. Heyns, and M. V. Fischetti, "Figure of merit for and identification of sub-60mV/decade devices", in *Applied Physics Letters*, vol. 102, p. 013510, 2013, doi: 10.1063/1.4773521.

[104] R. Gandhi, Z. Chen, N. Singh, K. Banerjee, and S. Lee, "CMOS-Compatible Vertical-Silicon-Nanowire Gate-All-Around p-Type Tunneling FETs With $\leq$ 50-mV/decade Subthreshold Swing," in *IEEE Electron Device Letters*, vol. 32, no. 11, pp. 1504-1506, Nov. 2011, doi: 10.1109/LED.2011.2165331.

[105] L. Knoll et al., "Inverters With Strained Si Nanowire Complementary Tunnel Field-Effect Transistors," in *IEEE Electron Device Letters*, vol. 34, no. 6, pp. 813-815, June 2013, doi: 10.1109/LED.2013.2258652.

[106] E. Memisevic, J. Svensson, M. Hellenbrand, E. Lind and L. -E. Wernersson, "Vertical InAs/GaAsSb/GaSb tunneling field-effect transistor on Si with S = 48 mV/decade and Ion = 10 A/m for Ioff = 1 nA/m at Vds = 0.3 V," in *2016 IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2016, pp. 19.1.1-19.1.4, doi: 10.1109/IEDM.2016.7838450.

[107] C. Wu, Q. Huang, Y. Zhao, J. Wang, Y. Wang and R. Huang, “A Novel Tunnel FET Design With Stacked Source Configuration for Average Subthreshold Swing Reduction," in *IEEE Transactions on Electron Devices*, vol. 63, no. 12, pp. 5072-5076, Dec. 2016, doi: 10.1109/TED.2016.2619694.

[108] W. Cheng et al., “Fabrication and Characterization of a Novel Si Line Tunneling TFET With High Drive Current," in *IEEE Journal of the Electron Devices Society*, vol. 8, pp. 336-340, 2020, doi: 10.1109/JEDS.2020.2981974.

[109] Y. Lu *et al.*, “Performance of AlGaSb/InAs TFETs With Gate Electric Field and Tunneling Direction Aligned," in *IEEE Electron Device Letters*, vol. 33, no. 5, pp. 655-657, May 2012, doi: 10.1109/LED.2012.2186554.

[110] H. W. Kim and D. Kwon, “Low-Power Vertical Tunnel Field-Effect Transistor Ternary Inverter," in *IEEE Journal of the Electron Devices Society*, vol. 9, pp. 286-294, 2021, doi: 10.1109/JEDS.2021.3057456.

[111] T. A. Ameen, H. Ilatikhameneh, G. Klimeck, and R. Rahman, “Few-layer Phosphorene: An Ideal 2D Material For Tunnel Transistors," *Scientific Reports*, vol. 6, no. 1, pp. 1-7, 2016, doi: 10.1038/srep28515

[112] A. Gupta and S. Saurabh, “Implementing a Ternary Inverter Using Dual-Pocket Tunnel Field-Effect Transistors," in *IEEE Transactions on Electron Devices*, vol. 68, no. 10, pp. 5305-5310, Oct. 2021, doi: 10.1109/TED.2021.3106618.

[113] P. C. Balla and A. Antoniou, “Low power dissipation MOS ternary logic family," in *IEEE Journal of Solid-State Circuits*, vol. 19, no. 5, pp. 739-749, Oct. 1984, doi: 10.1109/JSSC.1984.1052216.

[114] S. Lin, Y. Kim and F. Lombardi, “A novel CNTFET-based ternary logic gate design," in *2009 52nd IEEE International Midwest Symposium on Circuits and Systems*, Cancun, 2009, pp. 435-438, doi: 10.1109/MWSCAS.2009.5236063.

[115] S. Shin, E. Jang, J. W. Jeong, B. Park and K. R. Kim, “Compact Design of Low Power Standard Ternary Inverter Based on OFF-State Current Mechanism Using Nano-CMOS Technology," in *IEEE Transactions on Electron Devices*, vol. 62, no. 8, pp. 2396-2403, Aug. 2015, doi: 10.1109/TED.2015.2445823.

[116] Mouftah and Jordan, “Design of Ternary COS/MOS Memory and Sequential Circuits," in *IEEE Transactions on Computers*, vol. C-26, no. 3, pp. 281-288, March 1977, doi: 10.1109/TC.1977.1674821

[117] A. Heung and H. T. Mouftah, “Depletion/enhancement CMOS for a lower power family of three-valued logic circuits," in *IEEE Journal of Solid-State Circuits*, vol. 20, no. 2, pp. 609-616, April 1985, doi: 10.1109/JSSC.1985.1052354.

[118] A. Raychowdhury and K. Roy, “Carbon-nanotube-based voltage-mode multiple-valued logic design," in *IEEE Transactions on Nanotechnology*, vol. 4, no. 2, pp. 168-179, March 2005, doi: 10.1109/TNANO.2004.842068.

[119] S. Lin, Y. Kim and F. Lombardi, “CNTFET-Based Design of Ternary Logic Gates and Arithmetic Circuits," in *IEEE Transactions on Nanotechnology*, vol. 10, no. 2, pp. 217-225, March 2011, doi: 10.1109/TNANO.2009.2036845.

[120] H. W. Kim, S. Kim, K. Lee, J. Lee, B. G. Park and D. Kwon, “Demonstration of Tunneling Field-Effect Transistor Ternary Inverter," in *IEEE Transactions on Electron Devices*, vol. 67, no. 10, pp. 4541-4544, Oct. 2020, doi: 10.1109/TED.2020.3017186.

[121] Schuman, D. Catherine et al., “A Survey of Neuromorphic Computing and Neural Networks in Hardware," in *arXiv preprint arXiv:1705.06963*, 2017.

[122] K. Roy, A. Jaiswal and P. Panda, “Towards spike-based machine intelligence with neuromorphic computing", in *Nature*, vol. 575, no. 7784, pp. 607–617, 2019. [Online]. Available: doi: 10.1038/s41586-019-1677-2.

[123] J. Torrejon, M. Riou, F. Araujo et al., “Neuromorphic computing with nanoscale spintronic oscillators", in *Nature*, vol. 547, no. 7664, pp. 428–431, 2017, doi: 10.1038/nature23011.

[124] V. P. Hu, H. Lin, Y. Lin and C. Hu, “Optimization of Negative-Capacitance Vertical-Tunnel FET (NCVT-FET)," in *IEEE Transactions on Electron Devices*, vol. 67, no. 6, pp. 2593-2599, June 2020, doi: 10.1109/TED.2020.2986793.

[125] A. Gupta and S. Saurabh, “Unsupervised Learning in a Ternary SNN Using STDP," in *IEEE Journal of the Electron Devices Society*, vol. 12, pp. 211-220, 2024, doi: 10.1109/JEDS.2024.3366199.

[126] S. Garg and S. Saurabh, “Exploiting Within-Channel Tunneling in a Nanoscale Tunnel Field-Effect Transistor," in *IEEE Open Journal of Nanotechnology*, vol. 1, pp. 100-108, 2020, doi: 10.1109/OJ-NANO.2020.3031633.

Publications

1. **A. Gupta** and S. Saurabh, “Implementing a Ternary Inverter Using Dual-Pocket Tunnel Field-Effect Transistors,” in *IEEE Transactions on Electron Devices*, vol. 68, no. 10, pp. 5305-5310, Oct. 2021, doi: 10.1109/TED.2021.3106618.

2. **A. Gupta** and S. Saurabh, “Novel attributes of a dual pocket tunnel field-effect transistor,” in *Japanese Journal of Applied Physics*, vol. 61, no. 3, p. 035001, 2022. doi: 10.35848/1347-4065/ac3722.

3. **A. Gupta** and S. Saurabh, “An Energy-Efficient Ge-Based Leaky Integrate and Fire Neuron: Proposal and Analysis,” in *IEEE Transactions on Nanotechnology*, vol. 21, pp. 555-563, 2022, doi: 10.1109/TNANO.2022.3209078.

4. **A. Gupta** and S. Saurabh, “On-chip Unsupervised Learning using STDP in a Spiking Neural Network,” in *IEEE Transactions on Nanotechnology*, vol. 22, pp. 365-376, 2023, doi: 10.1109/TNANO.2023.3293011.

5. **A. Gupta** and S. Saurabh, “Unsupervised Learning in a Ternary SNN Using STDP," in *IEEE Journal of the Electron Devices Society*, vol. 12, pp. 211-220, 2024, doi: 10.1109/JEDS.2024.3366199

Brief Bio data of the Author

Abhinav Gupta was born in Delhi, India. He received his B.Tech Degree in Electronics and Communication Engineering from Delhi Technological University (DTU), Delhi, India, and his M.Tech Degree with specialization in Very Large Scale Integration and Embedded Systems from Delhi Technological University (DTU), Delhi, India. Currently, he is pursuing his Ph.D. with the Department of Electronics and Communications, Indraprastha Institute of Information and Technology, New Delhi, India. His research interests include Neuromorphic Computing, Nanoelectronics, Semiconductor Devices, Energy-Efficient Devices, and Circuits.