Synaptic plasticity in VLSI : a floating-gate approach

Gopalakrishnan, Roshan

2017

Gopalakrishnan, R. (2017). Synaptic plasticity in VLSI : a floating-gate approach. Doctoral thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/72073>

<https://doi.org/10.32657/10356/72073>

SYNAPTIC PLASTICITY IN VLSI: A FLOATING-GATE APPROACH

ROSHAN GOPALAKRISHNAN

SCHOOL OF ELECTRICAL AND ELECTRONIC ENGINEERING

2017

SYNAPTIC PLASTICITY IN VLSI: A FLOATING-GATE APPROACH

ROSHAN GOPALAKRISHNAN

School of Electrical & Electronic Engineering

A Thesis Submitted to the Nanyang Technological University
in partial fulfilment of the requirement for the degree of
Doctor of Philosophy

2017

Acknowledgements

You are the most special in my life.

You were always with me, to support me.

You created a whole new world for me.

Without you I am nothing.

I bend my head in gratitude, before you.

All I could give is deep hearted thanks as gifts

To my beloved, late father V.P Gopalakrishnan, my mother K. Premachandrika, my wife Neethu G.D, my sister Reshma Gopalakrishnan, my brother in law Dhanesh Das, my cousin Binu J.Raj and all other family members and friends.

To my supervisor Arindam Basu, who has opened a new life for me. I learned new perspective of understanding things from him. I am thankful to him for the valuable discussions with me and the time dedicated for me.

To Jennifer Hasler, who helped me during my journey.

To all my NTU friends and VIRTUS labmates for their valuable discussions and support.

To all the staff members of VIRTUS, IC Design Centre of Excellence for their technical and administrative support.

Finally, I want to acknowledge the financial support from NTU Research Scholarship, and from MOE through grant ARC 8/13.

Abstract

One of the major areas of research by neurobiologists is long term synaptic modification or plasticity of synapses. Synapses are the interconnections between neurons. Each synapse is associated with a weight quantity, which determines the amplitude of output excitatory postsynaptic current (EPSC) pulse from the respective synapse to the interconnected neuron. Neurons on the other hand are the basic computational unit. It integrates all the incoming current pulses from different synapses and fires a spike, if the integration reaches a threshold. Neurons and synapses are the major units of a biological neural network that are typically included in computational models; however, recently there has been an increasing focus on studying the role of astrocytes [1, 2] and may be a part of future large scale models.

It is generally believed that the ability to learn and execute difficult tasks in human brain is mediated by synapses. Likewise in artificial spiking neural networks (SNNs), synapses play a major role in the ability to carry out signal processing, classification, computations etc. Basically synapses play a fundamental role in learning through activity dependent modification of their efficacies. The mathematical model under which synapses modify their weights are termed as synaptic plasticity rule. Recently a learning rule known as spike timing dependent plasticity (STDP) that modifies weights based on the timing of presynaptic and postsynaptic spikes become very popular. The most commonly used rule posits weight change based on time difference be-

tween one presynaptic spike and one postsynaptic spike and is hence termed doublet STDP (D-STDP). Potentiation occurs when a postsynaptic spike succeeds a presynaptic spike; otherwise depression happens. However, D-STDP could not reproduce results of many biological experiments; a triplet STDP (T-STDP) that considers triplets of spikes as the fundamental unit has been proposed recently to explain these observations. Computationally T-STDP has advantages like replicating rate based plasticity experiments and detecting third order input correlations. As these rules are being proposed by neurobiologists, we as engineers try to implement them in silicon in order to employ them in complex real world applications.

We first propose a method to robustly perform doublet STDP rule in a single floating-gate (FG) transistor synapse. The experimental STDP plot of a FG synapse (change in weight against $\Delta t = t_{post} - t_{pre}$) from previous study shows a depression instead of potentiation at some range of positive values of Δt . To ameliorate this STDP graph of a FG synapse, a minimum hardware overhead solution is proposed. The measurement results from a FG synapse fabricated in AMS 0.35 μ m CMOS process design are presented to justify the claim.

Next, we propose a modification to the basic FG synapse to implement a more sophisticated and bio-realistic rule - triplet STDP (T-STDP). Compared to the previous work, here we propose a method to localize the effect of potentiation and depression on a FG synapse. The spike triplet affects the setting of drain voltage—we present a single pulse and a double pulse drain voltage method to obtain the desired dependence of weight on spike timing. We present a mathematical procedure to calculate the parameters of the drain voltage pulse to obtain results matched to the original theoretical T-STDP rule. We also show measurement results from a FG synapse fabricated in TSMC 0.35 μ m CMOS process in comparison with the biological experimental observations for (1) original doublet protocol, (2) two protocols of spike triplets, (3) frequency effects of pairing protocol, (4) quadruplet experiments and (5) replication of

BCM rule. Possible VLSI implementation of drain voltage waveform generator circuits are also presented with simulation results.

Finally, we showcase the computational advantages of T-STDP model apart from the experimental advantages compared to D-STDP model i.e. detecting third order input correlations and faster convergence of synaptic weights. A classic STDP experiment is chosen and implemented using both D-STDP model and T-STDP model in MATLAB. MATLAB simulations prove the faster convergence of synaptic weights in the case of T-STDP model. To understand the performance of a silicon neural network with integrate and fire neurons and the proposed FG synapses, we have performed SPICE simulations of such a system with a behavioural model of the FG device. The third order correlation model is simulated in SPICE and the results are matched with MATLAB simulations to lay the foundation for future silicon implementations.

In future, the work described in this thesis can be extended to make software simulations of larger neural networks with multiple neurons and synapses following the FG based T-STDP learning rules. VLSI implementations of the same may also be fabricated using low-power subthreshold analog techniques. Further, these models may be used to classify realistic spiking image datasets such as the one described in [3].

Contents

Acknowledgements	iv
Abstract	vi
List of Figures	xiv
List of Tables	xviii
List of Abbreviations	xix
1 Introduction	1
1.1 Motivation and Objective	1
1.2 Neural Network	3
1.2.1 Biological Neural Network	3
1.2.2 Artificial Neural Network	6
1.2.3 Spiking Neural Network	9
1.3 Neuromorphic Chips	11
1.4 Report Organization	12

2	Literature Review	14
2.1	Synaptic Plasticity: Learning Rules	15
2.1.1	Rate Based Models	16
2.1.2	Spike Based Models	17
2.2	Synaptic Plasticity: VLSI Circuits	21
2.2.1	Bi-Stable Synapse by Fusi	23
2.2.2	Bi-Stable Synapse by Indiveri	27
2.2.3	Bi-Stable Synapse by Srinjoy	32
2.2.4	Binary Synapse by Arthur	34
2.2.5	Synapse by Schemmel	35
2.2.6	Synapse by Bofill	37
2.2.7	T-STDP Synapse by Mostafa	39
2.2.8	Single Transistor Learning Synapse	41
2.2.9	Memristor Synapse	45
2.3	Comparison of Synapse Circuits	45
3	Robust Doublet STDP using floating-gate synapse	48
3.1	Floating Gate Synapse	49
3.1.1	Basic Operation	49
3.1.2	Theoretical Model of Learning	52

3.1.3	Proposed STDP protocol for FG synapse	59
3.2	Results	61
3.2.1	Experimental Setup	61
3.2.2	Experimental Results: Robustness of “non-STDP” Behavior	62
3.2.3	Simulation Results: Effect of Variations	64
3.3	Gate Voltage Generator: VLSI Implementation	64
3.4	Discussion	67
4	Triplet Spike Time Dependent Plasticity in a floating-gate synapse	68
4.1	STDP Synaptic Modification Rule	70
4.1.1	Triplet STDP (T-STDP) Model	71
4.2	Floating Gate Synapse	72
4.2.1	D-STDP Model on a Floating Gate Synapse (FG D-STDP)	77
4.2.2	T-STDP Model on a Floating Gate Synapse (FG T-STDP)	81
4.2.3	Parameter Translation from Theory to FG model	86
4.3	Results	88
4.3.1	FG D-STDP learning window	88
4.3.2	Failure of FG D-STDP	91
4.3.3	Success of FG T-STDP	93
4.4	Drain Voltage Generator: VLSI Implementation	98

4.4.1	Single-Pulsed Drain Voltage Waveform	98
4.4.2	Double-Pulsed Drain Voltage Waveform	99
4.5	Discussion	102
5	Comparison of Triplet-STDP with Doublet-STDP: A simulation study	106
5.1	Introduction	107
5.2	Model and Methods	108
5.2.1	Leaky Integrate and Fire (LIF) Neuron model	108
5.2.2	Synaptic Modification Rule	109
5.2.3	Parameter Extraction	109
5.3	Simulation Results: MATLAB	111
5.3.1	Single Wave Experiment (SWE): Mathematical Model	111
5.3.2	Single Wave Experiment (SWE): Floating Gate Model	113
5.3.3	Third Order Correlation (TOC) Experiment	117
5.4	Simulation Results: SPICE	119
5.4.1	Third Order Correlation (TOC) Experiment	119
5.5	Discussion	123
6	Conclusion and Future Work	125
6.1	Conclusion	125
6.2	Future Work	127

Publication	130
Bibliography	132

List of Figures

1.1	Biological Neuron Cell	4
1.2	Synaptic cleft between two neurons	5
1.3	Artificial Neural Network (ANN)	6
1.4	Block Diagram of a simple Neural Network	8
1.5	High level general block diagram of a Neuromorphic Chip	11
2.1	Theoretical implementation of Doublet STDP (D-STDP)	18
2.2	Triplet STDP model timing specification	20
2.3	Evolution of Hardware Synapse Circuits	23
2.4	Bi-stable Synapse Circuit by Fusi	24
2.5	Bi-stable Synaptic Dynamics	25
2.6	Bi-stable Synapse Circuit by Indiveri	27
2.7	Simulation Results	30
2.8	STD Results	31

2.9	SDSP Bi-stable Synapse Circuit	33
2.10	Stop Learning Circuit Simulation Result	34
2.11	STDP Circuit by Arthur	35
2.12	STDP Circuit by Schemmel	36
2.13	STDP Circuit by Bofill	38
2.14	T-STDP Circuit by Mostafa	39
2.15	STLS, STDP Curve and Voltage waveform	43
2.16	Layout and characteristics of Floating gate MOSFET:	44
3.1	Floating gate Mosfet	49
3.2	Parameter specifications of terminal voltage waveforms	50
3.3	Timing diagram for STDP in FG synapse	51
3.4	Weight variation in FG synapse obtained from theory for the two different gate voltage waveforms	55
3.5	Block Diagram of experimental setup	61
3.6	Experimental results for the STDP curve obtained from a FG device	63
3.7	Parametric variations of STDP curve from FG synapse	65
3.8	Circuit to generate modified gate voltage waveform and it's SPICE simulation result	66
4.1	Timing diagram of T-STDP rule	73

4.2	Floating Gate (FG) synapse	74
4.3	Timing diagram and voltage specifications of FG synapse	76
4.4	Different cases of drain voltage waveform	82
4.5	Experimental Setup	89
4.6	FG D-STDP learning window	89
4.7	Failure of FG D-STDP	90
4.8	FG T-STDP using single-pulsed drain waveform	94
4.9	FG T-STDP using double-pulsed drain waveform	95
4.10	FG T-STDP model can be mapped to BCM learning rule	97
4.11	VLSI implementation of single-pulsed and double-pulsed drain voltage waveform	100
4.12	SPICE Simulation results	101
4.13	System level architecture of SNN	103
5.1	Parameter Extraction: T-STDP experiment and T-STDP models on quadru- plets of spikes:	109
5.2	Single wave experiment	112
5.3	Postsynaptic Latency of SWE:Floating gate model:	114
5.4	Third order correlation experiment	118
5.5	Neural network consisting of 6 synapses and a neuron	120

5.6	Floating gate synapse model	121
5.7	Integrate and Fire Neuron model	122
5.8	Third Order Correlation Spice simulation	124

List of Tables

1.1	Comparison of Neuromorphic chips	12
2.1	Features of different synaptic plasticity rules	22
2.2	Hardware Synapse Circuit Comparison	47

List of Abbreviations

SNN	Spiking Neural Network
STLS	Single transistor learning synapse
STDP	Spike timing dependent plasticity
SDSP	Spike Driven Synaptic Plasticity
BCM	Bienenstock Cooper Munro
EPSC	Excitatory pre/post synaptic current
LTP	Long term potentiation
LTD	Long term depression
STD	Short term depression
CMI	Current mirror integrator
NMDA	N-methyl D-Aspartate
SRAM	Static Random Access Memory
FG	Floating Gate
HEI	Hot Electron Injection
FN	Fowler Nordheim
CMOS	Complementary Metal Oxide Semiconductor
MOSFET	Metal Oxide Semiconductor Field Effect Transistor

Chapter 1

Introduction

1.1 Motivation and Objective

Human brain's supremacy over traditional digital processors in pattern recognition and learning tasks is mainly due to parallel data processing and extremely low power consumption. Also, the performance of even insects with simple brains in tasks such as obstacle avoidance are better than state of the art UAV with sophisticated digital processors and algorithms. So understanding the principles of operation of the brain has always been an interesting field of research and this is regarded as one of the grand scientific challenges of the twenty-first century. Till now science could only understand the functionality of some parts of the brain – neuromorphic engineers try to mimic those behaviour of brain using electronics. So getting inspired by human learning, we devise mathematical models or algorithms for self-learning that are similar to the synaptic plasticity mechanisms observed in biology.

Our brain consists of millions of neurons which are interlinked together. Neurons are considered as the basic computational unit of the brain. Neuronal dynamics

depend on synaptic connections while synaptic efficacies or “weight” evolve according to the neuronal activity. These activities are revealed by the emission of “spikes” which is the output of neuron. These spike trains constitute a “neural code”. By analysing the temporal activities of neuron, computational neuroscientists have come up with possible learning or plasticity rules being employed in the brain. To realize these self learning rules using electronics, we have to take into account the different challenges in neuromorphic engineering [4] such as non volatile storage of weight, power consumption, silicon area occupation by neuron and synapse circuits etc as elaborated below.

If there are “ n ” neurons in the network, then the number of connections or synapses between these neurons will be of the order of n^2 . Since synapses are the main block for weight update or learning and their number is much more compared to neurons, it is extremely important to keep the synapse circuit compact while retaining the capability of implementing different learning or plasticity rules. Similar to compactness, low-power consumption is also a must and it can be achieved by utilizing the transistors operating deep in the subthreshold region or to minimize the number of active circuits (e.g. amplifiers). Another effective way to decrease power consumption of neuromorphic circuits is to optimize the neuron and synapse circuit bias parameters, as well as the width of spike pulses [4]. Finally, a third important criteria is that synapses must permanently store weight so that the effect of learning does not disappear in the absence of stimuli. Keeping the above points in mind, we create a robust self learning system with single transistor learning synapse (STLS) [5] which can implement different learning rules.

The STLS are modified EEPROM devices. STLS utilizes floating gate transistor’s quantum phenomena, such as hot electron injection and fowler nordheim tunneling to simultaneously provide long term storage (non-volatile) and adaptation of weight in a single device. This floating gate transistor is operated in subthreshold region to

reduce the power consumption. As the name suggest “single transistor”, the synapse circuit contain only one transistor which reduces the area effectively.

There are many different learning rules developed for spiking networks—we expect variations of the STLS can be used to implement each of these rules. Verifying and studying these rules on a single synapse level will make it easier to implement on a network level. In the long term, the goal would be to produce a novel, low-power, fault tolerant machine learning processor inspired from biological networks.

1.2 Neural Network

A neural network is an information processing paradigm that is inspired by the way in which brain process information to solve problems which cannot be rapidly tackled by conventional techniques [6]. Neuromorphic engineering is concerned with mimicking the structure, processes and functions of biological systems into electronic circuits. These electronic circuits are then combined to construct artificial neural networks (ANNs) which closely resemble biological neural networks (BNNs).

1.2.1 Biological Neural Network

The human brain contains more than 10^{10} interconnected neurons, with each having up to 10^3 synaptic connections [9]. Neuron is a processing block which typically integrates the incoming signals and transmits the processed data. A neuron typically operates slower than silicon logic gates. However, slower operating speed is compensated by the massive parallelism of the human brain. Fig.1.1 shows the typical structure of a biological neuron.

Figure 1.1: **Biological Neuron Cell:** The main parts of a typical neuron cell are Soma, Dendrites, Axon and Terminal Buttons. Figure is taken from [7].

A typical biological neuron contains a soma, an axon, dendrites and terminal buttons. The soma performs temporal integration of incoming synaptic transmissions. It has a typical resting membrane potential of -65mV . If these several synaptic input transmissions combine to increase the postsynaptic membrane potential (PSP), above a specific threshold voltage, then the neuron will generate or fire an action potential. This action potential propagates down the axon of the postsynaptic neuron. Fig.1.4(c) shows the generation of an action potential when successive synaptic inputs occur. After the generation of action potential, neuron enters into refractory period and usually lasts up to one millisecond. During refractory period, the membrane potential undershoots its resting potential, causing the neuron to enter a state of hyperpolarisation. The refractory period is due to inactivation mechanism in the voltage-gated

Figure 1.2: **Synaptic cleft between two neurons:** Synaptic vesicles and neurotransmitters and its receptors are shown. Figure is taken from [8].

sodium channels and a lag in the closing of potassium channels within the neuron. The membrane potential returns to its resting potential as the potassium channels begin to close, causing the refractory period to last up to milliseconds.

Transfer of information between each presynaptic neuron and postsynaptic neuron is achieved through chemical (or electrical) synapses. A chemical (or electrical) synapse is formed when a synaptic cleft is formed between the axon of a presynaptic neuron and the dendrite of a postsynaptic neuron, Fig.1.2. The axon terminal contains synaptic vesicles or neurotransmitter vesicles. Neurotransmitters are chemicals to amplify, modulate and relay signals between neurons. The neurotransmitters are released into the synaptic cleft and bind to neurotransmitter receptors located within the postsynaptic membrane causing ion channels to open or close. This modifies the potential of the postsynaptic membrane. Each synapse can be of two cat-

Figure 1.3: **Artificial Neural Network (ANN):** Single output, 'n' input, simple ANN is shown here with a summing junction and an activation block.

egories; Excitatory synapse, where neurotransmitters increase the potential of the postsynaptic membrane; Inhibitory synapse, where neurotransmitters decrease the postsynaptic membrane potential.

1.2.2 Artificial Neural Network

ANNs are computational representations to emulate the structure and operation of biological neural networks. ANNs are used to approximate functions that can depend on a large number of inputs and are generally unknown [10]. ANNs consist of several interconnected processing elements or neurons which exchange information between each other. These interconnections have numeric weights or efficacies that can be modified based on certain activities or experience, making the network capable of learning. Modification of the synaptic weight will alter the flow of information within the

network. The strength of the inputs to the neurons are altered such that the neurons weighted sum output is also changed. This forms the simplified basis for learning and synaptic plasticity. Fig.1.3 shows an example of artificial neuron which can be trained to solve a specified problem. This is achieved with the help of a learning rule which modifies the weighted synaptic connections between each neuron.

The output from the summing junction is given as;

$$V_k = \sum_{j=1}^n w_{kj} x_j \quad (\text{Eq. 1.1})$$

and the output from the neuron, Y_k , is;

$$Y_k = \phi(V_k) \quad (\text{Eq. 1.2})$$

$\phi()$ represents an activation function which determines when the neuron fires.

Artificial neural networks (ANNs) can broadly be classified into three generations [11]. The first generation models restricted the output signals to discrete '0' or '1' values by using a threshold activation function. These models consisted of McCulloch and Pitts neurons. McCulloch-Pitts neurons produce an output of 1 when the weighted sum of its inputs is $V_k > 0$. If the weighted sum is $V_k \leq 0$, an output of 0 is produced. Even though McCulloch-Pitts neurons produce a digital output, they have been successfully applied in networks of multilayer perceptron or Hopfield networks, where the overall output required is Boolean. The second generation models allowed the output to take values between '0' and '1' by using a continuous activation function. Here, threshold activation functions were replaced with continuous activation functions. Continuous activation functions, such as sigmoid or hyperbolic tangent functions etc were used. Second generation models are more suited to analog computations, at the same time requiring fewer neurons for digital computation than the first generation models [12]. These analog output between 0 and 1 could be considered

Figure 1.4: **Block Diagram of a simple Neural Network:** (a) There are two paths in the block diagram of neural network: one with the synapse between presynaptic neuron and postsynaptic neuron is the feedforward path while the other with the learning block gives a feedback to the synapse for its weight update according to the occurrence of presynaptic and postsynaptic spikes. (b) Current spike, I_{syn} is an exponential decay function at the occurrence of presynaptic spike. (c) (taken from [13]) Dynamics of membrane potential: when the membrane potential, $U(t)$ reaches a threshold, V_{thr} then neuron fires a spike and it is reset to the resting potential.

as normalized firing rates, which is often called a rate-coding scheme as it implies some averaging mechanism. In the second generation of ANN, learning rules based upon gradient-decent algorithms can be used to train the weights between neurons and aid in specific task.

1.2.3 Spiking Neural Network

Spiking neural networks (SNNs) belong to the third generation of artificial neural networks. Since the concept of time is inherent in the model, SNNs are more biologically plausible. Similar to biological neurons, they use spikes or pulses to represent information flow. Recent research shows the importance of time in that the biological neurons store information in the synapses with respect to the timing of spikes [11].

Block diagram in Fig.1.4(a) shows a simple network with a synapse connected between two neurons. This diagram has two important paths: one is the feedforward path which contains a presynaptic neuron, a synapse and a postsynaptic neuron. This path conducts the synapse current, I_{syn} from presynaptic neuron to postsynaptic neuron. During conduction, the input to synapse, V_{pre} is multiplied with synaptic weight, W to result in synaptic output, I_{syn} . This is expressed mathematically in Eq. 1.3. A pictorial example of I_{syn} is shown in Fig.1.4(b). On the other hand neuron sums up all the synaptic current inputs and fires a spike, V_{post} when the summed up current rises above a threshold, V_{thr} ; mathematical expression is given in Eq. 1.5. The second path (feedback path) contains the learning block which operates depending on the output spike of neuron (V_{post}) and the occurrence of presynaptic spike (V_{pre}). Depending on the temporal occurrence of V_{pre} and V_{post} , the learning block updates the synaptic weight, W , as given in Eq. 1.3. The function, f_{learn} , in the equation, depends on the plasticity rule used for learning and the details of the learning rules are given in the next chapter. Generalized mathematical expressions of feedforward and feedback signals of the neural network is given below.

$$\begin{aligned}
 I_{syn_i}(t) &= w_i(t)f_{syn}(V_{pre_i}(t)) \\
 V_{post}(t) &= f_{neu}(V_{post}(t-1), I_{syn_1}(t), I_{syn_2}(t), \dots) \\
 w_i(t) &= f_{learn}(V_{pre_i}(t), V_{post}(t))
 \end{aligned}
 \tag{Eq. 1.3}$$

where f_{syn} , f_{neu} and f_{learn} are the functions corresponding to synapse, neuron and learning respectively and 'i' denote a particular synapse to the neuron. $V_{pre}(t)$, as shown in Fig.1.4(b), as an example, is the summation of impulses at different time instances, $t_1 - t_6$. A popular choice for f_{syn} is an exponentially decaying function. With that choice, $I_{syn}(t)$ can be expressed as:

$$\begin{aligned} V_{pre1}(t) &= \sum_{t_i} \delta(t - t_i) \\ I_{syn1}(t) &= \sum_{t_i} w_1(t_i)k(t - t_i); \quad \text{where } k(t) = e^{\frac{-t}{\tau_{syn}}} \end{aligned} \quad (\text{Eq. 1.4})$$

For the neuronal dynamics, a popularly used model is the integrate and fire one. For that case, the mathematical expression depicting the operation of a neuron is given as,

$$\tau_{neu} \frac{dV_{mem}(t)}{dt} + V_{mem} = R_{mem} \sum_i I_{syn_i}(t) \quad (\text{Eq. 1.5})$$

where V_{mem} is the membrane potential of neuron. If $V_{mem}(t) > V_{thr}$ then $V_{post}(t) \Rightarrow 1$ and $V_{mem}(t) \Rightarrow 0$; else $V_{post}(t) = 0$.

After understanding these variants of neural networks as seen above, an usual query is why SNNs ? Brain activities are in the range of 1-10 Hz in average. Even with lower frequency, brain could do much complex tasks with ease (due to the parallel processing of brain). Hence, considering spike events, which are rare, make neural network behave close to brain in doing particular complex tasks. These spike events or pulses become more energy efficient for a neuromorphic chip. Moreover, SNNs can support a biologically plausible, unsupervised, stable learning rule, Spike Timing-Dependent Plasticity (STDP). Also for coding, SNNs can use both rate and temporal coding. Here, the control mechanism of a simple spiking neural network with a single synapse and two neurons are seen. Next, a bigger picture of a neuromorphic chip, which is a collection of these simple SNNs are shown.

Figure 1.5: **High level general block diagram of a Neuromorphic Chip:** Different blocks: activation unit, control unit, weights block and Neuron state block are shown. The activation block and control units are inserted in a dotted blue box which indicate that these blocks must be present on chip, while other blocks can be either on chip or off chip.

1.3 Neuromorphic Chips

Neuromorphic chips or neural network hardware are created for both general purpose or for a specific purpose. These chips can be made using analog circuits, digital circuits or a combination of both. However the main blocks, control units, inputs and outputs of a neuromorphic chip remains the same. A high level general architectural block diagram of a neuromorphic chip [14] is shown in Fig.1.5. The activation block and control unit are indicated in blue dotted box which indicate that these blocks must be implemented on chip where as other blocks can be either on chip or off chip. The activation block does the summation of products, i.e it sums up the product of inputs with corresponding weights. The weight block stores all the weights digitally or in analog form. The weight can be loaded statically before the processing of activation block

or updated dynamically either during the learning phase by the activation block or by the host. Weight block can be either on chip or off chip. The sum of products from the activation block is transformed through transfer function block before sending to neuron state block for processing the output. The control units are used to handle the neuromorphic chip by a host. The data flow between the blocks are controlled by the control parameters. Further the connectivity between neuromorphic chips are represented using the address-events, which is termed as Address Event Representation (AER) [15, 16, 17, 18]. AER is used to transmit spikes from an array of neurons on one chip to a particular location in an array on a second chip. Some of the examples of popular neuromorphic chips are in table 1.1.

Neuromorphic chips	Technology (nm)	Description	No: of Neurons	No: of Synapses	Total Chip Area per synapse per neuron (μm^2)	Total Energy Consumption per synaptic event (pJ)
HiCANN [19]	180	Mixed Signal. Each wafer has 384 chips. Neurons are analog. Synaptic weights and interchip communication are digital.	512 per chip	224 per neuron	436	198
SpiNNaker [20]	130	Completely digital. Consists of array of nodes. Each node has 18 ARM9 cores.	1000 per ARM9 core	10K per ARM9 core	-	-
Neurogrid [21]	180	Analog/Digital hybrid. 16 neurochips on full board.	256x256 per neurocore	500M per neurocore	0.63	119
GoldenGate [22]	45	Completely Digital. Synapse crossbar array. Processor in memory approach.	256	256K	16	1.9
TrueNorth [23]	28	Completely Digital. Event driven communication framework.	1 Million	256 Million	1.6	26
HRL chip [24]	90/180	Analog/Digital hybrid. Synaptic weights stored in memristors. 3.5 bit of weight storage per memristor.	576	73728 (70K)	1	-

Table 1.1: **Comparison of Neuromorphic chips:** Different popular neuromorphic chips are compared with different metrics.

1.4 Report Organization

The organization of this report is as follows.

Chapter 2 starts with a brief introduction on the different learning rules. Then we present the evolution of hardware synapse circuits till date. We explicitly mention

few of the synapse circuits which can implement learning rules. A comparison of these circuits are done as the conclusion to evolution of hardware synapses.

In chapter 3, we present our proposed floating-gate based synapse that can exhibit STDP learning rule robustly across different parameter regimes. Initially, a brief introduction to floating gate synapse is given. Then a “non-STDP” behaviour that is observed in the literature is shown. The result shows this “non-STDP” behaviour and an explanation of its occurrence is given. A remedy is proposed, which is then justified by measurement results from an AMS $0.35\mu m$ chip. Then the chapter is concluded with a possible VLSI implementation of the proposal.

Triplet spike time dependent plasticity on floating gate synapse is mentioned on chapter 4. A brief introduction to T-STDP model is given first. Then the D-STDP and T-STDP models on a floating gate synapse is derived along with the parameter translation from theory to floating gate model. The TSMC $0.35\mu m$ chip measurement result for failure of D-STDP model on FG synapse and success of T-STDP model on FG synapse is shown. Then the chapter is concluded with possible hardware implementations of drain voltage waveform generator and verified through SPICE simulation results.

Chapter 5 compares triplet spike time dependent rule with doublet spike time dependent rule. It starts with the model and methods used in the experiments, which explains an integrate and fire neuron. Then the MATLAB results for two of the classic STDP experiments are shown to prove the advantage of T-STDP over D-STDP. The chapter is concluded with the modelling of floating gate synapse using SPICE to replicate one of the classic STDP experiment, which is close enough to acknowledge the realization of a neuron with STDP models onto a neural network hardware.

Finally, chapter 6 concludes the thesis with a glimpse into the future work.

Chapter 2

Literature Review

In 1949, Donald Hebb [25], postulated that "When an axon of cell A is near enough to excite cell B or repeatedly or consistently takes part in firing it, some growth or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased." This created a base for the learning rules and later on several other learning rules came into being. Still neurobiologists are continuously trying to understand the nature of learning happening in brain and they devise new rules from their observations. Such important learning rules are explained in the upcoming section. Apart from helping us understand how the brain works, such learning rules can also be used to make "intelligent" machines that can possibly operate autonomously. Hence, neuromorphic engineers try to implement these observed learning rules on to low power circuits that can operate for long periods of time on a battery. Subsequent section to the upcoming section talks about the evolution of such learning circuits.

2.1 Synaptic Plasticity: Learning Rules

Synaptic plasticity rules depend on several factors such as the animal preparation studied, the area of the brain analyzed, the way the experiment is carried out and many other factors [26, 27]. These factors have created controversial results and led to the development of a large number of synaptic plasticity models ranging from abstract models to more detailed models. The abstract models of synaptic plasticity are typically based on the timing of the spikes produced by presynaptic neurons and postsynaptic neurons. Examples of such learning rules are doublet spike timing dependent plasticity(D-STDP) [28] and triplet spike timing dependent plasticity(T-STDP) [29]. On the other hand the more detailed models take into account the neuron and synapse state variables in addition to the timing of the spikes. Examples of such models include the spike driven synaptic plasticity(SDSP) model [30],the voltage-based STDP model [31],and membrane potential-based models [32]. When we want to use any of these models it is important to consider the target application and the level of complexity required. However, the two broad classification of synaptic plasticity models are

- (i) Phenomenological models
- (ii) Biophysical models

Phenomenological models try to replicate experimental findings at a very high level of abstraction. These models take in a set of variables as input to produce a change in synaptic efficacy as output. Biophysical models, in contrast to phenomenological models, concentrate on modeling the biochemical processes or ion channel dynamics that lead to the induction of synaptic plasticity in great details [33, 34, 35, 36]. Phenomenological models are further classified into rate based and spike based models.

2.1.1 Rate Based Models

In this model, the rate of pre and postsynaptic firing over some time period determines the sign and magnitude of the synaptic weight change [37]. Examples of rate based models are proposed by Linsker [38], plasticity model by Oja [39], and the popular BCM rule [40], which is summarized next.

2.1.1.1 BCM Rule

Bienenstock Cooper Munro (BCM) rule indicates the change in the efficacy of a synapse, not only depends on the instantaneous presynaptic and postsynaptic activities but also on a slowly varying time averaged value depending on the postsynaptic activity [40]. In this phenomenological rule, a presynaptic activity variable $g(t)$ is multiplied with a shifted postsynaptic activity variable $u(t)$ to define the change of synapse weight, m [4]:

$$\frac{dm(t)}{dt} = (u(t) - \phi_u) \cdot g(t) \quad (\text{Eq. 2.1})$$

where $m(t)$ is the synaptic weight, $u(t)$ is the neuron's membrane potential, ϕ_u is a threshold that separates LTP and LTD induction, and $g(t)$ is a conductance variable that is related to the postsynaptic current I_{psc} and therefore has its maximum value at the time of a presynaptic arrival and decays thereafter.

Synapse Dynamics: For the synaptic conductance $g(t)$, we use an exponential decay function in response to each presynaptic spike [32],

$$g(t) = G' \exp\left(\frac{-(t - t_j^{pre})}{\tau_g}\right), \quad t_j^{pre} \leq t < t_{j+1}^{pre} \quad (\text{Eq. 2.2})$$

where τ_g is the time constant of the decay and G' is the amplitude of the response. Setting the presynaptic conductance to its maximum G' at the beginning of each presy-

naptic pulse effectively makes the learning carried out in the local correlation plasticity (LCP) rule history-independent. This would be equivalent to a presynaptic nearest neighbour interaction in a standard STDP terminology [29].

Neuron Dynamics: For the membrane potential, $u(t)$ a spike response neuron model is chosen [13], consisting of a Dirac function for the action potential and an exponential decay function for the hyperpolarization after firing

$$u(t) = U_{p,n} \cdot \delta(t - t_n^{post}) + U_{refr} \cdot \exp \frac{-(t - t_n^{post})}{\tau_{refr}}, \quad t_n^{post} \leq t < t_{n+1}^{post} \quad (\text{Eq. 2.3})$$

where t_n^{post} denotes the n -th postsynaptic spike, $U_{p,n}$ is the area under the pulse curve and $U_{refr} < 0$ and τ_{refr} are the amplitude and time constant of hyperpolarization.

2.1.2 Spike Based Models

Increasing number of research experiments during the last decade, have shown synaptic weight changes with respect to the occurrence of presynaptic and postsynaptic spikes. This prompted the development of several spike timing based or spike timing dependent plasticity which is summarized below.

2.1.2.1 Doublet Spike Timing Dependent Plasticity (D-STDP)

In doublet STDP, potentiation occurs when a postsynaptic spike succeeds a presynaptic spike; otherwise depression happens. The weight changes can be governed by a temporal learning window. The temporal learning window for D-STDP can be

Figure 2.1: **Theoretical implementation of Doublet STDP (D-STDP):** The plot shows the implementation of equation Eq. 2.4 with the parameters $A^+ = 4.6m$ and $A^- = -3m$.

expressed as [41, 29]

$$\Delta w = \begin{cases} \Delta w^+ = A^+ e^{\left(\frac{-\Delta t}{\tau_+}\right)} & \text{if } \Delta t \geq 0 \\ \Delta w^- = -A^- e^{\left(\frac{\Delta t}{\tau_-}\right)} & \text{if } \Delta t < 0 \end{cases} \quad (\text{Eq. 2.4})$$

where $\Delta t = t_{post} - t_{pre}$ is the time difference between a postsynaptic spike and presynaptic spike, τ_+ and τ_- are the time constants of the learning window, and A^+ and A^- are the maximal weight changes for potentiation and depression, respectively. The theoretical graph for the above equation is simulated using MATLAB and is shown in Fig.2.1 with the parameters being obtained by data fitting as explained in [29]. As mentioned by Bi and Poo in [42], τ_+ and τ_- are taken as 16.8ms and 33.7ms respectively for the simulation.

2.1.2.2 Triplet Spike Timing Dependent Plasticity (T-STDP)

Previous studies shows that the doublet STDP fails to reproduce the experimental outcomes involving higher order spike patterns such as triplet and quadruplets of spikes [41, 29] and furthermore, fails to account for the observed dependence on repetition frequency of pairs of spikes. To resolve these issues, the doublet STDP was extended to include spike triplets resulted in triplet STDP(T-STDP) which could sufficiently reproduce physiological experiments. In a T-STDP model the basic building block of Long Term Potentiation (LTP) is not a pair, but a combination of 1 pre and 2 postsynaptic spikes. Similarly for the Long Term Depression (LTD), it is a combination of 2 pre and 1 postsynaptic spikes. Based on the triplet synaptic learning rule presented in [29], the postsynaptic spike arrival at time t_{post} triggers a change given by the following:

$$w(t) \rightarrow w(t) + e^{\left(\frac{-t}{\tau_+}\right)}(A_2^+ + A_3^+ e^{\left(\frac{-t}{\tau_y}\right)}) \quad \text{if } t = t_{post} \quad (\text{Eq. 2.5})$$

Similarly, a presynaptic spike at time t_{pre} triggers a change given by the following:

$$w(t) \rightarrow w(t) - e^{\left(\frac{-t}{\tau_-}\right)}(A_2^- + A_3^- e^{\left(\frac{-t}{\tau_x}\right)}) \quad \text{if } t = t_{pre} \quad (\text{Eq. 2.6})$$

This could be written in another way as given in [41] as

$$\Delta w(t) = \begin{cases} \Delta w^+ = e^{\left(\frac{-\Delta t_1}{\tau_+}\right)}(A_2^+ + A_3^+ e^{\left(\frac{-\Delta t_2}{\tau_y}\right)}) & \text{if } t = t_{post} \\ \Delta w^- = -e^{\left(\frac{\Delta t_1}{\tau_-}\right)}(A_2^- + A_3^- e^{\left(\frac{-\Delta t_3}{\tau_x}\right)}) & \text{if } t = t_{pre} \end{cases} \quad (\text{Eq. 2.7})$$

where A_2^+ and A_2^- denote the amplitude of the weight change whenever there is a pre-post pair or a post-pre pair respectively. Similarly, A_3^+ and A_3^- denote the amplitude of the triplet term for potentiation and depression, respectively as shown in Fig.2.2. $\Delta t_1 = t_{post} - t_{pre}$, $\Delta t_2 = t_{post(n)} - t_{post(n-1)} - \epsilon$ and $\Delta t_3 = t_{pre(n)} - t_{pre(n-1)} - \epsilon$ are time difference between combinations of pre and postsynaptic spikes. τ_- , τ_+ , τ_x and τ_y

Figure 2.2: Illustration of amplitude and decay time constant with respect to the pre and post synaptic spikes for doublet and triplet STDP, LTD and LTP cases. Figure is taken from [29].

are time constants and ϵ is a small positive value which selects the contribution to the weight change just before the final spike among the presented triplet [29].

2.1.2.3 Spike Driven Synaptic Plasticity (SDSP)

In the spike driven synaptic plasticity model, [30] changes in synaptic weight occur whenever a presynaptic spike arrives. At this time, if the postsynaptic membrane potential, V_{mem} is higher than a threshold voltage, V_{mth} , potentiation will occur, or if it is lower than the threshold voltage, V_{mth} , depression will occur, requiring the fact that the amount of calcium concentration in the postsynaptic site, $C(t)$ is within a predefined boundary. In short, the synaptic weight, W will change according to

$$\begin{cases} W = W + a, & \text{if } V_{mem}(t) > V_{mth} \quad \text{and} \quad \theta_{up}^l < C(t) < \theta_{up}^h \\ W = W - b, & \text{if } V_{mem}(t) \leq V_{mth} \quad \text{and} \quad \theta_{dn}^l < C(t) < \theta_{dn}^h \end{cases} \quad (\text{Eq. 2.8})$$

where a and b are the amount of potentiation and depression, respectively. In addition, $[\theta_{up}^l, \theta_{up}^h]$ and $[\theta_{dn}^l, \theta_{dn}^h]$ are the boundaries for the calcium concentration, $C(t)$ for the potentiation and depression respectively.

The increment and decrement of weight as in the above equation solely depend on the condition given. If the required conditions are not satisfied then there is no potentiation or depression. When there is no presynaptic spike occurrence, the synaptic weight, W will drift toward either high or low synaptic weight values. The direction of the drift will depend on the values of the weights at that specific time, which can be above or below a certain threshold θ_W . It is given by the slope equation

$$\begin{cases} \frac{dW(t)}{dt} = \alpha, & \text{if } W(t) > \theta_W \\ \frac{dW(t)}{dt} = -\beta, & \text{if } W(t) \leq \theta_W \end{cases} \quad (\text{Eq. 2.9})$$

This implies that whenever $W(t)$ is above a threshold, θ_W , then $W(t)$ moves towards high synaptic weight value with a positive slope of α . Similarly if $W(t)$ is less than the threshold, θ_W , then $W(t)$ moves towards low synaptic weight value with a negative slope of β .

2.2 Synaptic Plasticity: VLSI Circuits

It is believed that within the human nervous system each neuron can have up to 10^4 synaptic connections [44]. Therefore any synapses which are realized in hardware can potentially account for a large portion of the silicon required to implement neural networks in hardware. Therefore consideration to the functionality, size and power consumption while developing a synapse in silicon is necessary. Several silicon synapses have been proposed which aim to replicate the functionality of biological synapses in VLSI. Similar to the mechanisms of plasticity models explained earlier, Hasler [5] proposed that a silicon synapse should contain five main properties;

1. Synaptic weight should be stored permanently in the absence of learning.
2. Synaptic output should be a product of the neuronal input signal and synaptic

Features of different synaptic plasticity rules	
Learning Rule	Features
T-STDP	<ul style="list-style-type: none"> > Unsupervised Learning rule. > Explains synaptic changes due to triplets or quadruplets of spikes. > Create memories of temporal patterns of spikes. > Reproduce frequency effects. > Replicate rate based, BCM plasticity rule. > Detects third order input correlations. > Faster convergence of synaptic weights compared to D-STDP. > Needs more variables and coefficients, hence complicated implementation.
D-STDP	<ul style="list-style-type: none"> > Unsupervised Learning rule. > Create memories of temporal patterns of spikes. > Replicates rate based, BCM plasticity rule with specific condition. > Simplest form of D-STDP cannot learn patterns of mean firing rate. > Not sufficient to explain synaptic changes due to triplets or quadruplets of spikes. > Fails to reproduce frequency effects.
SDSP	<ul style="list-style-type: none"> > Supervised Learning rule. > Encode patterns of mean firing rates. > Learns to classify correlated patterns of spike trains. > Based on membrane potential. > Bistable Synaptic modifications.
BCM	<ul style="list-style-type: none"> > Create memories of rate based patterns of spikes. > Insensitive to temporal patterns of spike.

Table 2.1: **Features of different synaptic plasticity rules:** Features of different synaptic plasticity rules are summarized here.

weight.

3. Have a minimal area.
4. Have low power dissipation.
5. Capable of Hebbian or back-propagation learning such that the synaptic weight can be modified.

Fig.2.3 gives a detailed structure of classification of most of the synapse circuits in literature till date. This is classified based on the learning efficacy of synapse. Here only the details of the learning synapse circuits are explained. The explanation of each circuit is given in the subsequent subsections.

Figure 2.3: **Evolution of Hardware Synapse Circuits:** Here the focus is given to learning synapse circuits and are explained in detail. Non learning synapse circuit details are given in [43]. The dotted line under the learning synapse side denotes that more emerging nanotechnology synapse types can be included.

2.2.1 Bi-Stable Synapse by Fusi

The Bi-stable synapse circuit presented here is an implementation of the SDSP model explained earlier. The synaptic circuit charge or discharge a capacitor on short time scale and preserves the memory of two stable states on long time scale. The model of the synapse is implemented in aVLSI and consists of only 18 transistors. Synaptic

Figure 2.4: **Bi-stable Synapse Circuit by Fusi:** The circuit implements the plasticity model of SDSP. It has three different blocks, weight update block, refresh block and dendrite block. Figure is taken from [45].

transitions are triggered by elevated presynaptic rates; for low presynaptic rates there are essentially no transitions. The fact that the synapse is binary on long time scales, solves the stability problem of synaptic efficacy in absence of stimulation. Fig.2.4 shows the implemented synapse circuit [45]. Working of the circuit is explained along with the synaptic dynamics.

Synaptic dynamics: The synaptic dynamics is described in terms of the analog internal variable, $X(t)$, which in turn determines the synaptic efficacy, $J(t)$, defined as the change in postsynaptic depolarization or afferent current, induced by a single

Figure 2.5: **Bi-stable Synaptic Dynamics:** The figure clearly mentions all the variables, θ_v , a , b , α , β , θ_X used in the mathematical expression. The presynaptic spike denotes the spike arrival at t_k^{pre} . Figure is taken from [45].

presynaptic spike. Synaptic internal variable $X(t)$ can be expressed as

$$\begin{aligned} \frac{dX(t)}{dt} &= H(t) + R(t) \\ H(t) &= \sum_k F(V(t_k^{pre}))\delta(t - t_k^{pre}) \\ F(V(t)) &= a\Theta(V(t) - \theta_v) - b\Theta(\theta_v - V(t)) \\ R(t) &= -\alpha\Theta(-X(t) + \theta_X) + \beta\Theta(X(t) - \theta_X) \end{aligned} \quad (\text{Eq. 2.10})$$

The details of each variable given in above equations are explained along with the explanation of corresponding synaptic blocks.

$H(t)$ is the stimulus driven hebbian learning term. The Hebbian term depends on variables related to the activities of the two neurons connected by the synapse:

1. The spikes from the presynaptic neuron and
2. The instantaneous depolarization of the postsynaptic neuron.

Presynaptic spikes trigger the Hebbian dynamics: each spike induces a jump in X whose value depends on the instantaneous depolarization of the postsynaptic neuron. The sum is over all presynaptic spikes: spike number k arrives at the synapse at t_k^{pre} . The value of V_{post} upon arrival of the presynaptic spike determines the jump, $F(V_{post}(t_k^{pre}))$, in the synaptic variable. The sign of F determines whether the synapse is potentiated or depressed at the event. If presynaptic spike arrives when $V_{post} > \theta_V$, then the synaptic internal state is pushed up by 'a' toward the potentiated state, $X \rightarrow X + a$. If it arrives while $V_{post} < \theta_V$ then $X \rightarrow X - b$ where $b > 0$. The circuit implemented for the hebbian learning does the exact functionality as described by the equations above. Whenever a presynaptic spike arrive, both MOSFETs, M1 and M6 are 'ON' but waits for the 'Poststate' signal to charge or discharge the capacitor, C ; where either charging or discharging happens at an instant. Poststate signal depends on the function, F as given above. $V(\text{BiasP})$ and $V(\text{BiasN})$ controls the charging and discharging of the capacitor respectively which implies that variables 'a' and 'b' depends on $V(\text{BiasP})$ and $V(\text{BiasN})$ respectively. Fig.2.4 shows the synapse circuit.

$R(t)$ is the refresh term responsible for memory preservation. It is implemented using a latch with a comparator and a current source inverter or can be called as a low slew-rate amplifier in positive feedback configuration. Between spikes the dynamics is dominated by the term $R(t)$, which is characterized by drift down of rate α , or drift up with rate β , depending on whether X is below or above a synaptic threshold θ_X . $\Theta()$ is the heavyside function. In the refresh block, the internal variable, $X(t)$ is compared with θ_X to enable or disable a pMOS. $X(t) > \theta_X$, output of comparator is '0' and pMOS is enabled. $X(t) < \theta_X$, output of comparator is '1' and pMOS is disabled. Enabling pMOS results in the flow of current Refrp , controlled by $V(\text{Refrp})$, to charge the capacitor, C , this is denoted by the drift up rate β in the equation of $R(t)$. It implies that drift up rate β depends on current, Refrp . Similarly disabling the PMOS results in the flow of current Refrn , controlled by $V(\text{Refrn})$, discharges the capacitor, C , this is denoted by the drift down rate α in the equation of $R(t)$. It implies that drift down rate

α depends on current, Refrn. Fig.2.5 shows the explanation given above.

The synaptic efficacy, $J(t)$ has two values: it is potentiated ($J(t) = J_1$) when the internal variable $X(t)$ is above the threshold θ_X and depressed ($J(t) = J_0$) if $X(t) < \theta_X$. X is restricted to the interval $[0,1]$. This dynamics is explained in dendritic block and is enabled only during the arrival of a presynaptic spike. Disadvantage of this synapse is that it does not integrate input spikes into continuous output currents.

2.2.2 Bi-Stable Synapse by Indiveri

Figure 2.6: **Bi-stable Synapse Circuit by Indiveri:** The circuit implements the plasticity model of STDP. It has four different blocks, bistability, STDP block, STD block and CMI block. Figure is taken from [46].

One of the key properties of biological synapse is their ability to exhibit short term and long term plasticity. The excitatory synaptic circuit shown in Fig.2.6 implements both types of plasticity mechanisms. In this circuit, long term plasticity is implemented using the "STDP" and "bistability" blocks, while short term plasticity is implemented by the short term depression, "STD" block. The integrating property of the synapse is achieved by using a current mirror integrator (CMI) circuit [46]. In this circuit STDP rule is implemented rather than a SDSP rule implemented in the previous subsection. The circuit explanation is done along with the synaptic dynamics.

Synaptic dynamics: In order to cope with the problem of long term storage of analog values in CMOS technology, this bistable block also drives the synaptic weight to one of two possible states on long time scales. The functionality of both the circuits are exactly similar. The only difference is the dependency of bias current on slopes. In this circuit both drift up and down slopes are controlled by same bias current, $I(V_{leak})$ while in previous circuit it was with two different bias currents. The bistability circuit drives V_{w0} in two ways, depending on the difference between the value of V_{w0} and the asymptote: if $|V_{w0} - V_a| > 4U_T$, the bistability circuit drives toward V_a linearly, where V_a represents either V_{high} or V_{low} , depending on the sign of $V_{w0} - V_{thr}$. As V_{w0} gets close to the asymptote and $|V_{w0} - V_a| < 4U_T$, V_{w0} begins to approach the asymptote exponentially. These dynamics of the bistability block is given below as

$$\begin{aligned}
 &|V_{w0} - V_a| > 4U_T \\
 &\begin{cases} V_{w0}(t) = V_{w0}(0) + \frac{I_{leak}}{C_w}t, & \text{if } V_{w0} > V_{thr} \\ V_{w0}(t) = V_{w0}(0) - \frac{I_{leak}}{C_w}t, & \text{if } V_{w0} < V_{thr} \end{cases} \\
 &|V_{w0} - V_a| < 4U_T \\
 &\begin{cases} V_{w0}(t) = V_{high} - V_{w0}(0) \exp\left(-\frac{I_{leak}}{C_w U_T}t\right), & \text{if } V_{w0} > V_{thr} \\ V_{w0}(t) = V_{low} + V_{w0}(0) \exp\left(-\frac{I_{leak}}{C_w U_T}t\right), & \text{if } V_{w0} < V_{thr} \end{cases}
 \end{aligned} \tag{Eq. 2.11}$$

where C_w is the capacitor used in the circuit. The signal, V_{thr} is a threshold voltage that can be set externally. Operating in subthreshold, I_{leak} is given as

$$I_{leak} = I_0 \exp[(\kappa V_{leak} - V_{low})/U_T] \quad (\text{Eq. 2.12})$$

The STDP circuit increase or decrease the analog voltage V_{w0} , depending on the relative timing of the pulses ‘pre’ and ‘post’. The STDP circuit implemented is fully symmetric: upon the arrival of a presynaptic pulse, pre, a waveform (for potentiation) $V_{pot}(t)$ is generated. Similarly, upon the arrival of a postsynaptic pulse, a complementary waveform (for depotentiation), $V_{dep}(t)$ is generated. Both waveforms have a sharp onset and decay linearly with time, at a rate set by V_{tp} and V_{td} respectively. The pre- and postsynaptic pulses also switch on two gates that allow the currents I_{pot} and I_{dep} to flow, resulting in a weight increase or decrease. The bias voltages V_p and V_d set an upper bound for the amount of current that can be injected into or removed from the capacitor C_w . As those 6 middle transistors operate in subthreshold regime [47] the expression for $I_{pot}(t)$ and $I_{dep}(t)$ are given as

$$\begin{aligned} I_{pot}(t) &= \frac{I_0}{e^{-(\kappa/U_T)V_{pot}(t-t_{pre})} + e^{-(\kappa/U_T)V_p}} \\ I_{dep}(t) &= \frac{I_0}{e^{-(\kappa/U_T)V_{dep}(t-t_{post})} + e^{-(\kappa/U_T)V_d}} \end{aligned} \quad (\text{Eq. 2.13})$$

where t_{pre} and t_{post} are the times at which the pre- and postsynaptic spikes are emitted. The change in synaptic efficacy is ΔV_{w0} then

$$\begin{cases} \Delta V_{w0}^+ = \frac{I_{pot}(t_{post})}{C_p} \Delta t_{spk}, & \text{if } t_{pre} < t_{post} \\ \Delta V_{w0}^- = -\frac{I_{dep}(t_{pre})}{C_d} \Delta t_{spk}, & \text{if } t_{pre} > t_{post} \end{cases} \quad (\text{Eq. 2.14})$$

where Δt_{spk} is the pre- and postsynaptic spike width, C_p and C_d are the parasitic capacitances of nodes V_{pot} and V_{dep} , respectively.

Figure 2.7: **Simulation Results:** The STDP curve variation with different circuit parameters are shown. This shows the advantage of circuit in terms of adaptability of STDP curve. The curves in the left plot were obtained for six different values of V_{tp} and V_{td} , while the curves in the right plot were obtained for six different values of V_p and V_d . Results are taken from [46].

Experimental data showing how ΔV_{w0} changes as a function of $\Delta t = t_{pre} - t_{post}$, for different values of V_{td} and V_{tp} are plotted in Fig.2.7(a). Similarly, Fig.2.7(b) shows plots of ΔV_{w0} versus Δt for three different values of V_p and three different values of V_d . As there are four independent control biases, it is possible to set the maximum amplitude and temporal window of influence independently for positive and negative changes in V_{w0} .

Over long time scales, the dynamics of V_{w0} are governed by the bistability circuit, while on short time scales they are governed by the STDP circuits and the precise timing of presynaptic and postsynaptic spikes.

As biological experiments suggest the short term depression (STD) circuit is used to implement local gain control for stimulus specific adaptation and for nonlinear temporal summation. The circuit shown in Fig.2.6 acts on the synaptic weight voltage, V_w . At steady state, V_w settles to the voltage V_{w0} , set by the STDP circuit. During the arrival of presynaptic spike on the 'pre' node, the synaptic weight is gradually

Figure 2.8: **STD Results:** Transient responses to a (a) 10 Hz regular spike train and a (b) 10Hz Poisson distributed train. The input spikes are shown at the bottom of each plot. In (a), the amplitude of the EPSP decreases with each incoming input spike, clearly showing the effect of synaptic depression. In (b), the EPSP amplitude depends on the occurrence of the previous spike. Results are taken from [46].

reduced at a rate set by the bias voltage, V_{std} . As a consequence the synapse has a high-pass response: its weight is maximum at the onset of the stimulation, and gradually decreases throughout the duration of the stimulation. More details of the short term synaptic depression is given in [48]. The results which depict the short term depression is shown in Fig.2.8 which is taken from [48].

The current mirror integrator (CMI) circuit operates in the following way: during an input spike arrival on the node 'pre', an amount of charge set by the synaptic weight voltage, V_w is discharged from the integrating capacitor and a mirrored current, I_{inj} is increased accordingly. In the absence of spikes, the integrating capacitor charges through the diode connected transistor and the synaptic current slowly goes to zero. A thorough analysis of this compact and elegant non linear integrator circuit is given in [49].

2.2.3 Bi-Stable Synapse by Srinjoy

The synapse circuit proposed here is a combination of both the circuits proposed above. One of the most interesting feature of this circuit is it's ability to implement spike driven learning algorithms, which has been successfully used to classify complex patterns in a semi supervised fashion. During training, patterns of activity are sequentially imposed on the input neurons, and an additional instructor signal drives the output neurons toward the desired activity. The synapses are bistable, and they are modified by the arrival of presynaptic spikes. The sign of the change is determined by both the depolarization and the state of a variable that integrates the postsynaptic action potentials. Following the training phase, the instructor signal is removed, and the output neurons are driven purely by the activity of the input neurons weighted by the plastic synapses. In the absence of stimulation, the synapses preserve their internal state indefinitely [30]. Fig.2.9 shows the implementation of synapse circuit. In the circuit the main block is the "stop learning module" which decides whether potentiation or depression to happen along with control in the learning.

NMDA Comparator block implements short term synaptic depression explained in the previous section. In some sense, even not exactly, short term plasticity is due to the hyperpolarization. The membrane voltage V_{mem} is compared for hyperpolarization against V_{nmda} . If hyperpolarized, then maximum current flows through the MOSFET with gate voltage, V_{nmda} of the differential pair compared to the other MOSFET. This create a depression in weight, V_{w0} during the period of occurrence of presynaptic pulse, "pre". During this time I_{syn} will be decaying exponentially to zero. When $V_{mem} > V_{nmda}$, then maximum current flows through M12 compared to M7 reducing voltage, V_{syn} . This will suddenly increase I_{syn} , whose amplitude depends on voltages V_{w0} and V_{mem} . This will only happen during the pulse width of "pre", after that I_{syn} decay exponentially to zero, till the occurrence of another "pre" pulse.

Figure 2.9: **SDSP Bi-stable Synapse Circuit:** (a) The synapse circuit which has four different blocks: bi-stability circuit, weight update block, NMDA comparator and CMI (b) The stop learning module is shown which produces two signals, V_{pot} and V_{dep} to control potentiation and depression respectively. Figures are taken from [50].

2.2.3.1 Stop learning module

The stop learning module is composed of a standard comparator, a linear integrator, and two low-power dual comparators. Low power dual comparator is nothing but a current comparator or a winner take all circuit [50].

The amplifier compares the membrane potential V_{mem} to the threshold V_{mth} and produces a digital signal V_{cmp} used to enable or disable the comparators, P and D blocks. The linear integrator, integrates the incoming postsynaptic spikes to generate the voltage, V_{Ca} , functionally equivalent to the calcium concentration in real neuron. The integrated signal, V_{Ca} is harnessed to stop learning when the neuron responds correctly i.e. either neuron fires at higher rate or at lower rate. When the teacher signal and the current generated by the synaptic input disagree, the postsynaptic neuron fire at intermediate frequencies, far from the minimum and the maximum. In this condition V_{Ca} fluctuates in an intermediate range indicating that learning must occur.

Figure 2.10: **Stop Learning Circuit Simulation Result:** The dynamics of the stop learning block is clearly shown. V_{pot} and V_{dep} are the output voltage signals obtained for different conditions of dual comparators. The signals are given to the weight update block for potentiation and depression to occur. This result is taken from [50].

Specifically, the dual threshold comparators P and D compare the signal, V_{Ca} to two threshold signals for defining when the synapse should be allowed to increase its weight or when it should be allowed to decrease it. The dynamics of the stop learning block is shown in Fig.2.9(b). The simulation result of the stop learning block is shown in Fig.2.10 which is taken from [50].

2.2.4 Binary Synapse by Arthur

The binary synapse implemented here is to learn stimulated patterns of activity and to compensate for variability in excitability. The decay and integrator blocks [51] as shown in Fig.2.11 are used to implement potentiation and depression. They are used

Figure 2.11: **STDP Circuit by Arthur:** (a) The STDP circuit consists of three sub blocks: Decay, Integrator and SRAM (b) the decay circuit and (c) integrator circuit. Figures are taken from [51].

in a symmetric fashion in the STDP circuit. The SRAM, a cross-coupled inverter pair, holds the current binary state of the synapse. For potentiation, the decay remembers the last presynaptic spike. The capacitor is charged when the presynaptic spike occur and discharges linearly thereafter. A postsynaptic spike samples the remaining charge on the capacitor, passes it through an exponential function, and dumps the resultant charge into the integrator. Again this charge decays linearly thereafter. At the time of the postsynaptic spike, SRAM reads the voltage on the integrator's capacitor. If it exceeds a threshold, SRAM switches state from depressed to potentiated (LTD goes high and LTP goes low). The depression side of the STDP circuit is exactly symmetric, except that it responds to postsynaptic activation followed by presynaptic activation and switches the SRAM's state from potentiated to depressed. When SRAM is in the potentiated state, presynaptic spike activates the principal neuron's synapse, otherwise the spike has no effect.

2.2.5 Synapse by Schemmel

Fig.2.12 shows the STDP circuit [52] with only the causal part shown, as circuit is symmetrical with respect to the presynaptic and postsynaptic inputs. The working of

Figure 2.12: **STDP Circuit by Schemmel:** The circuit implements the plasticity model of STDP. Figure is taken from [52].

the block is as follows: Consider the postsynaptic pulse arrival, then the MOSFETs M_1 and M_3 present in the latch will be switched on. M_6 isolates the three series connected transistors M_8 to M_{10} from the supply. M_7 connects the capacitor, C_1 to the supply through the current source, M_{11} . The gate voltage of M_{13} is kept high by M_{11} . If both the row and column inputs are held inactive, then C_2 holds its charge. The voltage on C_2 represents the stored weight for STDP causal event.

When a presynaptic pulse arrives, the transistors M_5 and M_{12} will be switched on for the period of presynaptic pulse width. Switching on M_{12} charges C_1 to V_1 initially. The MOSFETs, M_4 and M_2 of the latch will be switched on as M_5 turned on. This will switch on M_6 connecting the capacitor, C_1 to the supply through MOSFETs M_8 to M_{10} . Since the MOSFETs are operated in subthreshold region, the capacitor, C_1 charges exponentially from the initial voltage, V_1 . At this time M_7 is in off state. Any presynaptic

spike arrival will restart the whole process once again. This explains the measurement phase of the causal side. Suppose a postsynaptic pulse happened after Δt time, then the exponential charging of the capacitor, C_1 ceases as M_6 switches off and linear charging of C_1 towards V_{dd} starts through current source, M_{11} , as M_7 turned on. The voltage at the gate of M_{13} for a time span roughly proportional to the remaining charge on C_1 will activate M_{13} to remove a small amount of charge from C_2 . This is the charge accumulation phase of causal side. The present value of C_2 can be read-out with the help of source follower, M_{16} and enable switch, M_{17} . This stored weight is compared against a programmable threshold and changes the digital weight storage accordingly. After this the row and column reset are enabled to reset the value stored in C_2 .

2.2.6 Synapse by Bofill

The feedforward network of neurons with STDP synapse implemented in hardware is used to investigate the detection and amplification of spike timing synchrony. The implemented STDP synapse circuit is shown in Fig.2.13. The signals preLong, post_bp, postLong and pre as shown in Fig.2.13 are generated using pulse waveform generation circuit [53] (which is not shown here). Weight change circuit for causal correlations between pre and postsynaptic spikes is shown in Fig.2.13(a). When a presynaptic neuron fires, a “preLong” pulse switches on MOSFET, N8. Note that the duration of the pulses “preLong” and “postLong” is much more compared to the time duration of “pre” or “post_bp”. Once N8 is on, the capacitor, C_{pot} is charged to the gate voltage created by I_{bpot} on N5. When N8 switches off, the capacitor, C_{pot} discharges through N1 and the discharge is controlled by the voltage, V_{bpot} . Thus the voltage across C_{pot} gives a measure of the time elapsed since the last presynaptic spike. Now when the pulse, post_bp arrives, a quantity of charge is discharged from C_w through N2 and N3. This discharge depends on the gate voltage of N2 and pulse duration of post_bp. This results in decrease in voltage, V_w and thus potentiation happens. P5, P6, N7

Figure 2.13: **STDP Circuit by Bofill**: (a) The STDP circuit as shown in [53]. It is shown as three blocks: one for potentiation, weight change circuit (a), other for depression, acausal correlation circuit (b) and another for EPSC generation, synapse (c). Figures are taken from [53].

and N6 introduce a weight dependency on the weight potentiation. P5 is a narrow transistor whereas P6 is a wide transistor operating in weak inversion. When V_w is close to V_{dd} , a negligible current flows through P5 and P6. If V_w is not close to V_{dd} , the current through P5 and P6 increases and P5 enters into linear region of operation. This linearizes the voltage to current curve of P5. Finally the weight dependent current flowing through P5, P6 and N7 is mirrored and subtracted from I_{bpot} . The operation of acausal correlation circuit, Fig.2.13(b) is also similar to the weight change circuit but without the weight dependent structure. This circuit is connected to weight change

Figure 2.14: **T-STDP Circuit by Mostafa**: The circuit implements the plasticity model of T-STDP. Figure is taken from [54]. The red dashed boxes shows the potentiation where as the blue dashed boxes represent depression.

circuit through a current mirror controlled by “pre” signal. Synapse circuit shown in Fig.2.13(c) as EPSC generation circuit implements a voltage to current converter. The amplitude of the current pulses are proportional to the weight value. P4 is a narrow transistor maintained in linear region for a wide range of V_w values by a wide transistor, P2.

2.2.7 T-STDP Synapse by Mostafa

The T-STDP circuit implemented in Fig.2.14, builds upon an earlier study mentioned in Synapse by Bofill. Eight parameters in the full T-STDP model, as in equation Eq. 2.7, can be tuned in the proposed circuit, by controlling eight bias currents as follows: I_{dep1} , I_{pot1} , I_{dep2} and I_{pot2} represent the amplitude of synaptic weight changes for postpre (A_2)

and prepost (A_2^+) spike pairs, and prepostpre (A_3) and postprepost (A_3^+) combinations of spike triplets, respectively and the required time constants in the model for postpre (τ_-), prepost (τ_+), prepostpre (τ_x) and postprepost (τ_y) spike patterns can be adjusted using I_{td1} , I_{tp1} , I_{td2} and I_{tp2} respectively.

The proposed circuit works as follows, as given in [54]: M9 and M15 are switched on upon the arrival of a pre-synaptic pulse, $V_{pre}(n)$. The first potentiation capacitor, C_{pot1} is charged through M9 to the voltage of V_{pot1} by I_{pot1} . Once $V_{pre}(n)$ is finished, V_{pot1} starts to decay linearly through M11 and with a rate proportional to I_{tp1} . Now, when a post-synaptic pulse, $V_{post}(n)$ arrives at M13 in the decaying period of V_{pot1} , when M12 is still active, the weight capacitor, C_W , will be discharged through M12-M13 transistors and a potentiation occurs (prepost combination of spikes). Additionally, if a postsynaptic spike has arrived at M19, soon before the current presynaptic spike at M15, the weight capacitor can be charged through M14-M15 transistors and a depression happens(postpre combination of spikes). The amount of depression depends on V_{dep1} , which itself can be tuned by the relevant amplitude parameter I_{dep1} . Also, the activation interval of M18 can be modified by changing the related time constant parameter I_{td1} . Furthermore, another contribution to depression can occur if a previous presynaptic pulse, $V_{pre}(n-1)$, has arrived at M26 soon enough before the current presynaptic happens at M15 and also before a postsynaptic pulse happens at M19. In this situation, the weight capacitor can be charged again through M14-M15 by an amount proportional to an effect of both V_{dep2} and V_{dep1} , simultaneously. This triplet interaction leads to the required non-linearity mentioned in the triplet learning rule. A similar description holds for the situation when a postsynaptic pulse occurs at M13 and M19 transistors. But this time one depression will take place in the result of charging the weight capacitor up through M14-M15 and because of arriving a postsynaptic spike at M19 before a presynaptic spike at M15. Besides, two potentiation events can happen if an appropriate situation is provided to discharge the weight capacitor because of a pre-post or a post-pre-post combination of spikes. Note that,

in this implementation, the synaptic strength is inversely proportional to the voltage stored on the weight capacitor, C_W .

2.2.8 Single Transistor Learning Synapse

The single transistor learning synapse (STLS) utilizes a floating gate (FG) device to permanently store synaptic weight. Fig.2.15(a) shows the floating gate synapse. The adaptation of the weight happens with the help of hot electron injection (HEI) and fowler nordheim (FN) tunneling phenomena. Necessary conditions for injection to happen are: 1, sufficient current should flow through the channel of floating gate MOSFET and 2, sufficiently large drain source voltage, V_{ds} . If above two conditions are met then electrons from channel jump to floating gate node. This reduces the floating gate voltage, V_{fg} and hence weight increases. Similarly for tunneling to happen the voltage across tunneling capacitor, C_{tun} must be high. i.e. V_{tun} must be much higher than V_{fg} . Then electrons from floating gate node will be transferred to C_{tun} . Along with this weight adaptation mechanism, another functionality of synapse is excitatory postsynaptic current (EPSC) generation. As the biological observations suggest, it is better to keep the EPSC as a pulse with fast rise time and slow exponentially decaying fall time. To achieve this an asymmetric triangular waveform, with fast falling slope and a slow rising slope, is given to gate voltage, V_g . V_{tun} is kept as a pulse waveform with sufficient amplitude for tunneling to happen, during that pulse period. Similarly V_d is also an inverted pulse waveform which satisfies the conditions of injection. These different terminal voltage waveforms are shown in Fig.2.15(c). As voltages are defined for necessary condition, next we will see the timing of voltage waveforms. V_g occurs when a presynaptic spike happens while V_{tun} and V_d occurs, When postsynaptic spike arrives.

For LTP to happen, V_{tun} must be in the initial value; Whereas LTD to hap-

pen, V_d must be in the initial value. For STDP: as V_g waveform is triangular, it gives a measure of time elapsed since the occurrence of presynaptic spike. Now when a postsynaptic spike happens then V_d occurs, which samples the gate voltage, V_g exactly at the time of occurrence of postsynaptic spike. Thus the current flowing through the floating gate will be proportional to the gate voltage, which in turn create potentiation. Thus potentiation depends on the time difference between post and presynaptic spikes, $\Delta t = t_{post} - t_{pre}$. During this potentiation, as V_{tun} occurs simultaneously with V_d , even depression occurs at same time. Usually tunneling will be less compared to injection as the voltage difference, $V_{tun} - V_{fg}$ is less. Still the parameters of the voltage waveforms can be modelled so that injection dominates over tunneling to get the proper STDP curve. Similarly when postsynaptic spike happens first and then presynaptic spike, V_{tun} and V_d occurs first, then V_g occurs creating high voltage difference across C_{tun} . Thus tunneling happens. Here the pulse width of drain voltage waveform is much smaller compared to tunneling waveform hence injection is neglected. Thus a brief explanation of STDP using STLS is given. More details of operation of STLS can be found in next chapter.

Fig.2.15(b) given is the STDP curve obtained from the STLS. This curve shows a depression in the positive side of $t_{post} - t_{pre}$. It is undesirable to have depression, as biological experiments suggest potentiation in the positive x-axis. A simple remedy for this problem is proposed in the next chapter.

The layout of the floating gate MOSFET is shown in Fig.2.16(a). The layout components are shown in different colours, with each component mentioned on the left hand side of the figure. The input terminal voltages are also shown as in Fig.2.15(a). Synaptic current increases as injection or potentiation happens, while the floating gate voltage reduces and synaptic current reduces with tunneling or depotentiation, while the floating gate voltage is increased. This is shown as the characteristics of floating gate MOSFET in Fig.2.16(b), as output current versus input voltage characteristics

Figure 2.15: **STLS, STDP Curve and Voltage waveform:** (a) Floating gate synapse with different terminals V_g , V_{tun} and V_d . (b) STDP curve obtained using a constant tunneling pulse with $V_{tun} = 15V$, $V_d = 4.2V$, $T_{tun} = 200ms$, $T_{inj} = 100\mu s$. (c) Different terminal voltage waveforms. Figure is taken from [55].

of floating gate MOSFET for different mechanisms of injection and tunneling. Note that when injection happens, the curve of transistor drain current (red curve) shifts to the right of the normal transistor curve (dotted curve) that is without any mechanism of injection or tunneling, whereas when tunneling happens, the curve of drain current (blue curve) shifts to the left of normal transistor curve. This also implies the shift in the threshold voltage of the floating gate transistor as a result of the quantum mechanisms.

(a)

(b)

Figure 2.16: **Layout and characteristics of Floating gate MOSFET:** (a) Floating gate synapse layout with different colours of layout components and the corresponding terminal voltages are shown. (b) The output current versus input voltage characteristics of floating gate MOSFET for different mechanisms of injection and tunneling.

2.2.9 Memristor Synapse

Leon Chua postulated the existence of fourth basic electrical component as memristor around four to five decades back. Recently, memristors became promising, novel device technology for realization of non-volatile memories with complex multilevel and plastic behaviors, which made them a good candidate for the implementation of artificial synapses in neuromorphic engineering. Memristive effects rely on diverse physical phenomena, and their behaviours differ strongly from one technology to another. However, [56] shows that memristors can be used for implementing several plasticity rules.

Similar to memristors there are other nano-technology devices like Resistive Random Access Memories (R-RAMs), Spin-Transfer Torque Magnetic Random Access Memories (STT-MRAMs) and Phase Change Memories (PCM) used for the implementation of artificial synapses. Integration of these nano-technology devices to CMOS is a challenge and usually reduce the advantage of having better synaptic density of neural on-chip architecture. It is observed that, for the implementation of artificial synapses using any of the technologies like floating gate MOSFET, memristors or very recent spin devices, understanding the device physics is the key. Among these devices, floating gate devices are far ahead in CMOS implementation [57] and showed promising plasticity behaviours whereas, other devices lack CMOS integration and are still under basic research prior to commercialisation.

2.3 Comparison of Synapse Circuits

Table 2.2 shows the comparison of different synapse circuits with respect to different important metrics. The important metrics to be compared are power consumption,

silicon area and weight storage. Power consumption is not mentioned in most of the references. So we could not take it into much account for comparison. Silicon area comparison is further simplified with the help of synapse density [57]. Synapse density is defined as the ratio of synapse size to the square of the design technology. From the table it is clear that floating gate synapse has much lesser silicon area compared to other circuits. The metric, weight storage in the case of floating gate synapse is also non volatile compared to other synapses, but has a weight resolution more than 10 bits. Since these floating gates can be operated with few nano amperes of current the power consumption by these devices can be reduced. Thus these comparisons clearly gives a priority for floating gate synapse among other synapse circuits. The weight adaptation and the operation of a floating gate synapse is explained in next chapter.

Note that the estimated area and power consumption data in this table only reflect the reported data in the related papers. These numbers are dependent on many parameters, including the synaptic plasticity rule implemented, the synaptic plasticity parameters, the weight storage techniques, and the network stimulation pattern and protocols. Since, in some papers, the exact power consumption and area requirement of the synapse is not available, the total power and area of the chip are divided by the number of synapses and neurons on the chip, to calculate a rough value of the size and power requirement of the synapse [4].

Comparison of Hardware Synapse Circuits with different metrics

Reference	Process technology (nm)	Synapse area (μm^2)	Normalized Area ¹	Weight storage (precision)	Plasticity rule	Chip measurement
This paper	350	6337	51730	Floating gate (> 10 bits)	T-STDP	Yes
[54]	350	-	-	Capacitor (Analog)	T-STDP	No
[58]	350	133	1088	Floating gate (> 10 bits)	STDP	Yes
[59]	350	400	3265	Capacitor (Analog)	STDP	Yes
[55]	350	100	816	Floating gate (> 10 bits)	STDP	Yes
[60]	250	5000	80000	Capacitor (Analog)	STDP	Yes
[61]	600	72000	200000	Capacitor (Analog)	STDP	Yes
[46]	800	4495	7023	Capacitor (bistable)	STDP	Yes
[62]	250	238	3810	SRAM (1 bit)	STDP	Yes
[52]	180	108	3338	SRAM (4 bits)	STDP	Yes
[50]	350	3000	24490	Capacitor (bistable)	SDSP	Yes

Table 2.2: **Hardware Synapse Circuit Comparison:** Different hardware synapse circuits are compared with different metrics.

¹Normalized area is termed as the ratio of the synapse area to the square of the minimum transistor dimension in that process technology.

Chapter 3

Robust Doublet STDP using floating-gate synapse

The experimental STDP plot (Fig.2.15) of a floating gate (FG) synapse (change in weight against $\Delta t = t_{post} - t_{pre}$) from previous study [55] shows a depression instead of potentiation at some range of positive values of Δt for a wide set of parameters. To ameliorate this STDP graph of a FG synapse, a minimum hardware overhead solution is proposed as shown in Fig.3.1 – this is the major contribution described in this chapter. The solution contains a modified gate voltage waveform in red dashed line as shown in Fig.3.1 while the gate voltage waveform proposed in previous work [55] is shown in black line in the same figure. This simple solution based on changing control gate waveforms of the FG device makes the weight change conform closely with biological observations over a wide range of parameters. We show results from a theoretical model to illustrate the effects of the modified waveform. The experimental results from a FG synapse fabricated in AMS 0.35 μ m CMOS process design are also presented to justify the claim. The contents of this chapter is based on our work given in [63, 64].

Figure 3.1: **Floating gate Mosfet:** The different terminal voltages V_{tun} , V_g , V_d as shown above are used to obtain the desired plasticity rule. The gate voltage waveform, V_g shown in black is the original voltage waveform which is as in [55] and the gate voltage waveform, V_g in red dashed line is the novel modified technique.

3.1 Floating Gate Synapse

3.1.1 Basic Operation

The equation for drain current of a subthreshold saturated pFET whose well is tied to V_{dd} is given by [55]

$$I_d = I_{s0} e^{\kappa(V_{dd} - V_{fg})/U_T} \quad (\text{Eq. 3.1})$$

Figure 3.2: **Parameter specifications of terminal voltage waveforms:** The different parameters are given as $T_g = 34ms$, $T_{tun} = 300ms$, $T_d = 300us$, $V_{tun_max} = 12V$, $V_{tun_min} = 5.6V$, $V_{g_max} = 3.3V$, $V_{g_min} = 2.5V$, $V_{d_max} = 5V$ and $V_{d_min} = 0V$.

where U_T is the thermal voltage and κ is the gate coupling coefficient. Due to the exponential relationship between the gate voltage and drain current of the MOS transistor, the gating voltage to the synapse has to be a triangular waveform which decreases from its maximum value [55]. The current at the maximum gate voltage is nearly zero. In order to generate EPSCs similar to biological observations, we apply triangular waveform with unequal slopes at the input [55]. The fast decreasing part of the input triangular waveform with slope S_1 results in a quick rise of the synaptic current while the slowly increasing part with slope S_2 determines the exponential decay in the output current. In this paper we modify the triangular waveform present at the input without harming the generation of EPSCs similar to biological observation.

The synaptic weight adaptation for obtaining different plasticity rule depends on the signals on the FG mosfet. The terminal voltages of the floating gate transistor shown in the Fig.3.1 are the gate voltage V_g , drain voltage V_d and tunneling voltage

Figure 3.3: **Timing diagram for STDP in FG synapse.** (a) The plot shows the time instances of pre- and post- synaptic spike occurrence for the case of $\Delta t = (t_{post} - t_{pre}) > 0$. Note the timescale which helps while integration during +ve Δt of the STDP curve. (b) The plot shows the time instances of pre- and post- synaptic spike occurrence for the case of $\Delta t = (t_{post} - t_{pre}) < 0$. Note the timescale which helps while integration during -ve Δt of the STDP curve.

V_{tun} . The specifications of terminal voltage waveforms are given in Fig.3.2.

Synaptic weight modification in a FG mosfet uses a combination of quantum mechanisms such as hot-electron injection (HEI) and Fowler-Nordheim tunneling [55]. HEI adds electrons on to the floating gate node, which reduces the voltage on the floating gate thus resulting in more current through the transistor hence increasing the weight of the synapse. On the other hand tunneling takes away electrons from the FG node, which increases the voltage of floating gate node thus reducing the drain current of the transistor and the synaptic weight. At every pre-synaptic spike, a triangular gate voltage waveform is generated while at every post-synaptic spike, a triangular tunneling voltage and a drain voltage pulse is generated as illustrated in the Fig.3.3(a)

and (b). Hence, injection and tunneling currents (and the resultant weight change) also depend on the time difference between pre- and post-synaptic spike times. The governing equations for injection and tunneling are given as [65, 55]

$$\begin{aligned} I_{inj} &= I_{inj0} (I_d / I_{s0})^\alpha e^{-\Delta V_{ds} / V_{inj}} \\ I_{tun} &= I_{tun0} e^{(V_{tun} - V_{fg}) / V_{ox}} \end{aligned} \quad (\text{Eq. 3.2})$$

where I_d is the drain current, $\alpha = 1 - U_T / V_{inj}$, V_{ox} and V_{inj} are process dependent parameters.

3.1.2 Theoretical Model of Learning

If we compare the result obtained from the experimental measurement using original gate voltage waveform in Fig.7(a) of [55] with the theoretical graph of STDP in Fig.2.1, we notice that there is depression happening instead of potentiation on the positive x-axis of the experimental curve of STDP. To understand this effect, we first develop a mathematical model of the change in weight for a FG synapse. The weight of the synaptic device is defined as [55]:

$$w = e^{\frac{-\kappa \overline{V_{fg}}}{U_T}} \quad (\text{Eq. 3.3})$$

The floating gate voltage, V_{fg} is defined as, $V_{fg} = V_{fg0} + \Delta V_{fg} + \overline{V_{fg}}$, where V_{fg0} is the constant bias part, ΔV_{fg} is the fast time scale change due to capacitive coupling and $\overline{V_{fg}}$ denotes the change on a slow time scale due to the quantum mechanics of floating gate MOSFET.

Hence, equations to predict the change in FG voltage effectively predict the

change in weight. i.e. Eq. 3.3 can be modified as,

$$\frac{w_2}{w_1} = e^{\frac{-\kappa(\overline{V_{fg2}} - \overline{V_{fg1}})}{U_T}} \quad (\text{Eq. 3.4})$$

subtracting 1 from both side of the equation and approximating exponential term, we get

$$\begin{aligned} \frac{w_2 - w_1}{w_1} &= \frac{-\kappa(\overline{V_{fg2}} - \overline{V_{fg1}})}{U_T} \\ \text{i.e. } \frac{\Delta w}{w_1} &= \frac{-\kappa \Delta \overline{V_{fg}}}{U_T} \end{aligned} \quad (\text{Eq. 3.5})$$

Similarly, considering Eq. 3.1, we can modify it as,

$$\begin{aligned} \frac{I_{d2}}{I_{d1}} &= e^{\frac{-\kappa(\overline{V_{fg2}} - \overline{V_{fg1}})}{U_T}} \\ \frac{\Delta I_d}{I_{d1}} &= \frac{-\kappa \Delta \overline{V_{fg}}}{U_T} \end{aligned} \quad (\text{Eq. 3.6})$$

This implies that the normalized change of current in the floating gate MOSFET is equal to the normalized weight change of the floating gate synapse.

The following assumptions are made in deriving the theoretical equations:

1) To ensure small change in weight at very large negative and positive values of Δt , V_{g_init} has to be high enough so that $V_{tun_max} - V_{g_init}$ is small enough for negligible tunneling. Similarly $\Delta V_g = V_{g_init} - V_{g_min}$, should be small enough so that $V_{tun_init} - V_{g_min}$ is small enough for negligible tunneling.

2) Strong coupling from gate to floating gate node where as a weak coupling from tunneling node to floating gate node. This is justified since typically $C_g \gg C_{tun}$.

3) The gate voltage waveform falls to its minimum value instantaneously. In other words, $S_1 \gg S_2, S_3$ as shown in Fig.3.2. Also, we use T_g to denote the temporal duration of gate voltage with the understanding that $T_g = T_{g_ori}$ for the original

waveform proposed in [55] while $T_g = T_{g_mod}$ for our proposed modification.

We can now derive the slow time scale equation [55] for change in FG voltage due to tunneling and injection as:

$$C_T \frac{d\overline{V_{fg}}}{dt} = I_{tun} - I_{inj} = C_T \frac{d\overline{V_{fg2}}}{dt} - C_T \frac{d\overline{V_{fg1}}}{dt} \quad (\text{Eq. 3.7})$$

where C_T is the total capacitance on the FG node and $\overline{V_{fg}}$ denotes change on a slow time scale. Where, $\overline{V_{fg1}}$ is the slow time scale change in V_{fg} due to injection only and $\overline{V_{fg2}}$ is the slow time scale change in V_{fg} due to tunneling only.

3.1.2.1 Case 1: $\Delta t > 0$

First, we consider the case of $\Delta t > 0$ i.e the positive axis of STDP curve and combine equations (Eq. 3.2) and (Eq. 3.1) to get:

$$C_T \frac{d\overline{V_{fg1}}}{dt} = -I_{inj0} (e^{\kappa(V_{dd}-V_{fg})/U_T})^\alpha e^{-\Delta V_{ds}/V_{inj}} \quad (\text{Eq. 3.8})$$

where $\overline{V_{fg1}}$ is the slow time scale change in V_{fg} due to injection only. Since change in V_{fg} on the RHS happens due to coupling from the gate voltage, we can write:

$$V_{fg} = V_{fg_min} + \frac{C_g}{C_T} S_2 t \quad (\text{Eq. 3.9})$$

where S_2 is positive slope of V_g and C_g is the capacitance connected between the gate terminal and the floating gate terminal. Here V_{fg_min} is not same as V_{g_min} but due to initial charge stored on the FG. Substituting Eq. 3.9 in Eq. 3.8, we get

$$C_T \frac{d\overline{V_{fg1}}}{dt} = -A e^{-Xt} \quad (\text{Eq. 3.10})$$

Figure 3.4: **Weight variation in FG synapse obtained from theory for the two different gate voltage waveforms.** (a) shows the effect of injection on ΔV_{fg} for the case $\Delta t = (t_{post} - t_{pre}) > 0$, (b) shows the effect of tunneling on ΔV_{fg} for the case $\Delta t = (t_{post} - t_{pre}) > 0$, (c) shows the effect of tunneling on ΔV_{fg} for the case $\Delta t = (t_{post} - t_{pre}) < 0$, (d) shows the effect of both injection and tunneling on $\Delta w/w$ which is equivalent to ΔV_{fg} . (e) shows the result as in (d) but for a different parameter of $V_{tun_max} = 11$ V while (f) shows the result for $V_{tun_max} = 11$ V but $\Delta V_d = 4.5$ V.

where

$$A = I_{inj0} e^{\alpha \kappa (V_{dd} - V_{fg_min}) / U_T} e^{-\Delta V_{ds} / V_{inj}}$$

$$X = \frac{\alpha \kappa C_g S_2}{C_T U_T}$$
(Eq. 3.11)

Referring to Fig.3.3(a), significant amount of injection happens in the time from Δt to $\Delta t + T_d$, where ΔV_{ds} is constant and significant. Also, since T_d is very small compared to T_g , we can assume that V_g is constant during the drain pulse. Hence, we finally get:

$$C_T \Delta \overline{V_{fg1}} = -AT_d e^{-X\Delta t} \quad (\text{Eq. 3.12})$$

The above equation for $\Delta \overline{V_{fg}}$ is a function of Δt and is shown in Fig.3.4(a) (plotted in red and marked as “Original V_g ”).

Now let us analyze the contribution of tunneling to $\Delta \overline{V_{fg}}$ denoted as $\Delta \overline{V_{fg2}}$. With reference to Fig.3.3(a) and from assumption 1, the effect of tunneling is significant only from Δt to T_g . Hence, we have:

$$C_T \int_0^{\Delta \overline{V_{fg2}}} d\overline{V_{fg}} = \int_{\Delta t}^{T_g} I_{tun} dt \quad (\text{Eq. 3.13})$$

Substituting I_{tun} from Eq. 3.2, we further get:

$$C_T \Delta \overline{V_{fg2}} = \int_{\Delta t}^{T_g} I_{tun0} e^{(V_{tun} - V_{fg})/V_{ox}} dt \quad (\text{Eq. 3.14})$$

where

$$V_{tun} = \begin{cases} V_{tun.init} & \text{if } 0 < t < \Delta t \\ V_{tun.max} + S_3(t - \Delta t) & \text{if } \Delta t < t < T_g \end{cases} \quad (\text{Eq. 3.15})$$

where S_3 is the negative slope of V_{tun} . Thus substituting Eq. 3.15 and Eq. 3.9 into Eq. 3.14, we finally get:

$$\begin{aligned} C_T \Delta \overline{V_{fg2}} &= B e^{\frac{-S_2 \Delta t}{V_{ox}}} \int_{\Delta t}^{T_g} e^{Yt} dt \\ &= B' (e^{YT_g} - e^{Y\Delta t}) e^{\frac{-S_2 \Delta t}{V_{ox}}} \end{aligned} \quad (\text{Eq. 3.16})$$

where

$$\begin{aligned}
 Y &= \frac{S_3 - \frac{C_g S_2}{C_T}}{V_{ox}} \\
 B &= I_{tun0} e^{\frac{V_{tun,max} - V_{fg,min}}{V_{ox}}} \\
 B' &= B/Y
 \end{aligned} \tag{Eq. 3.17}$$

The maximum value of potentiation, A^+ is obtained from the sum of $\Delta \overline{V_{fg1}}$ and $\Delta \overline{V_{fg2}}$ at $\Delta t = 0$ (ideal case). Hence we get,

$$\begin{aligned}
 A^+ &= \frac{-k \Delta \overline{V_{fg}}|_{\Delta t=0}}{U_T} \\
 &= \frac{-k(\Delta \overline{V_{fg1}} + \Delta \overline{V_{fg2}})|_{\Delta t=0}}{U_T} \\
 &= \frac{-k(-AT_d + B'(e^{YT_g} - 1))}{C_T U_T}
 \end{aligned} \tag{Eq. 3.18}$$

From this we can conclude that A^+ depends on free parameters ΔV_{ds} and $V_{tun,max}$. Similarly τ^+ can be calculated from the equation $\frac{-k \Delta \overline{V_{fg}}}{U_T} = 0.1\%$ of A^+ . Δt obtained after solving the equation gives the value of τ^+ . Intuitively we could see that τ^+ depends mainly on T_g . The above equation for $\Delta \overline{V_{fg}}$ is also a function of Δt and is shown in Fig.3.4(b).

3.1.2.2 Case 2: $\Delta t < 0$

Now we consider the case of $\Delta t = (t_{post} - t_{pre}) < 0$ i.e the negative axis of STDP curve. Similar to what we have done above, let us see the effect of tunneling and injection separately.

For the contribution of injection to $\overline{V_{fg}}$, we could see that injection happens only during the initial small period, T_d of time axis where the drain current of the MOS is almost zero. So we can completely neglect the effect of injection on $\overline{V_{fg}}$ in this case.

Now let us consider the contribution of tunneling to $\overline{V_{fg}}$. Similar to the analysis above, using assumption 1, we have:

$$C_T \int_0^{\Delta \overline{V_{fg2}}} dV_{fg} = \int_{-\Delta t}^{T_{tun}} I_{tun} dt \quad (\text{Eq. 3.19})$$

While performing the integration, there arises two cases for the upper limit of integration based on whether $-\Delta t + T_g$ is greater or lesser than T_{tun} .

Also, similar to the case of positive Δt , here we have:

$$\begin{aligned} V_{fg} &= V_{fg.min} + \frac{C_g}{C_T} S_2(t - (-\Delta t)) \\ V_{tun} &= V_{tun.max} + S_3 t \end{aligned} \quad (\text{Eq. 3.20})$$

Substituting equations (Eq. 3.20) and (Eq. 3.2) into equation (Eq. 3.19), we get

case 1: $-\Delta t + T_g < T_{tun}$

$$\begin{aligned} C_T \Delta \overline{V_{fg2}} &= B e^{\frac{-C_g S_2 \Delta t}{C_T V_{ox}}} \int_{-\Delta t}^{-\Delta t + T_g} e^{Yt} dt \\ &= B' (e^{Y(T_g - \Delta t)} - e^{-Y\Delta t}) e^{\frac{-C_g S_2 \Delta t}{C_T V_{ox}}} \end{aligned} \quad (\text{Eq. 3.21})$$

case 2: $-\Delta t + T_g > T_{tun}$

$$\begin{aligned} C_T \Delta \overline{V_{fg2}} &= B e^{\frac{-C_g S_2 \Delta t}{C_T V_{ox}}} \int_{-\Delta t}^{T_{tun}} e^{Yt} dt \\ &= B' (e^{YT_{tun}} - e^{-Y\Delta t}) e^{\frac{-C_g S_2 \Delta t}{C_T V_{ox}}} \end{aligned} \quad (\text{Eq. 3.22})$$

where Y, B and B' are as given above.

Here A^- is obtained from $\Delta \overline{V_{fg}}$ at $\Delta t = 0$ (ideal case). Hence we get,

case 1: $-\Delta t + T_g < T_{tun}$

$$\begin{aligned} A^- &= \frac{-k\Delta\overline{V_{fg}}|_{\Delta t=0}}{U_T} \\ &= \frac{-k(B'(e^{YT_g} - 1))}{C_T U_T} \end{aligned} \quad (\text{Eq. 3.23})$$

case 2: $-\Delta t + T_g > T_{tun}$

$$\begin{aligned} A^- &= \frac{-k\Delta\overline{V_{fg}}|_{\Delta t=0}}{U_T} \\ &= \frac{-k(B'(e^{YT_{tun}} - 1))}{C_T U_T} \end{aligned} \quad (\text{Eq. 3.24})$$

From this we can conclude that A^- depends on free parameters T_{tun} and V_{tun_max} . Similarly τ^- can be calculated from the equation $\frac{-k\Delta\overline{V_{fg}}}{U_T} = 0.1\%$ of A^- . Δt obtained after solving the equation gives the value of τ^- . Intuitively we could see that τ^- depends mainly on T_{tun} .

The above equation for $\Delta\overline{V_{fg}}$ is a function of Δt which can be plotted on the graph and is shown in Fig.3.4(c). Both original and modified cases of gate voltage waveform is plotted in the figure. The combined effect of tunneling and injection is shown in Fig.3.4(d). It is clearly seen that the theoretical model does predict a depression of weight for some range of positive values of Δt . Next, we shall analyze the reason behind this and suggest a method to rectify this.

3.1.3 Proposed STDP protocol for FG synapse

Since we want there to be only potentiation for positive values of Δt , we have to design the parameters of the FG synapse so that injection dominates over tunneling in this entire range. On the other hand, we want tunneling to dominate over injection for all

negative values of Δt . From Fig.3.4(d) we see that for a range of positive values of Δt , tunneling dominates over injection. The reason for this can be realized by studying the injection and tunneling characteristics separately in Fig.3.4(a) and (b) respectively. We can see that the effect of injection reduces to zero at $\Delta t \approx 10$ ms (Fig.3.4(a)) while the effect of tunneling persists till ≈ 30 ms (Fig.3.4(b)). This results in tunneling dominating for $\Delta t > \approx 10$ ms creating depression.

The reason for the sharp decline of the potentiation curve in Fig.3.4(a) is that injection depends directly on drain current which reduces exponentially fast with Δt (Eq. 3.12). Thus there is a large part of the gate waveform (the part for which $V_g > \approx V_{g_min} + 0.2$ V) which does not create significant post-synaptic current or potentiation. Intuitively, we can eliminate this part and get a modified gate control voltage waveform as shown in Fig.3.3. The effect of this modification on the weight change curve is shown in Fig.3.4(d) (blue curve marked “modified V_g ”). As predicted, the magnitude of depression for positive values of Δt is much reduced now. For certain sets of parameters (e.g. if V_{tun_max} is much lower or V_{g_min} is higher), the original gate waveform also does not produce significant depression for $\Delta t > 0$. As an example, Fig.3.4(e) plots the STDP curve for parameter, $V_{tun_max} = 11$ V showing no appreciable depression for positive Δt . However, the modified waveform provides a more bio-realistic STDP curve for a much wider set of parameters and hence is more robust. As an example, Fig.3.4(f) plots the STDP curve with $V_{tun_max} = 11$ V but with a reduced $\Delta V_d = 4.5$ V compared to $\Delta V_d = 5$ V in the earlier case. It can be seen that the original gate voltage waveform again creates a depression while the proposed one is free from this artifact. We shall present results to prove this fact in the next section.

Figure 3.5: **Block Diagram of experimental setup:** The different terminal voltages V_{tun} , V_g , V_d of floating gate MOSFET is generated by FPAA board, which is controlled by MATLAB code. The picoammeter is used for measuring the current through MOSFET which in turn gives the weight change measurement in the floating gate MOSFET.

3.2 Results

We present experimental measurements from a floating gate synapse designed in AMS $0.35\mu\text{m}$ CMOS process. The gate voltage waveform shown has two cases as discussed in the earlier section. The experimental and simulation results for both the cases of gate voltage waveform discussed in the earlier section are given. For higher accuracy, the simulation results are obtained from solving the original differential equations for ΔV_{fg} and not the simplified theoretical model presented earlier. The effect of process variations and different biasing conditions are also presented based on extensive simulations.

3.2.1 Experimental Setup

Block diagram of the experimental setup is shown in Fig 3.5. Floating gate MOSFET is interfaced with the Field Programmable Analog Array (FPAA) board [66]. DAC circuitry

in the FPAA board generates the necessary waveforms needed for the implementation of plasticity in floating gate MOSFET. MATLAB code is used to control the FPAA board for precise time sequence generation for experiments. As normalised current change in the floating gate MOSFET is equal to the normalised weight change (Eq. 3.5 and Eq. 3.6), the weight change of the synapse is identified with the help of change in current flowing through the floating gate MOSFET. A picoammeter is used for measuring lower range of current from floating gate MOSFET. It is connected in series with the drain of MOSFET.

3.2.2 Experimental Results: Robustness of “non-STDP” Behavior

In order to justify the robustness of proposed STDP protocol in suppressing the “non-STDP” phenomenon, two parameters, V_{tun_max} and T_{tun} have been varied and the STDP curve has been measured. All other parameters are fixed to the values mentioned in Fig.3.1. The experimental result for the parameters $V_{tun_max} = 12V$ and $T_{tun} = 300ms$ is taken first and are plotted in Fig.3.6(a). It can be seen that the modified V_g case is indeed able to suppress the “non-STDP” behavior compared to the original case. Next the parameter $V_{tun_max} = 12V$ is changed to $V_{tun_max} = 11.9V$ keeping other parameter, $T_{tun} = 300ms$ unchanged and the corresponding STDP curve is plotted in Fig.3.6(b). In this case, we can see that even the original gate voltage waveform, V_g does not exhibit “non-STDP” behavior. However, when we change $T_{tun} = 300ms$ to $T_{tun} = 400ms$ keeping $V_{tun_max} = 11.9V$, the “non-STDP” behavior appears again for the original V_g but not for the proposed V_g as shown in Fig.3.6(c). These results indicate that the “non-STDP” behavior does exist for a range of parameters in the original FG synapse and our proposed method is able to robustly reduce that effect.

Figure 3.6: **Experimental results for the STDP curve obtained from a FG device for (a) $V_{tun_max} = 12V$ and $T_{tun} = 300ms$ (b) $V_{tun_max} = 11.9V$ and $T_{tun} = 300ms$ (c) $V_{tun_max} = 11.9V$ and $T_{tun} = 400ms$.** Along with the experimental result, simulation results are also included for comparison. These curves are plotted for both original and modified cases of gate voltage waveform, V_g .

3.2.3 Simulation Results: Effect of Variations

Since the measurement of a STDP curve is very time consuming, we next performed simulations of the differential equations, Eq. 3.7 to explore the effect of parametric variations in more details. The effect of statistical variations between FG devices on a chip is first considered by varying V_{inj} and V_{ox} from their nominal values by a maximum of 5%. For each parameter set, we find B^+ which is defined as the least magnitude of $\Delta w/w$ for $\Delta t > 0$. For a bio-realistic STDP curve, $B^+ \approx 0$ while for “non-STDP” behavior, $B^+ < 0$. From the results shown in Fig.3.7(a), we can conclude that B^+ for the modified V_g case is always closer to the bio-realistic result. Furthermore, to show that these device parameter variations are significant, we also plot the same results against A^+ (maximum potentiation) and A^- (maximum depression) in Fig.3.7(b). It can be seen that the variations of A^+ and A^- are quite large proving that this robustness for modified V_g holds for significant variations. Figures 3.7(c) and (d) show similar plots of B^+ when bias parameters V_{tun_max} and T_{tun} are varied. Again, we see that B^+ for modified V_g is always closer to 0 than that for the original gate waveform. Further, we can see from Fig.3.7(d) that the original gate waveform will have $B^+ \approx 0$ for any bias condition for which $|A^+| \gg |A^-|$. This corresponds to the measurement results shown in Fig.3.6(b).

3.3 Gate Voltage Generator: VLSI Implementation

Fig.3.8(a) shows a possible circuit to implement the modified gate voltage waveform. It has very little overhead compared to the gate waveform generator circuit in [55]. The pre-synaptic pulse is input to the current starved inverter composed of M1, M3 and M4. When the input pulse arrives, the output of the inverter, V_{out} , falls quickly to V_{g_min} . Next, when the input becomes zero the output rises slowly to V_{dd} . The time

Figure 3.7: **Parametric variations of STDP curve from FG synapse.** Plot of simulation results obtained for B^+ due to variations in the device parameters V_{inj} and V_{ox} . The data is plotted as a function of (a) V_{inj} and V_{ox} and (b) A^+ and A^- . Plot of simulation results obtained for B^+ due to variations in the bias parameters V_{tun_max} and T_{tun} . The data is plotted as a function of (c) V_{tun_max} and T_{tun} and (d) A^+ and A^- .

Figure 3.8: (a) **Circuit to generate modified gate voltage waveform.** The threshold voltage for sudden rise of the waveform to V_{dd} is controlled using V_{thr} . (b) The SPICE simulation result of the circuit for different V_{thr} voltage exhibit different thresholds.

constant for charging is set by the bias voltage V_{biasp} and a FG transistor can be used here to correct for mismatch [58]. This part is very similar to the original circuit in [55]. To create the modified waveform, V_{out} has to be pulled up quickly to V_{dd} whenever V_{out} crosses a critical voltage higher than V_{g_min} . For example, in the earlier measurement results this critical voltage is approximately $V_{g_min} + 0.2V$. We add the transistors M5, M6 and M2 to implement this function. M5 and M6 form a common source amplifier that quickly pulls up V_{out} through the positive feedback action of M2 when V_{out} reaches critical value. This critical value can be set by the voltage V_{thr} or the bias voltage V_{bias} that sets the bias current I_{bias} in the common source amplifier. Simulation results for this circuit are shown in Fig.3.8(b) and operates as expected.

3.4 Discussion

The detailed description of D-STDP and a solution to the “non-STDP” behaviour is presented in this chapter. However, D-STDP could not reproduce results of many biological experiments; a triplet STDP (T-STDP) that considers triplets of spikes as the fundamental unit has been proposed recently to explain these observations. Next chapter describes the compact implementation of a synapse using single floating-gate (FG) transistor that can demonstrate T-STDP learning rule by modifying drain voltages according to triplets of spikes. We describe a mathematical procedure to obtain control voltages for the FG device for T-STDP and also show measurement results from a FG synapse fabricated in TSMC 0.35 μm CMOS process to support the theory.

Chapter 4

Triplet Spike Time Dependent Plasticity in a floating-gate synapse

Over the past ten years, numerous experimental studies [67, 42, 68, 26] have shown that the synaptic strength varies as a function of the precise spike timing difference $\Delta t = t_{post} - t_{pre}$ between the firing times t_{pre} and t_{post} of the presynaptic and postsynaptic neurons respectively. This synaptic plasticity rule, called Spike Time-Dependent Plasticity (STDP), has evolved as one of several unsupervised plasticity rules that play an important role in learning and memory in the brain. The mathematical model of STDP based on a pair of pre- and post-synaptic spike is referred as doublet STDP (D-STDP) while the one based on triplet of synaptic spikes [69, 70, 71, 72] i.e either pre-post-pre synaptic spike or post-pre-post synaptic spike is referred to as triplet STDP (T-STDP). Several variants of these rules have also been proposed for pattern classification tasks [73, 50].

Experimental results [29] indicate that D-STDP model based on pairs of spikes are not sufficient to explain synaptic changes due to triplets or quadruplets of spikes. D-STDP model also fails to reproduce frequency effects. However, T-STDP

model can reproduce frequency effects along with the explanation of synaptic changes due to triplets and quadruplets of spikes. It is also important to be able to replicate rate based plasticity experiments. It has been demonstrated that the Bienenstock-Cooper-Munro (BCM) learning rule [40] based on firing rates can be obtained from D-STDP when pre-synaptic and post-synaptic neurons fire uncorrelated or weakly correlated Poisson spike trains, and only nearest-neighbor spike interactions are taken into account [74]. However, it is not possible to make a strict theoretical mapping from the nearest-spike interactions D-STDP models to the BCM rule [29] whereas T-STDP allows for such theoretical mapping. Apart from these experimentally observed benefits of T-STDP model, there are some computational advantages as well [75]. It has been shown that T-STDP can detect input correlations higher than the second order ones to which D-STDP is sensitive. Hence, it has been shown to drive direction and orientation selectivity [75]. Further, it can be shown to reproduce a more generalized version of BCM overcoming the limitations of the original one. Though there are several models of plasticity with varying degrees of bio-realism [76], T-STDP is a good compromise between simplicity and richness of function. Its simplicity also lends itself to easy analysis making it a good choice for a plasticity rule with functionality beyond D-STDP.

Recently STDP became so popular in computational neuroscience that neuromorphic engineers who try to emulate brain function using VLSI have also tried to emulate this behaviour in silicon. However, implementing a compact learning synapse continues to be one of the big challenges in the field [57]. Several recent papers have reported D-STDP implementations [62, 46, 77, 61, 78] and T-STDP implementation [54]; however, these synapses could either only store states in a transient fashion (using charge on a capacitor) or only hold two states in the long term. The size of these synapses are also large hindering scalability of these designs. A promising solution for non-volatile analog weight storage in very small area is provided by a floating-gate (FG) device [57, 58, 79, 63, 64, 80, 81, 82, 55]. This concept was utilized recently

to show weight storage and adaptation due to quantum phenomena based on input signal timing [55].

Compared with other work, here we demonstrate for the first time the implementation of the T-STDP rule in a FG synapse by appropriately modifying the drain voltage pulse based on spike triplets. From chip measurement results, we show that FG synapse in Fig.4.2(b) can reproduce (1) D-STDP learning window and (2) T-STDP results when appropriate control signals are applied on its terminals. Some initial results for triplet experiment based on this work were presented in [83]. In this chapter, we present results for quadruplet experiments and frequency effects as well as a new drain voltage generation scheme. We also present circuits for VLSI implementation of the drain voltage waveforms.

The chapter is organized as follows: Section 4.1 reviews a brief explanation of STDP model – T-STDP. Section 4.2 introduces the working of floating gate synapse, relation between FG synapse and D-STDP rule and then, finally arrive at the relation between FG synapse phenomenon and T-STDP rule. The experimental results are included in section 4.3. Section 4.4 presents a hardware implementation of the drain voltage waveform generator along with the simulation results. Finally the chapter is concluded with discussions on consolidated work. The contents of this chapter is based on our work given in [83, 84].

4.1 STDP Synaptic Modification Rule

In biology, synapses are specialized structures that permit the transfer of signals between two neurons with an associated synaptic strength or weight. Learning typically implies the modification of synaptic weight due to the activities of the pre- and post-synaptic neurons. As explained in chapter 2, the triplet STDP model is again reviewed

in detail next.

4.1.1 Triplet STDP (T-STDP) Model

Previous studies [85, 71] show that the D-STDP model fails to reproduce the experimental outcomes involving higher order spike patterns such as triplet and quadruplets of spikes and furthermore, fails to account for the observed weight dependence on repetition frequency of pairs of spikes. To resolve the above mentioned issues, the D-STDP model was extended in [29] to include spike triplets resulting in T-STDP model which could sufficiently reproduce physiological experiments.

The T-STDP rule is written as a function of difference in spike timings as, [54, 29]

$$\begin{aligned}\Delta w^+ &= e^{\left(\frac{-\Delta t_1}{\tau_+}\right)}(A_2^+ + A_3^+ e^{\left(\frac{-\Delta t_2}{\tau_y}\right)}) \quad \text{if } t = t_{post} \\ \Delta w^- &= -e^{\left(\frac{\Delta t_1}{\tau_-}\right)}(A_2^- + A_3^- e^{\left(\frac{-\Delta t_3}{\tau_x}\right)}) \quad \text{if } t = t_{pre}\end{aligned}\tag{Eq. 4.1}$$

where A_2^+ and A_2^- denote the amplitude of the weight change whenever there is a pre-post pair or a post-pre pair respectively. Similarly, A_3^+ and A_3^- denote the amplitude of the triplet term for potentiation and depression, respectively. $\Delta t_1 = t_{post}(n) - t_{pre}(n)$, $\Delta t_2 = t_{post}(n) - t_{post}(n-1)$ and $\Delta t_3 = t_{pre}(n) - t_{pre}(n-1)$ are time difference between combinations of pre and post-synaptic spikes as shown in Fig.4.1. τ_- , τ_+ , τ_x and τ_y are time constants for the above spike pairings.

In [29], though the T-STDP rule above is introduced first, it is shown later that not all terms are needed to explain biological data. Thus two different minimal models are defined later: (1) $A_2^+ = 0$ and $A_3^- = 0$ for visual cortex data and (2) $A_3^- = 0$ for hippocampal culture data set. Hippocampal culture data set [71] is used for obtaining the results for triplets of spikes whereas visual cortex data [85] is used for showing the

frequency effects of T-STDP rule. For visual cortex data set, Eq. 4.1 simplifies to,

$$\Delta w(t) = \begin{cases} \Delta w^+ = A_3^+ e^{\left(\frac{-\Delta t_2}{\tau_y}\right)} e^{\left(\frac{-\Delta t_1}{\tau_+}\right)}; & t = t_{post} \\ \Delta w^- = -e^{\left(\frac{\Delta t_1}{\tau_-}\right)} (A_2^-); & t = t_{pre} \end{cases} \quad (\text{Eq. 4.2})$$

On the other hand, for hippocampal culture data set, Eq. 4.1 simplifies to,

$$\Delta w(t) = \begin{cases} \Delta w^+ = e^{\left(\frac{-\Delta t_1}{\tau_+}\right)} (A_2^+ + A_3^+ e^{\left(\frac{-\Delta t_2}{\tau_y}\right)}); & t = t_{post} \\ \Delta w^- = -e^{\left(\frac{\Delta t_1}{\tau_-}\right)} (A_2^-); & t = t_{pre} \end{cases} \quad (\text{Eq. 4.3})$$

For both cases, Δw^- is exactly same as the case of long term depression (LTD) in D-STDP as shown in equation (Eq. 2.4). Hence, to implement the triplet rule in circuits, we only need to modify pre-existing FG design to add the extra term in the potentiation case.

4.2 Floating Gate Synapse

Fig.4.2(a) shows the architecture of a single floating gate synapse in prior work [55] or in chapter 3. It has three main terminals for programming as shown in the dashed box. The terminals are named as gate, drain and tunnel terminals with the respective voltages denoted as V_g , V_d and V_{tun} . A “non-STDP” behaviour seen in this work is ameliorated in our previous work [63, 64] or chapter 3 along with detailed analysis of the operation of D-STDP learning rule in a floating gate synapse. In previous works [55, 63, 64], the quantum mechanism of tunneling is spread across a larger time scale (illustrated in Fig.4.3(a) and (b)) which makes it difficult to analyze the effect of triplets and quadruplets of spikes in a floating gate synapse. Similar to an approach in [81, 82], the effect of tunneling can be localized at the occurrence of pre-synaptic spikes with the modification (red blocks) shown in the architecture of Fig.4.2(b) making

Figure 4.1: **Timing diagram of T-STDP rule:** The timing specifications of data points (x-axis) for triplet experiments: protocol 1 and 2. A comparison of the temporal notations in this chapter and [29] (mentioned as Pfister Paper in figure) is shown.

it easier to mathematically analyze the weight change for triplets. In the new architecture, whenever a pre-synaptic spike occurs, a triangular gate voltage waveform is generated which will create an exponential excitatory post-synaptic current (EPSC), similar to biology, because of the exponential relationship between the gate voltage and drain current of the MOS transistor in subthreshold region. The current at the maximum gate voltage is nearly zero. Similarly, whenever a post-synaptic spike arrives, a global triangular tunnel voltage waveform and an inverted pulse drain voltage waveform is generated. The global triangular tunnel voltage waveform is then sampled at the occurrence of pre-synaptic spike with the help of pulse extender block and the multiplexer. This creates a voltage waveform, V_{tun_eff} at the tunnel terminal. Thus, in the new architecture, during pre-synaptic spike, a gate voltage, V_g and a tunnel volt-

Figure 4.2: **Floating Gate (FG) synapse:** (a) The architecture used in previous work. (b) The architecture of a FG synapse used in this work with different input terminal voltage waveforms. The additional blocks used in the architecture is shown in red.

age, V_{tun_eff} waveforms are generated and during post-synaptic spike, a drain voltage, V_d waveform is generated. Whereas, in the previous architecture, a gate voltage waveform is generated during pre-synaptic spike and both drain voltage and tunnel voltage waveforms are generated during post-synaptic spike.

The equation for drain current of a subthreshold saturated pFET whose well is tied to V_{dd} is given by [55]

$$I_d = I_{s0} e^{\kappa(V_{dd}-V_{fg})/U_T} \quad (\text{Eq. 4.4})$$

where U_T is the thermal voltage and κ is the gate coupling coefficient.

As explained in chapter 3, weight modification in a FG synapse uses a combination of hot-electron injection (HEI) and Fowler-Nordheim tunneling [55]. HEI adds electrons on to the floating gate node, which reduces the floating gate voltage resulting in more current through the transistor hence increasing the weight of the synapse. On the other hand tunneling removes electrons from the FG node to reduce the synaptic weight. In the previous chapter, for the case of D-STDP, a gate voltage waveform is generated at every pre-synaptic spike while at every post-synaptic spike, a tunneling voltage and a drain voltage is generated as illustrated in Fig.4.3(a). This shows that at every pre-synaptic spike only tunneling happens while at every post-synaptic spike there is both injection and tunneling. The occurrence of both tunneling and injection at the post-synaptic spike arrival necessitates the effect of injection to be greater than tunneling to obtain potentiation as in the traditional D-STDP learning window. Moreover since the effect of tunneling is spread over a long time, mathematical analysis or intuitive understanding of the effect due to multiple pre- and post-synaptic spikes becomes difficult. In contrast, the mathematical versions of both STDP models have potentiation and depression of weights localized at pre- and post-synaptic events. Hence, to relate the mathematical analysis of FG synapse with the T-STDP model easily, we also decided to localize the effect of tunneling and injection at pre- and post-synaptic pulses respectively. A similar technique has also been used in [81, 82].

Figure 4.3: **Timing diagram and voltage specifications of FG synapse:** (a) D-STDP model, cases of long term potentiation (LTP) and long term depression (LTD), where effect of injection and tunneling is present in LTP where as only tunneling effect is present for LTD. These waveforms are used in the previous chapter (b) The details of timing and voltage specifications used for testing the FG synapse for both triplet protocols is shown here. The modification of tunneling voltage, V_{tun_eff} (shown in red) sampled from V_{tun} is shown.

An illustration for this is shown in Fig.4.3(b) where a tunneling voltage waveform is still created at every post synaptic pulse as in earlier work [63, 55] but is not applied directly to the tunneling junction of the FG. Instead, it is sampled by the pre-synaptic pulse through a multiplexer to create a new waveform V_{tun_eff} , which is applied to the FG (refer to Fig.4.2). Figure 4.3(b) also shows the specifications of terminal voltage waveforms for the case of triplets of spike. The well for the high voltage pMOS in the multiplexer can be shared with tunneling junction; nevertheless, this incurs an area penalty motivating us to look at possibilities of using non-localized tunneling for T-STDP in future. The governing equations for injection and tunneling are given as [55, 65]

$$\begin{aligned} I_{inj} &= I_{inj0} (I_d / I_{s0})^\alpha e^{-\Delta V_{ds} / V_{inj}} \\ I_{tun} &= I_{tun0} e^{(V_{tun} - V_{fg}) / V_{ox}} \end{aligned} \quad (\text{Eq. 4.5})$$

where I_d is the drain current, $\alpha = 1 - U_T / V_{inj}$, V_{ox} and V_{inj} are process dependent parameters.

4.2.1 D-STDP Model on a Floating Gate Synapse (FG D-STDP)

The weight of the synaptic device can be defined as [55]:

$$w = e^{\frac{-\kappa V_{fg}}{U_T}} \quad (\text{Eq. 4.6})$$

Hence, equations to predict the change in FG voltage effectively predict the change in weight. The following assumptions are made in deriving the theoretical equations as in chapter 3 [63]:

- (i) To ensure small change in weight at very large negative and positive values of Δt , V_{g_init} has to be high enough so that $V_{tun_max} - V_{g_init}$ is small enough for negligible tunneling. Similarly $\Delta V_g = V_{g_init} - V_{g_min}$, should be small enough so

that $V_{tun_init} - V_{g_min}$ is small enough for negligible tunneling.

- (ii) Strong coupling from gate to floating gate node where as a weak coupling from tunneling node to floating gate node. This is justified since typically gate capacitance, $C_g \gg$ tunneling capacitance, C_{tun} .
- (iii) The gate voltage waveform falls to its minimum value instantaneously. In other words, $S_1 \gg S_2, S_3$ as shown in Fig.4.3(a).

One difference from the earlier case in [63] is that now we only have injection for $\Delta t > 0$ and only tunneling for $\Delta t < 0$ due to the localization of tunneling effects.

We can now derive the slow time scale equation [55] for change in FG voltage due to tunneling and injection as:

$$C_T \frac{d\overline{V_{fg}}}{dt} = I_{tun} - I_{inj} = C_T \frac{d\overline{V_{fg_tun}}}{dt} - C_T \frac{d\overline{V_{fg_inj}}}{dt} \quad (\text{Eq. 4.7})$$

where C_T is the total capacitance on the FG node and $\overline{V_{fg}}$ denotes change on a slow time scale.

4.2.1.1 Case 1: $\Delta t > 0$

First, we consider the case of $\Delta t > 0$ i.e the positive axis of STDP curve and combine equations (Eq. 4.5) and (Eq. 4.4) to get:

$$C_T \frac{d\overline{V_{fg_inj}}}{dt} = -I_{inj0} (e^{\kappa(V_{dd}-V_{fg})/U_T})^\alpha e^{-\Delta V_{ds}/V_{inj}} \quad (\text{Eq. 4.8})$$

where $\overline{V_{fg.inj}}$ is the slow time scale change in V_{fg} due to injection only. Since change in V_{fg} on the RHS happens due to coupling from the gate voltage, we can write:

$$V_{fg} = V_{fg.min} + \frac{C_g}{C_T} S_2 t \quad (\text{Eq. 4.9})$$

where S_2 is positive slope of V_g and C_g is the capacitance connected between the gate terminal and the floating gate terminal. Here $V_{fg.min}$ is not same as $V_{g.min}$ due to initial charge stored on the FG. Substituting equation (Eq. 4.9) in equation (Eq. 4.8), we get

$$C_T \frac{d\overline{V_{fg.inj}}}{dt} = -A e^{-Xt} \quad (\text{Eq. 4.10})$$

where

$$\begin{aligned} A &= I_{inj0} (e^{\alpha\kappa(V_{dd}-V_{fg.min})/U_T}) (e^{-\Delta V_{ds}/V_{inj}}) \\ X &= \frac{\alpha\kappa C_g S_2}{C_T U_T} \end{aligned} \quad (\text{Eq. 4.11})$$

Referring to Fig.4.3(a), significant amount of injection happens in the time from Δt to $\Delta t + T_d$ (circled in red), where ΔV_{ds} is constant and significant. Also, since T_d is very small compared to T_g , we can assume that V_g is constant during the drain pulse. Hence, we finally get:

$$C_T \Delta \overline{V_{fg.inj}} = -A T_d e^{-X\Delta t} \quad (\text{Eq. 4.12})$$

For more details of the derivation, refer to chapter 3.

With reference to Fig.4.3(b), since we have modified the tunneling voltage waveform from V_{tun} to $V_{tun.eff}$ the effect of tunneling is not present in $\Delta t > 0$. Hence, we have $\Delta \overline{V_{fg}} = \Delta \overline{V_{fg.inj}}$.

4.2.1.2 Case 2: $\Delta t < 0$

Now we consider the case of $\Delta t = (t_{post} - t_{pre}) < 0$ i.e the negative axis of STDP curve. Similar to what we have done above, let us see the effect of tunneling and injection separately.

For the contribution of injection to $\overline{V_{fg}}$, we could see that injection happens only during the initial small period, T_d of time axis where the drain current of the MOS is almost zero. So we can completely neglect the effect of injection on $\overline{V_{fg}}$ in this case.

Now let us consider the contribution of tunneling to $\overline{V_{fg}}$. Similar to the analysis above, using assumption 1, we have:

$$C_T \int_0^{\Delta \overline{V_{fg.tun}}} dV_{fg} = \int_{-\Delta t}^{T_{tun}} I_{tun} dt \quad (\text{Eq. 4.13})$$

Also, similar to the case of positive Δt , here we have:

$$\begin{aligned} V_{fg} &= V_{fg.min} + \frac{C_g}{C_T} S_2(t - (-\Delta t)) \\ V_{tun} &= V_{tun.max} + S_3(t - T_{tun.delay}) \\ V_{tun.eff} &= \begin{cases} V_{tun}; & -\Delta t < t < -\Delta t + T_{tun.pulse} \\ V_{tun.init}; & \text{for other values of } t \end{cases} \end{aligned} \quad (\text{Eq. 4.14})$$

Substituting equations (Eq. 4.14) and (Eq. 4.5) into equation (Eq. 4.13), we get

$$\begin{aligned} C_T \Delta \overline{V_{fg.tun}} &= B e^{\frac{-C_g S_2 \Delta t}{C_T V_{ox}}} \int_{-\Delta t}^{-\Delta t + T_{tun.pulse}} e^{Yt} dt \\ &= B' (e^{Y(T_{tun.pulse} - \Delta t)} - e^{-Y\Delta t}) e^{\frac{-C_g S_2 \Delta t}{C_T V_{ox}}} \end{aligned} \quad (\text{Eq. 4.15})$$

where

$$\begin{aligned}
 Y &= \frac{S_3 - \frac{C_g S_2}{C_T}}{V_{ox}} \\
 B &= I_{tun0} \left(e^{\frac{V_{tun-max} - S_3 T_{tun-delay} - V_{fg-min}}{V_{ox}}} \right) \\
 B' &= B/Y
 \end{aligned} \tag{Eq. 4.16}$$

Since injection is negligible for $\Delta t < 0$, we have $\Delta \overline{V_{fg}} = \Delta \overline{V_{fg,tun}}$.

4.2.2 T-STDP Model on a Floating Gate Synapse (FG T-STDP)

Comparing equation (Eq. 2.4) with equation (Eq. 2.7), the extra term in T-STDP compared to the D-STDP model is $A_3^+ e^{(\frac{-\Delta t_2}{\tau_y})} e^{(\frac{-\Delta t_1}{\tau_+})}$, which occurs at the arrival of post-synaptic spike. Intuitively, for the case of triplet of spikes, we can see that there should be some modification to the drain voltage pulse compared to the doublet case since it is related to potentiation and is generated at the post-synaptic spike. There are two ways in which we can have more potentiation due to injection: (1) increase ΔV_{ds} or (2) increase injection pulse width T_d . Based on this, we propose the following two drain voltage waveform that accounts for triplet spike interactions:

- (i) Single-pulsed V_d : This case implements the entire potentiation term in equation (Eq. 4.3) by a single pulse of width T_d at t_{post} (Fig.4.4 – however, the voltage V_d depends on the time difference Δt_2 between successive post spikes. So from Fig.4.4 expression for V_d becomes:

$$V_d = \begin{cases} V_{d,min} - \Delta V_d^s(\Delta t_2); & t_{post}(n) < t < t_{post}(n) + T_d \\ V_{d,init}; & \text{for other values of } t \end{cases} \tag{Eq. 4.17}$$

Note that to satisfy the results of the doublet case, $\Delta V_d^s(\Delta t_2) \rightarrow 0$; as $\Delta t_2 \rightarrow \infty$.

- (ii) Double-pulsed V_d : The double-pulsed V_d waveform (Fig.4.4) implements the dou-

Figure 4.4: **Different cases of drain voltage waveform:** V_d for different experiments like pairing protocol, triplets and quadruplets.

blet term with the first pulse of width T_d and amplitude same as in FG D-STDP. It then creates the extra triplet term with the following pulse again of width T_d but with different amplitude. The entire waveform is given by:

$$V_d = \begin{cases} V_{d_min}; & t_{post}(n) < t < t_{post}(n) + T_d \\ V_{d_min} - \Delta V_d^d(\Delta t_2); & t_{post}(n) + T_d < t < t_{post}(n) + 2 * T_d \\ V_{d_init}; & \text{for other values of } t \end{cases} \quad (\text{Eq. 4.18})$$

Again, note that to satisfy the results of the doublet case, $V_{d_min} - \Delta V_d^d(\Delta t_2) \rightarrow V_{d_init}$; as $\Delta t_2 \rightarrow \infty$ i.e second pulse in double-pulsed V_d vanishes as $\Delta t_2 \rightarrow \infty$.

As can be seen above, both cases of drain voltage waveform will require

some different method to calculate $\Delta V_d^s(\Delta t_2)$ and $\Delta V_d^d(\Delta t_2)$ to match results from the T-STDP model with desired parameters. So, next we do a comparison of the FG phenomenon and T-STDP model to understand the kind of modification to be done. Here, we consider only the FG phenomenon of injection that happens at post-synaptic pulses, since our modification has localized tunneling to pre-synaptic events. Then, we can equate this to the weight change from T-STDP model for any specific case (e.g. hippocampal culture data set in [29]) to extract the exact parameters of the pulse drain waveform. Here, we will express the weight change for T-STDP normalized to D-STDP for ease of understanding; this is intuitively good since T-STDP has modifications on top of D-STDP. Let us consider this case by case.

(i) Case 1: FG T-STDP for spike doublet inputs

The FG phenomenon on doublets of spike is exactly similar to the derivation shown in previous subsection 4.2.1 for single and double pulsed V_d cases. We consider the case of $\Delta t > 0$ i.e the positive axis of STDP curve and combine equations (Eq. 4.5) and (Eq. 4.4) to get:

$$C_T \frac{d\overline{V_{fg_inj}}}{dt} = -I_{inj0} (e^{\kappa(V_{dd}-V_{fg})/U_T})^\alpha e^{-\Delta V_{ds}/V_{inj}} \quad (\text{Eq. 4.19})$$

where $\overline{V_{fg_inj}}$ is the slow time scale change in V_{fg} due to injection only. Since change in V_{fg} on the RHS happens due to coupling from the gate voltage, we can write:

$$V_{fg} = V_{fg_min} + \frac{C_g}{C_T} S_2 t \quad (\text{Eq. 4.20})$$

where S_2 is positive slope of V_g and C_g is the capacitance connected between the gate terminal and the floating gate terminal. Here V_{fg_min} is not same as V_{g_min} due to initial charge stored on the FG. Substituting equation (Eq. 4.20) in

equation (Eq. 4.19), we get

$$C_T \frac{d\overline{V_{fg.inj}}}{dt} = -I_{inj0} (e^{\alpha\kappa(V_{dd}-V_{fg.min})/U_T}) (e^{-\Delta V_{ds}/V_{inj}}) (e^{-Xt}) \quad (\text{Eq. 4.21})$$

where

$$X = \frac{\alpha\kappa C_g S_2}{C_T U_T} \quad (\text{Eq. 4.22})$$

Referring to Fig.4.3(a), significant amount of injection happens in the time from Δt to $\Delta t + T_d$ (circled in red). Integrating both sides,

$$C_T \int_0^{\Delta \overline{V_{fg.inj}^{doublet}}} dV_{fg.inj} = \int_{\Delta t}^{\Delta t + T_d} -I_{inj0} (e^{\alpha\kappa(V_{dd}-V_{fg.min})/U_T}) (e^{-\Delta V_{ds}/V_{inj}}) (e^{-Xt}) dt \quad (\text{Eq. 4.23})$$

where ΔV_{ds} is constant and equal to $V_{d.min} - V_{dd}$. Therefore,

$$\Delta \overline{V_{fg.inj}^{doublet}} = A e^{-X\Delta t} e^{(V_{dd}-V_{d.min})/V_{inj}} \quad (\text{Eq. 4.24})$$

where

$$A = \frac{I_{inj0} (e^{\alpha\kappa(V_{dd}-V_{fg.min})/U_T}) (e^{-XT_d} - 1)}{C_T X} \quad (\text{Eq. 4.25})$$

$$X = \frac{\alpha\kappa C_g S_2}{C_T U_T}$$

(ii) Case 2: FG T-STDP for spike triplet inputs

(a) Single-pulsed V_d : Similar to the mathematics in case 1, we have:

$$C_T \frac{d\overline{V_{fg.inj}}}{dt} = -I_{inj0} (e^{\alpha\kappa(V_{dd}-V_{fg.min})/U_T}) (e^{-\Delta V_{ds}/V_{inj}}) (e^{-Xt}) \quad (\text{Eq. 4.26})$$

Referring to Fig.4.3(b), significant amount of injection happens in the time from $t_{post}(n)$ to $t_{post}(n) + T_d$ which is same as Δt to $\Delta t + T_d$. Integrating both

sides,

$$C_T \int_0^{\overline{\Delta V_{fg.inj}^{triplet}}} dV_{fg.inj} = \int_{\Delta t}^{\Delta t + T_d} -I_{inj0} (e^{\alpha \kappa (V_{dd} - V_{fg.min}) / U_T}) (e^{-\Delta V_{ds} / V_{inj}}) (e^{-Xt}) dt \quad (\text{Eq. 4.27})$$

where ΔV_{ds} is constant and equal to $V_{d.min} - \Delta V_d^s(\Delta t_2) - V_{dd}$. Therefore,

$$\overline{\Delta V_{fg.inj}^{triplet}} = A e^{-X \Delta t} e^{\frac{\Delta V_d^s(\Delta t_2)}{V_{inj}}} e^{\frac{(V_{dd} - V_{d.min})}{V_{inj}}} \quad (\text{Eq. 4.28})$$

where A and X are same as above. The ratio of change in weight due to FG T-STDP on triplets and doublets of spike is given as,

$$Y_{fg} = \frac{\Delta W_{inj}^{triplet}}{\Delta W_{inj}^{doublet}} = \frac{\overline{\Delta V_{fg.inj}^{triplet}}}{\overline{\Delta V_{fg.inj}^{doublet}}} = e^{\frac{\Delta V_d^s(\Delta t_2)}{V_{inj}}} \quad (\text{Eq. 4.29})$$

(b) Double-pulsed V_d : Again, similar to the mathematics in above case, we get:

$$\overline{\Delta V_{fg.inj}^{triplet}} = \overline{\Delta V_{fg.inj}^{doublet}} (1 + e^{-X T_d} e^{\frac{\Delta V_d^d(\Delta t_2)}{V_{inj}}}) \quad (\text{Eq. 4.30})$$

where X is same as above. Here, the term $e^{-X T_d}$ arises since the gate voltage has changed a little bit at the start of the second pulse compared to the first. Since $e^{-X T_d} \approx 1$ due to small value of T_d , the ratio of change in weight due to FG T-STDP on triplets and doublets of spike can be written as:

$$Y_{fg} = \frac{\Delta W_{inj}^{triplet}}{\Delta W_{inj}^{doublet}} = \frac{\overline{\Delta V_{fg.inj}^{triplet}}}{\overline{\Delta V_{fg.inj}^{doublet}}} = 1 + e^{\frac{\Delta V_d^d(\Delta t_2)}{V_{inj}}} \quad (\text{Eq. 4.31})$$

(iii) T-STDP rule for spike doublets and triplets

For the case of T-STDP and D-STDP models the ratio Y can be obtained directly from equation (Eq. 4.3) as:

$$Y_{theory} = \frac{\Delta W_{triplet}^+}{\Delta W_{doublet}^+} = 1 + \frac{A_3^+ e^{\frac{(-\Delta t_2)}{\tau_y}}}{A_2^+} \quad (\text{Eq. 4.32})$$

Now, we can equate the Y_{fg} values in equations (Eq. 4.29) and (Eq. 4.31) with the Y_{theory} value in equation (Eq. 4.32) to get the desired drain voltage parameters.

(a) Single-pulsed V_d :

$$\begin{aligned}
 Y_{fg} &= Y_{theory} \\
 \implies e^{\Delta V_d^s(\Delta t_2)/V_{inj}} &= 1 + \frac{A_3^+ e^{(-\frac{\Delta t_2}{\tau_y})}}{A_2^+} \\
 \implies \Delta V_d^s(\Delta t_2) &= V_{inj} \ln\left(1 + \frac{A_3^+ e^{(-\frac{\Delta t_2}{\tau_y})}}{A_2^+}\right)
 \end{aligned} \tag{Eq. 4.33}$$

(b) Double-pulsed V_d :

$$\begin{aligned}
 Y_{fg} &= Y_{STDP} \\
 \implies e^{\Delta V_d^d(\Delta t_2)/V_{inj}} &= \frac{A_3^+ e^{(-\frac{\Delta t_2}{\tau_y})}}{A_2^+} \\
 \implies \Delta V_d^d(\Delta t_2) &= V_{inj} \ln\left(\frac{A_3^+ e^{(-\frac{\Delta t_2}{\tau_y})}}{A_2^+}\right)
 \end{aligned} \tag{Eq. 4.34}$$

4.2.3 Parameter Translation from Theory to FG model

Given values of parameters such as A_2^+ or A_3^+ of the theoretical model, it should be possible to use the above equations (Eq. 4.33) and (Eq. 4.34) to find the values of ΔV_d^s and ΔV_d^d . However, in our case, the measurement setup restricted the temporal width of voltage waveforms to a maximum value of 300ms (note the maximum time duration used for T_{tun}). Hence, compared to the learning window τ^+ (refer to Fig.2.1) shown in [29], our learning window τ_{fg}^+ (Fig.4.6) has lesser width. τ_{fg}^+ can be defined as the time decay constant of STDP learning window on the positive x-axis. This implies that we need to apply a compression factor on the time scale to match our results with those

in [29]. We define the compression factor r as follows:

$$r = \frac{\tau^+}{\tau_{fg}^+} \quad (\text{Eq. 4.35})$$

For the present case, we get the value of $r \approx 2$. Thus, to replicate the weight changes in [29], we need to reduce all the temporal dimensions in our case by the factor of r . This can be formally written as:

$$\begin{aligned} \Delta V_d^s(\Delta t_2/r) &= V_{inj} \ln\left(1 + \frac{A_3^+ e^{\left(\frac{-\Delta t_2}{\tau_y}\right)}}{A_2^+}\right) \\ \Delta V_d^d(\Delta t_2/r) &= V_{inj} \ln\left(\frac{A_3^+ e^{\left(\frac{-\Delta t_2}{\tau_y}\right)}}{A_2^+}\right) \end{aligned} \quad (\text{Eq. 4.36})$$

Next, we show this translation explicitly for the three experimental protocols considered in [29].

(i) Triplet experiments:

For the different parameters of hippocampal culture data set from table 4 in [29] i.e $A_3^+ = 9.1 \times 10^{-3}$, $A_2^+ = 4.6 \times 10^{-3}$ and $\tau_y = 48\text{ms}$, we can obtain from equation (Eq. 4.33) for the case of single-pulsed V_d waveform $\Delta V_d^s(\Delta t_2) \approx 0.24\text{V}$ for the case of $\Delta t_2 = 10\text{ms}$ and $\Delta V_d^s(\Delta t_2) \approx 0.21\text{V}$ for $\Delta t_2 = 20\text{ms}$. Similarly, for the case of double-pulsed V_d waveform, we can obtain from equation (Eq. 4.34); $\Delta V_d^d(\Delta t_2) \approx 0.12\text{V}$ for the case of $\Delta t_2 = 10\text{ms}$ and $\Delta V_d^d(\Delta t_2) = 66\text{mV}$ for $\Delta t_2 = 20\text{ms}$.

The definitions of $(\Delta t_1$ and $\Delta t_2)$ here are similar to [54] and are different compared to the notations given in [29]. For clarity, the difference is shown in Fig.4.1. As an example, the data points chosen in protocol 2 are given by $(\Delta t_1/r, \Delta t_2/r)$ where $(\Delta t_1, \Delta t_2) \Rightarrow (-5, 5), (-10, 10), (-5, 15)$ and $(-15, 5)$ ms corresponds respectively to data points in [29].

(ii) Frequency effects of pairing protocol:

In [29], Δt used for the case of frequency effects is 10ms which is halved to 5ms for obtaining the measurement results in this chapter. Also, ρ Hz mentioned in frequency effects of pairing protocol results of [29] will be equal to $r \times \rho$ Hz in our measurement results.

(iii) Quadruplet experiments:

Quadruplet experiments contain two temporal variables Δt and T as shown in Fig.4.4. $\Delta t = 5\text{ms}$ in [29]—hence, we have used $\Delta t = 2.5\text{ms}$ for measurements. Then the chip measurement values for quadruplets are taken for different values of T/r up to 80ms. For comparing these measurement results with mathematical model, we simulated the mathematical model for double the value of T i.e up to 160ms.

4.3 Results

The measurement results shown in this section is obtained from a single FG synapse fabricated in TSMC $0.35\mu\text{m}$ CMOS process with $C_g = 5.5$ pF. Though not optimized for synaptic density, the results here do serve as a proof of concept for our FG T-STDP implementation. The experimental setup is shown in Fig.4.5. The details of the FPAA board used in the experimental setup can be found in [66]. Detailed block diagram and description of the experimental setup is given in chapter 3, section 3.2.

4.3.1 FG D-STDP learning window

The first set of testing on FG synapse is to check whether the FG T-STDP can reproduce D-STDP learning window. Fig.4.6 shows the D-STDP learning window obtained

Figure 4.5: **Experimental setup:** consists of floating-gate transistor test chip, board with microcontroller and digital-analog converters for generating voltage waveforms, picoammeter for measuring current and power supply.

Figure 4.6: **FG D-STDP learning window:** The curve shows learning window obtained from FG synapse using T-STDP rule. Here, $A_{fg}^+ \approx 0.26$, $A_{fg}^- \approx -0.15$, $\tau_{fg}^+ \approx 25\text{ms}$ and $\tau_{fg}^- \approx 22\text{ms}$

Figure 4.7: **Failure of FG D-STDP:** In all four subgraphs, black bars denote experimental data (taken from [29]), blue bars and lines, mentioned as D-STDP correspond to mathematical model proposed in [29] and red lines correspond to measurement results. (a) The bar graphs shows the comparison of results obtained from FG D-STDP (red) with the biological experimental results (black) and mathematical model (blue) for two cases of triplets, case 1: Pre-Post-Pre ($|\Delta t_1|/r, \Delta t_3/r$) and (b) case 2: Post-Pre-Post ($\Delta t_1/r, \Delta t_2/r$). FG D-STDP model (red bars) cannot replicate the experimental results shown in black bars. (c) FG D-STDP fails to reproduce quadruplet experiments. (d) Weight change in FG D-STDP rule as a function of frequency, $r \times \rho$. Note that, thick line correspond to $\Delta t = 5$ ms and dash line correspond to $\Delta t = -5$ ms.

from FG synapse. Thus for obtaining the result, we set V_{d_min} as 0.3V (here, $\Delta V_d(\Delta t_2) = 0$) w.r.t the mathematical analysis done in the section 4.2.2 and other voltage specifications are as given in Fig.4.3(b) i.e $V_{dd} = 5.2V$, $V_{g_init} = 3.3V$, $V_{g_min} = 2.5V$, $T_g = 100ms$, $V_{tun_init} = 5.4V$, $V_{tun_max} = 16.5V$, $T_{tun} = 300ms$, $T_{tun_delay} = 1ms$, $T_{tun_pulse} = 2ms$, $V_{d_init} = 5V$ and $T_d = 500us$.

4.3.2 Failure of FG D-STDP

We applied D-STDP rule on FG synapse to the pre- and post-synaptic spike pairs as done in [29]. For obtaining the measurement results in Fig.4.7, we set the voltage and timing parameters as given in Fig.4.3(b) i.e, same as mentioned in subsection 4.3.1.

4.3.2.1 FG D-STDP fail to reproduce frequency effects

For obtaining measurement results, we have set $\Delta t = \pm 5ms$ as shown in Fig.4.4. The measurement results are plotted as weight change in D-STDP rule as a function of frequency, ρ . The measurement results (red lines) shown in Fig.4.7(d), almost follows the mathematical D-STDP model (blue lines) in [29]. As mentioned in [29], this can be attributed to the following reasons. First, As pointed out in [85], at low repetition frequency, ρ , there is no potentiation. This cannot be captured by FG D-STDP, because pulsed drain voltage waveform, V_d at the occurrence of a post-synaptic spike after a pre-synaptic spike by few milliseconds induces LTP. Second, In experiments, for $\Delta t > 0$, potentiation increases when frequency increases. This trend can also not be reproduced by FG D-STDP. In D-STDP model, as soon as the frequency increases, the pre-post spike pairs approach each other and the post-synaptic spike of the first spike pair interact with the pre-synaptic spike of the subsequent spike pair. With increase in frequency, the post-pre spike interaction increases and therefore depress

the synapse, which is not observed in experiments.

4.3.2.2 FG D-STDP fail to reproduce triplet experiments

In triplet experiments, as shown in Fig.4.7(a) and (b), there is a clear asymmetry between two protocols mentioned (black bars). 60 repetitions of pre-post-pre triplet yields less weight change, whereas 60 repetitions of post-pre-post triplet yields a weight change of $\sim 30\%$ (black bars). However, FG D-STDP shown in red bars, predicts almost same result for both protocols, because the mechanism of potentiation and depression happening due to the generation of different voltage waveforms at pre- or post-synaptic spikes are similar in both protocols. Therefore, triplet results cannot be explained by a sum of pre-post injection mechanism and a post-pre tunneling mechanism.

4.3.2.3 FG D-STDP fail to reproduce quadruplet experiments

The asymmetry present in the quadruplet experiments, as shown in Fig.4.7(c), also causes some problem for FG D-STDP. A quadruplet consists of a pre-post-post-pre sequence where, $T < 0$ or a post-pre-pre-post sequence where, $T > 0$ as shown in Fig.4.4. Here, $|T|$ denotes the interval between the first and last pair of spikes within the quadruplet. Sequence, pre-post-post-pre consists of two pre-post interactions and a post-pre interaction whereas; for the sequence, post-pre-pre-post, the opposite occurs i.e two post-pre interactions and only one pre-post interaction. This clearly leads to an asymmetry which is not seen in experiments [29].

4.3.3 Success of FG T-STDP

FG T-STDP is implemented with the help of a drain voltage waveform generator, explained in next section 4.4. As mentioned in the previous section 4.2.2, the mathematical analysis provides an intuitive understanding of utilization of V_d pulse waveform to obtain the extra triplet term in the T-STDP rule (equation (Eq. 2.7)). As the extra triplet term can be achieved with the two cases of proposed pulse drain voltage waveform, here, we have shown measurement results for both the cases of V_d waveform. The measurement results obtained for both the case of drain voltage waveform satisfy the behavior seen in experimental results [29]. The timing specifications for frequency effects of pairing protocol, triplets of spike and quadruplets are shown in Fig.4.4.

4.3.3.1 FG T-STDP can reproduce triplet and quadruplet experiments

T-STDP model on floating gate synapse does not only reproduce the learning window (Fig.4.6), but also it can reproduce most of the triplet and quadruplet experiments as shown in Fig.4.8(a),(b) and (c) and Fig.4.9(a),(b) and (c). The voltage and timing parameters for the case of single-pulsed drain voltage waveform are given in Fig.4.3(b) i.e, same as mentioned in subsection 4.3.1. Here, $V_{d.min} = 0.3V$ is set with respect to $\Delta V_d(\Delta t_2)$ value (equation (Eq. 4.33)), for parameters of Nearest-Spike minimal model, hippocampal culture data set, from table 4 in [29] i.e $A_3^+ = 9.1 \times 10^{-3}$, $A_2^+ = 4.6 \times 10^{-3}$ and $\tau_y = 48ms$. For the case of double-pulsed drain voltage waveform, voltage and timing parameters are also same as in the case of single pulse but, $\Delta V_d(\Delta t_2)$ is calculated using equation (Eq. 4.34). In the triplet measurement results the red bars can almost capture the experimental black bars i.e more potentiation in the case of protocol 1 and less potentiation in the case of protocol 2. The blue bars shows the result of the mathematical model in [29]. The quadruplet measurement results also replicate the symmetry seen in the experimental results [29].

Figure 4.8: **FG T-STDP using single-pulsed drain waveform:** In all four subgraphs, black bars denote experimental data (taken from [29]), blue bars and lines, mentioned as T-STDP correspond to Nearest-Spike minimal triplet model proposed in [29] and red lines correspond to measurement results. (a) The bar graphs shows the comparison of results obtained from FG T-STDP (red) with the biological experimental results (black) and mathematical model (blue) for T-STDP protocol 1:Pre-Post-Pre ($|\Delta t_1|/r, \Delta t_3/r$). Similar to (a), results for T-STDP protocol 2:Post-Pre-Post ($\Delta t_1/r, \Delta t_2/r$) is shown in (b). FG T-STDP (red bars) can replicate the experimental results shown in black bars. (c) FG T-STDP can also reproduce quadruplet experiments. (d) Weight change in T-STDP rule as a function of frequency, $r \times \rho$. FG T-STDP can reproduce frequency effects compared to FG D-STDP. Note that, thick line correspond to $\Delta t = 5$ ms and dash line correspond to $\Delta t = -5$ ms.

Figure 4.9: FG T-STDP using double-pulsed drain waveform: In all four subgraphs, black bars denote experimental data (taken from [29]), blue bars and lines, mentioned as T-STDP correspond to Nearest-Spike minimal triplet model proposed in [29] and red lines correspond to measurement results. (a) The bar graphs shows the comparison of results obtained from FG T-STDP (red) with the biological experimental results (black) and mathematical model (blue) for T-STDP protocol 1:Pre-Post-Pre ($|\Delta t_1|/r, \Delta t_3/r$). Similar to (a), results for T-STDP protocol 2:Post-Pre-Post ($\Delta t_1/r, \Delta t_2/r$) is shown in (b). FG T-STDP (red bars) can replicate the experimental results shown in black bars. (c) FG T-STDP can also reproduce quadruplet experiments. (d) Weight change in T-STDP rule as a function of frequency, $r \times \rho$. FG T-STDP can reproduce frequency effects compared to FG D-STDP. Note that, thick line correspond to $\Delta t = 5$ ms and dash line correspond to $\Delta t = -5$ ms.

4.3.3.2 FG T-STDP model can reproduce frequency effects

FG T-STDP shown in red lines in Fig.4.8(d) and Fig.4.9(d) ameliorated the results obtained with FG D-STDP (Fig.4.7(d)). The trend of increase in potentiation with frequency is seen for both cases of pulse drain voltage waveform: single and double pulse. The voltage and timing parameters for the case of single-pulsed drain voltage waveform are also same as mentioned in subsection 4.3.1 but, with a difference of $V_{d,min} = 0.5V$ instead of $0.3V$. $V_{d,min}$ is set with respect to the $\Delta V_d(\Delta t_2)$ value, (equation (Eq. 4.33)), for parameters of $A_3^+ = 28.7 \times 10^{-3}$, $A_2^+ = 4.6 \times 10^{-3}$ and $\tau_y = 48ms$. Note that, we have used $A_3^+ = 28.7 \times 10^{-3}$, which is almost thrice compared to $A_3^+ = 9.1 \times 10^{-3}$ of the hippocampal culture data set, to ensure more potentiation with increase in frequency. Also note that, the measurement results for frequency effect of pairing protocol is obtained with the help of hippocampal culture data set instead of visual cortex data set mentioned in [29]. Again, for the case of double-pulsed drain voltage waveform, voltage and timing parameters are same as in the case of single pulse but, $\Delta V_d(\Delta t_2)$ is calculated using equation (Eq. 4.34), for parameters of $A_3^+ = 28.7 \times 10^{-3}$, $A_2^+ = 4.6 \times 10^{-3}$ and $\tau_y = 48ms$. As in [29], here also we have a limitation i.e the absence of potentiation at low frequency is not observed in the measurement results in the case of single-pulsed and double-pulsed drain waveform (Fig.4.8(d) and Fig.4.9(d)). This is because the single pulse at the first post-synaptic spike itself is enough to generate some injection.

4.3.3.3 FG T-STDP model can be mapped to BCM learning rule

Another advantage of T-STDP is to map the rate based BCM learning rule. BCM rule is explained in detail in chapter 2. The weight change in BCM learning rule is based on the postsynaptic firing rate. Lower the postsynaptic firing rate, depotentiation happens where as for higher postsynaptic firing rates potentiation happens. There is

Figure 4.10: **FG T-STDP model can be mapped to BCM learning rule:** The time based T-STDP learning rule successfully maps the rate based BCM learning rule, showing the advantage of T-STDP rule to modify synaptic weights based on both rate as well as timing of spikes. Chip measurement results for two parameters shows different thresholds for BCM rule.

a threshold at which the weight change is zero, known as the sliding threshold. The chip measurement results to substantiate the advantage is shown in Fig.4.10. The change in synaptic weight is plotted against the postsynaptic firing frequency. It can be seen that lower firing frequency results in depression while higher frequency results in potentiation. The point between these transition is the threshold point of BCM rule, where the change in weight is zero. The threshold can be easily controlled using the V_{tun_max} parameter in a floating gate synapse.

4.4 Drain Voltage Generator: VLSI Implementation

The drain voltage waveform generator is the important block for generating the voltage pulses according to spike timing as shown in Fig.4.2. The input to the generator is a post-synaptic pulse from a neuron as shown in Fig.4.13. The output pulse from the generator is fed back to the drain terminal of FG synapse. The number of drain waveform generators in a system depend on the number of neurons present in that neural network architecture. We propose circuits for single- and double-pulsed drain voltage waveform generator below along with their SPICE simulation results.

4.4.1 Single-Pulsed Drain Voltage Waveform

From equation (Eq. 4.33), we need to create an exponentially decaying voltage trace for $\Delta V_d^s(\Delta t_2)$ for large values of Δt_2 . Also, V_{d_min} is the default value of V_d when there is no post synaptic pulse for a long time ($\Delta t_2 \rightarrow \infty$). In order to create the exponential voltage trace, we can use a capacitor, C and a switched capacitor resistor, R_{sc} , where the capacitor charges from the lowest voltage, $V_{d_min} - \Delta V_{dmax}$ to V_{d_min} through the resistor (Fig.4.11(a)). Here, ΔV_{dmax} is given by the equation (Eq. 4.33), where $\Delta t_2 \rightarrow 0$. The operation of the circuit is as follows: at every post-synaptic pulse denoted as CLK in Fig.4.11(a), the voltage across the capacitor V_c^s is sampled as V_d through the multiplexer. At other times, the multiplexer enforces $V_d = V_{d_init}$. A delayed pulse CLK_d is also generated after the clock pulse such that it does not overlap with CLK. At this pulse, V_c^s is pulled down to $V_{d_min} - \Delta V_{dmax}$. After this, V_c^s decays back to V_{d_min} by discharging through the resistor R_{sc} . The decay constant of this circuit is set by the parameter τ_y in the equation (Eq. 4.33). Thus, the resistance of switched capacitor and capacitance, C are designed as $R_{sc} C = \tau_y$.

The simulation result of $\Delta V_d^s(\Delta t_2)$ obtained from the VLSI circuit is shown in

Fig.4.12(a) with comparison to the ideal theoretical plots. For simulation results, we have used $C = 1\text{pF}$, $V_{d,min} = 0.3\text{V}$, $V_{d,init} = 5\text{V}$ and $V_{d,min} - \Delta V_{d,max} = 25\text{mV}$ for $A_3^+ = 9.1 \times 10^{-3}$ and $A_2^+ = 4.6 \times 10^{-3}$. For designing switched capacitor resistor, R_{sc} we have:

$$\begin{aligned} R_{sc} &= \frac{T_{sc}}{C_{sc}} \\ R_{sc}C &= \tau_y \\ \Rightarrow \frac{C}{C_{sc}} &= \frac{\tau_y}{T_{sc}} \end{aligned} \quad (\text{Eq. 4.37})$$

Here, C_{cs} is the switched capacitor and $1/T_{sc}$ is the frequency of non overlapping clocks used in switched capacitor. T_{sc} is limited by the desired time resolution for Δt_2 , which is around 2ms. Thus, $C/C_{sc} = 24$ and $C_{sc} \approx 42\text{fF}$ for $\tau_y = 48\text{ms}$. It can be seen that the circuit simulation does not exactly match the theory for moderate values of Δt_2 due to our approximation of equation (Eq. 4.33) by an exponential.

4.4.2 Double-Pulsed Drain Voltage Waveform

Fig.4.11(b) shows the circuit implementation of double-pulsed drain voltage waveform according to equations (Eq. 4.34) and (Eq. 4.18). The circuit operation is exactly similar to single pulsed drain voltage waveform generator except that $\Delta V_d^d(\Delta t_2)$ is linearly dependent on Δt_2 . Hence, the resistor is replaced with a current source, I_p . Since, we need to create double pulse, an extra clock is used here. CLK1 is to create the first pulse, during which $V_d = V_{d,min}$ and CLK2 is to sample the voltage, V_c^d across capacitor, C on to the output node, V_d in order to create the second pulse. Whenever there is no post-synaptic spike, V_d is held at $V_{d,init}$ through the switch controlled by NOR of CLK1 and CLK2. For deciding the value of I_p ; simplifying equation (Eq. 4.34), we get:

$$\Delta V_d^d(\Delta t_2) = V_{inj} \ln\left(\frac{A_3^+}{A_2^+}\right) - V_{inj} \frac{\Delta t_2}{\tau_y} \quad (\text{Eq. 4.38})$$

Figure 4.11: **VLSI implementations:** Circuit implementation of (a) single-pulsed drain voltage waveform and (b) double-pulsed drain voltage waveform.

also from capacitor charging,

$$C \frac{dV}{dt} = I_p \quad (\text{Eq. 4.39})$$

Thus equating both the slopes of equations (Eq. 4.38) and (Eq. 4.39) , we get:

$$\frac{I_p}{C} = - \frac{V_{inj}}{\tau_y} \quad (\text{Eq. 4.40})$$

(a)

(b)

Figure 4.12: **SPICE Simulation results:** (a) The simulation result of the single-pulsed drain voltage waveform circuit. The result shows the variation of $\Delta V_d^s(\Delta t_2)$ with respect to Δt_2 . The ideal MATLAB simulation result is also included for comparison. (d) The simulation result of the double-pulsed drain voltage waveform circuit.

Hence, $I_p = -5.2$ pA for $\tau_y = 48$ ms and $V_{inj} = 0.25$ V. The simulation result of $\Delta V_d^d(\Delta t_2)$ obtained from the VLSI circuit is shown in Fig.4.12(b) with comparison to the ideal MATLAB plots. For simulation results, we have used $C = 1$ pF, $V_{d,min} = 0.3$ V, $V_{d,init} = 5$ V and $V_{d,min} - \Delta V_{d,max} = 125$ mV for $A_3^+ = 9.1 \times 10^{-3}$ and $A_2^+ = 4.6 \times 10^{-3}$. It is easily seen that there is a much better match between the circuit and MATLAB simulation in this case due to the simpler functional form of ΔV_d^d .

4.5 Discussion

An important consideration for synapse designs is scalability to large arrays. Fig.4.13 shows the system level architecture of a neuromorphic hardware device with N neurons, M inputs and $M \times N$ synapses (number of synapses outnumbered compared to number of neurons). It shows the connection between FG synapses and neurons and gives an idea about the number of different voltage waveform generators to be used for generating the terminal voltages for the FG synapse shown in Fig.4.2. Here, for the entire system shown, we need only N drain voltage, N tunnel voltage and M gate voltage waveform generators. Thus the synaptic area overhead compared to earlier implementations [58] is only the added multiplexer for switching tunneling voltages.

Table 2.2 compares this work with other reported implementations of plastic synapses—a detailed review of these circuits can be found in [4]. It can be seen that our work is the first that combines high resolution non-volatile storage with sophisticated plasticity rules. The term normalized area is used to denote the ratio of the synapse area to the square of the process technology. It is a normalized metric to compare the size of a synapse circuit independent of process technology—smaller numbers refer to more compact designs. The floating-gate device used in our test chip is quite large. However, this is not a fundamental problem since we have earlier demonstrated STDP

Figure 4.13: **System level architecture:** The system level implementation of a network with $M \times N$ synapses and N neurons.

in much smaller floating-gate devices [58]—the only difference of this work with our earlier one is in terms of peripheral circuits to control drain and tunnel waveforms. Hence, after this proof of concept work, we can make a dedicated chip with floating-gates occupying $\approx 100\mu m^2$ area. Some emerging devices like memristors are also showing promise as a compact learning synapse [86, 56] for spiking systems. Though we are not yet aware of reports of dense arrays of memristive STDP synapses integrated with

CMOS neurons in hardware, this seems like a promising area in future when scaling of flash memory or floating gates become limited.

One of the important aspects of a circuit implementation of a learning rule is the ease with which its parameters can be tuned. We showed in Section 4.2.2 how the parameter A_3^+ can be tuned based on ΔV_d . Other learning rule parameters are also directly related to parameters of the control voltage waveforms. For example, τ_+ can be modified using T_g , τ_- using T_{tun} and A_2^- using T_{tun_pulse} and V_{tun_max} . This is evident from the mathematical analysis of section 4.2. Similarly, from equation (Eq. 4.33) and equation (Eq. 4.34), A_2^+ , A_3^+ and τ_y can be tuned using ΔV_d^s or ΔV_d^d , keeping two of them constant at a time. Some results for the variation of learning window to different parameters can be seen in our previous chapter. Also, other minor changes to the learning rule can be done by modifying the circuits at the neuron and periphery that generate the gate, drain and tunnel waveforms.

After this proof of concept, we will continue the work by simulating a spiking neural network (SNN) using SPICE simulations to understand the difference between T-STDP and D-STDP learning rules and then extend our work in future by fabricating chips with thousands of neurons and millions of learning synapses to do tasks like rapid and robust pattern recognition [87, 88]. Such neuromorphic chips that incorporate noise and heterogeneity are useful to understand the principles used by our brain to compute using imprecise elements [89, 21] as well as for accelerated simulations of neural networks [77, 20]. Moreover, we hope to use neuromorphic systems as the “brain” for real-time behaving systems like robots [89] where both low-power dissipation and real-time operation are necessary. In these cases, using a traditional computer for the implementation is inefficient due to the mismatch between Von-Neumann computing model of digital computers and the massively parallel analog computing of the brain where memory and computing are closely intermixed [90]. Hence, it is useful to be able to mimic biological neural networks closely in circuits to enable experimental

paradigms as well as low-power intelligence.

Chapter 5

Comparison of Triplet-STDP with Doublet-STDP: A simulation study

Synaptic plasticity rules are modelled for the formation of memory or the development of neuronal selectivity. There are different variants of these rules based on rate and temporal aspects of synaptic spikes. Previous chapters looked into both D-STDP and T-STDP rules in detail and their VLSI implementations using floating-gate transistors. Computationally T-STDP has advantages like replicating rate based plasticity experiments and detecting input correlations higher than second order. This chapter explores the advantages of T-STDP model compared to D-STDP in a spiking neural network from several benchmark experiments. The results include MATLAB and SPICE simulations of experiments to compare T-STDP model with D-STDP model.

5.1 Introduction

T-STDP has been shown to correlate better with experiments than D-STDP, as shown in earlier chapter. Apart from these experimentally observed benefits of T-STDP model, there are some computational advantages as well. It has been shown that T-STDP can detect input correlations higher than the second order ones to which D-STDP is sensitive [75]. In [91], it has been shown that, neurons tune to the earliest spike through STDP (used only D-STDP). It is shown that STDP is able to solve a very difficult computational problem: to localize a repeating spatio-temporal spike pattern embedded in equally dense distractor spike trains. The same difficult computational problem is extended with multiple STDP neurons and multiple repeating patterns in continuous spike trains [92]. There are more experiments designed to exploit the computational ability of D-STDP [93]. Here, we have taken two experiments, each from above papers,

- (i) Single Wave Experiment (SWE) [91]: implemented using both T-STDP and D-STDP to prove the advantage of T-STDP
- (ii) Third Order Correlation (TOC) experiment [75]: implemented in both MATLAB and SPICE.

Intuitively, early convergence of weight or quick learning is expected in the case of T-STDP model. Here we prove the computational advantage of T-STDP model over D-STDP model and we implement the TOC experiment using SPICE simulations, which is one step closer to hardware realization.

This chapter is an extension of our previous work [83, 84]. This chapter is organized as follows: Section 5.2 provides a brief explanation of Neuron and Synapse models. Section 5.3 implements the classic STDP experiments using D-STDP as

well as T-STDP. A simple third order correlation experiment is proved both in MATLAB simulations and SPICE simulations in section 5.4. Finally the chapter is concluded with discussions on consolidated work.

5.2 Model and Methods

5.2.1 Leaky Integrate and Fire (LIF) Neuron model

Neuron sums the weighted exponentially decaying excitatory postsynaptic current (EPSC) due to each incoming spike. When reaching the threshold of neuron, the postsynaptic neuron fires. This will trigger a time based learning rule for its first action potential, and then sets its potential back to a resting level of 0. Refractory period of 2 ms is used. The EPSC as a result of an input spike is given by

$$I_{syn} = I_0 e^{\left(\frac{-t}{\tau_{syn}}\right)} \quad (\text{Eq. 5.1})$$

where τ_{syn} is the decay constant of the synaptic current. I_0 is the peak value of the weighted synaptic current.

The membrane voltage V_{mem} dynamics of the neuron [94, 95] is modelled as

$$\tau_{neu} \frac{dV_{mem}}{dt} = \frac{\tau_{neu} I_{syn}}{C} - V_{mem} \quad (\text{Eq. 5.2})$$

where τ_{neu} is the decay constant of the neuron membrane potential. C is the neuron capacitance. The membrane voltage trace in a random time window is shown in Fig.5.2(c).

5.2.2 Synaptic Modification Rule

In biology, synapses are specialized structures that permit the transfer of signals between two neurons with an associated synaptic weight. Learning typically implies the modification of synaptic weight due to the activities of the pre- and post-synaptic neurons. The STDP models are explained in the previous chapters [64, 84, 63, 83].

5.2.3 Parameter Extraction

Figure 5.1: **Parameter Extraction: T-STDP experiment and T-STDP models on quadruplets of spikes:** The T-STDP floating gate model and mathematical model together with the T-STDP floating gate chip measurement results to show an instance of parameter extraction. Blue line models mathematical model of T-STDP, red line is for T-STDP floating gate chip measurement result and black line for T-STDP floating gate model.

The important parameters to be extracted from floating gate MOSFET are V_{inj} during potentiation part and V_{ox} during depotentiation part of the chip measurement. The procedure to extract V_{inj} is (1). Fix the voltage levels of floating gate

tunnelling terminal, V_{tun} and gate terminal, V_g to avoid any tunnelling in the floating gate transistor. (2). Then floating gate MOSFET current, I_d is measured for different values of V_d at drain terminal. (3). From I_d , we can get the change in floating gate voltage, ΔV_{fg} (Eq. 3.6) and this ΔV_{fg} is used to get I_{inj} (Eq. 3.7) and is used in Eq. 3.2 to extract V_{inj} parameter. Similarly, the procedure to extract V_{ox} is (1). Fix the voltage levels of floating gate drain terminal, V_d and gate terminal, V_g to avoid any injection in the floating gate transistor. (2). Then floating gate transistor current, I_d is measured for different values of V_{tun} at tunnelling terminal. (3). From I_d , we can get the change in floating gate voltage, ΔV_{fg} and this ΔV_{fg} is used to get I_{tun} and is used in Eq. 3.2 to extract V_{tun} .

We have shown an example of using these extracted parameters in Fig.5.1. It contains the data of measurement result (red line) together with the T-STDP floating gate model (black line) and T-STDP mathematical model (blue line) for quadruplets of spikes, where the floating gate model uses these extracted parameters. For the SPICE simulations of third order correlation experiment in this chapter, these extracted parameters are used in the behavioural model along with the peripheral programming circuits.

The chip measurement results in Fig.5.1 shows some non-ideal effects with respect to the floating gate model. This could be due to the non-idealities in the experimental setup. We are using a picoammeter for measuring current in the lower range of pico ampere. There can be non-idealities from picoammeter measurements. The mathematical model and the floating gate model results also show some difference. These could be due to the inherent dependence of change in synaptic weight (Floating Gate voltage) on the excitatory postsynaptic current (Floating Gate drain current). Whereas, this dependence is not present in T-STDP mathematical model. These differences will also affect the measurement and the theoretical results apart from other hardware non-idealities as shown in measurement results in previous chapters.

5.3 Simulation Results: MATLAB

5.3.1 Single Wave Experiment (SWE): Mathematical Model

A simple version of the single wave experiment illustrated in [91] is replicated here. Single wave experiment, as the name suggests, consists of many afferent neurons but with only one input presynaptic spike within an epoch. These afferent neurons are connected to a single output neuron through STDP synapses. The output neuron follows integrate and fire neuron model. The MATLAB simulation results for a single postsynaptic neuron with D-STDP and T-STDP learning rule and 10 afferent neurons are shown in Fig.5.2. Here, we are not considering the effects of jitter and spontaneous activities in the input spike samples. For the triplet and doublet STDP model, the parameters used are those of minimal model [29] i.e. $A_2^+ = 1 \times 10^{-3}$, $A_2^- = -1 \times 10^{-3}$, $A_3^+ = 5 \times 10^{-3}$, $\tau_+ = 20ms$, $\tau_- = 22ms$, $\tau_y = 20ms$.

Observation: Each afferent neuron carry a spike into the postsynaptic neuron within a time duration of 50ms as shown in Fig.5.2(a). Initial iteration produces just a single postsynaptic spike. After each presentation of the input spikes the postsynaptic latency start to fall down. For some starting number of presentations, roughly about 100 iterations, it can be seen that D-STDP and T-STDP follows the same path, as there is only one postsynaptic activity at the output neuron. Once the number of postsynaptic spikes start to increase, the T-STDP rule will increase the change in synaptic weight compared to D-STDP rule. This will make the T-STDP rule to converge faster than D-STDP rule. That is, as shown in Fig.5.2(b), the postsynaptic latency curve of T-STDP falls down steeper when more weight increase happens compared to the postsynaptic latency curve of D-STDP model. Sum of weights of all the synapses in the network gives an understanding of convergence of weight as the weights are bounded between maximum of 1 and minimum of 0, this is shown in Fig.5.2(d). Fig.5.2(e) and

Figure 5.2: **Single wave experiment:** (a) Raster plot of 10 input afferent neurons and its spike count is shown. (b) Postsynaptic latency vs presentation shows the number of presentation needed for the postsynaptic neuron to converge. T-STDP converges faster. (c) Membrane voltage trace during a random time window for the cases of D-STDP and T-STDP rules. (d) shows the sum of weight, which gives a gist of convergence of synaptic weight. (e) & (f) shows the evolution of weights in the case of D-STDP and T-STDP rules. Note that the afferent neurons are arranged in the increasing order of latency.

(f) shows the dynamics of weight change of each synapses connected between afferent neuron and output neuron after each iteration. Note that the afferent neurons are arranged in the ascending order of input spike time latency for better clarity. That is, afferent neuron 6 receives the first input spike in an epoch, whereas afferent neuron 9 receives the last input spike in an epoch, thus synapses are arranged in the order of spike arrival, from bottom to top. The result gives a better visualization of early convergence of weight of synapses which implies that the respective neurons tune to the earliest input spikes. The neurons at bottom tune to the earliest input spikes and their respective synapses converge to the maximum value of weight (coloured in yellow), whereas the neurons at the top tune to the delayed input spikes and their respective synapses converge to the minimum value of weight (coloured in blue). The number of neurons to be in yellow coloured or blue coloured zones are determined by the point of convergence of postsynaptic latency.

Conclusion: The plot for the postsynaptic latency over number of input presentations shows a good result indicating the advantage of faster convergence of T-STDP rule over D-STDP rule. Fig.5.2(e) and (f), convergence of synaptic weights, clearly shows that the neuron tune to the earliest input spikes through D-STDP and T-STDP respectively. Note that the afferent neurons are arranged in the order of input spike latency for better clarity.

5.3.2 Single Wave Experiment (SWE): Floating Gate Model

Experiment 1: Along with building any experiment using mathematical model of D-STDP and T-STDP, it is equally important to build such an experiment using floating gate model to evaluate the non-ideal effects of the learning system in hardware. This software learning system using floating gate model can be used to evaluate the effectiveness of the developed plasticity rule, in terms of accuracy and complexity. Hence,

Figure 5.3: **Postsynaptic Latency of SWE: Floating gate model:** (a) Three different cases are shown. Green line models mathematical model of T-STDP, red line is for T-STDP ideal Vd case of floating gate model and black line for T-STDP non-ideal Vd case of floating gate model. (b) The postsynaptic latency for process variations among synapses.

in this subsection, we do the same for the SWE described earlier. In Fig.4.12(a), it can be seen that there is a non-ideality or a difference between the generation of single pulsed drain voltage waveform electronically compared to theoretical model. This implies that the proposed circuit of single pulsed drain voltage generator is an approximate implementation of Eq. 4.33. Thus, for the purpose of evaluation of effectiveness, we have considered two different single pulsed drain voltage waveform for the case of T-STDP model, one is modelled using the mathematical model as in Eq. 4.33 – named as ‘FG ideal Vd’ and another is modelled using the circuit implementation as in Eq. 4.37 – named as ‘FG non-ideal Vd’. The parameters of floating gate model is obtained by matching the potentiation and depotentiation of the mathematical model as mentioned in the previous chapters. The mathematical T-STDP model of SWE result is compared with ‘FG ideal Vd’ and ‘FG non-ideal Vd’ T-STDP models to verify the effectiveness of

the learning system in Fig.5.3(a).

Experiment 2: We have added some process variations among different synapses to understand the optimization of learning. Variables V_{inj} and V_{ox} are chosen for the analysis of process variations. The values of these variables are randomly varied among different synapses. The values to be chosen are distributed in Gaussian with their nominal values as mean and deviated by a maximum of 5%. Fig.5.3(b) shows the result of the learning system, the convergence of postsynaptic latencies over different epochs. Here, we have reduced the number of presentations during each simulation, as we have done the experiment of process variations around 10 times and all the results are lying around the plot of mathematical model which is shown as thick green line.

For the floating gate model, the parameters used are $V_{inj} = 250 \times 10^{-3}$, $V_{ox} = 150 \times 10^{-3}$, $T_g = 100\text{ms}$, $T_{tun} = 1500\text{ms}$, $T_{tun_pulse} = 2\text{ms}$, $T_{tun_delay} = 1\text{ms}$, $T_d = 500\text{us}$, $V_{tun_max} = 16.3\text{V}$, $V_{tun_min} = 5.4\text{V}$, $\Delta V_g = 0.8\text{V}$, $V_{d_max} = 5\text{V}$ and $V_{d_min} = 0.3\text{V}$.

Observation: Results of all the models are converging to the same postsynaptic latency but takes different paths depending on the change in weight during each epochs or iterations. The weight change during both trajectories of red and black lines differ due to the difference in drain voltage generation hence, there will be a slight change in the trajectories of red line and black line as in Fig.5.3. As mentioned in previous section, the steeper the postsynaptic latency curve implies that the change in weight increases faster in each iterations compared to a less steeper curve. It can be seen that T-STDP FG non-ideal Vd curve (black curve) converges little faster or more steeper than the T-STDP FG ideal Vd curve (red curve). This is in fact true that the single pulsed drain voltage generated by circuit implementation in Fig.4.12(a) induces more potentiation during each iteration compared to the theoretical implementation of single pulsed drain voltage, which results in more steepness. Another observation from Fig.5.3(a) is that the T-STDP mathematical model (green line) is following

a slightly different trajectory compared to the T-STDP Floating Gate models (red and black lines). There are two reasons for this as mentioned below.

- (i) In the mathematical model of learning system, the exponentially decaying excitatory postsynaptic current (EPSC) is multiplied with the respective synaptic weight and these products from all the synapses are integrated in the output neuron. Whereas, in the floating gate model of learning system, the multiplication is embedded along with the generation of EPSC by floating gate synapse. The EPSC generated in the case of floating gate model is slightly different compared to the mathematical model.
- (ii) Another major difference in floating gate model and mathematical model of learning system is the inherent dependence of change in synaptic weight (Floating Gate voltage) on the excitatory postsynaptic current (Floating Gate drain current).

Thus, making floating gate model an approximation to the mathematical model. The final value or the convergence point of postsynaptic latency is important for the accuracy of learning system whereas trajectory shows the speed in reaching the point of convergence. Hence, from Fig.5.3, it can be seen that these differences as mentioned above do not affect the learning system significantly.

The observation from experiment 2 is almost same as in experiment 1. The only difference is that the result shows more plots of T-STDP FG models from different simulations. Fig.5.3(b) shows that the T-STDP FG models take different paths to reach the convergence point of postsynaptic latency. In some simulations, the trajectory taken is steeper (red and black line converges faster than green line) whereas in other simulations it is much slower (red and black line converges slower than green line). This is due to the addition of process variations among synapses. These process variations or variations in $V_{in,j}$ and V_{ox} changes the amount of potentiation and

depotential among synapses and affects the change of weight in each epoch. This is seen as steeper trajectory of red and black lines in Fig.5.3(b), when the amount of potentiation is more than the nominal amount as in mathematical model in each epoch. Whereas less steep in trajectory implies less potentiation in each epoch.

Conclusion: The plot for the postsynaptic latency over number of input presentations shows a good result indicating the effectiveness of the developed plasticity rule in terms of accuracy. This result also depicts the non-ideality in the circuit implementation of the drain voltage waveform generator in a learning system. The results of experiment 2 shows that the learning system converges properly even though there are process variations among the synapses in the system. This implies that the process variations among different synapses does not impede the convergence or optimization of learning.

5.3.3 Third Order Correlation (TOC) Experiment

We have replicated exactly the same simulations illustrated in [75]. The feedforward network consists of six afferent neurons connected to a single postsynaptic output neuron. Two correlation-based patterns are presented to the network, which contains input of same firing rates and same pairwise correlations, but differed in the presence or absence of third-order correlations in half of the inputs (refer to Fig.5.4(a)). First pattern contains third-order correlations in inputs 1 to 3 (mentioned as group 1) and no third-order correlations in inputs 4 to 6 (mentioned as group 2). Second pattern contains third-order correlations in inputs 4 to 6 of group 2 and no third-order correlations in inputs 1 to 3 of group 1. These patterns are presented to the network with a fixed probability; that is out of 10 pattern presentations, pattern 1 was presented on average eight times (with probability 0.8). Three of the afferent neuron contain only 20% of third order correlation and another three afferent neurons contain 80%

Figure 5.4: **Third order correlation experiment:** (a) Raster plot of 6 input afferent neurons and the postsynaptic spikes. (b) EPSC, membrane potential and postsynaptic spikes (c) shows the evolution of weights in the case of T-STDP. (d) shows the evolution of weights in the case of D-STDP.

of third order correlation as shown in raster plot of spikes (Fig.5.4(a)). LIF model of neuron is used with minimum triplet model of STDP. EPSC, membrane potential and postsynaptic spikes are shown for the first presentation of input spikes in Fig.5.4(b).

Here for the triplet model, the parameters used are those of minimal model [29] i.e. $A_2^+ = 0$, $A_2^- = -6.5 \times 10^{-3}$, $A_3^+ = 7.1 \times 10^{-3}$, $\tau_+ = 16.8ms$, $\tau_- = 33.7ms$, $\tau_y = 114ms$

Observation: First inference is that the evolution of weights of group 1 synapses with higher percentage of input correlation patterns tend to the maximum weight whereas the group 2 synapses with lesser percentage of input correlation patterns tend to minimum weight. Fig.5.4(c) shows that the group 1 neurons, i.e. afferent neurons 1 to 3 converge to the maximum weight and is coloured in yellow, whereas group 2 neurons, i.e. neurons 4 to 6 converge to the minimum weight and is coloured in blue. Fig.5.4(d) shows that the evolution of weights of group 1 as well as group 2 tend to the minimum weight, which is coloured in blue.

Conclusion: The result depicts that the postsynaptic neuron with T-STDP synapses tend to capture the third order correlations in input, whereas postsynaptic neuron with D-STDP synapses cannot capture the inputs with third order correlation. After convergence, the postsynaptic neuron will respond to those input patterns with third order correlations from group 1 afferent neurons.

5.4 Simulation Results: SPICE

5.4.1 Third Order Correlation (TOC) Experiment

In this section, we show SPICE simulations of circuits that integrate few synapse circuits with the circuit of an integrate and fire neuron, bringing us a step closer to chip realizations of the same. A simple neural network with six synapses connected to a neuron is shown in Fig.5.5, same as the one used in the MATLAB simulation. The floating gate synapse is shown in Fig.5.6. The simulation model of floating gate transistor is explained in [96, 97, 98]. The quantum mechanisms - hot electron injection and fowler nordheim tunneling is modelled using VERILOGA. The voltage waveform generators for gate terminal, tunneling terminal and drain terminal used in the interface

Figure 5.5: **Neural network consisting of 6 synapses and a neuron** The input and output spikes are shown.

circuitry for floating gate synapse are discussed in previous chapters. Double pulsed drain voltage waveform generator is used for the SPICE simulations. The circuit implementation of the integrate and fire neuron model as explained in section 5.2.1 is shown in Fig.5.7 [99]. C_{mem} integrates the incoming current to create a membrane potential. The membrane threshold voltage of the neuron is set by the threshold of the inverter. If C_{mem} voltage reaches the threshold, the neuron fires an output pulse. The positive feedback is created by M1-M5 transistors, which pulls the membrane voltage quickly to high potential. It reduces the short circuit current in the inverter, thus reducing power dissipation. After a spike is fired, REQ signal indicates spike emission, whereas ACK signal is received to reset the membrane potential to ground potential

through M_{rst} . The refractory period is set in the circuit using voltage, V_{ref} and the starved inverter marked in red. I_{syn} is the excitatory postsynaptic current input to the neuron.

Figure 5.6: **Floating gate synapse model:** The injection and tunneling blocks are modelled in VERILOGA. The interface circuitry is as explained in the previous chapters.

A similar experimental framework explained in section 5.3.3 is also used for the SPICE simulations. The maximum values of potentiation and depression is not exactly the same as in MATLAB simulation. Change in synaptic weight is directly propor-

Figure 5.7: **Integrate and Fire Neuron model:** Consists of a capacitor for integration (green), a positive feedback loop (purple) for spike generation and a refractory circuit (red) for setting refractory period of neuron and a reset (black) transistor for discharging the membrane capacitor to ground. A typical membrane voltage trace is shown in inset.

tional to the change in the floating gate voltage (Eq. 3.5) [55]. Hence, synaptic weights can be directly obtained from measuring the floating gate voltage of each synapse. 6 different floating gate voltages are shown in Fig.5.8. For the floating gate model, the parameters used are $V_{inj} = 250 \times 10^{-3}$, $V_{ox} = 150 \times 10^{-3}$, $T_g = 100 \text{ms}$, $T_{tun} = 300 \text{ms}$,

$T_{tun_pulse} = 2\text{ms}$, $T_{tun_delay} = 1\text{ms}$, $T_d = 500\text{us}$, $V_{tun_max} = 16.5\text{V}$, $V_{tun_min} = 5.6\text{V}$, $\Delta V_g = 0.8\text{V}$ and $V_{d_max} = 5\text{V}$. These parameters are extracted from the measurement results in the previous chapter as shown in section 5.2. The T-STDP floating gate SPICE model is generated by the floating gate MOSFET synapse using the VERILOGA model for the quantum mechanics along with the voltage generators to generate the floating gate peripheral programming voltages with above parameters.

Observation: The SPICE result matches exactly to the MATLAB simulation result. Fig.5.8 shows that the postsynaptic neuron with T-STDP synapses tend to capture the input third order correlations. Decrease in floating gate voltage implies more synaptic weight. Thus evolution of weights of first three synapses tend to the maximum weight whereas the last three synapses with lesser percentage of correlation tend to minimum weight.

5.5 Discussion

Computational advantage of T-STDP such as detecting third order input correlations and faster convergence of synaptic weights are shown here. Advantages are proved using MATLAB simulations. The effectiveness of the developed plasticity rule in terms of accuracy and complexity is evaluated using the difference between the generation of drain voltage theoretically and electronically in single wave experiment. The process variations among different synapses do not impede the convergence or optimization of learning. Modelling floating-gate synapses connected to an integrate and fire neuron using SPICE brings us closer to chip realizations of learning neural networks. The third order correlation experiment is satisfactorily proved both in MATLAB as well as SPICE simulations. Future work is to fabricate a chip with the similar neural network topology and to test the results obtained.

Figure 5.8: **Third Order Correlation Spice simulation:** The floating gate voltages of 6 synapses are shown. The first three synapses have maximum third order input correlation (green waveforms) whereas the last three are least correlated (black waveforms).

Chapter 6

Conclusion and Future Work

6.1 Conclusion

Synaptic plasticity is a major area of research in neuromorphic engineering. Several plasticity rules had been emerged from experimental understanding of the brain. Among them spike timing dependent plasticity became very popular. In this work, different aspects of spike timing dependent plasticity are well understood and a floating-gate approach is taken for the VLSI implementation of plasticity rule.

Chapter 2 contains different plasticity rules, different synapse circuits and their implementation in silicon. The neuromorphic challenges during implementation of different synapse circuits on silicon is portrayed by comparing different synapse circuits against different metrics like plasticity rule, synapse area and precision of weight storage. Chapter 2 is completed after the comparison, giving the advantage of usage of floating gate synapse.

Chapter 3 presented a robust STDP implementation using single FG transistor acting as a learning synapse for VLSI spiking neural networks. Compared to

earlier work in [55], our proposed method allows the STDP curve of the FG device to be similar to the biological one over a much wider range of parameters. This is achieved by simply modifying the gate voltage waveform—hence, the hardware overhead compared to [55] is minimal. A mathematical model for learning is presented to illustrate the benefit of our proposed method. We also present measurement results from a $0.35\mu m$ chip to justify the claim. Finally, a circuit implementation of the modified gate voltage generator is also shown.

Simulation and hardware implementation of triplet STDP (T-STDP) on a floating gate device is presented in chapter 4. At lower frequencies (1Hz - 5Hz) of presynaptic and postsynaptic spike pairings, the doublet STDP rule shows potentiation and depression correctly. However at higher frequencies ($>25\text{Hz}$) of pre-post pairings, the potentiation reduces, this is because in addition to potentiation within pre-post pairs there is depression due to the interaction of pairs itself, i.e. postsynaptic spike of the first pair interact with the presynaptic spike of the second pair to cause depression. This phenomena of reduction in potentiation at higher frequencies of pre-post synaptic spike pairs is against the observations made in experiments. Thus triplet STDP rule [29], where potentiation and depression happens with respect to spike triplets (1 presynaptic and 2 postsynaptic spikes or 2 presynaptic and 1 postsynaptic spikes) instead of spike doublets, came into being. Thus implementation of triplet STDP rule on a floating gate device is important compared to doublet STDP.

We have presented a spike triplet based learning rule using a single FG transistor as the synapse for VLSI spiking neural networks. The spike triplet affects the setting of drain voltage—we presented a single pulse and a double pulse drain voltage method to obtain the desired dependence of weight on spike timing. We presented a method to calculate the parameters of the drain voltage pulse to obtain results matched to the original theoretical T-STDP rule. We also show FG measurement results in comparison with the biological experimental observations for (1) original

doublet protocol, (2) two protocols of spike triplets, (3) frequency effects of pairing protocol, (4) quadruplet experiments and (5) replication of BCM rule. The failure of FG D-STDP rule in replicating the biological results is also included. Possible hardware implementations of drain voltage waveform generator are also proposed and verified through SPICE simulation results. It was shown that the voltage waveform for double pulse case can be generated more accurately due to its simplistic nature.

Apart from the experimental advantages of T-STDP over D-STDP, computationally T-STDP has advantages like replicating rate based plasticity experiments, detecting third order input correlations and faster convergence of synaptic weights. Chapter 5 shows a proof of these advantages using MATLAB simulations. Further, we show SPICE simulations of a simple network with FG synapse and integrate and fire neuron circuits. The third order correlation experiment is simulated in SPICE and shows similar results as those obtained in MATLAB. This gives us confidence about the correct operation of future chips with networks of neurons and synapses.

The next section talks about the work to be done in the future.

6.2 Future Work

Similar to the categorization of work done, the future work can be categorized into the following:

- (i) Explore T-STDP implementation without using multiplexer in each cell.

Even though implementation of T-STDP with multiplexer is easier to control the weight modification, multiplexer produces hardware overhead. The number of synapse circuits outnumber the number of neurons in a neural network architecture. As these multiplexers are used with each and every floating-gate transistor

synapses irrespective of the peripheral interface circuitry for voltage waveforms, these multiplexers produce sufficient hardware overhead. The percentage hardware area overhead per synapse is around 175.5%. Thus, an important avenue for future research is to explore the implementation of plasticity rules without using multiplexer in every synapse.

- (ii) Implementation of plasticity rule on network of synapses and neurons to generate a self learning system.

Once it is well understood how to implement certain plasticity rule (D-STDP or T-STDP) for synapses, it is equally important to develop a self learning system out of several such plastic synapses. Hence, it is important to next develop a software simulation of a simple neural networks with floating-gate synapses trained by D-STDP and T-STDP. Possible simple simulations would be single and multi-layer perceptrons for classifying spike patterns such as the ones obtained from spiking retinas [3].

- (iii) Fabricate IC with an array of FG device with T-STDP and spiking neurons.

Once the software framework is verified, the aim is to implement the network architecture in a chip. This chip could be then used to demonstrate several classification or pattern recognition tasks in real time.

- (iv) Explore the implementation of synaptic plasticity rules using emerging nano-technology devices

Similar to memristors there are other nano-technology devices like Resistive Random Access Memories (R-RAMs), Spin-Transfer Torque Magnetic Random Access Memories (STT-MRAMs) and Phase Change Memories (PCM) used for the implementation of artificial synapses. Integration of these nano-technology devices to CMOS is a challenge. The advantages should still stand when using implementation devices other than Floating Gates. However, one must still

understand and characterize the device physics of the implementation devices to develop the synaptic plasticity rule. The procedure for weight update as well as the hardware interface circuitry will change with any of the implementation devices.

Publication

Conference Papers

- (i) Shaista Hussain, Roshan Gopalakrishnan, Arindam Basu and Shih-Chii Liu, "Morphological Learning: Increased Memory Capacity of Neuromorphic Systems with Binary Synapses Exploiting AER Based Reconfiguration" accepted at *IEEE International Joint Conference on Neural Networks (IJCNN 2013)*, Dallas, TX, USA 2013.
- (ii) Roshan Gopalakrishnan and Arindam Basu, "Robust Doublet STDP in a Floating-Gate Synapse" accepted at *IEEE International Joint Conference on Neural Networks (IJCNN 2014)*, Beijing, China 2014.
- (iii) Roshan Gopalakrishnan and Arindam Basu, "Triplet Spike Time-Dependent Plasticity in a Floating-Gate Synapse" accepted at *IEEE International Symposium on Circuits and Systems (ISCAS 2015)*, Lisbon, Portugal 2015.

Journal Papers

- (i) Roshan Gopalakrishnan and Arindam Basu, "On the "non-STDP" Behavior and

its Remedy in a Floating-Gate Synapse” accepted in *IEEE Transactions on Neural Networks and Learning Systems* (TNNLS 2015).

- (ii) Roshan Gopalakrishnan and Arindam Basu, “Triplet Spike Time-Dependent Plasticity in a Floating-Gate Synapse” accepted in *IEEE Transactions on Neural Networks and Learning Systems* (TNNLS 2016).

Bibliography

- [1] G Perea, M Navarrete, and A Araque. Tripartite synapses: astrocytes process and control synaptic information. *Trends in Neurosciences*, 32:421–431, Aug. 2009.
- [2] S. G. Tewari and K. K. Majumdar. A mathematical model of the tripartite synapse: astrocyte-induced synaptic plasticity. *Journal of Biological Physics*, 38:465–496, Jun. 2012.
- [3] Garrick Orchard, Ajinkya Jayawant, Gregory K. Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in Neuroscience*, 9:437, 2015.
- [4] Mostafa Rahimi Azghadi, Nicolangelo Iannella, Said F Al-Sarawi, Giacomo Indiveri, and Derek Abbott. Spike-Based Synaptic Plasticity in Silicon: Design, Implementation, Application, and Challenges. *Proceedings of the IEEE*, pages 1–21, 2014.
- [5] B. A. Minch P. Hasler, C. Diorio and C. Mead. Single transistor learning synapses with long term storage. volume 3, pages 1660–1663, 1995.
- [6] https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html.
- [7] <https://en.wikipedia.org/wiki/Neuron>.
- [8] <https://en.wikipedia.org/wiki/Synapse>.
- [9] David H Goldberg, Gert Cauwenberghs, and Andreas G Andreou. Probabilistic

- synaptic weighting in a reconfigurable network of VLSI integrate-and-fire neurons. *Neural Networks*, 14(6):781–793, 2001.
- [10] https://en.wikipedia.org/wiki/Artificial_neural_network.
- [11] Ankur Gupta. *Efficient and scalable biologically plausible spiking neural networks with learning applied to vision*. PhD thesis, The Pennsylvania State University, Pennsylvania, USA, 2010.
- [12] J. Vreeken. Spiking neural networks, an introduction. Technical Report UU-CS-2003-008, Institute for Information and Computing Sciences, Utrecht University, 2002.
- [13] Wulfram Gerstner and Werner M Kistler. *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge University Press, 2002.
- [14] I. Aybay, S. Cetinkaya, and U. Halici. Classification of neural network hardware. *Neural Network World*, 6(1):11–29, 1996.
- [15] M. Mahowald. *VLSI analogs of neuronal visual processing: A synthesis of form and function*. PhD thesis, California Inst. Technol., Pasadena, CA, 1992.
- [16] M. Sivilotti. *Wiring considerations in analog VLSI systems, with application to field-programmable networks*. PhD thesis, California Inst. Technol., Pasadena, CA, 1991.
- [17] M. Mahowald. *An Analog VLSI Stereoscopic Vision System*. Norwell, MA: Kluwer, 1994.
- [18] K. A. Boahen. Point-to-point connectivity between neuromorphic chips using address events. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 47(5):416–434, May 2000.
- [19] J. Schemmel, D. Briiderle, A. Griibl, M. Hock, K. Meier, and S. Millner. A wafer-scale neuromorphic hardware system for large-scale neural modeling. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pages 1947–1950, May 2010.
- [20] E. Painkras, L. A. Plana, J. Garside, S. Temple, F. Galluppi, C. Patterson, D. R.

- Lester, A. D. Brown, and S. B. Furber. SpiNNaker: A 1-W 18-core system-on-chip for massively-parallel neural network simulation. *IEEE Journal of Solid-State Circuits*, 102(5):699–716, Apr. 2013.
- [21] B. V. Benjamin, P. Gao, E. McQuinn, S. Choudhary, A. R. Chandrasekaran, J-M. Bussat, R. Alvarez-Icaza, J. V. Arthur, P. A. Merolla, and K. Boahen. Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations. *Proceedings of the IEEE*, 48(8):1943–1953, Aug. 2014.
- [22] J. s. Seo, B. Brezzo, Y. Liu, B. D. Parker, S. K. Esser, R. K. Montoye, B. Rajendran, J. A. Tierno, L. Chang, D. S. Modha, and D. J. Friedman. A 45nm cmos neuromorphic chip with a scalable architecture for learning in networks of spiking neurons. In *2011 IEEE Custom Integrated Circuits Conference (CICC)*, pages 1–4, Sept 2011.
- [23] J. s. Seo, B. Brezzo, Y. Liu, B. D. Parker, S. K. Esser, R. K. Montoye, B. Rajendran, J. A. Tierno, L. Chang, D. S. Modha, and D. J. Friedman. A 45nm cmos neuromorphic chip with a scalable architecture for learning in networks of spiking neurons. In *2011 IEEE Custom Integrated Circuits Conference (CICC)*, pages 1–4, Sept 2011.
- [24] Jose M Cruz-Albrecht, Timothy Derosier, and Narayan Srinivasa. A scalable neural chip with synaptic electronics using cmos integrated memristors. *Nanotechnology*, 24(38):384011, 2013.
- [25] D O Hebb. *The Organization of Behavior: A Neuropsychological Study*. Wiley, 1949.
- [26] L. F. Abbott and S. B. Nelson. Synaptic plasticity: taming the beast. *Nature neuroscience*, 3:1178–1183, Nov. 2000.
- [27] W.Gerstner H.Markram and P. J. Sjöström. Spike-timing-dependent plasticity: A comprehensive overview. *Frontiers in Synaptic Neuroscience*, 4, 2012.
- [28] K D Miller S Song and L F Abbott. Competitive Hebbian learning through spike-timing-dependent synaptic plasticity. *Nature neuroscience*, 3:919–926, 2000.

- [29] J.-P. Pfister and W. Gerstner. Triplets of spikes in a model of spike timing-dependent plasticity. *The Journal of Neuroscience*, 26(38):9673–9682, Sep. 2006.
- [30] Walter Senn Joseph M Brader and Stefano Fusi. Learning real-world stimuli in a neural network with spike-driven synaptic dynamics. *Neural Computation*, 19:2881–2912, 2007.
- [31] Eleni Vasilaki Claudia Clopath, Lars Busing and Wulfram Gerstner. Connectivity reflects Coding: A Model of Voltage-based Spike-Timing-Dependent-Plasticity with Homeostasis. *Nature neuroscience*, 13:344–352, 2010.
- [32] Johannes Partzsch Christian Mayr, Marko Noack and Rene Schuffny. Replicating experimental spike and rate based neural learning in CMOS. pages 105–108, 2010.
- [33] W. Senn, H. Markram, and M. Tsodyks. An Algorithm for Modifying Neurotransmitter Release Probability Based on Pre- and Post-Synaptic Spike Timing. *Neural Computation*, 13:35–67, 1999.
- [34] W. Senn. Beyond spike timing: the role of nonlinear plasticity and unreliable synapses. *Biological Cybernetics*, 87(5), 2002.
- [35] Harel Z. Shouval, Mark F. Bear, and Leon N. Cooper. A unified model of nmda receptor-dependent bidirectional synaptic plasticity. *Proceedings of the National Academy of Sciences*, 99(16):10831–10836, 2002.
- [36] Neel T. Shah, L. C. Yeung, Leon N. Cooper, Y. Cai, and Harel Z. Shouval. A biophysical basis for the inter-spike interaction of spike-timing-dependent plasticity. *Biological Cybernetics*, 95(2), 2006.
- [37] http://www.scholarpedia.org/article/Models_of_synaptic_plasticity.
- [38] R. Linsker. From basic network principles to neural architecture: emergence of spatial-opponent cells. *Proceedings of the National Academy of Sciences*, 83(19):75087512, Oct. 1986.
- [39] R. Kempter, W. Gerstner, and J. L. van Hemmen. Hebbian learning and spiking

- neurons. *Physical Review E*, 59:4498–4514, 1999.
- [40] L. N. Cooper E. L. Bienenstock and P. W. Munro. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 2(1):32–48, Jan. 1982.
- [41] M. R. Azghadi, O. Kavehei, S. Al-Sarawi, N. Iannella, and D. Abbott. Novel vlsi implementation for triplet-based spike-timing dependent plasticity. In *Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2011 Seventh International Conference on*, pages 158–162, Dec 2011.
- [42] G.-Q. Bi and M.-M. Poo. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *The Journal of Neuroscience*, 18(24):10464–10472, Dec. 1998.
- [43] Chiara Bartolozzi and Giacomo Indiveri. Synaptic dynamics in analog VLSI. *Neural computation*, 19:2581–2603, 2007.
- [44] Andrew Smith. *Synaptic weight modification and storage in hardware neural networks*. PhD thesis, University of Liverpool, 2012.
- [45] M Annunziato S Fusi and D Badoni. Spike-driven synaptic plasticity: theory, simulation, VLSI implementation. *Neural Computation*, 2000.
- [46] G. Indiveri, E. Chicca, and R. Douglas. A VLSI array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity. *IEEE Transactions on Neural Networks*, 17(1):211–221, Jan. 2006.
- [47] G. Indiveri T. Delbruck S.-C. Liu, J. Kramer and R. Douglas. *Analog VLSI: Circuits and Principles*. Cambridge Ma: MIT Press, 2002.
- [48] Pascal Suter Malte Boegerhausen and Shih-Chii Liu. Modeling short-term synaptic depression in silicon. *Neural computation*, 15:331–348, 2003.
- [49] Giacomo Indiveri Elisabetta Chicca and Roudney Douglas. An adaptive silicon synapse. volume 1, pages I–81–I–84, 2003.
- [50] S. Mitra, S. Fusi, and G. Indiveri. Real-time classification of complex patterns us-

- ing spike-based learning in neuromorphic vlsi. *IEEE Transactions on Biomedical Circuits and Systems*, 3(1):32–42, Feb. 2009.
- [51] J. V. Arthur and Kwabena Boahen. Learning in silicon: Timing is everything. *Advances in neural information processing systems*, pages 75–82, 2006.
- [52] J. Schemmel, A. Grubl, K. Meier, and E. Mueller. Implementing synaptic plasticity in a VLSI spiking neural network model. In *Proceedings of the International Joint Conference on Neural Networks*, pages 1–6, Vancouver, BC, Jul. 2006.
- [53] A. Bofill i Petit and A. F. Murray. Synchrony detection and amplification by silicon neurons with STDP synapses. *IEEE Transactions on Neural Networks*, 15(5):1296–1304, Sep. 2004.
- [54] M. R. Azghadi, S. Al-Sarawi, D. Abbott, and N. Iannella. A neuromorphic vlsi design for spike timing and rate based synaptic plasticity. *Neural Networks*, 45:70–82, 2013.
- [55] S. Ramakrishnan, P. Hasler, and C. Gordon. Floating gate synapses with spike-time-dependent plasticity. *IEEE Transactions on Biomedical Circuits and Systems*, 5(3):244–252, Jun. 2011.
- [56] S. Saghi, C. Mayr, T. Serrano-Gotarredona, H. Schmidte, G. Lecerf, J. Tomas, J. Grollier, S. Boyn, A. F. Vincent, D. Querlioz, S. La Barbera, F. Alibart, D. Vuillaume, O. Bichler, C. Gamrat, and B. Linares-Barranco.
- [57] J. Hasler and B. Marr. Finding a roadmap to achieve large neuromorphic hardware systems. *Frontiers in neuroscience*, 7(00118), Sep. 2013.
- [58] P. Hasler, S. Ramakrishnan, R. Wunderlich, A. Basu, S. Brink, S. Nease and B. Degnan. A learning-enabled neuron array ic based upon transistor channel models of biological phenomenon. *IEEE Transactions on Biomedical Circuits and Systems*, 7:71–81, Feb 2013.
- [59] S.A. Bamford, A.F. Murray, and D.J. Willshaw. Spike-timing-dependent plasticity with weight dependence evoked from physical constraints. *IEEE Transactions on Biomedical Circuits and Systems*, 6(4):385–398, Aug. 2012.

- [60] H. Tanaka, T. Morie, and K. Aihara. A CMOS spiking neural network circuit with symmetric/asymmetric STDP function. *IEICE Transactions on Fundamentals of Electronics Communications and Computer Sciences*, E92-A(7):1690–1698, Jul. 2009.
- [61] T. J. Koickal, A. Hamilton, T. S. Lim, J. A. Covington, J. W. Gardner, and T. C. Pearce. Analog VLSI circuit implementation of an adaptive neuromorphic olfaction chip. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 54(1):60–73, Jan. 2007.
- [62] J. Arthur and K. Boahen. Synchrony in silicon: The gamma rhythm. *IEEE Transactions on Neural Networks*, 18(6):1815–1825, Nov. 2007.
- [63] R. Gopalakrishnan and A. Basu. Robust doublet STDP in a floating-gate synapse. In *Proceedings of the International Joint Conference on Neural Networks*, pages 4296–4301, Beijing, China, Jul. 2014.
- [64] R. Gopalakrishnan and A. Basu. On the Non-STDP Behavior and Its Remedy in a Floating-Gate Synapse. *IEEE Transactions on Neural Networks and Learning Systems*, 26(99):2596–2601, Feb. 2015.
- [65] P. Hasler, A. Basu, and S. Koziol. Above threshold pfet injection modeling intended for programming floating-gate systems. In *IEEE International Symposium on Circuits and Systems*, pages 1557–1560, New Orleans, LA, USA, May. 2007.
- [66] A. Basu, S. Brink, C. Schlottmann, S. Ramakrishnan, C. Petre, S. Koziol, F. Baskaya, C. M. Twigg, and P. Hasler. A floating-gate based field programmable analog array. *IEEE Journal of Solid-State Circuits*, 45(9):1781–1794, Sep. 2010.
- [67] H. Markram, J. Lübke, M. Frotscher, and B. Sakmann. Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science*, 275(5297):213–215, 1997.
- [68] L. Zhang, H. Tao, C. Holt, W. Harris, and M. M. Poo. A critical window for cooperation and competition among developing retinotectal synapses. *Nature neuroscience*, 395:37–44, Sep. 1998.

- [69] G.-Q. Bi and H.-X. Wang. Temporal asymmetry in spike timing-dependent synaptic plasticity. *Physiology and Behavior*, 77:551–555, 2002.
- [70] R. Froemke and Y. Dan. Spike-timing-dependent synaptic modification induced by natural spike trains. *Nature*, 416:433–438, 2002.
- [71] H. Wang, R. Gerkin, D. Nauen, and G. Bi. Coactivation and timing-dependent integration of synaptic potentiation and depression. *Nature neuroscience*, 8:187–193, 2005.
- [72] R. Froemke, I. Tsay, M. Raad, J. Long, and Y. Dan. Contribution of individual spikes in burst-induced long-term synaptic modification. *Journal of Neurophysiology*, 95(3):1620–1629, 2006.
- [73] J.J. Wade, L.J. McDaid, J.A. Santos, and H.M. Sayers. SWAT: A Spiking Neural Network Training Algorithm for Classification Problems. *IEEE Transactions on Neural Networks*, 21(11):1817–1830, Nov. 2010.
- [74] E. Izhikevich and N. Desai. Relating STDP to BCM. *Neural Computation*, 15(7):1511–1523, 2003.
- [75] J. Gjorgjieva, C. Clopath, J. Audet, and J.-P. Pfister. A triplet spike-timing dependent plasticity model generalizes the Bienenstock-Cooper-Munro rule to higher-order spatiotemporal correlations. *Proceedings of the National Academy of Sciences*, 108(48):19383–19388, 2011.
- [76] C. Mayr and J. Partzsch. Rate and pulse based plasticity governed by local synaptic state variables. *Frontiers in Computational Neuroscience*, 2(33), 2010.
- [77] J. Wijekoon and P. Dudek. VLSI circuits implementing computational models of neocortical circuits. *Journal of Neuroscience Methods*, 210(1):93 – 109, Sep. 2012. Special Issue on Computational Neuroscience.
- [78] B. Linares-Barranco, T. Serrano-Gotarredona, A. L. Camuas-Mesa, A. J. Perez-Carrasco, C. Zamarreo-Ramos, and T. Masquelier. On spike-timing-dependent-plasticity, memristive devices, and building a self-learning visual cortex. *Frontiers in Neuroscience*, 5(26), 2011.

- [79] M. Pankaala, M. Laiho, and P. Hasler. Compact floating-gate learning array with STDP. In *Proceedings of the International Joint Conference on Neural Networks*, pages 2409–2415, Atlanta, GA, Jun. 2009.
- [80] A. M. Haas. Compact circuits and adaptation techniques for implementing adaptive neurons and synapses with spike timing dependent plasticity (STDP). *Patent US20090292661*, Nov. 26, 2009.
- [81] S.-C. Liu and R. Mockel. Temporally learning floating-gate VLSI synapses. In *IEEE International Symposium on Circuits and Systems*, pages 2154–57, Seattle, WA, May. 2008.
- [82] A.W. Smith, L.J. McDaid, and S. Hall. A compact spike-timing-dependent-plasticity circuit for floating gate weight implementation. *Neurocomputing*, 124:210 – 217, 2014.
- [83] R. Gopalakrishnan and A. Basu. Triplet spike time dependent plasticity in a floating-gate synapse. In *IEEE International Symposium on Circuits and Systems*, pages 710–713, Lisbon, Portugal, May. 2015.
- [84] R. Gopalakrishnan and A. Basu. Triplet spike time-dependent plasticity in a floating-gate synapse. *IEEE Transactions on Neural Networks and Learning Systems*, PP(99):1–13, Dec. 2015.
- [85] P. Sjostrom, G. Turrigiano, and S. Nelson. Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. *Neuron*, 32(6):1149–1164, Dec. 2001.
- [86] Z.Q. Wang, H.Y. Xu, X.H. Li, H. Yu, Y.C. Liu, and X.J. Zhu. Synaptic learning and memory functions achieved using oxygen ion migration/diffusion in an amorphous ingazno memristor. *Advanced Functional Materials*, 22(13):2759–2765, 2012.
- [87] J. Hu, H. Tang, K.C. Tan, H. Li, and L. Shi. A spike-timing-based integrated model for pattern recognition. *Neural Computation*, 25(2):450–472, Feb. 2013.
- [88] Y. Qiang, H. Tang, K.C. Tan, and H. Li. Rapid feedforward computation by temporal encoding and learning with spiking neurons. *IEEE Transactions on Neural Networks and Learning Systems*, 24(10):1539–1552, Oct. 2013.

- [89] E. Chicca, F. Stefanini, C. Bartolozzi, and G. Indiveri. Neuromorphic Electronic Circuits for Building Autonomous Cognitive Systems. *Proceedings of the IEEE*, 102(9):1367–1387, Sep. 2014.
- [90] C. Mead. Neuromorphic electronic systems. *Proceedings of the IEEE*, 78:1629–1636, Oct. 1990.
- [91] R. Guyonneau, S.J. Thorpe, and R. VanRullen. Neurons tune to the earliest spikes through stdp. *Neural Computation*, 17:859–879, 2005.
- [92] T. Masquelier, R. Guyonneau, and S.J. Thorpe. Competitive STDP-based spike pattern learning. *Neural Computation*, 21(5):1259–1276, 2009.
- [93] M. Gilson, T. Masquelier, and E. Hugues. Stdp allows fast rate-modulated coding with poisson-like spike trains. *PLoS ONE*, 7:e1002231, 2011.
- [94] A. van Schaik. Building blocks for electronic spiking neural networks. *Neural Networks*, 14(67):617 – 628, 2001.
- [95] G. Indiveri, B. Linares-Barranco, T. J. Hamilton, A. van Schaik, R. Etienne-Cummings, T. Delbruck, S.-C. Liu, P. Dudek, P. Hfliger, S. Renaud, J. Schemmel, G. Cauwenberghs, J. Arthur, K. Hynna, F. Folowosele, S. Saghi, T. Serrano-Gotarredona, J. Wijekoon, Y. Wang, and K. Boahen. Neuromorphic silicon neuron circuits. *Frontiers in Neuroscience*, 5(73), 2011.
- [96] K. Rahimi, C. Diorio, C. Hernandez, and M. D. Brockhausen. A simulation model for floating-gate mos synapse transistors. In *IEEE International Symposium on Circuits and Systems. ISCAS 2002.*, volume 2, pages II–532–II–535 vol.2, 2002.
- [97] C. Diorio, P. Hasler, B. A. Minch, and C. Mead. A single-transistor silicon synapse. *IEEE Transactions on Electron Devices*, 43(11):1972–1980, 1996.
- [98] C. Diorio, P. Hasler, B. A. Minch, and C. A. Mead. A floating-gate mos learning array with locally computed weight updates. *IEEE Transactions on Electron Devices*, 44(12):2281–2289, Dec 1997.
- [99] G. Indiveri. A low-power adaptive integrate-and-fire neuron circuit. In *IEEE International Symposium on Circuits and Systems. ISCAS 2003.*, volume 4, pages

IV-820-IV-823 vol.4, May 2003.