Bielefeld University

Faculty of Technology

CITEC

VLSI implementation of a calcium-based plasticity learning model

Frank Lucio Maldonado Huayaney

A thesis
submitted in partial fulfilment of the requirements
for the degree of

Doctor of Engineering (Dr.-Ing.)

April 2018

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

Bielefeld, April 2018

_____ (Signed)

Frank Lucio Maldonado Huayaney (Name of student)

Abstract

A key feature of autonomous systems is the ability to solve computationally intensive tasks while adapting to changes in the environment; therefore, in these systems learning is needed to predict the responses of the environment to the system actions, thus guiding the system to achieve its goals. However, the learning capabilities required for this feature are underdeveloped in artificial systems, especially when compared to those of humans and animals.

Highly-computational processors are embedded in chip technology (i.e. CPU and GPU) which every year uses lower dimension transistors yielding high speed, low leakage power, and low cost per transistor. However, the conventional approach to computation, based on the von Neumann architecture with separate units for information storage and processing, is still outperformed in energy efficiency by biological nervous systems in cognitive tasks, such as classification and prediction, where the input data is characterized by ambiguity and uncertainty. In this sense neuromorphic engineering solves specific tasks which are easily performed by biological systems using computational models discovered in biological organisms and where classical processors' architecture would have difficulties.

This thesis aims at the implementation of biologically inspired learning algorithm to be embedded in full-custom VLSI spiking neural networks with the goal of constructing compact real-time low-power learning systems with potential application in computational neuroscience basic research investigation, and applications where input data is ambiguous such as in patter recognition.

The starting point of this research is based on recent studies that demonstrated a key role of calcium ions for long term synaptic plasticity. These experimental results have inspired mathematical models and hardware implementations of calcium based learning algorithms. Here I present two

prototypes of a novel Very-large-scale Integration (VLSI) implementation of a recently proposed calcium-based learning algorithm, its circuital and computation model simulation results and comparison with the mathematical model. The second improved circuit corrects errors observed in the first chip and it is connected to a low-power neuron in a small array.

The elaboration of this learning system embedded in a chip provides insight and significant progress in the complex task to understand how to build brain-like integrated systems. This system can be used also as a tool for validating hypotheses arising from experimental observations of biological systems and computational models.

Acknowledgements

Dedico el presente trabajo a mi madre Agliberta Huayaney que desde el cielo siempre me da fuerza y me genera las oportunidades para seguir aprendiendo y mejorando mis conocimientos y habilidades. Gracias a mi padre Lucio Maldonado por su esfuerzo innegable en educarme y siempre estar a mi lado brindandome sus consejos y cariño. Muchas gracias a mis hermanos Ronald y Carol por proveerme su soporte durante muchos años de mis estudios y por mantenerme siempre presente a pesar de la distancia. Gracias a mi segunda madre Elva, mis tios Daniel, Libia, Yolanda, Pablo Huayaney y quienes ya no estan físicamente presente por sus sabios consejos e innegable apoyo a lo largo de mi vida. Gracias igualmente a mis primos y grandes amigos que conocí en los diferentes lugares en que viví por los cosejos y mensajes de motivación. Los amigos que conocí en Bielefeld hicieron mi estadía placentera en Alemania; gracias a la comunidad latina con la pude tener un pedazo de Perú en este lindo país.

I want to express my gratitude to my supervisor Elisabetta Chicca and my colleague Stephen Nease who guided me in the development of this project from its theoretical concept passing to its circuit design and finally fabrication and test of the prototype.

I am also grateful to the members of the Neuromorphic Behaving System laboratory who provided me their friendship, help and made my stay comfortable and funny when working in the CITEC department (and of course outside of work also!).

Many thanks to the friends I met in Bielefeld University, being in such a nice community provided me a good environment to grow as a person and learn from different life experiences. Vielen Dank Deutschland!!!

Cuando despiertan mis ojos y veo que sigo viviendo contigo Perú
emocionado doy gracias al cielo por darme la vida contigo Perú
(...) sobre mi pecho yo llevo tus colores
y están mis amores contigo Perú.

Acronyms

AER Address-Event Representation

AMPA α-Amino-3-Hydroxy-5-Methyl-4-Isoxazolepropionic Acid Receptor

ASIC Application Specific Integrated Circuit

CAD Computer Aided Tool

CAM Content-Addressable Memory

CHP Communicating Hardware Processes

CMOS Complementary Metal Oxide Semiconductor

DAC Digital-to-Analog Converter

DP Depression-Potentiation

DPD Depression-Potentiation-Depression

DPI Diff-Pair Integrator

DRAM Dynamic Memory Technology

DRC Design Rule Check

EDA Electronic Design Automation

EPSC Excitatory Postsynaptic Current

EPSP Excitatory Postsynaptic Potential

FDSOI Fully Depleted Silicon On Insulator

FSMs Finite State Machines

GABA γ-Aminobutyric Acid

IF Integrate-and-Fire

IC Integrated Circuit

ISI Interspike Interval

LTD Long-Term Depression

LTP Long-Term Potentiation

MIM Metal-Insulator-Metal

NMDA N-Methyl-D-Aspartate Receptor

OTA Operational Transconductance Amplifier

PTP Posttetanic Potentiation

QDI Quasi-Delay-Insensitive

RAM Random-Access Memory

ROLLS Reconfigurable On-Line Learning Spiking

SNNs Spiking Neural Networks

SoC System-on-Chip

SpiNNaker Spiking Neural Network Architecture

SRAM Static Random Access Memory

SRM Spike Response Model

STDP Spike Time Dependent Plasticity

sWTA Soft Winner-Take-All

TFS Source-Follower-Circuit-with-Transconductance-Amplifier

VLSI Very-Large-Scale Integration

WTA Winner-Take-All

Contents

1 Introduction **1**
- 1.1 Microprocessor Evolution and Technology Challenges 1
- 1.2 Learning in Autonomous Systems 4
- 1.3 The Neuromorphic Approach 4
- 1.4 Outline of this Thesis 7
- 1.5 Acknowledgement to the Contributors 8

2 Models of Synaptic Plasticity **9**
- 2.1 The Spike Response Model 10
- 2.2 Synaptic Plasticity 11
- 2.3 The Simplified Calcium-based Learning Model 17
- 2.4 Discussion 23

3 Neuromorphic Circuits Blocks **27**
- 3.1 CMOS Operation in Inversion Region 29
- 3.2 MOSFET Characterization 32
- 3.3 Mismatch 35
- 3.4 The Diff-Pair Integrator Circuit (DPI) 41
- 3.5 The Operational Transconductance Amplifier (OTA) 43
- 3.6 The Winner-take-all Circuit 46
- 3.7 Discussion 48

4 First Synapse Circuit Implementation **51**
- 4.1 The Calcium Synapse Circuit 54
  - 4.1.1 Simulation Results 60
- 4.2 Hardware measurement results 69
- 4.3 Discussion 73

5 Second Synapse Circuit Implementation **77**
- 5.1 The Calcium Circuit 77
- 5.2 The Synapse Core and The Bistability Circuits 79
- 5.3 The Linearizer 80
- 5.4 The Configurable Bias Current Generator 81

5.5 The Neural Network Block 86
5.6 Simulation Results 87
5.6.1 Bistability 88
5.6.2 Potentiation and Depression 90
5.6.3 STDP Waveform 91
5.6.4 Configurable Bias Circuit 91
5.7 Hardware measurement results 92
5.7.1 STDP Measurement Results 92
5.7.2 Potentiation and Depression 94
5.7.3 Bistability 95
5.7.4 Linearizer 96
5.8 Discussion 97

6 Network Operation **99**
6.1 Single Synapse Learning 101
6.2 Simple Perceptron 107
6.3 Discussion 111

7 Mismatch Characterization **113**
7.1 The Calcium Circuit 115
7.2 The Synapse Circuit 117
7.3 The Bistability Circuit 118
7.4 Discussion 119

8 Conclusions **121**
8.1 Future Work 123

Bibliography **127**

Publications **143**

1 Introduction

1.1 Microprocessor Evolution and Technology Challenges

Over the past few decades a rapid evolution in microprocessor performance indicators [1] (Fig. 1.1) such as speed, power consumption and components integration (this last as Moore's law predicted [2]) was achieved mainly due to three factors: transistor scaling, core microarchitecture techniques and cache memory [3]. This microprocessor evolution has driven drastic improvements in electronic hardware with high computational power.

The transistor scales down by 30% (0.7x) every two years having as basis to keep the electric field constant everywhere; the benefits of this are the increase of energy efficiency (MIPPS/watt) by the cube of the scaling factor due to increased speed (40% faster) and reduced power consumption (50% lower) [4]. In 1974, Dennard [5] described this scaling principle using the scale factor as the only parameter by transforming three variables: dimension (insulator thickness, junction depth, channel length and width), voltage applied to the device and substrate doping concentration. However, when transistors reach node technologies below $65nm$ [6] (a process node denotes a specific semiconductor manufacturing process, and in general it characterizes the minimum transistor feature size) more challenges need to be overcome; one is that the voltage scaling, which is limited by threshold voltage, produces considerable sub-threshold leakage currents; similar problem occurs when scaling the gate oxide thickness, here tunneling current is a considerable percentage of power consumption because of its small dimensions [7]. Therefore, the end of Dennard scaling poses a serious problem for computing's status quo. Post-Dennard scaling yields limited energy efficiency gains in each new device generation, which results in a significant amount of underutilized, or "dark" silicon [8]; furthermore, supply-voltage scaling increases considerably device variability which could leads to unreliable systems. Some progress in this field has been achieved by using Fully Depleted Silicon On Insulator (FDSOI) [9] and Fin Field Effect Transistor (FinFET) [10] technologies which appear below $22nm$ node process. FinFET is a 3D struc-

ture gate that envelopes the transistor channel; however, its use increases significantly the design complexity. FDSOI which consists of a substrate underneath each transistor as well as a shallower channel provides dynamic control of the transistor threshold by polarizing its substrate however its production cost is high. Nowadays CMOS processes have reached even $10nm$ node technology; however, forecast predict to reach in 2020 the scaling limit of 2-3-nanometre [11].

Microarchitecture techniques such as multicore increases computational throughput, however their benefits are at the expense of energy efficiency where an excessive amount of cores added to a high frequency operation could reach prohibitive power consumption levels [12]; it is also important to highlight that processors should implement parallelism among their cores in order to spread computation tasks and therefore gain speed. Complementary to multicore, customization can be used to reduce execution latency, this strategy increases computational performance by exploiting hardwired for data movement, therefore reducing the number of instructions per operation. The challenges in parallelism and customization for future microprocessor will be to reduce the energy expended for data movement (keep data locally as much as possible) as well as reducing processors synchronization.

Dynamic Memory Technology (DRAM) density has doubled nearly every two years; however, the access time to store/readout data has improved slower giving as result a gap between processor and DRAM speed which is now the primary obstacle to improve computer system performance. For instance a processor spends 75% of its time in memory operations and if the clock of a system increases, the processor could spend even larger fraction of application time waiting for memory processing. The unification of logic and DRAM on a single chip provides potential improvements by provinding higher bandwidth, low latency and better energy efficiency, which leads to considerable cost savings from removing unnecessary memory and reducing board area [13].

It is clear that radical innovation is necessary in either device technology or system architecture to continue historical performance improvements. A potential source of inspiration for new directions comes from neuroscience [8]. Conventional von Neumann architectures, which separate memory and computation, are outperformed by biological systems for typical cognitive tasks. These tasks (inference, classification, goal-based control, etc.) are often

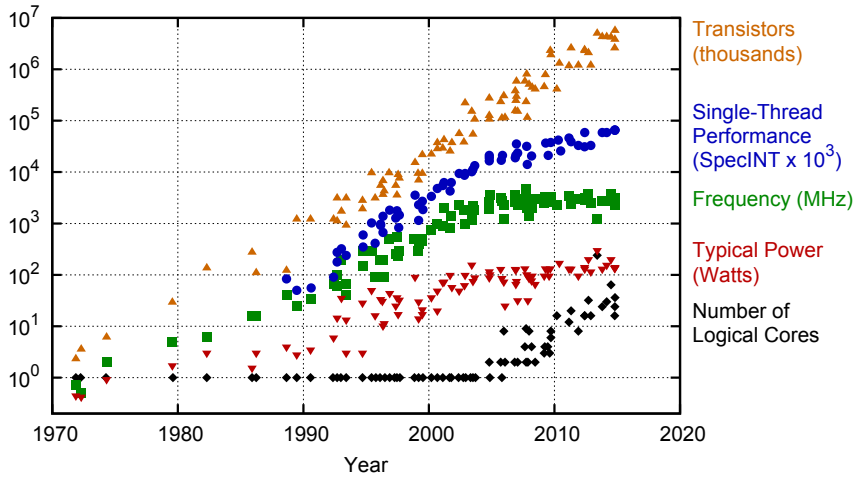


Fig. 1.1: 40 Years of Microprocessor Trend Data. Original data up to the year 2010 collected and plotted by M.Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, ad C. Balten; New plot and data collected for 2010-2015 by K. Rupp. A slow down evolution in the performance parameters is observed in the last decade.

performed in environments featuring uncertainty and ambiguity. Evolution has tuned neural architectures and learning structures to perform highly efficiently in these environments. This efficiency comes from the brain's highly parallel co-location of computation and memory and its ability to learn the statistics of its environment.

Furthermore, in digital processors, circuit blocks are implemented by using transistors as switch devices (digital gates), ignoring the analog nature and hence advantages of this device. On the other hand, the macroscopic conductance of the voltage-gate ion channels, which set the permiability of the neuron cell membranes [14] has the same physical principles applied to the conductance of a transistor operated in weak inversion regime (we will describe this analog operation region in a following chapter) as an exponential dependence on the applied voltage; therefore, this characteristic could be exploited.

Neuroscience can also take advantage of this full-custom analog Very-Large-Scale Integration (VLSI) approach to test hypothesis concerning the computation performance in the brain. It is believed that this computation is organized by a finite set of primitives [15] and if we are able to reproduce them we will progress in the complex task to understand how to build brain-like integrated systems.

1.2 Learning in Autonomous Systems

One of the most amazing properties of our brain is its capability to progressively learn from experiences and consequently predict responses of the environment to system actions. This mechanism is usually described as plasticity and take place in the nervous systems specifically in the modification of the synaptic strength which set the information flow among neurons [16].

In order to build autonomous cognitive systems, it is required to equip them with plasticity characteristics that let them adapt to a constantly-changing environment. This adaptability requires significant computational resources devoted to learning, and current artificial systems are lacking in these resources when compared to humans and animals. It is here that Neuromorphic engineering as a multidisciplinary field aims to build such cognitive agents which feature learning structures similar to those in biology, with the goal of achieving the performance and efficiency of natural systems. Here we present a novel VLSI implementation of a calcium-based synaptic plasticity model, comparisons between the model and circuit simulations, and measurements of the fabricated circuit. Applications of this model into a small neural network are also presented.

1.3 The Neuromorphic Approach

Neuromorphic engineering's goal [17] is to create electronic systems which emulate both the architecture and functional primitives of the nervous system such as interconnections betweens synapses and neurons whereupon it intends also to reduce the gap between computing performance and technology scaling in the fields of parallelism, energy, memory and reliability [18].

The neuromorphic approach proposed by C. Mead in the late eighties led to the development of subthreshold VLSI chips which typically feature parallel and distributed computation, asynchronous event-driven communication, and the co-location of computation and memory via the interconnection of artificial synapses and neurons; however, few of them integrate the ability to learn and adapt to the environment through synaptic plasticity.

To equip autonomous systems with cognitive capabilities comparable to those of biological systems, particular efforts have to be made to mimic their ability to learn from experience and adapt to changing environmental conditions. The theoretical neuroscience research community has produced a plethora of learning models, and some of them have been translated into analog VLSI circuits [19–22].

Furthermore, some large-scale neuromorphic systems have already been implemented, they emphasize diverse characteristics such as neuron model, synapse model and communication architecture which lead to strengths in flexibility or low-power consumption as described below.

The Spiking Neural Network Architecture (SpiNNaker) [23] is a System-on-Chip (SoC) hardware that runs in biological real time, it was designed with digital Electronic Design Automation (EDA) tools (Synopsys and Silistix) and each chip (node) uses 18 ARM968 cores each one with local and shared RAM running at 200MHz. One of the cores is in charge of system management tasks, 16 of the other cores are used for neuromorphic computation, and an extra core is available to improve the manufacturing yiel. The fabrication process was UMC $130nm$ CMOS and a node dissipates up to 1W [24]. The system is programmed in high level description language with PyNN [25]. Each core is able to model 1000 neurons each with 10 000 inputs synapses. The internode communication in the SpiNNaker is via packets that transmit the spike events. In order to address each package, each chip integrates a router component that uses Address-Event Representation (AER) [26] protocol to transmit them. If the router finds a packet delayed (two time phases old), this is move to a garbage collection mechanism.

The TrueNorth chip [27] is a brain inspired processor that consumes 65mW and operates in real-time, highly-parallel and is scalable. It contains 4096 cores and each core includes 256 input axons, 256 neurons and 64k synaptic crossbars. The design methodology for all the communication and control circuits is asynchronous whereas for computation it is synchronous. The asynchronous circuits which use request and acknowledge handshaking bits to transmit data without clock, allows to save power consumption by only execute switching activity when there is a required operation. The clock signals for the synchronous circuits are generated in each core by an asynchronous control circuit thereby reducing the number of clock transitions and therefore minimizing power consumption. For asynchronous design, they selected a

Quasi-Delay-Insensitive (QDI) approach which are circuits almost invariant to the delays of any wire or component and are described using Communicating Hardware Processes (CHP) description language [28]. From here they manually decomposed in production rules and then into transistors netlist by academic tools. The TrueNorth chip, was fabricated using Samsung's 28nm LPP CMOS process technology [29]. However, one drawback of this design is that it does not include plasticity, so this is performed off-chip.

Other neuromorphic hardwares include the Reconfigurable On-Line Learning Spiking (ROLLS) neuromorphic processor [30] which consists of 265 neurons and 126K synapses with a bi-stable spike-based plasticity mechanisms that provides on-line learning abilities, the Neurogrid chip [31] designed with a mixed-signal approach, the BrainScaleS chip [32] which operates in accelerated mode, and the event-based neural network with asynchronous programmable synaptic memory chip [33] which consist of Integrate-and-Fire (IF) neurons and excitatory and inhibitory synapses where the synaptic strength is stored in a Static Random Access Memory (SRAM) module (off-chip plasticity).

This thesis presents a novel implementation of a scalable aVLSI neural network which integrates theoretical models of synaptic plasticity such as learning algorithms with local event-based mechanisms of weight update, which makes them especially suited for neuromorphic implementations. In particular, I explore the calcium-based model because it reproduces a variety of experimental protocols not explained by phenomenological models [34]. Here I describe the initial steps of the project consisting of a computational model implementation of the synapse up to its behaviour in a fabricated small VLSI neural network array that comprises low power IF neurons and calcium-based plasticity synapses communicated through the AER protocol. The blocks were designed using fully analog Computer Aided Tool (CAD) tools and the implemented learning circuit is based on a recently proposed computational model [35] that accounts for plasticity behaviours evoked by different features of the pre- and post-synaptic activities (e.g. spike timing as in Spike Time Dependent Plasticity (STDP), firing rate as in Hebbian learning).

1.4 Outline of this Thesis

The work in this thesis covers the main stages of a VLSI chip project, from the conception of the theoretical model going through establishing the electrical and layout characteristics for a successful tapeout up to the chip measurement and performance characterization in specific tasks.

In the second chapter, I present a general overview of neuron and synapse biological characteristics as well as previous learning computational models such as classical STDP, finally I describe the important role of calcium ions Ca^{2+} to generate Long-Term Depression (LTD) and Long-Term Potentiation (LTP) in biology and how computational models take in account this variable to come up with generalized learning models.

The third chapter, deals with VLSI topics starting from explaining the principles of weak inversion operation model in CMOS transistors. This operation region characterizes for low-power consumption and provides similar characteristics to its biological counterpart, transmitter ions, such as exponential I-V behaviour. Later I describe the electrical and mismatch CMOS characteristics for the selected technology AMS $0.18\mu m$. The last part of this chapter presents the basics circuit topologies that are used to build neuromorphic learning block.

In the fourth chapter, I present in detail the first fabricated VLSI calcium-based plasticity learning circuit. This synapse consists of three main blocks calcium, synapse core and bistability; and the electrical characteristics for each of them are explained. Finally, measurement results are presented with discussions about trade-off considerations and ultra-low power techniques.

In the fifth chapter, I present the analysis of a second improved VLSI chip which overcomes the observed problems of the first learning circuit and in addition contains a small neural network. Here, additional circuit blocks are included such as a synaptic weight linearizer whose target is reducing single synapse current contribution in the network, and a bandgap which improves bias control compared to direct fixed voltages supply in the parameter signals. For all of them, circuit architecture, simulation results and measurements are presented.

In the sixth chapter, I present and analyse neural network operations using the calcium-based learning circuit of the second chip. These experiments include a single synapse connected to a neuron as well as two synapses connected to one neuron (single perceptron).

In the seventh chapter, I describe an improved version of the second chip focusing on mismatch reduction which ended up being the main constraint of the neural network. Here I estimate variability of each model parameter.

The conclusions and future work are presented in the last chapter based on the analysis of the previous ones.

1.5 Acknowledgement to the Contributors

For the development of the project presented in this thesis and in related publications [36, 37], the contributions of Prof. Elisabetta Chicca and Dr. Stephen Nease were insightful. Throughout this thesis I use the term "**we**" to refer the three of us. Despite all of us contributed in every stage of the project, I would like to highlight the support of Prof. Elisabetta Chica in the theoretical neuroscience and neuromorphic circuit background. Likewise, Dr. Stephen Nease provided meaningful ideas in low power consumption techniques, he also together with the Neuromorphic Cognitive Systems lab at the Institute of Neuroinformatics (University of Zurich and ETH Zurich) implemented the top-level configuration and layout of the testchips which include bias generator circuits and bound-pads routing; in addition, he designed the PCB setup for the testchips. My contribution to this work was providing support in the layout routing of top circuit blocks, PCB verification and components soldering.

2 Models of Synaptic Plasticity

Synaptic plasticity is considered the essential element in learning and memory; therefore, in order to understand its behaviour we must figure out how experience and training modify synapses and how their modifications change neural firing patterns. The starting point was proposed by Donald Hebb [38] who concluded that if the input of one neuron contributes to the firing of a second one, the synapse from the first to the second neuron should be strengthened. The original Hebb's suggestion was generalized and currently includes also decreases in synaptic strength. Subsequent measurements in brain regions such as in hippocampus and neocortex showed that constant amplitude stimulation in the synapse can produce changes on it that last for more than 15 min [39]. In order to reach stability when interconnecting neurons, synaptic strengths need to be scaled and upper and lower bonds set. Typically synaptic plasticity occurs only if the difference in the pre- and postsynaptic spike times falls within a window time lower than $\pm$ tens of milliseconds. More recently experiments reported that synaptic modification as a function of these spikes timing (STDP) is just one mechanism for the induction of long-term changes and that biological plasticity is significantly more complex [40]; consequently, the design of biological synapse models can considerably increase the amount of information per memory [41] as well as the number of storable memories [42]; however, these types of synapses are hard to implement in silicon and the area occupied can be so wide that larger number of simple synapse would end up been more efficient [43]. Synaptic rules are represented as differential equations describing the synaptic weight variation as a function of pre- and post- spikes (although other parameters can also be added) which generally model a slow process that gradually modifies synaptic weights. Neuromorphic circuits which are implemented with Complementary Metal Oxide Semiconductor (CMOS) transistors in VLSI take advantage of these formulations to emulate the electro-physiological properties of biological neurons in hardware.

I start this chapter by describing the spiking neuron model which later will be used to implement neural networks and whose outputs are used to generate synaptic modifications; later I describe two synapse models starting from

the classical STDP and then explaining in detail a more complex model "calcium-based learning" which is the target of this work.

2.1 The Spike Response Model

Action potentials are the result of currents that flow through ion channels in the cell membrane. Hodgkin and Huxley [44] measured these currents and described their dynamics in terms of differential equations. However, the system of equations that they proposed was too complex to analyse given that it consisted of four dimensional nonlinear differential equations which make the variables' waveform hard to visualize. Simplified models [45] aim to reduce the number of differential equations which consequently reduces their processing time when implementing them as algorithms. The Spike Response Model (SRM) shown in Fig. 2.1 is a generalization of the phenomenological leaky IF neuron model [46]. While in the IF model the potential is expressed as a function of a voltage, in the SRM parameters depend on the relative time from the last output spike; this model presents a formulation of the membrane potential by using an integral over the past which is the explicit solution of the differential equations, thus reducing computation complexity.

The neuronal signals consist of short electrical pulses which are called action potentials or spikes which are generated whenever the membrane potential u crosses a threshold v from below [47]. The waveform of this signal does not vary through its propagation along the axon. We define the moment a membrane potential u crosses a threshold v as the firing time $t^{(f)}$.

$$t^{(f)} = u(t^{(f)}) = v \quad and \quad \left.\frac{du(t)}{dt}\right|_{t=t^{(f)}} > 0. \tag{2.1}$$

The mathematical expression of the membrane potential is given in eq. 2.2. Here ϵ describes its evolution for incoming spikes, η defines the decrease after the membrane reaches the threshold and u_{rest} defines the resting potential at steady state [45]. In circuit design η can be implemented as a current source charging a RC or C circuit which results in an exponential waveform in the voltage as function of time.

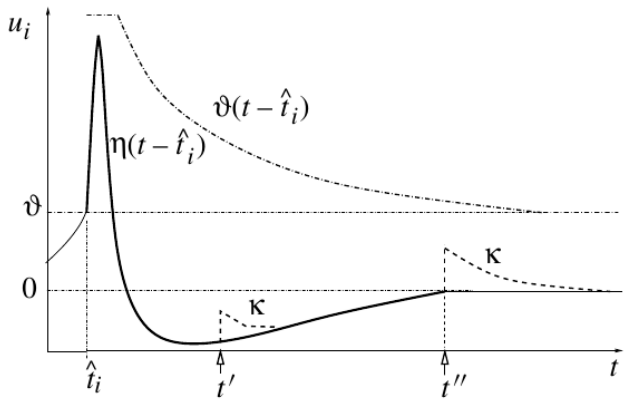


Fig. 2.1: Membrane potential dynamics for spike response model. η describes its response for an input spike, ϑ defines its decrease after reaching a threshold (extracted from [45]).

$$u_i(t) = u_{rest} + \eta(t-\hat{t}_i) + \sum_j w_{ij} \sum_f \epsilon\left(t-\hat{t}_i, t-t_j^{(f)}\right), \tag{2.2}$$

where $\eta(t-\hat{t}) = -\eta_0 e^{-(t-\hat{t})/\tau_{refr}} \mathcal{H}(t-\hat{t})$, $t_j^{(f)}$ are spikes of presynaptic neurons j and w_{ij} the synaptic efficacy; $\mathcal{H}$ denotes the heaviside step function. All the terms depend on $t-\hat{t}_i$, the time since the last output spike, w_{ij} is the synaptic strength and determines the amplitude of the postsynaptic response to an incoming action potential. w_{ij} is the term that provides plasticity to a neural network, the modification of this factor leads to the learning process which is the main topic of this thesis. Nevertheless, this neuron model provides insightful information about how spikes are originated providing also considerable details to resemble its biological counterpart. In a following chapter network experiments will be presented which are implemented by using synapses and neurons in hardware.

2.2 Synaptic Plasticity

In the brain there are two main classes of synapses: electrical and chemical. The chemical synapse is more common given that this is the principal mediator of targeted neuronal communication [48]. Its advantages are that it can produce either excitatory or inhibitory actions and it can amplify neuronal signals, allowing a small presynaptic nerve terminal to modify the potential of a large postsynaptic cell [14].

The synaptic transmission mechanism consists of a presynaptic spike that depolarizes the synaptic terminal which produces calcium ions flow through the presynaptic calcium channels, causing vesicles of neurotransmitters to be released into the synaptic cleft. The neurotransmitters bind temporally to postsynaptic channels, opening them and allowing ionic current to flow across the membrane as shown in Fig. 2.2 [49]. Typical excitatory receptors are α-Amino-3-Hydroxy-5-Methyl-4-Isoxazolepropionic Acid Receptor (AMPA) and N-Methyl-D-Aspartate Receptor (NMDA) neurotransmitters while γ-Aminobutyric Acid (GABA) receptors are inhibitory [43].

LTP induction is obtained by simultaneous presynaptic neurotransmitter release and postsynaptic depolarization. Presynaptic stimulation at a low frequency produces Excitatory Postsynaptic Potential (EPSP) which does not change its magnitude when it is followed by postsynaptic depolarization [50, 51]. On the other hand, a high presynaptic rate with simultaneous postsynaptic hyperpolarization leads to a persistent potentiation (Posttetanic Potentiation (PTP)). Only when synaptic input is paired with postsynaptic depolarization is LTP induced. In addition, the induction of LTP requires activation of NMDA receptors which are directly gated by both voltage and neurotransmitter, so that they let current flow only when the membrane is depolarized sufficiently to relieve a block by magnesium ions. Synaptic depression occurs when there is a decrease in the probability of transmitter release as result of a moderate increase of postsynaptic Ca (low frequency stimulation of afferents [52]); in this case the amount of depression depends not on the number of presynaptic action potentials but on the number of vesicles released [53].

Some experiments have also demonstrated that the induction of synaptic potentiation and depression depends on the timing between pre- and postsynpatic spikes. When this timing is positive ($t_{post} - t_{pre} > 0$) we can obtain potentiation, and similarly in the opposite case we obtain depression [54]. In addition, modification in synaptic strength occur when pre- and post- spikes are close enough each other. Two experimental protocols to induce LTP one for high frequency and the other for different timing $t_{post} - t_{pre}$ are shown in Fig. 2.3 and Fig. 2.4 respectively.

In Fig. 2.3, a pre- spike is applied to measure the synaptic strength of the postsynaptic response as shown in (A). This pulse generates postsynaptic potential but not action potential. Later an input spike train with high

frequency is supplied which generates postsynaptic firing as shown in (B). Finally again the neuron is stimulated with the pre- spike and a considerable increase of the postsynaptic potential is observed as shown in (C) [55].

Fig. 2.4 shows the experimental results obtained in [54] in terms of the induced synaptic modification depending on the correlated timing between pre- and post- synaptic spikes. As can be seen if the presynaptic spike occurs before a postsynaptic action potential, a LTP is generated; however, if a postsynaptic spike occurs before a presynaptic action potential, a LTD is generated. A common approximation of this data is an exponential decay with positive and negative coefficients for potentiation and depression respectively.

Computational models of biological synapses were successfully described by Destexhe [56]. He used a first-order kinetic equation for the neurotransmitters' dynamic of a synapse, obtaining hence exponential functions to fit Excitatory Postsynaptic Current (EPSC).

$$R + T \underset{\beta}{\overset{\alpha}{\rightleftharpoons}} TR^*, \tag{2.3}$$

where R and TR^* are the unbound and bound forms of the post-synaptic receptor, α and β are the forward and backward rate constants for transmitter binding. Considering that the change in neurotransmitter concentration T in the cleft occurs in a brief pulse, and defining r as the fraction of receptors in the activated state, we obtain a first-order differential equation of the kinetic model.

$$\frac{dr}{dt} - \alpha[T](1-r) - \beta r, \tag{2.4}$$

The electrical current that results from the release of a unit amount of neurotransmitter at time t_s is

$$I_{syn}(t) = \overline{g_{syn}}\, r(t)\, (V(t) - E_{syn})\,, \tag{2.5}$$

where $\overline{g_{syn}}\, r(t)$ is the synaptic conductance change in the postsynaptic membrane because of the effect of a transmitter binding to an opening postsynaptic receptors and $\overline{g_{syn}}$ is its amplitude. V(t) is the voltage across the postsynaptic membrane and E_{syn} is the reversal potential of the ion channels that mediate the synaptic current. Simple waveforms are used to describe $r(t)$ since the arrival of a presynaptic spike which include alpha functions, single exponential decays and dual exponential functions [48]; the former is the most precise

representation and it is stated in Eq. 2.6. In a synaptic model, the weight w_{ij} could be used as a scaling factor for the maximum postsynaptic receptor conductance $\overline{g_{syn}}$.

$$r(t) = \frac{t - t_s}{\tau} exp\left(-\frac{t - t_s}{\tau}\right), \tag{2.6}$$

Multiple learning models have been proposed for explaining synaptic changes and some have also a corresponding implementation in VLSI circuits. However, hardware models that show how the dynamics of the postsynaptic calcium alone determine the outcome of synaptic plasticity are still uncovered. Recently "A calcium-based plasticity model" that determine a large diversity of spike timing-dependent plasticity by varying the parameters that define the calcium dynamics has been proposed [35].

In computational models we define the synaptic strength as w_{ij} which connects neuron j to neuron i and by modifying its value we can optimize a neural network performance. Learning consist of modifying this value and the function that correlate this variation with respect to the input spikes is called learning rule. The most common phenomenological learning model is STDP which defines the variation of the synaptic strength as a function of the timing between pre and post spikes by exponential functions. Mathematically we can simplify the synaptic dynamics as stated in Eq. 2.7.

$$\begin{aligned} \frac{d}{dt} w_{ij}(t) = & S_j(t) \left[\int_0^\infty W^{pre,post}(s) S_i(t-s)\, ds \right] + \\ & S_i(t) \left[\int_0^\infty W^{post,pre}(-s) S_j(t-s)\, ds \right], \end{aligned} \tag{2.7}$$

where $S_j(t) = \sum_f \delta(t - t_j^{(f)})$ and $S_i(t) = \sum_f \delta(t - t_i^{(f)})$ are pre- and postsynaptic spike trains, respectively. The terms containing $W(s)$ describe the form of the "learning window". The kernel $W^{post,pre}$ gives the amount of weight change when a presynaptic spike is followed by a postsynaptic action potential with delay. The kernel $W^{pre,post}$ describes the amount of change if the timing is the other way round [45].

The learning waveform $W(t)$ is defined by:

$$W(t) = \begin{cases} A_+ e^{-(t - t_{pre})/\tau_1} & for \quad t_{post} > t_{pre} \\ A_- e^{-(t - t_{post})/\tau_2} & for \quad t_{pre} > t_{post} \end{cases} \tag{2.8}$$

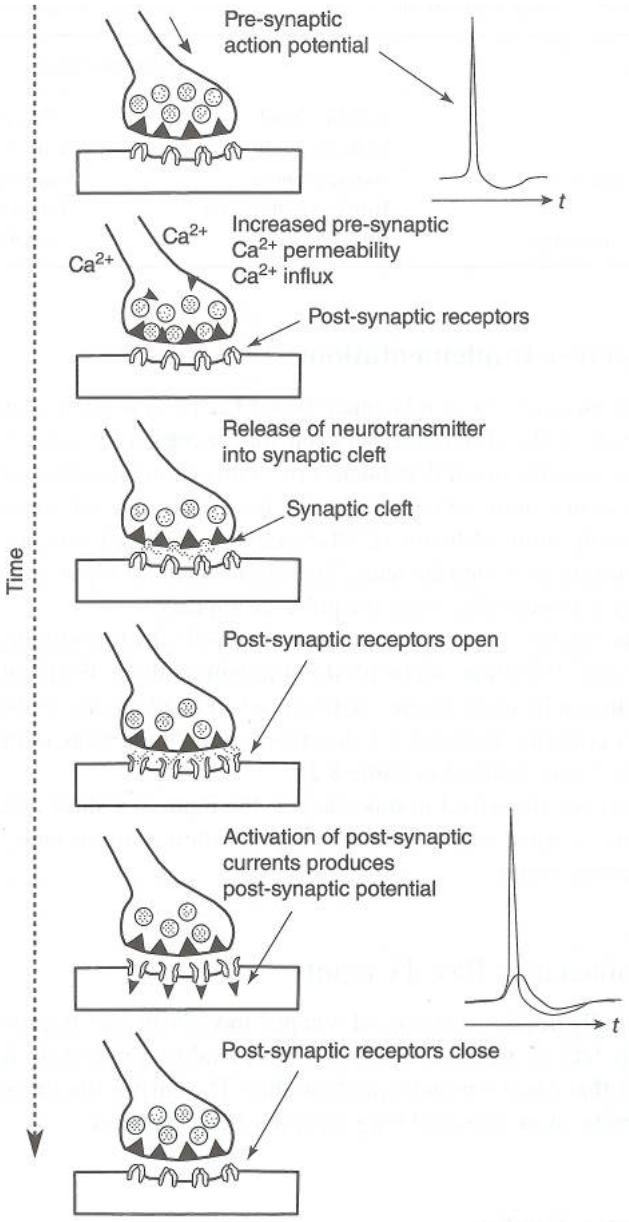


Fig. 2.2: Schematic illustrating the signaling cascade underlying synaptic tranmission. In response to a presynaptic action potential, calcium enters the presynaptic terminal via voltage-gated calcium channels and triggers the release of glutamate-containing vesicles. Glutamate diffuses into the synaptic cleft and activates postsynaptic AMPA and NMDA receptors, ionotropic receptors that act via opening of an ion channel permeable to sodium, potassium and calcium, givin rise to a fast excitatory postsynaptic current (EPSC) (figure extracted from [43]).

with constants $A_+ > 0$, $A_- < 0$ that represent the maximum increment and decrement of the synaptic strength respectively, and $\tau_{1,2}$ which denotes the time decay of the exponential function.

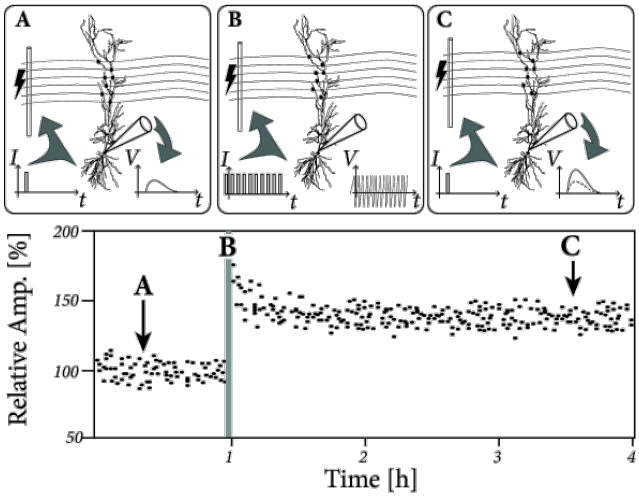


Fig. 2.3: LTP protocol. In (A) a single spike is injected to generate postsynaptic potential. In the second step in (B) a spike train with enough high frequency stimulates the neuron to induce postsynaptic firing. In (C) the postsynaptic response is compared with the one observed in (A) and an increase is observed (figure extracted from [55]).

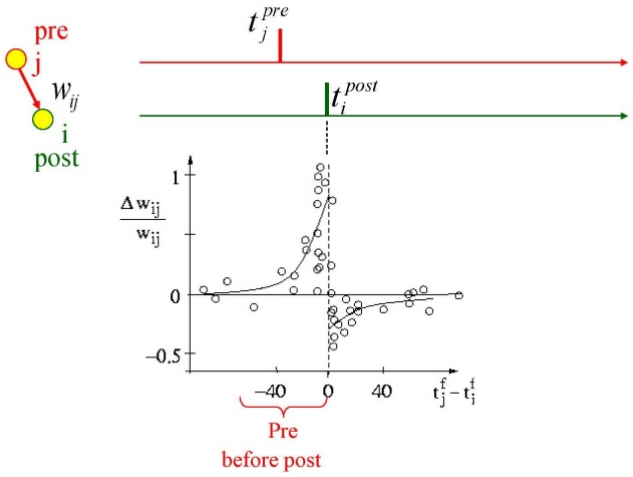


Fig. 2.4: Spike-Timing-Dependent Plasticity. During this protocol both neurons are stimulated to fire at a fixed time. After the spikes pair, the presynaptic neuron is stimulated again to compare the initial synaptic potentiation with the initial. A synaptic change Δw_{ij} is obtained when the pre- and post- spikes are close enough, and LTD is observed when the pre- spike fires after the post- spike, similarly LTP is obtained when post- spike fires after pre- spike (figure reproduced STDP article on scholarpedia, which is based on original from [54]).

The STDP rule is a phenomenological model which is successfully used to resolve many computational task implying neural networks; however, it lacks to reproduce biophysical properties of the synapse and therefore is not suitable for exploring the role of calcium concentration in information processing.

An algorithm that incorporates the STDP learning rule in a neural network simulation was proposed in [57]; this pair-based STDP rules is implemented with two local variables for the low-pass filtered presynaptic and postsynaptic spikes respectively as shown in Eq. 2.9 and Eq. 2.10. A decrease of the synaptic weight is set when y_i is sampled for t_j pre-spikes; similarly, an increase of the synaptic weight is set when x_j is sampled for t_i post-spikes which is expressed in Eq. 2.11. The simulation results of [57] are reproduced using Matlab in Fig. 2.5 and its learning waveform is shown in Fig. 2.6. The figures confirm an exponential increase/decrease of the synaptic weight as a function of the pre- and post- spike timing.

$$\frac{dx_j}{dt} = -\frac{x_j}{\tau_x} + \sum_{t_j^f} \delta\left(t - t_j^f\right), \tag{2.9}$$

$$\frac{dy_i}{dt} = -\frac{x_i}{\tau_x} + \sum_{t_i^f} \delta\left(t - t_i^f\right), \tag{2.10}$$

$$\frac{dw_{ij}}{dt} = -F_-(w_{ij})\; y_i(t)\; \delta\left(t - t_j^f\right) + F_+(w_{ij})\; x_j(t)\; \delta\left(t - t_i^f\right), \tag{2.11}$$

where $F_\pm(w_{ij})$ describes the dependence of the update on the current weight of the synapse.

2.3 The Simplified Calcium-based Learning Model

In 2001, Liesman proposed that modifications in synaptic plasticity depends on the amount of calcium Ca^{2+} concentration [58]; some experiments later demonstrated that high Ca^{2+} elevation triggers LTP, moderate Ca^{2+} elevation triggers LTD and lower level Ca^{2+} do not change the synaptic strength [59]. In addition, long term modification can be observed when the density of NMDA receptors is modified, consequently varying the calcium flux [60]. Fig. 2.7 shows this effect by partially blocking the calcium-permeable NMDA receptors from an initial LTP originated by a high calcium flux to an induced LTD originated by a lower calcium flux.

An initial biological model that tried to mimics the synapse dependence on Ca^{2+} was proposed by Shouval et al. [40] and later its simplification as a

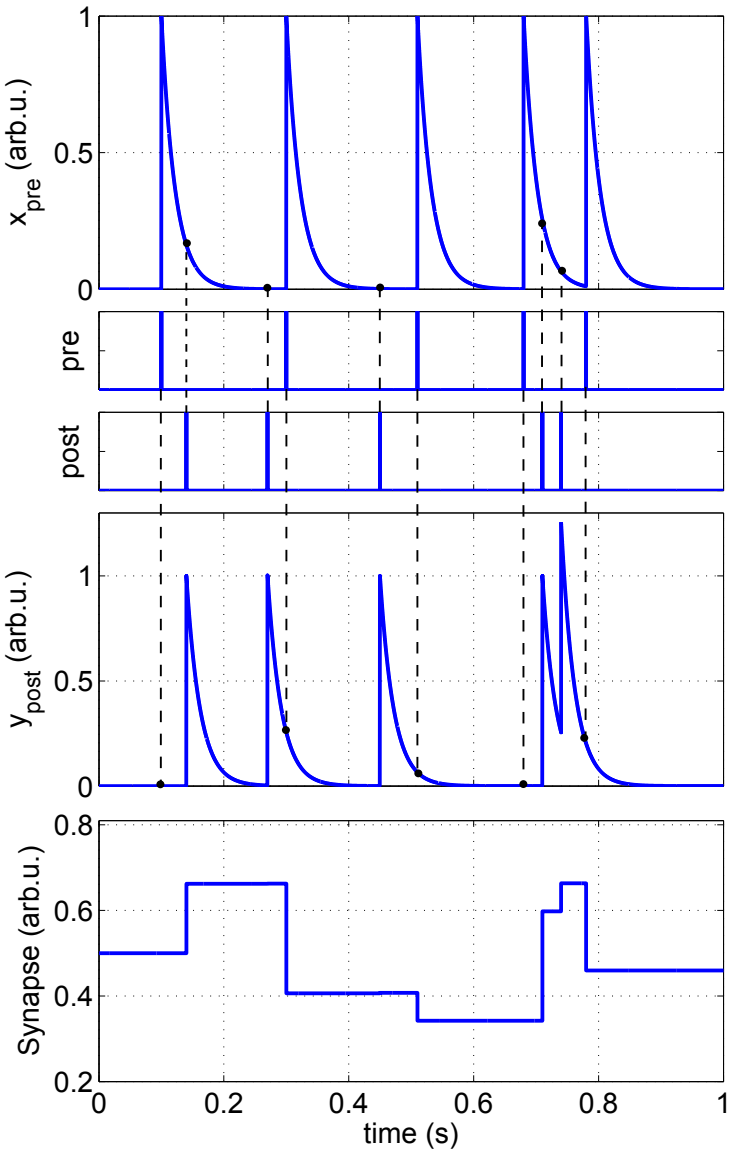


Fig. 2.5: Simulation results for pair-based STDP rule proposed in [57]. Two local variables are used to trace pre- spike (x_{pre}) and post- spike (y_{post}), the synapse decreases when x_{pre} is sampled for each pre- spike and similarly it increases when y_{post} is sampled for each post- spike. $t_{pre} - t_{post} > 0$ is observed for the first spike pair and $t_{pre} - t_{post} < 0$ in the second spike pair, additionally overlap effect when consecutive spikes arrive is shown for post-spikes around 0.78s. For this simulation results the chosen parameters are: $\tau_x = 22ms$, $\tau_y = 22ms$, $F_+ = 1$, and $F_- = 1$.

phenomenological model was proposed by M. Graupner et al. [35]. My work is based on this last model which we will refer as Calcium-based plasticity learning model. The advantage of this model is that it can explain a plethora of learning waveforms that are found in the different areas of the brain [61], these waveforms are shown in fig. 2.8 and include only depression (D), only potentiation (P) and a mix of them DP, DPD and PDP. The selection of one

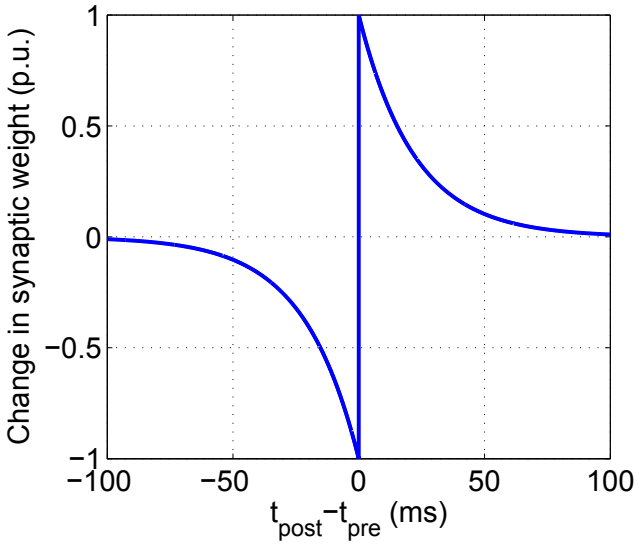


Fig. 2.6: Learning waveform $W(t_{pre} - t_{post})$ for the spike-based STDP rule, as observed an increase of synaptic weight is obtained when $t_{post} - t_{pre} > 0$, and a decrease for $t_{post} - t_{pre} < 0$. The waveform is normalized with respect to the maximum value of Δw_{ij} instead of the maximum w_{ij}. Parameter values here are the same than in Fig. 2.5.

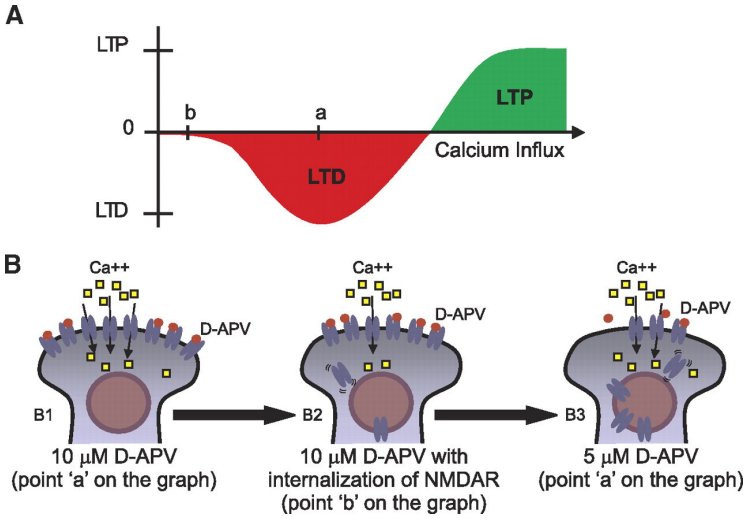


Fig. 2.7: Interaction between NMDA receptors and calcium scale of synaptic plasticity. In (a) large calcium influx generate potentiation, a reduction of this turns to depression and finally a small amount do not produces change. (b) Blocking NMDA receptors reduces the influx of calcium and induces depotentiation B1 and point a, in B2 the number of NMDA is reduced which reduces the influx of calcium and is expressed as a movement to point b, to get more depotentation, the concentration of nmda receptor antagonist must be decreased (B3) which increases the calcium flux (extracted from [60]).

of these waveforms depends on the assigned calcium variables which are explained below.

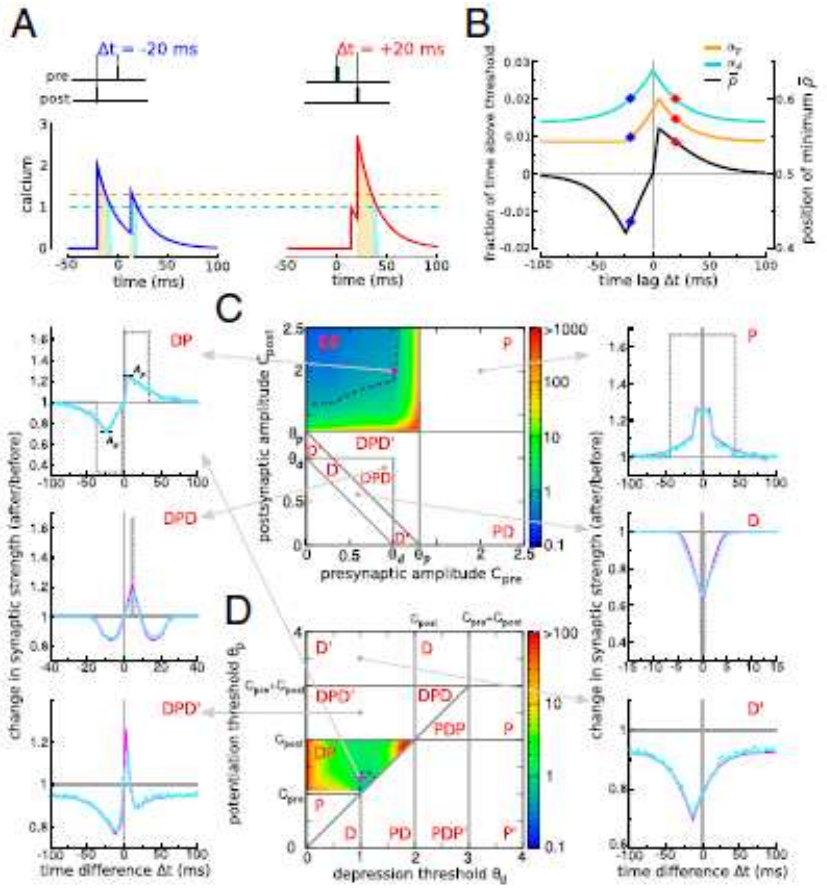


Fig. 2.8: Diversity of STDP curves in response to spike pair Stimulation (extracted from [35]). (A) Compound calcium transients evoked by a pair of pre-nad postsynaptic spikes for two values of Δt. (B)Fraction of time spent above the depression and potentiation. (C and D) The shape of STDP curves varies as a function of the pre- and postsynaptic calcium amplitudes C_{pre} and C_{post} (extracted from [35]).

The model describes the temporal dynamics of the synaptic efficacy w by using a first order differential equation as a function of potentiation and depression terms that depend on calcium concentration [62]. Pre- and post-synaptic pulses produce calcium peaks that decay exponentially in the absence of spikes. The long-term memory property is achieved through a bistability factor [63] which drives w to only two possible values. The temporal dynamics of the synaptic efficacy w are described by the following equation:

$$\begin{aligned} \tau\frac{dw}{dt} &= -w(1-w)(w_*-w) + \gamma_p(1-w)\mathcal{H}\Big[c(t)-\theta_p\Big] \\ &\quad -\gamma_d(w)\mathcal{H}\Big[c(t)-\theta_d\Big] + Noise(t), \end{aligned} \tag{2.12}$$

where τ is the time constant of the synaptic efficacy w, w_* is the boundary between the basins of attraction of the two stable states, γ_p and γ_d are the potentiation and depression coefficients respectively, θ_p and θ_d are the calcium potentiation and depression thresholds respectively, $\mathcal{H}$ denotes the Heaviside function, and $c(t)$ is the calcium concentration. Given that our main goal is comparing the theoretical and hardware synaptic dynamics, the noise term included in [35] is omitted here. The bistable term $-w(1-w)(w_*-w)$ pushes the synaptic efficacy w towards zero or one at a rate that depends on $1/\tau$ and w. When w is lower than w_*, w tends towards zero, and when w is greater than w_*, w tends towards one. The rate at which w moves towards these bistable states is approximately zero as w approaches toward zero, w_*, or one. w_* is an unstable fixed point, so noise or any significant synaptic activity will push w away from w_*. The second term $\gamma_p(1-w)\mathcal{H}\Big[c(t)-\theta_p\Big]$ implements LTP. The synaptic efficacy increases by a factor $\gamma_p(1-w)$ whenever the calcium concentration $c(t)$ is greater than the threshold θ_p. This factor approaches zero as the synaptic efficacy is close to one. The third term $\gamma_d(w)\mathcal{H}\Big[c(t)-\theta_d\Big]$ implements LTD. The synaptic efficacy decreases by a factor $\gamma_d(w)$ when $c(t)$ is above θ_d. This last factor approaches zero as the synaptic efficacy is close to zero.

The simulation results obtained for Graupner and Brunel in [35] are shown in Fig 2.8, where we shows the calcium dynamics for $t_{post} < t_{pre}$ and $t_{post} > t_{pre}$ in A, and the learning waveforms obtained as results of different parameter values such as C_{pre}, C_{post}, θ_p, θ_d. Classical STDP (Depression-Potentiation) is given when $c_{pre} < \theta_d < c_{post} < \theta_p$.

We designed a novel analog circuit based on the plasticity model proposed in [35], so as a first step we simplified the model to make it more suitable for hardware implementation. The original model imposes soft bounds on the weight by multiplying the potentiation and depression factors by $(1-w)$ and (w), respectively. We replaced these soft bounds with hard bounds [64], which more accurately model the behavior of CMOS circuits, because they often feature slewing behaviour up to a power rail. This simplification can be expressed as the following:

$$\begin{aligned} \tau\frac{dw}{dt} &= -k_{bs}w(1-w)(w_*-w)+\gamma_p\mathcal{H}\Big[c(t)-\theta_p\Big] \\ &\quad -\gamma_d\mathcal{H}\Big[c(t)-\theta_d\Big], \\ &\quad \begin{cases} w>1 \rightarrow w=1 \\ w<0 \rightarrow w=0 \end{cases} \end{aligned} \tag{2.13}$$

where k_{bs} is a constant that scales the bistability dynamics, and the hard bounds are implemented by the conditional expression. We compare this mathematical model with our circuit simulations in Sec. 4.1.1. With this equation, potentiation and depression are independent of the synapse value except near saturation values 0 and 1. Similarly, k_{bs} sets the bistability slope independently of τ, γ_p, and γ_d. The introduction of this variable allows a more direct comparison with the circuit design. k_{bs} can be in fact interpreted as the bistability slew rate set by the bias current of a wide range transconductance amplifier (see Sec. 4.1). Furthermore, the simplified dynamics for the calcium variable proposed in [35] are considered for the hardware implementation:

$$\frac{dc}{dt} = -\frac{c}{\tau_{Ca}} + C_{pre}\sum_i \delta(t - t_i - D) + C_{post}\sum_j \delta(t - t_j), \tag{2.14}$$

where c is the total calcium concentration, τ_{Ca} is the calcium decay time constant, C_{pre} and C_{post} are the pre- and post-synaptic calcium amplitudes, D is the delay in the response to pre-synaptic spikes, t_i and t_j are the pre- and post-synaptic spikes, and δ denotes the Dirac delta function. The temporal derivative of $c(t)$ is given by the sum of three terms. The first term $-\frac{c}{\tau_{Ca}}$ describes the decay of the calcium concentration. In the absence of pre- or post-synaptic spikes the solution of the differential equation is a decaying exponential function. The second term $C_{pre}\sum_i \delta(t - t_i - D)$ produces an instantaneous rise in the calcium variable after a time D from the time of occurrence of a pre-synaptic spike (t_i). The third term $C_{post}\sum_j \delta(t - t_j)$ produces an instantaneous rise in the calcium variable at the time of occurrence of a post-synaptic spike (t_j).

In Eq. 2.14, pre- and post-synaptic spikes are idealized as Dirac impulses. In physical systems, pulses have a duration, and particularly in analog circuits the pulse width is important because it can be used to define the time window in which a current flow occurs. This current usually charges a capacitor, and a saturation occurs if the time window is too long (the integrated voltage hits the supply rail). In the more realistic assumption of finite duration pulses, the calcium dynamics can be described as follows:

$$\begin{aligned} c(t) = C_{pre}\int_0^\infty F(s - D)\sum_i P_i(t - s - D)ds + \\ C_{post}\int_0^\infty F(s)\sum_j P_j(t - s)ds, \end{aligned} \tag{2.15}$$

where $F(t) = e^{-\frac{t}{\tau_{Ca}}}\mathcal{H}(t)$ models an exponential decay after one spike and can also be described as the solution of Eq. 2.14 for one post-synaptic spike with calcium amplitude 1, and $P_i(t) = \mathcal{H}\Big[t - t_i\Big] - \mathcal{H}\Big[t - (t_i + \Delta t_{pw})\Big]$ is the i-th pulse with duration Δt_{pw}. We compare this mathematical model with our circuit simulations in Sec. 4.1.1.

The simulation result for the simplified calcium-based model is shown in Fig. 2.9 which demonstrates that we can generate STDP by setting parameter values of an intermediate calcium variable in specific ranges. Here pseudo random pre- and post- spikes were generated to demonstrate the effect of plasticity due to timing (near 0.4s in the graphic) and frequency (near 0.8s in the graphic). In addition, the negative and positive slopes generated by the bistability are observed in the synaptic weight plot confirming that the system can only reach two possible values for long periods of time.

2.4 Discussion

I started this chapter by describing the computational model of the biological synapse; later I showed former experiments that demonstrate that synaptic strength can be modified by the pre- and post-synaptic spikes timing (STDP) as well as by the input spike rate. Finally, I presented a novel model obtained by simplifying the Calcium-based learning model proposed in [35] which is capable of reproducing timing and rate dependency in the synapse by implementing an intermediate calcium variable. This simplified version in addition to reproduce a plethora of learning rules depending on its parameters setup is also suitable for VLSI implementations.

This simplified calcium model uses hard bounds which are more accurately modelled by CMOS circuits; however, it is also important to highlight that the dynamics of the CMOS transistor depend on its operation point. For example in the case of a NMOS transistor in saturation, if we reduce V_{ds}, its current I_{ds} will slightly decrease (because of channel length modulation effect) and eventually the operation region will move to triode when $V_{ds} < 4U_T$ (for weak inversion). In this latter region a stepper decay in the current occurs when further reducing V_{ds}, reaching leakage current levels at $V_{ds} \approx 0$. As we will see in the next chapter, the synaptic strength is modified by sourcing or sinking a current flow through a capacitor by using a PMOS or NMOS as the bias device respectively.

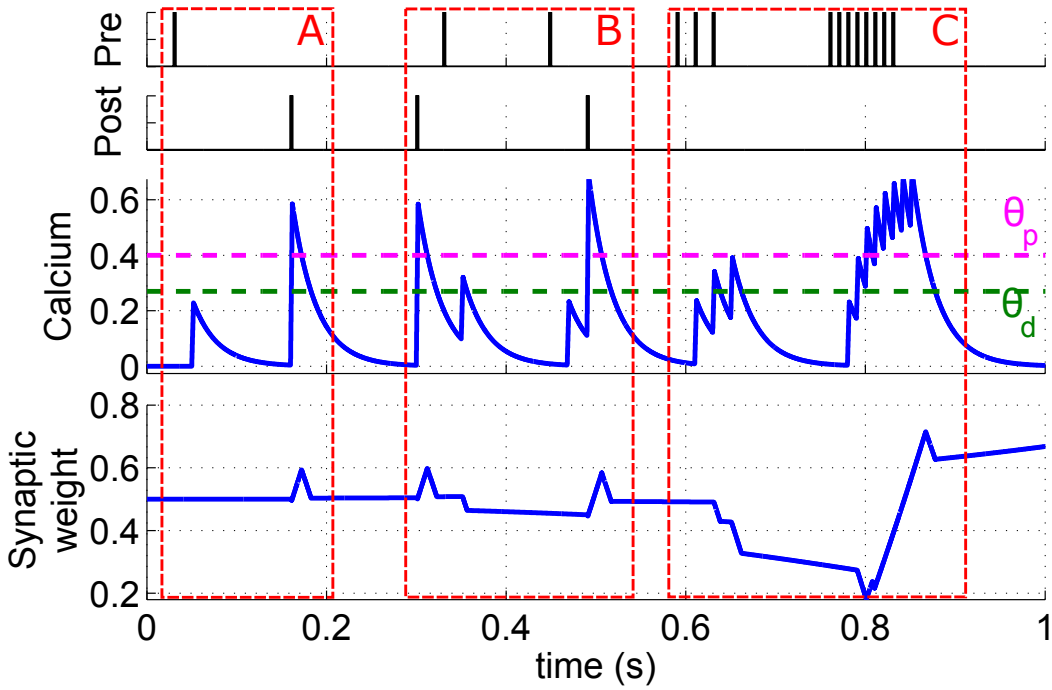


Fig. 2.9: Simulation results for the simplified calcium-based model programmed in Matlab. Pre- and post- spikes are generated pseudo randomly to show the effects of spikes timing and rate. When only a pre- spike appears, the calcium signal rises after a delay D and it does not reach any of the threshold θ_p and θ_d values (see graphic near the time $0.05s$). In the case of a only post- spike, the calcium signal rises instantly and reaches values above both thresholds, C_{post} and γ_p/γ_d are set in the way that the time that the calcium signal spends above θ_p and θ_d balance the total potentiation and depression without modifying the final synaptic strength (see graphic near the time $0.2s$). In the block (A) the pre- and post- spikes are far each other therefore they do not generate an overlap in the calcium signal and the synaptic weight remains constant. In the block (B) the pre- and post- spikes are enough close each other and their calcium signal components are overlapped, for the first pre-post pair the timing $t_{post} - t_{pre}$ is negative thus depression is generated in the synaptic weight, on the contrary in the last pair $t_{post} - t_{pre}$ is positive therefore generating potentiation. In the group (C) the effect of increasing the spike frequency is observed, for the first group the calcium signal remains above θ_d but below θ_p generating only depression, when the pre- spike frequency increases the calcium signal rises to values higher than the threshold θ_p generating more potentiation than depression. Bistability effect occurs at any synaptic weight value, in the graphic its effect is more noticeable for synaptic weight values around 0.3 or 0.7 (see graphic around $0.8s$), for lower values than 0.5, the synaptic goes down slowly to 0, and for values higher than 0.5 it goes up to 1.

In order to compare the soft and hard bounds effect in the synaptic modification lets define the synaptic strength as w in a range of $[0-1]$. In the case of soft bounds when potentiation or depression occurs, w is modified as $w \rightarrow w + q_+(w)$ or $w \rightarrow w - q_-(w)$ respectively, where the step size q_+

and q_- are variable and defined as $q_+(w) = \alpha(1 - w)$ and $q_-(w) = \alpha w$. On the other hand, in the case of hard bounds we define a constant step size for potentiation and depression independent on the value of synaptic state, then $q_+(w) = q_-(w) = \alpha$. A more detailed explanation of bounded synapses is described in [64].

By comparing the previous bounds definitions with the CMOS dynamics, we can argue that using soft bound near the power rails ($|V_{ds}| < 4U_T$) resembles a triode region operation in CMOS; however, for the most of the swinging range, the operation is reasonably approximated to hard bounds.

3 Neuromorphic Circuits Blocks

I start this chapter by giving a theoretical analysis of the CMOS transistor transfer function followed by its detailed characterization for the technology used in this work. This provides a solid background for the following sections, in which I use the obtained results such as threshold voltage, current values for weak and strong inversion and leakage current levels to implement the desired waveforms.

The MOS transistor consists of a low lightly doped semiconductor called body substrate (p-type for NMOS and n-type for PMOS) and two high doped semiconductors called drain and source diffusions implanted in the substrate (n-type for NMOS and p-type for PMOS) separated by a distance L. Throughout this distance an insulator with oxide thickness (t_{ox}), width W and length L is formed above the substrate. A polysilicon layer is grown above the oxide generating a fourth terminal called gate.

In order to describe the physical effects in CMOS transistors some definitions are stated in this paragraph. The flatband voltage (V_{fb}) is defined as the required gate-body voltage (V_{gb}) to keep the semiconductor everywhere neutral by cancelling the effects of the contact potentials and the parasitic charge. The surface potential (ψ_s) is defined as the total potential drop across the region defined from the surface to a point in the bulk outside the depletion region as depicted in Fig. 3.1 [65].

If V_{gb} increases above the V_{fb}, the total charge on the gate (Q_g) becomes positive[1] and this difference in the charge on the gate per unit area (Q'_g) [2] is balanced by a negative change in the charge of the semiconductor under the oxide per unit area (Q'_c). This positive change in V_{gb} causes also an increase in the potential drop across the oxide (ψ_{ox}) and ψ_s. For values slightly higher than V_{fb}, the holes are driven away from the surface leaving a depletion region. As V_{gb} increases further, ψ_s becomes sufficiently positive to attract a significant number of free electrons to the surface and eventually, with a

[1] To be more precise a slight parasitic charge (Q_0) which is mainly located at the oxide-semiconductor interface should also be considered. At flatband condition Q_g=-Q_0, however for the matter of simplicity we neglect this term here.

[2] An apostrophe (′) after Q_x denotes charge per unit area.

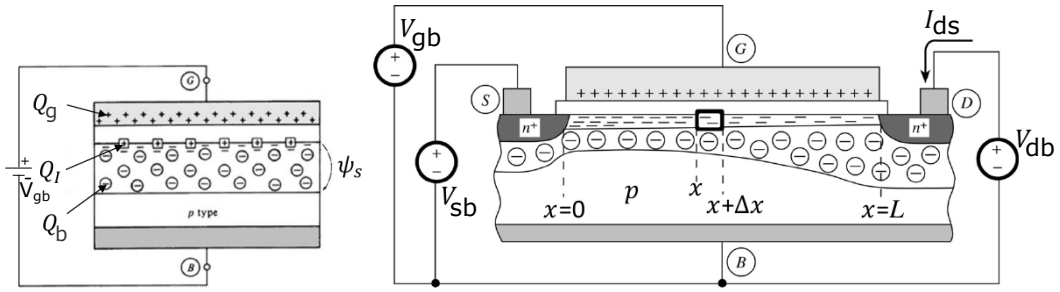


Fig. 3.1: MOS terminal structures in inversion region (extracted from [66]). (a) A MOS two-terminal (MOS capacitor), if V_{gb} is higher enough inversion (charge Q_I) and depletion (charge Q_b) regions are formed, this leads to a potential ψ_s between the surface and a point outside the depletion region. (b) A four-terminal MOS transistor, the added drain (D) and source (S) terminals unbalance the $Q_b^{'}$ and $Q_I^{'}$ charges and the potential ψ_s with the distance, the depletion region formed by the reverse bias V_{ds} is greater than the one formed by V_{sb} since $V_{dg} > V_{sb}$.

sufficiently high V_{gb} voltage, the density of electrons can exceed the one of holes at the surface.

The inversion layer in MOS transistors is created when V_{gb} is high enough to repel majority carriers in the substrate and even attract minority carriers to its surface giving as a result a path of same carriers that allows current flow between the substrate surface and the drain and source terminals. Here $Q_c^{'}$ consists of the deplexion charge $Q_b^{'}$ plus the inversion $Q_I^{'}$ charge. The inversion region is divided into weak, moderate and strong inversion. In weak inversion practically all the charge below the oxide is due to $Q_b^{'}$ and the surface potential reaches smaller values than two times the Fermi potential (ϕ_F) which allows to approximate $Q_I^{'}$ to a exponential function of ψ_s. In strong inversion the surface potential ψ_s is assumed to be constant (independent of V_{gb} values); this leads to a simplified linear relationship between $Q_I^{'}$ and V_{gb}. In moderate inversion none of the previous simplifications is valid; therefore, the $Q_I^{'}(V_{gb})$ is neither a straight line nor an exponential [66].

A similar characteristics to the exponential function of the transistor in weak inversion configuration is found in the ions conductance of neuron cells with respect to its membrane potential [44]. Therefore, this configuration is suitable in the neuromorphic field to mimic conductance dynamics [67]. Furthermore when CMOS are operated in the subthreshold domain they draw small currents producing low power consumption. The disadvantages of the subthreshold region are that mismatch effects are stronger and transistors cannot operate at high frequencies [68].

3.1 CMOS Operation in Inversion Region

For a better understanding of MOS transistor operation, first a two-terminal structure usually named MOS capacitor (Fig. 3.1a) is explained and then the results are extended to a four terminals structure which is the real fabricated device in silicon (Fig. 3.1b). In this latter, the current flow is derived as a function of the voltages in its terminals.

Mathematically the charge $Q_c^{'}$ as a function of ψ_s is calculated by using the Poisson's equation considering dopant ions and electron contribution in the surface (for the general derivation refer to [66]). The expressions for $Q_c^{'}$, $Q_b^{'}$ and $Q_I^{'}$ in inversion region are given in Eqs. 3.1, 3.2 and 3.3, where $q = 1.602 \times 10^{-19} C$ is the electron charge, $\epsilon_s = 1.05 \times 10^{12} F/cm$ is the permittivity for silicon, N_A is the acceptor concentration ($\approx 10^{17} - 10^{18} cm^{-3}$), U_T is the thermal potential ($U_T = 25mV$ at room temperature) and γ is the body effect coefficient.

$$Q_c' = -\sqrt{2q\epsilon_s N_A}\sqrt{\psi_s + U_T e^{\frac{\psi_s - 2\phi_F}{U_T}}} \tag{3.1}$$

$$Q_b' = -\sqrt{2q\epsilon_s N_A}\sqrt{\psi_s} \tag{3.2}$$

$$Q_I' = Q_c' - Q_b' = -\sqrt{2q\epsilon_s N_A}\left(\sqrt{\psi_s + U_T e^{\frac{\psi_s - 2\phi_F}{U_T}}} - \sqrt{\psi_s}\right) \tag{3.3}$$

$$\psi_s \approx \left(-\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} + V_{gb} - V_{fb}}\right)^2, \text{ weak inversion} \tag{3.4}$$

$$\psi_s \approx \phi_0 = 2\phi_F + \Delta\phi, \text{ strong inversion} \tag{3.5}$$

These equations are plotted as a function of ψ_s in Fig 3.2. For ψ_s values lower than $2\phi_F$ practically all the surface charge is caused by the charge in the depletion region. As ψ_s increases above $2\phi_F$, $|Q_I^{'}|$ starts to become significant and strong inversion takes place from $\phi_0 = 2\phi_F + \Delta\phi$, ($\Delta\phi$ is considered a constant between $5U_T$ to $6U_T$). In weak inversion region $\mid Q_I' \mid << \mid Q_b' \mid$ is considered and Q_b' is approximated to a constant value along the channel. In strong inversion the depletion region charge is assumed to have reached a maximum value of $Q_{b0}^{'} = -\sqrt{2q\epsilon_s N_A}\sqrt{\phi_0}$.

By adding the drain and source terminals to the MOS capacitor and connecting them to different voltages as depicted in Fig. 3.1b, the attractiveness of

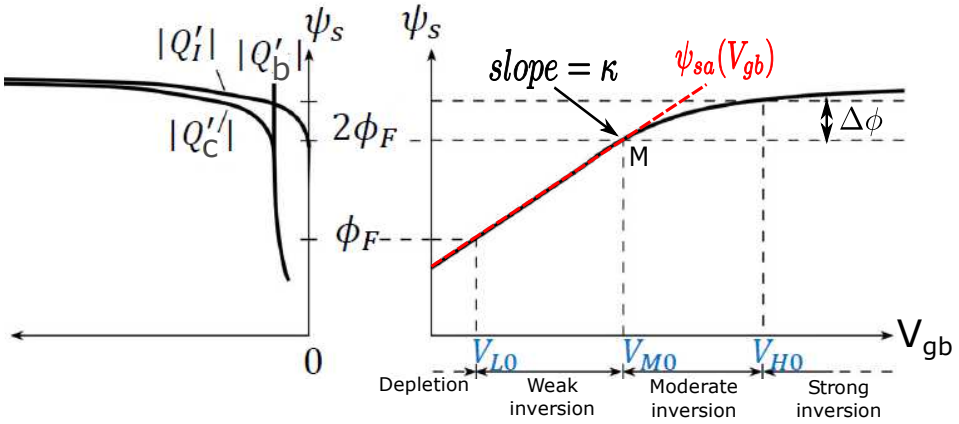


Fig. 3.2: Surface potential ψ_s and charges $Q_I^{'}$, $Q_b^{'}$ and $Q_c^{'}$ vs. gate-body voltage V_{gb} for two-terminals MOS transistor. For low ψ_s values a considerable change $\Delta\psi_s$ is required to modify $Q_c^{'}$. For high ψ_s values a slight $\Delta\psi_s$ generates considerable variation in $Q_c^{'}$ due to the steepness of $Q_I^{'}$. In weak inversion the slope $d\psi_s/dV_{gb}$ is nearly constant and in strong inversion ψ_s is practically constant (extracted from [66]).

the surface for the electrons depends on how large ψ_s is in comparison to V_{cb}, where $V_{cb}(0) = V_{sb}$ and $V_{cb}(L) = V_{db}$; this leads to a vertical shift of ψ_s in Fig. 3.2 (right plot) by V_{cb}. In addition, a horizontal field component (much smaller than the vertical one) is generated.

The current flow between the drain and source terminals in inversion region consist of drift and diffusion components as stated in Eq. 3.6. To obtain a simple explicit solution for I_{ds}, ψ_{sa} is approximated by a linear function in weak inversion where source (source-referenced model) [69, 70] or body (body-referenced model) [71] are chosen as the reference voltage. In the case of strong inversion, ψ_{sa} is considered constant and the source-referenced model is generally used.

$$I_{ds} = \frac{W}{L}\left[\underbrace{\int_{\psi_{s0}}^{\psi_{sL}} \mu(-Q_I^{'})d\psi_s}_{drift} + \underbrace{U_T \int_{Q_{I0}^{'}}^{Q_{IL}^{'}} \mu dQ_I^{'}}_{diffusion}\right] \tag{3.6}$$

In weak inversion, due to the constant depletion region, the electric field has a zero horizontal component so no drift current is generated. Since $\phi_s < 2\phi_F$, the Eq. 3.3 can be simplified to an exponential function of ψ_{sa}. The ψ_{sa} expression given in Eq. 3.5 is also approximated to a linear function; one approach is to expand this function from the point M (top weak inversion limit) which leads to a more precise model as explained in [66]. Other approach is to expand ψ_{sa} from $V_{gb} = 0$ [71]; this last approximation is more

commonly assumed in the neuromorphic field. With this two simplifications $Q_I^{'} = Q_M^{'} e^{\frac{\kappa_M (V_{gb} - V_{mb})}{U_T}}$ for expansion in M or $Q_I^{'} = Q_0^{'} e^{\frac{(\kappa_0 V_{gb} - V_{sb})}{U_T}}$ for expansion in $V_{gb} = 0$. The solution of Eq. 3.6 with the simplified charge function using the body-referenced model and expanding ψ_{sa} linearly from $V_{gb} = 0$ gives the explicit mathematical expression for I_{ds} shown in Eq. 3.7, where $I_0 = qDN_0 e^{-q\frac{\phi_0}{KT}} \frac{W}{L}$ is the leakage current when $V_{gs} = 0$. Saturation region is considered if $V_{ds} > 4U_T$, this condition leads to the simplified Eq. 3.8 where the term $e^{\frac{-V_{ds}}{U_T}}$ is neglected.

$$I_{ds} = I_0 e^{\frac{kV_{gb} - V_{sb}}{U_T}} \left(1 - e^{\frac{-V_{ds}}{U_T}}\right), \tag{3.7}$$

$$I_{ds} = I_0 e^{\frac{kV_{gb} - V_{sb}}{U_T}} \tag{3.8}$$

In strong inversion Eqs. 3.3 and 3.5 are rewritten including drain and source voltages as $\psi_s(x) = \phi_0 + V_{cb}(x)$ and $Q_I' = -C_{ox}^{'}(V_{gb} - V_{cb}(x) - V_{T0})$. Thus, the drain current is assumed to be due to drift. Using the source-reference model, the solution of Eq. 3.6 results in Eq. 3.9

$$I_{ds} = \mu C_{ox} \frac{W}{L} \left[(V_{gs} - V_{th}) V_{ds} - \frac{1}{2} V_{ds}^2 \right], \tag{3.9}$$

where

$$V_{th} = V_{th0} + \gamma \left(\sqrt{|\phi_0 + V_{sb}|} - \sqrt{|\phi_0|} \right) \tag{3.10}$$

is the threshold voltage with $V_{th0} = V_{FB} + \phi_0 + \gamma\sqrt{\phi_0}$. If V_{ds} is slightly greater than $V_{gs} - V_{th}$, then the inversion layer stops at a distance $x \leq L$ (pinch-off) so Eq. 3.9 is taken until $L - \Delta L$ where $V_{ds} = V_{gs} - V_{th}$. A firs-order relationship between $\Delta L/L$ and V_{ds} is commonly assumed giving as result the Eq. 3.11; this effect is called channel-length modulation where λ is its coefficient [70]. Another important effect occurs when the source voltage is different from the substrate; if the gate voltage increases from an initial value equal to the substrate, the depletion region in the body also increases ($Q_b^{'}$). Therefore, in order to get the initial $Q_b^{'}$, the voltage in the gate should increase; this phenomenon is called body-effect and it is represented as an increase of the threshold voltage as stated in Eq. 3.10 where γ is the representative coefficient. The terms μC_{ox} and $\mu C_{ox} \frac{W}{L}$ are usually referred as K and β respectively.

$$I_{ds} = \frac{1}{2} \mu C_{ox} \frac{W}{L} (V_{gs} - V_{th})^2 (1 + \lambda V_{ds}) \tag{3.11}$$

3.2 MOSFET Characterization

In order to use the deduced model equations of the previous section (quadratic and exponential I_{ds} functions), appropriate parameters values need to be estimated; the procedure to obtain this data is called characterization. In the previous equations, some parameters do not have an exact theoretical value i.e. ϕ_0, V_{th} and γ, so the most suitable values for them depend on the desired operation point. In a good model the parameter values are reasonable close to the physical ones. The technology used in our circuits implementation is AMS $0.18\mu m$; therefore, the characterization has been performed in this node.

Differences between the model and real data occur because ideal situations were considered in the model i.e. the previous equations were obtained for constant doping substrate, however small technologies like $0.18\mu m$ uses halo implants which creates a higher doping concentration in the substrate near the source and drain which decreases the mobility in these areas, inaccurate flatband voltage and/or empirical parameters assumption and approximation also leads to mismatch errors. However, in computational models a constant lateral doping is assumed. Another discarded effect is that the transistor dimensions' W and L are the electrical channel width and length, which are slightly smaller than the corresponding layout versions.

In order to set the parameter values, an optimization process is used to minimize the error between modelled and measured values. A common approach is to minimize drain current errors through least mean square error fit (linear regression), although more complex approaches using least mean square error fit with weighting coefficients are used in computational models. Since wide and long transistors are closer to the ideal behaviour, they are generally used to extract the parameters. Another option is to extract parameters from transistor's dimensions and biases that are important for the circuit operation (in our case $W/L = 1\mu m/0.5\mu m$) [66].

The Eqs. 3.7- 3.11 are useful for hand calculation; however, in computational models equations are more complex [72,73]. A first step to set the parameters is to define the inversion regions. The onset of weak, moderate and strong inversion are denoted by V_L, V_M and V_H in Fig. 3.3(d). Despite V_{th} appears in the strong inversion, the MOS transistor is not in strong inversion at $V_{gb} = V_{th}$, the beginning of strong inversion is taken after V_{th} ($\approx 2\%$ error between the

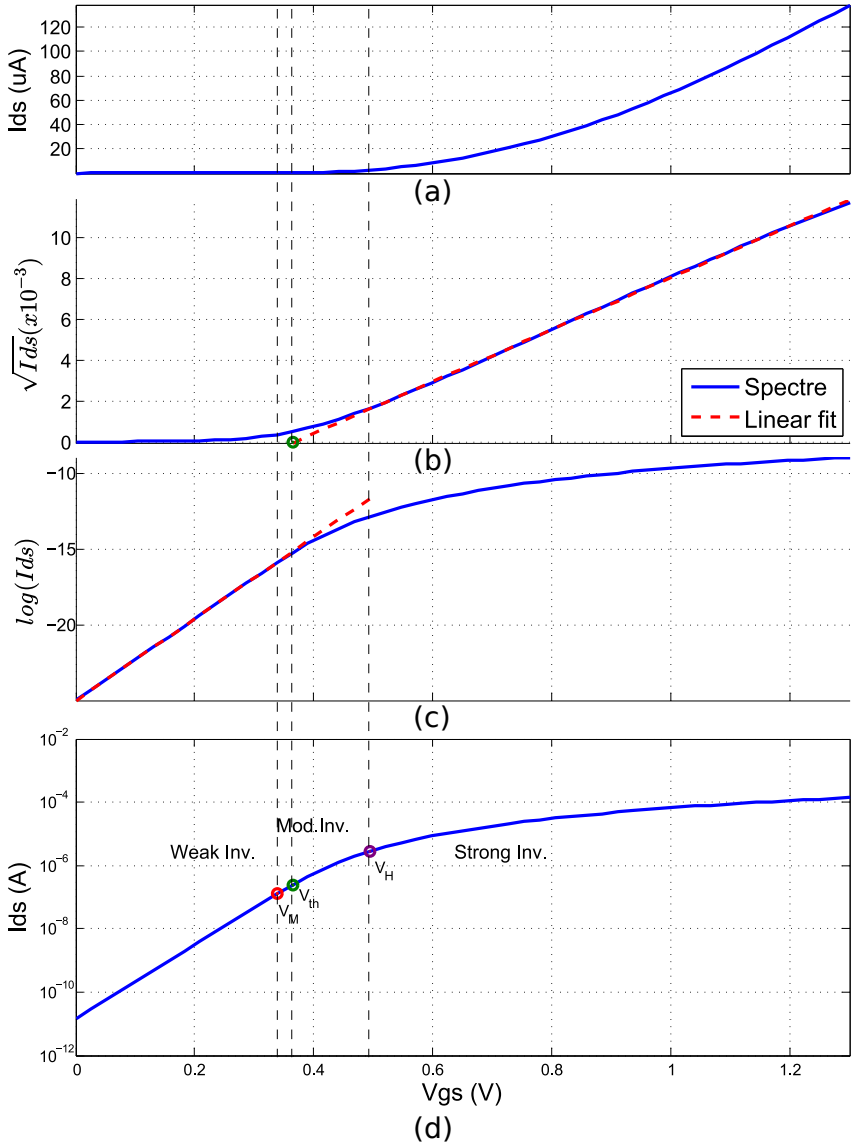


Fig. 3.3: NMOS inversion limits and characterization for $V_S = V_B = 0$, $V_D = 1.3V$ and $W/L = (1/0.5)\mu m$, results obtained for strong inversion are $V_{th} = 0.364V$, $K = 1.63 \times 10^{-4} A/V^2$; in weak inversion $I_0 = 14.1pA$ and $\kappa = 0.7$, inversion limits are $V_H = 0.494V$ and $V_M = 0.338V$. (a) I_{ds} takes off in strong inversion in which the current has values from $\approx 3.2\mu A$. (b) In strong inversion, the parameters values are obtained with a linear regression on $\sqrt{I_{ds}}$ from $V_{gs} = V_H$, the intersection point of the regression with the origin determines the threshold V_{th} value, (c) In weak inversion the linear regression is evaluated with $log(I_{ds})$ until $V_{gs} = V_M$. (d) V_M and V_H are calculated when the regression and simulation data start to diverge, logarithmic plot of Ids is useful to observe current values in weak inversion.

regression and data in Fig. 3.3(b) [69]). In weak inversion I_{ds} has low values and is exponential dependent on V_{gs}; thus, $log(I_{ds})$ depends linearly on V_{gs}. By using a linear regression from $V_{gs} = 0$ until V_M, parameters κ and I_0 are obtained as shown in Fig. 3.3(c).

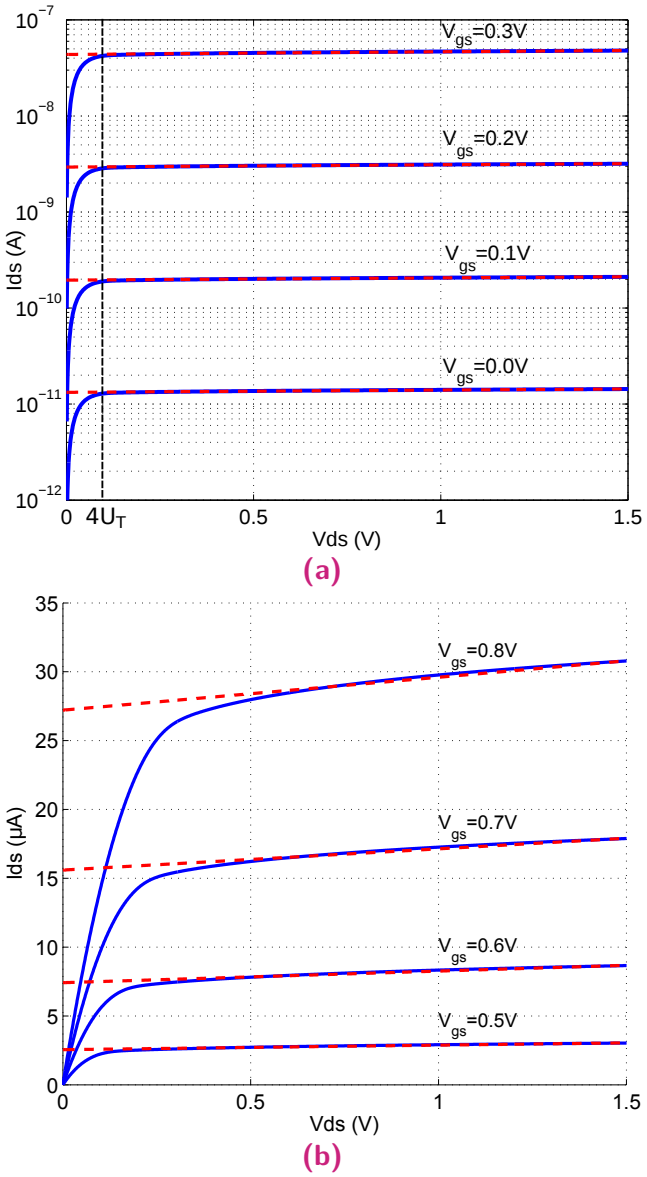


Fig. 3.4: Channel length modulation for (a) weak inversion and (b) strong inversion. NMOS parameters are: $(W/L) = (1\mu m/0.5\mu m)$. Characterization results are shown in Table 3.1.

One approach to calculate the threshold voltage is setting V_{gs} to a high value so that the transitor is configured in strong inversion and V_{ds} (for $V_{sb} = 0$) should be set to a small value to considerably reduce the effective mobility. Under these assumptions Eq. 3.9 can be simplified to $I_{ds} = \frac{W}{L}\mu_0 C_{ox}(V_{gs} - V_{th})V_{ds}$ and the extrapolation of the linear fit for I_{ds} in $I_{ds} = 0$ gives the V_{th} value (more precisely the value $V_{th} - 0.5V_{ds}$). Another technique is plotting $\sqrt{I_{ds}}$ vs. V_{gs} in strong inversion and saturation (Eq. 3.11) which predicts also a straight line. Since triode is less common than saturation operation, the second technique is preferred; however, the drawback is that in saturation, V_{th} decreases considerable for short channel devices due to drain-induced barrier lowering [74] and velocity saturation [75]; therefore, a general V_{th}

Table 3.1: Approximate ideal current $I_{ds} = I_{sat}$ and channel-length modulation coefficient λ for Fig. 3.4. V_{gs} voltages around $0.3 - 0.5V$ belong to moderate inversion therefore they do not follow a cuadratic nor exponential waveform. For weak inversion the channel-effect is almost negligible.

Inversion	$V_{gs}(V)$	$I_{sat}(A)$	$\lambda(V^{-1})$
Weak	0	12.9p	0.0577
	0.1	190p	0.0577
	0.2	2.86n	0.0586
	0.3	42.4n	0.0692
Strong	0.5	2.51μ	0.131
	0.6	7.43μ	0.113
	0.7	15.8μ	0.0974
	0.8	27.9μ	0.0857

value without considering slight variations because of physical dimensions or voltages configuration is obtained in triode while a specific threshold is estimated for saturation in Fig 3.3(b).

Two second order effects, "channel-length modulation" and "body effect", are characterized in Figs. 3.4 and 3.5, and Table 3.1 describes the approximate ideal current and the channel-length coefficient. In order to obtain these values, linear regressions in saturation region for different V_{gs} voltages were calculated where the raw data was gotten from the spectre simulator. The ideal $I_{ds} = I_{sat}$ was calculated in the operation point $V_{ds} = V_{gs} - V_{th} + \Delta V$, where ΔV is a small value that ensures saturation region with an approximated linear function in the current I_{ds}.

3.3 Mismatch

Due to the nature of fabrication process, a statistical variation in the CMOS transistor operation is expected. Mismatch analysis allows to estimate this by calculating CMOS parameters fluctuation. Pioner research in [76] deduced that the main source of mismatch in the threshold voltage V_{th} was the variation in the depletion charge density $\Delta Q_b'$; likewise, variations in the dimensions, channel mobility and gate oxide capacitance were related to mismatch in β. Lakshmikumar et al. also proposed the classical $\sqrt{WL}$ relationship for V_{th} and I_{ds}. Pelgrom et al. [77] included body effect coefficient, and based on Fourier analysis they described the relationship of $\sigma(V_{th})$ and

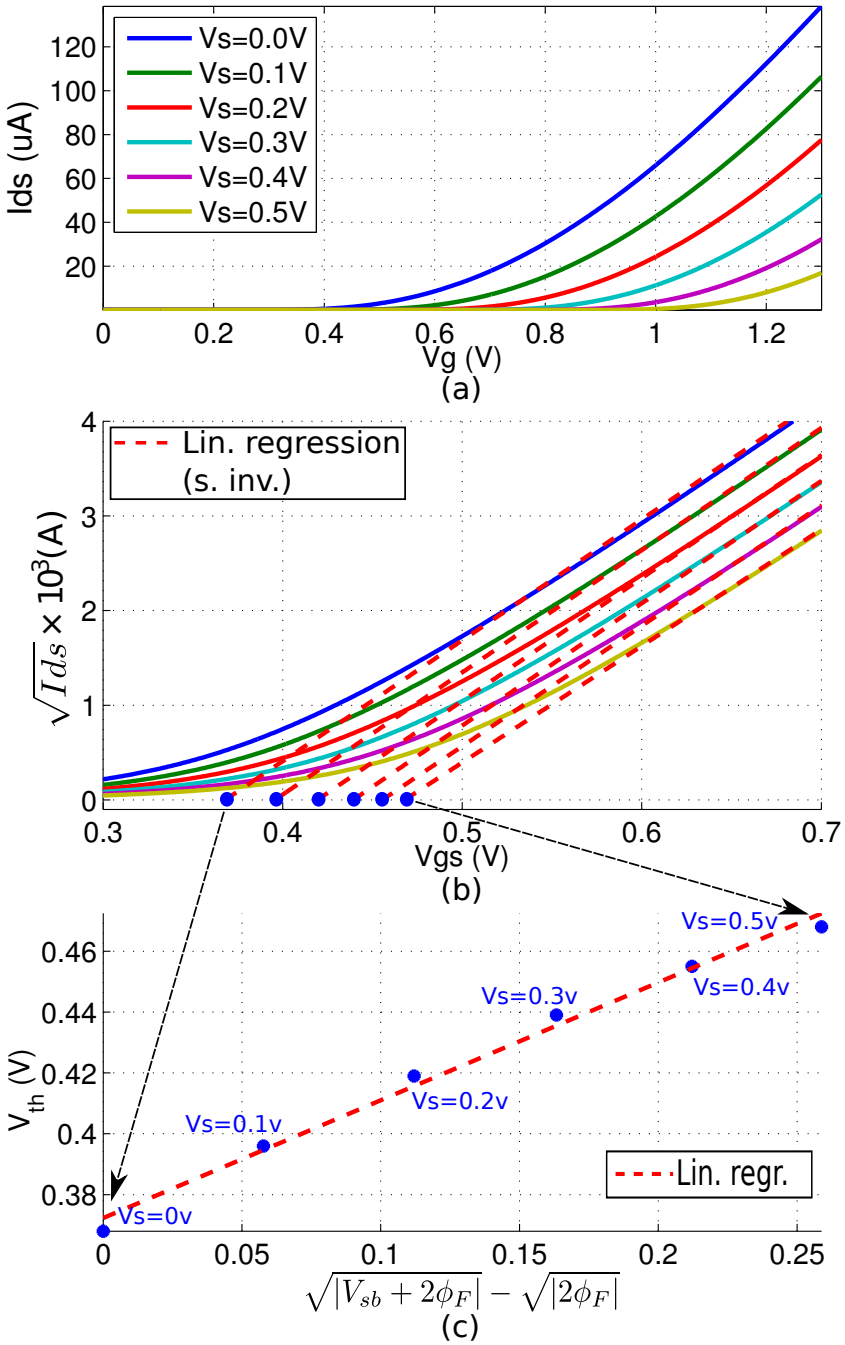


Fig. 3.5: Body effect as result of variations of V_{sb}, as observed V_{th} increases when V_{sb} also increases, the obtained results are $V_{th0} = 0.37V$ and $\gamma = 0.39V^{1/2}$, the parameter values for V_D, V_B and (W/L) are the same than in Fig. 3.3, additionally $\phi_F = 0.35$. (a) Ids is shown for different values of V_g and V_s (data obtained with spectre simulation). (b) Previous data in (a) was processed to show $\sqrt{Ids}$ as function of V_{gs}, then we apply linear regression to the points in the curve that are in strong inversion to obtain the threshold points for each V_{sb}. (c) The obtained threshold points in (b) are plotted as function of $\sqrt{|V_{sb}+2\phi_F|}+\sqrt{|2\phi_F|}$ to obtain the approximated parameters of body effect γ and V_{th0}.

$\sigma(I_{ds})$ with $\sqrt{WL}$ and the distance between transistors as stated in Eqs. 3.12 and 3.13.

$$\sigma^2(\Delta V_{th}) = \frac{A_{V_{th}}^2}{WL} + S_{V_{th}}^2 D^2 \tag{3.12}$$

$$\left(\frac{\sigma\Delta\beta}{\beta}\right)^2 = \frac{A_\beta^2}{WL} + S_\beta^2 D^2 \tag{3.13}$$

where $A_{V_{th}}$, A_β are the area proportionality constants for parameters V_{th} and β. $S_{V_{th}}$ and S_β describe the variation of parameters V_{th} and S_β with the spacing. All the previous constants are process-related.

The most important contribution to $A_{V_{th}}$ is the fluctuation number of doping atoms in the depletion layer ($A_{Vth}\,\alpha\,\Delta Q_B'/C_{ox}$); therefore, this coefficient can be reduced by decreasing t_{ox}; on the other hand, a higher substrate doping level leads to a larger $A_{V_{th}}$. $\Delta\beta$ is related mainly to the mobility variation. The reduction of A_{vth} by scaling down the technology provides better matching for devices considering the same dimensions (W/L); however, this advantage is vanished due to power supply scaling. When considering minimal size device of technology nodes, the transistor area is reduced quadratically with the feature size while the reduction in $A_{v_{th}}$ is only linear; therefore, the matching of the the minimal size device (W/L) degrades with scaling [78]. Despite V_{th} and β share some common process parameters, experimental data shows a low correlation between them and is normally accepted as independent random variables [79] although other results argue that neglecting this correlation can lead to an overestimated factor as large as two [80].

Previous mismatch equations were obtained based on transistors in strong inversion and then extending the same relationships to weak inversion, so it is expected that Eqs. 3.12 and 3.13 are less precise outside strong inversion. A mismatch model that consider all regions operation was proposed in [78] deriving Eqs. 3.14 and 3.15. Parameters definition and approach are based on [81].

$$\left(\frac{\sigma\Delta I_{ds}}{I_{ds}}\right)^2 = \left(\frac{\sigma(\Delta\beta)}{\beta}\right)^2 + \left(\frac{g_m}{\overline{I_{ds}}}\right)^2 \sigma^2(\Delta V_{th}) \tag{3.14}$$

$$\sigma^2(\Delta V_{gs}) = \sigma^2(\Delta V_{th}) + \frac{1}{(g_m/\overline{I_{ds}})^2}\left(\frac{\sigma(\Delta\beta)}{\beta}\right)^2 \tag{3.15}$$

where gm is the transistor transconductance, $\overline{I_{ds}}$ is the mean I_{ds} of the sampled data

These expressions provide insight to get error values when designing bias transistors in circuit blocks. In the case of voltage biased pair (current mirror), the ΔI_{ds} error is obtained; and in current bias pair (differential pair), the ΔV_{gs}.

In the case of wide transistors, the main cause of mismatch is the high channel dopant concentration compared to the bulk dopant concentration which defines the threshold voltage of the transistor, so neglecting distance effect and for typical bias voltages, mismatch in threshold voltage V_{th} is greater than β mismatch. Considering a bias value $(g_m/I_{ds})_x$ where the condition $(g_m/I_{ds}) = A_\beta/A_{Vth}$ is fulfil, this leads to a equal mismatch contribution of the factors $\sigma(\Delta\beta/\beta)^2$ and $(g_m/I_{ds})^2\sigma^2(\Delta V_{th})$; therefore, using bias values higher than $(V_{gs} - V_{th})_m$ provides small mismatch; however, $(V_{gs} - V_{th})_m$ is limited by the power supply. This effect is explained when the gate overdrive voltage $V_{od} = V_{gs} - V_{th}$ increases; in this case the parameters that affect V_{th} have less impact on the I_{ds} mismatch [78].

Fig. 3.6 shows mismatch simulation results of a nMOS transistor for technology AMS $0.18\mu m$ considering the effect of local and global sources. A single transistor with different geometry dimensions was configured to bias voltage and current. As expected for wide and long channel and high voltage/current mismatch decreases. The mismatch is reduced considerable for $W > 1\mu m$ and $L > 1\mu m$.

By considering the type of CMOS transistor, the lower mobility of pMOS requires larger $|V_{gs}|$ values to generate the same reference current, therefore pMOS provides less mismatch than nMOS in the case of current bias. However, in the case of voltage, it is not possible to set a relationship.

Most mismatch models are based on simple MOS transistor equations (level 1 model) in the saturation region, so it is arguable that its extension to other regions can lead to considerable error. In addition, smaller technologies than $0.18\mu m$ uses halo implant ions which modifies original length and width channels dimensions for effective ones. A more complex model that deals with these cases (proposed in [80, 82]) is compatible with SPICE and can be used for different geometry and inversion region conditions, including

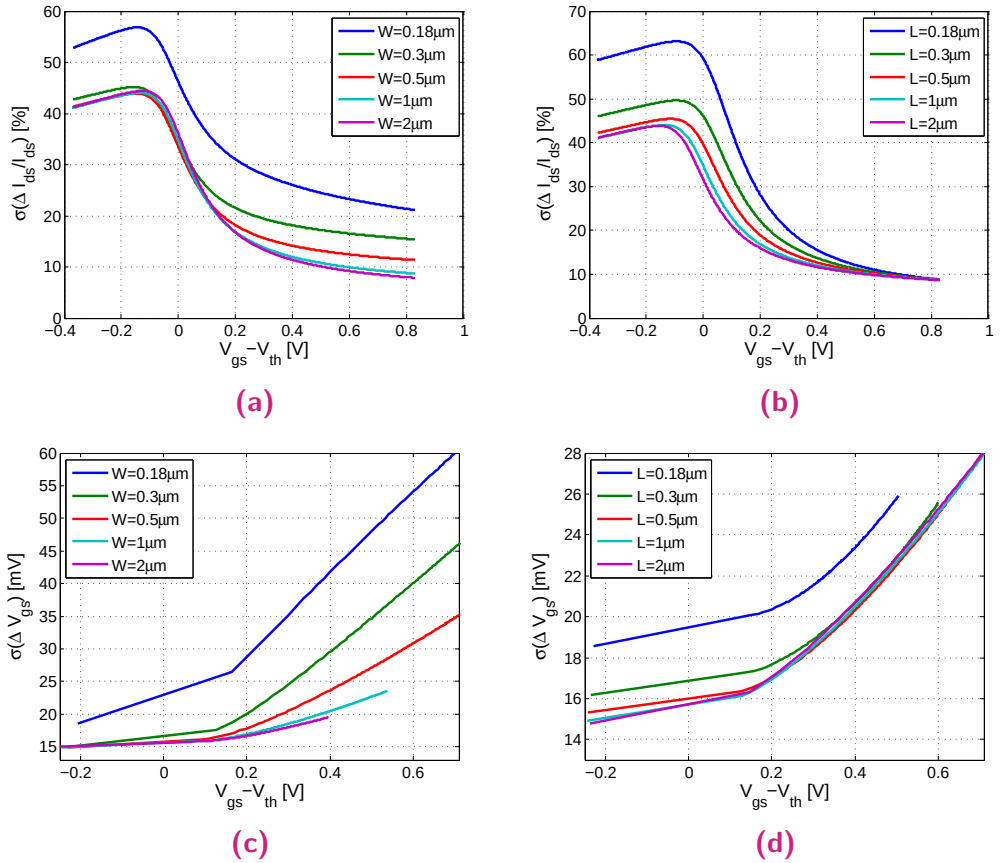


Fig. 3.6: Mismatch estimation for a single transistor configured as bias device using AMS $0.18\mu m$ [78]. (a)-(b) Voltage biasing where transistor variations result in variations in the drain-source currents, figures plot $\sigma(\Delta I_{ds}/I_{ds})$ vs. V_{gs} for $L = 1\mu m$ and $W = 1\mu m$ respectively. (c)-(d) Current biasing where transistor variations result in variations in the gate-source voltages, figures plot $\sigma(\Delta V_{gs})$ vs. V_{gs} for $L = 1\mu m$ and $W = 1\mu m$ respectively. Simulations were obtained using Virtuoso ADEXL with Montecarlo Analysis for variation in process and mismatch, random sampling method with 200 number of points was used. For voltage biasing the transistor was connected to $V_{ds} = 1.3V$ to ensure saturation operation and V_{gs} was swept from 0 to 1.2V. For current biasing the transistor source terminal was connected to an ideal current source swept from 0 to $25\mu A$ for $L = 1\mu m$ and from 0 to $90\mu A$ for $W = 1\mu m$. Simulation results show that by increasing the channel length/width, the mismatch is reduced.

phenomena such as source/drain series resistance, body bias effects, short and narrow channel effects, mobility degradation and graded-channel effects.

Variations in the current I_{ds} are modelled considering local and global error sources. Local parameters have short correlation with the distance; therefore, the error sources are considered independent (no correlation between them) and its variance depends on the transistors size (W and L). Among the process that affect its mismatch are ion implantation, oxide growth

and lithography [83]. Interdie variation is simulated by using montecarlo analysis.

For global error sources (long correlation distance), parameters variation are related to device fabrication steps that occur radial to the wafer such as gate oxide growth and polysilicon etching. Therefore, placing i.e. two transistors in the direction of the parameter gradient increases the mismatch effect; on the other hand, if the transistor pair is placed orthogonally, the global error source does not contribute to mismatch. However, we can not predict the direction so a uniform placement simulates situations in which the position of the die on the wafer is unknown, and in that case the variance can be simplified to a quadratic relationship with the distance between transistors as in Eqs. 3.12 and 3.13 [83]. Die to die variations is simulated by corner analysis using best and worst cases.

In addition to process variables, systematic mismatch (deterministic variation) is produced depending on the layout style. Therefore, it is important to consider layout techniques like cross couple, strip pair, adding dummy devices and place devices with same orientation [84]. Alternative techniques like interdigitated waffle, common centroid and finger demonstrate considerable decrease of mismatch [85]. Another strategy to consider is using transistor multipliers or fingers; in this case the process parameter variance component increases by a factor of n given that each MOSFET has its own local parameter variation; on the other hand, the squared density decreases by a factor of n^2 because each device has less impact of the current; consequently, the total $\sigma_{I_{ds}}$ decreases by a factor of $\sqrt{n}$ [80]. An additional systematic mismatch effect is produced due to the different thermal expansion coefficients between the substrate and the silicon die (strain mismatch) given that the substrate has higher thermal expansion coefficient than the die which generates stress and strain on the substrate bending the structure; this mechanical stress causes variation of the carrier mobility and therefore in the I_{ds} current, nonetheless strain mismatch is lower in the center of the die [83]. The well proximity is another systematic variation that affect the threshold voltage of MOSFETs; during the implant process, some atoms scatter laterally from the edge of the photoresist mask and insert in the silicon surface around the vicinity of the well edge; therefore, the well surface concentration changes with the lateral distance from the mask edge; this non-uniform well doping causes MOSFET threshold voltages variations depending on the distance of the transistor to the well edge [86].

3.4 The Diff-Pair Integrator Circuit (DPI)

The Diff-Pair Integrator (DPI) originally proposed in [87] is a current-mode circuit that has log-domain filter properties [88]; the schematic and simulation results of this circuit are shown in Figs. 3.7 and 3.8 respectively. The output current I_{out} of the circuit responds exponentially to an input spike spk_{in}. During the time in which the signal is active (charge phase) a current $(I_{in} - I_{\tau})$ charges the internal capacitor C with high slew rate giving as result a linear decrease of V_c and an exponential increase in I_{out} (M_6 configured in weak inversion). Similarly, when the pulse ceases the capacitor C discharges linearly (discharge phase) through M_5 which returns V_c to its initial state and decreases the output current I_{out} exponentially until reaching zero. The DPI circuit includes a scaling factor that can be used to amplify (or attenuate) the charge phase response amplitude [89]. Simulation results are correlated with the theoretical equations although slight variations occur because of second order effects in the transistors such as channel length modulation and body effect.

The mathematical expressions for I_{out} considering all transistors in weak inversion is obtained as follow:

In charge phase:

The current in M_4 is obtained analysing the transistors M_1, M_2, M_3 and M_4

$$I_{in} = \frac{I_w}{1 + e^{\frac{\kappa(V_{thr} - V_c)}{U_T}}} = \frac{I_w}{1 + \frac{I_{out}}{I_{gain}}}, \tag{3.16}$$

where $I_{gain} = I_0 e^{-\frac{\kappa(V_{thr} - V_{dd})}{U_T}}$.

The current in M_5 is set to be considerable lower than I_{in} so that I_{M6} can implement an ideal linear system. Considering the substrate connected to the source, then

$$I_{\tau} = I_0 e^{\frac{\kappa(V_{dd} - V_{\tau})}{U_T}} \left(1 - e^{\frac{-(V_{dd} - V_c)}{U_T}}\right) = I_{\tau 0} \left(1 - e^{\frac{-(V_{dd} - V_c)}{U_T}}\right) \tag{3.17}$$

For $V_c < (V_{dd} - 4U_T)$ we can consider that M_5 is in saturation, so I_{τ} is simplified to the constant value

$$I_{\tau} = I_0 e^{\frac{\kappa(V_{dd} - V_{\tau})}{U_T}} \tag{3.18}$$

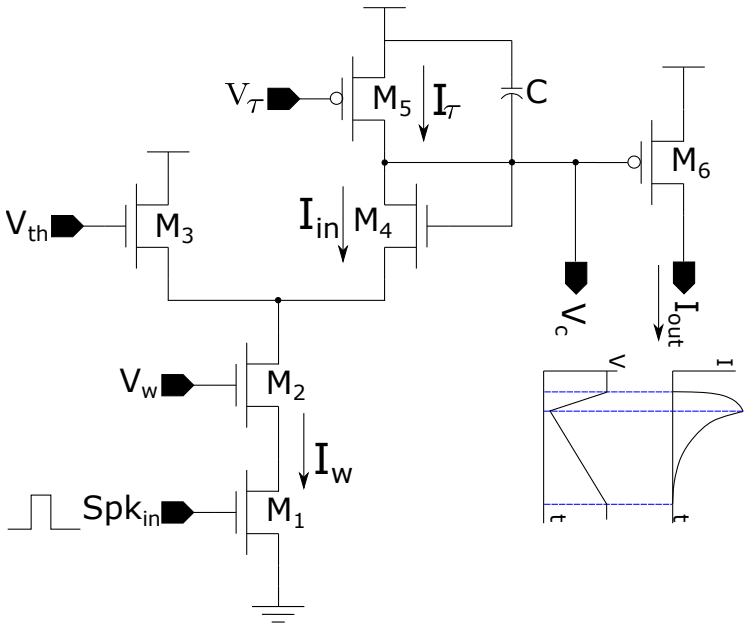


Fig. 3.7: Diff-pair integrator circuit. The circuit consist of two transistors in stack M_1 and M_2 operating as a bias current source for the differential pair $M_3 - M_4$, the M_4 and M_5 transistors are connected to a C capacitor to charge and discharge it linearly respectively. The transistor M_6 converts the linear voltage waveform V_c in exponential when it operates in weak inversion or quadratic when it operates in strong inversion. When spk_{in} is in high level, I_w current flow is generated with a value depending on V_w, $(I_{in} - I_w)$ current charges C reducing the V_c voltage. The DPI circuit is set for $I_w >> I_\tau$. For $V_c > V_{th}$, $I_{in} \approx I_w$, and the V_c saturation value is set by V_{th} voltage.

The total current flowing in the capacitor C is

$$C\frac{dV_c}{dt} = -(I_{in} - I_\tau) \tag{3.19}$$

For any pMOS transistor in weak inversion and saturation the differential equation 3.20 can be demonstrated

$$\frac{dI_{sd}}{dt} = -I_{sd}\frac{\kappa}{U_T}\frac{dV_g}{dt} \tag{3.20}$$

Then by replacing $\frac{dV_c}{dt}$ (Eq. 3.19) in $\frac{dV_g}{dt}$ (Eq. 3.20) we obtain

$$\tau\frac{dI_{out}}{dt} + I_{syn} = \frac{I_w}{I_\tau}\frac{I_{out}}{1 + \left(\frac{I_{out}}{I_{gain}}\right)} \tag{3.21}$$

Considering $I_\tau << I_{in}$ and $I_{gain} << I_{out}$, a simplified differential equation is obtained

$$\tau\frac{dI_{out}}{dt} + I_{out} = \frac{I_w I_{gain}}{I_\tau} \tag{3.22}$$

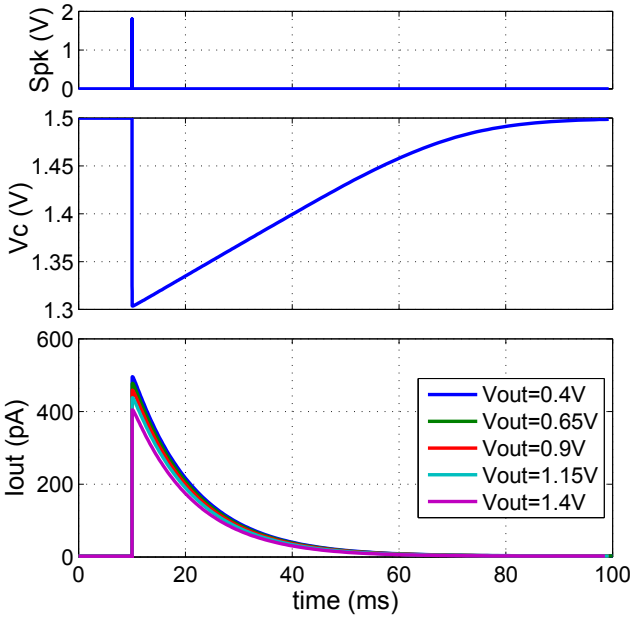


Fig. 3.8: Simulation results for the DPI circuit, $(W/L)_p = (W/L)_n = (1\mu m/0.5\mu m)$, M_5 and M_6 source voltage are connected to $Vddll = 1.5V$, M_1 source is connected to $gndll = 0.3V$ to reduce leakage current, M_3 drain is connected to $V_{dd} = 1.8V$, $V_\tau = 1.35V$, $V_{th} = 0.5V$, the spike pulse width is $20\mu s$. I_{out} pin is connected to a Vdc source which is swept from $V_{out} = 0.4V$ to $V_{out} = 1.4V$. During the input spike V_c decreases (and I_{out} increases) from its initial resting voltage $1.5V$ to a lower value (higher value for I_{out}) depending on the set Vw value, after the spike ceases V_c returns to its resting voltage linearly and $Iout$ exponentially. As observed, non-ideal effect takes place when V_{out} varies giving slight $Iout$ amplitude differences.

with solution

$$I_{out}(t) = \frac{I_{gain} I_w}{I_\tau}\left(1 - e^{-\frac{(t-t_i^-)}{\tau}}\right) + I_{out}^- e^{-\frac{t-t_i^-}{\tau}} \tag{3.23}$$

In discharge phase:

Eq. 3.22 can be solved for $I_w = 0$ giving as result

$$I_{out}(t) = I_{out}^+ e^{-\frac{t-t_i^+}{\tau}} \tag{3.24}$$

3.5 The Operational Transconductance Amplifier (OTA)

A transconductance amplifier is a circuit that converts a differential input voltage into an output current; this circuit also provides high rejection to the

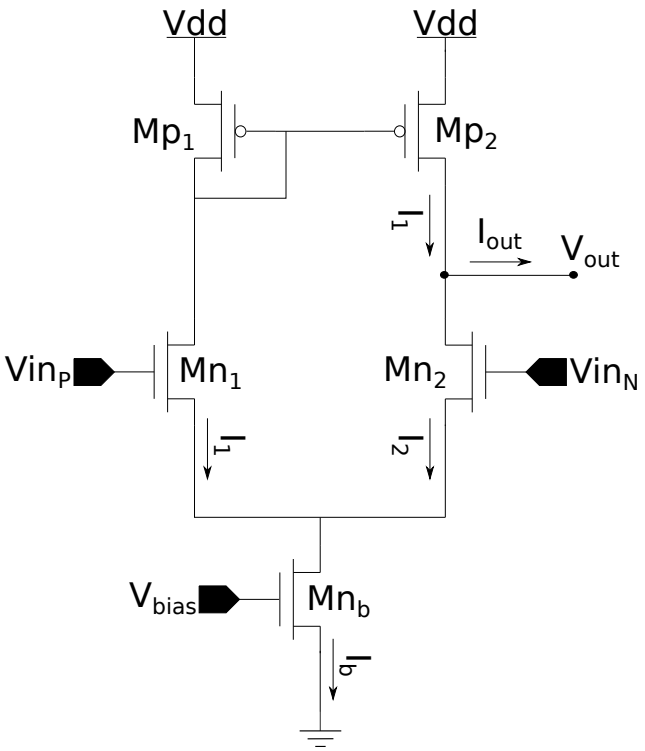


Fig. 3.9: Operational transconductance amplifier. Transistor Mn_b implements a current bias for the differential pair $Mn_1 - Mn_2$, the current I_1 generated in Mn_1 is mirrored by he current mirror configuration $Mp_1 - Mp_2$, finally $Iout$ is obtained as the different of the current flows I_1 and I_2. The values I_1 and I_2 have a quadratic (strong inversion) or exponential (weak inversion) relationship with respect to the voltages Vin_P and Vin_M

voltage supply noise and to the common mode of the input terminals [69]. The OTA circuit shown in Fig. 3.9 subtracts the two currents generated in a differential pair I_1 and I_2; this subtraction is generated by mirroring one of the differential pair currents to the complementary one. The terminal V_{out} is connected to a load, in our case a capacitor. Simulation results for this circuit are shown in Fig. 3.10; here both currents I_1 and I_2 together with their difference I_{out} are presented. The waveforms in this graphic approximate to hyperbolic functions when the transistors are configured in weak inversion.

Considering $Vin_P > Vin_N$ and same transistor dimensions (W/L) for Mn_1 and Mn_2. Then, since $(V_{gs} - V_{th})_{Mn1} > (V_{gs} - V_{th})_{Mn2}$ the majority part of the current I_b will flow through Mn_1. Given that in strong inversion the relationship between I_{ds} and V_{gs} is quadratic while in weak inversion it is exponential, a small variation in the differential voltage causes a considerable variation in the difference of the currents.

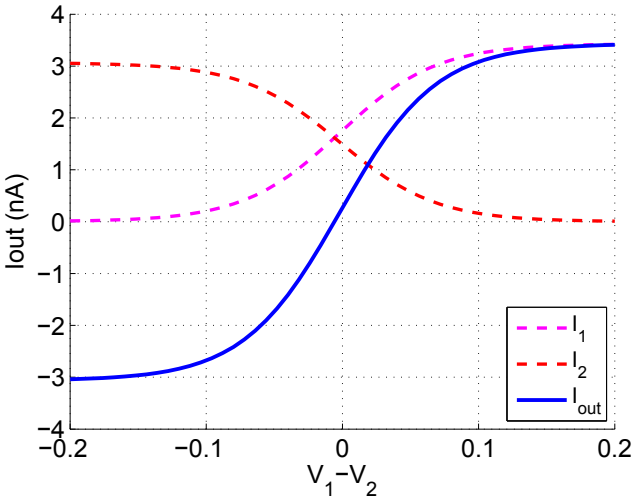


Fig. 3.10: Simulation results for the transconductance amplifier in weak inversion. In the figure I_1, I_2 and $I_{out} = I_1 - I_2$ are plot as a function of $V_{ID} = V_1 - V_2$. Parameter values are $V_b = 0.2V$, $V_{cm} = 0.9V$, $Vdd = 1.8V$, $(W/L)_p = (W/L)_n = 1\mu m/0.5\mu m$, $Vout$ pin is connected to a Vdc source with value $V_{out} = 0.9V$. As observed the direction of the current flow $Iout$ is given for the higher value between V_1 and V_2.

The swing output voltage is limited in the range $Vsat_{Mnb} + Vsat_{Mn2} < V_{out} < V_{dd} - Vsat_{Mp2}$; if more swing range is required a wide-range OTA can be used [71].

Considering the input diferential voltage $V_{ID} = V_1 - V_2$ and the input common mode voltage $V_{cm} = \frac{V_1+V_2}{2}$; it can be demonstrated that for strong inversion the relationship between $Iout$ and V_{ID} [69] is:

$$I_{out} = I_b \left(\frac{\beta V_{ID}^2}{I_b} - \frac{\beta^2 V_{ID}^4}{4 I_b^2} \right)^{1/2} \tag{3.25}$$

The previous relationship is only useful for $V_{ID} < 2\left(I_b/\beta\right)^{1/2}$, since outside of this range the transistors move to triode region.

For weak inversion, the I_{out} can be simplified to a hyperbolic relationship respect to V_{ID} as follow:

$$I_{out} = I_b \times tanh\left(\frac{\kappa_n}{2V_T} V_{ID} \right) \tag{3.26}$$

For small differential voltages Eq. 3.26 is approximately linear:

$$I_{out} = g_m V_{ID}, \tag{3.27}$$

where

$$g_m = \frac{I_b \kappa_n}{2V_T}$$

For large differential voltage I_{out} saturates to $\pm I_b$ values.

3.6 The Winner-take-all Circuit

The Winner-Take-All (WTA) circuit was originally proposed by Lazzaro et al. [90]. This circuit (shown in Fig. 3.11) compares two input currents Iin_1 and Iin_2 (or voltages Vin_1 and Vin_2) and generates two output currents $Iout_1$ and $Iout_2$ where one of them is practically zero (if the input signals are different) and the other one has the same value than the bias current I_b. The branch where the current flows is decided by the higher input voltage; therefore, this circuit works as a comparator with output current. Transistors Mp_1, Mp_2 and Mn_b generate ideal current sources where the first two are the input signals and the last one the bias current of the circuit. Simulation results for this circuit are shown in Fig. 3.12 which shows practically only two possible current values at each output terminal.

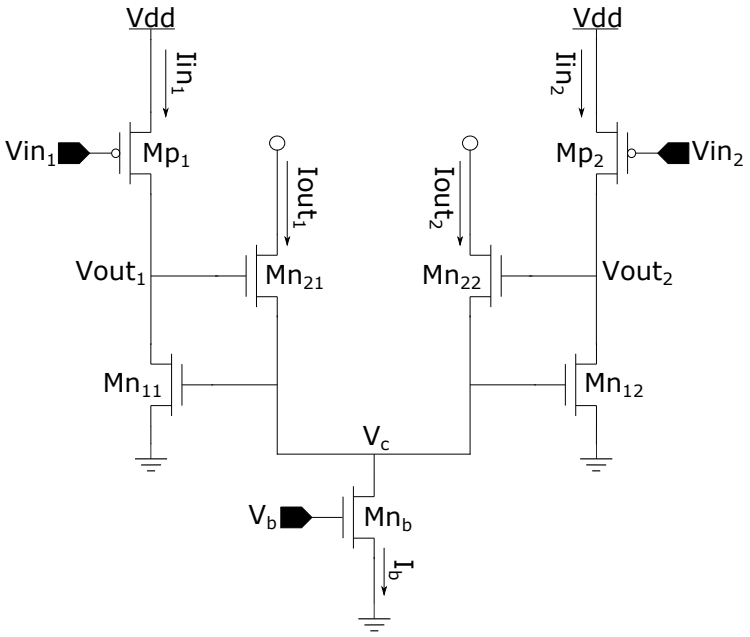


Fig. 3.11: Schematic for the WTA circuit. Mp_1 and Mp_2 implement input current sources which are inversely proportional to Vin_1 and Vin_2 voltages, the bias current for the differential pair $Mn_{21} - Mn_{22}$ is implemented by the transistor Mn_b. For a considerable difference between the Iin_1 and Iin_2 input currents, only one output voltage $Vout_1$ or $Vout_2$ is high and the bias current I_b flows fully through the transistor with higher gate voltage in the differential pair.

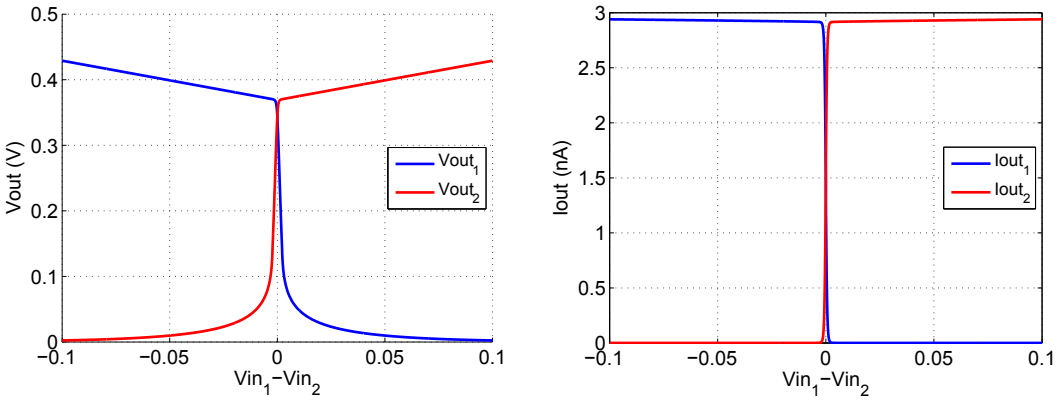


Fig. 3.12: Simulation result for the circuit in Fig. 3.11. The parameters are $V_{cm} = 1.6V$, $V_b = 0.2V$, $Vdd = 1.8V$, $(W/L)_p = (W/L)_n = 1\mu m/0.5\mu m$, the source of transistors V_{21} and V_{22} are tied to $Vdd = 1.8V$. As result of the common mode $1.6V$ and the swept in the differential mode, Vin_1 swept from $1.55V$ to $1.65V$ and Vin_2 from $1.65V$ to $1.55V$. Fig. (a) shows V_{out} vs $V_{ID} = V_1 - V_2$, $Vout$ reaches just after few millivolts of difference in V_{ID} 0 or a linear region. Fig. (b) shows I_{out} vs V_{ID}, here also just after few millivolts $Iout$ reaches two values 0 or I_b.

Considering first $Vin_1 = Vin_2$ (or $Iin_1 = Iin_2 = I_m$), then transistors Mn_{11} and Mn_{12} have the same voltages in their terminals, so the currents generated at the output terminals are $Iout_1 = Iout_2 = I_b/2$.

Considering all transistors in weak inversion. Then in Mn_{21} and Mn_{22} the output voltage can be expressed as:

$$V_{out} = \frac{V_t}{\kappa^2} ln\left(\frac{I_m}{I_0}\right) + \frac{V_t}{\kappa} ln\left(\frac{I_b}{2I_0}\right) \tag{3.28}$$

This equation differs from the original work [90] since the MOS physical model used here is the body-reference while in Lazzaro et al. the source-reference model is used.

If now Iin_1 is increased to $I_m + \delta I$, V_c also increases; however, the transistor Mn_{12} should sink I_m current for the same increased voltage; in order to drive I_m current Mn_{12} has to move from saturation to triode region decreasing $Vout_2$ voltage. Since now we have that $Vout_1 > Vout_2$, the current flow in the differential pair $Mn_{21} - Mn_{22}$ is unbalanced while more current flows in Mn_{21}. For a considerable difference between $Vout_1$ and $Vout_2$ we can consider that all the current I_b will flow for only one of the two branches of the winner-take-all.

The output voltage and current as function of the difference of the input voltages $Vin_1 - Vin_2$ is plot in Fig. 3.12a and Fig. 3.12b respectively. When $Vin_1 > Vin_2$ an inverse relationship is obtained in the input currents $Iin_2 > Iin_1$ because PMOS are used as current sources. This difference of currents set a high voltage in $Vout_1$ which also generates a current flow $Iout_1$ in Mn_{21}. When Vin_2 starts to increase and eventually turns greater than Vin_1, $Vout_1$ decreases until sourcing only leakage current. The output voltage contains a small slope for considerable differences in the input voltages because a raise in $Vout_1$ is driven by an increment in Vc to keep $V_{gs}(Mn_{21})$ and therefore $Iout_1$ is constant (the current flow in Mn_{21} is approximated to I_b when Iin_1 is considerable greater than Iin_2).

3.7 Discussion

In this chapter I introduced the basic principles of CMOS operation and I characterized it for the technology used to fabricate our chips. Special focus

was given to subthreshold regime (only weak inversion) which is a region that has similar characteristics to its biological counterparts such us exponential dependency on the gate voltage.

An important aspect when replicating same circuit cells to obtain a matrix of blocks is avoiding excessive mismatch; therefore, variations in a single transistor as function of its size and bias values were also presented to understand trade-off considerations at the moment of implementing the layout. In addition, good layout techniques to avoid systematic mismatch were discussed.

Furthermore, I presented the essential circuits such as DPI, OTA and WTA that are used to build up the synapse block. Those circuits were modified from their original proposal version to obtain our desired operations. For instance in the DPI circuit an additional branch was added to have two discharge paths that are activated for the pre- and post- synaptic spikes, or an additional stage was included in the Operational Transconductance Amplifier (OTA) to reduce the headroom voltage thus obtaining a better output swing range. Other modifications will be shown in the second fabricated version chip which was improved according to previous measurement results.

Finally, it is important to mention that the mathematical formulation for each circuit neglected second order effects as well as mismatch contribution; therefore, they are ideal results which deviate from the measured ones and vary in different degree depending on design constraints.

4 First Synapse Circuit Implementation

Model choice is a critical step in the design of neuromorphic systems. Designers must decide which model details are critical for the application and which are superfluous. For example, we choose synaptic plasticity models which can be implemented with local event-based weight updates because this allows for extremely low-power systems that easily fit into the neuromorphic paradigm. As we described before, VLSI technology that uses CMOS transistors in weak inversion shares the same primitives of neural computing, we can recognize that these models easily fit into hardware and are compatible with biology.

Implementation of a model into a physical system poses similar constrains than its biological counterpart such as finite power supply, optimization of energy consumption and the need for space optimization when wiring computational structures, these constraints do not exist in simulation; therefore, they force the researcher to be more biological realistic. Most modern neuromorphic systems [23, 30, 91] implement some form of the STDP [54] model, which states that synaptic weight change is a function of the timing between the pre- and post-synaptic spikes. LTP is an increase of the synaptic weight induced when the post-synaptic spike follows the pre-synaptic spike, and its magnitude is an exponentially-decaying function of the time difference; LTD is a decrease of the synaptic weight induced when the order is reversed and is also an exponential function of the time difference. While this is an attractive model because of its simplicity and success in performing useful computations in simulation, it does not explain certain aspects of biological synaptic plasticity. For example, as the rate of stimulation increases, LTP tends to be induced regardless of spike timing. Therefore, more recent research has emphasized that synaptic plasticity is a multi-factor phenomenon, depending on several parameters.

The change in the synaptic strength as a function of the spike time difference $t_{post} - t_{pre}$ is considered a form of Hebbian learning because presynaptic neurons that are active slightly before the postsynatic neuron are those which take part in firing it, while those that fire later do not contribute to the post-

synaptic action potential [45]. Alternative learning rules in nature are also possible such as in the electric fish which implement an opposite dependence on the timing between presynaptic and postsynaptic spike generally called as anti-Hebbian plasticity.

VLSI circuits used in neuromorphic implementations offer the advantage of dedicated hardware that perform massively parallel computation; design conditions can also be implemented to obtain real-time and low power operations by using long-time constants and weak inversion configurations respectively. On the other hand, VLSI systems require a long development time and high costs. Ideally one should figure out the most general model so that the implemented hardware can model the plethora of learning phenomenologies observed in the neural system. Some computational models implement biopysically realistic synapses with the intention to capture the multiple learning mechanims that coexist in a single cell including STDP and its dependence on stimulation frequency by considering the location of the synapse along the dendrite [92] and homeostatic process [93]; however, given their mathematical complexity which is translated into high power consumption and wide silicon area in VLSI, they are not suitable for hardware implementation. The work presented in [21] is closer to the synapse model implemented here because the weight update is triggered by the presynaptic spike and depends on the post-synaptic membrane potential. A calcium concentration variable only depending on the post-synaptic activity is used in [22] to decide when the neuron should stop learning [94]. In contrast, the calcium concentration in our circuits depends on both pre- and post-synaptic spiking activity and is crucial for determining the sign and strength of the change in weight.

I present here an analog VLSI circuit based on one multi-factor model [34] which shows how plasticity's dependence on these factors (i.e. rate and timing) could be explained by the behavior of a single variable, calcium concentration. This differs from most of the VLSI learning models presented in the past, which typically belong to one of two classes: models which explicitly measure the time difference between pre- and post-synaptic spikes [19, 20, 95] and models which compute synaptic weight change by using some variable besides spike timing (such as membrane voltage or calcium concentration) [20–22, 30, 96, 97]. The proposed circuit offers several advantages over these previous designs. Its advantage over the first class of models is that it reproduces certain biophysical properties of synaptic behavior not captured by those models, such as the dependence of plasticity on stimulation

frequency. Pfister et al. [98] showed that STDP rules which are based purely on the difference between the arrival times of a pair of pre- and post- synaptic spikes do not reproduce biological data as the frequency of stimulation increases. As opposed to other models [19, 20], synaptic plasticity in the calcium-based model emerges from the calcium variable's dynamics, and any spike can generate a change in the calcium concentration. This approach could yield more biologically-realistic network behavior. Our circuit's benefit over both model classes is that it can produce several types of STDP learning profiles also observed in different brain regions and across layers within one region [61], expanding the available selection of learning behaviors. The second class of models uses the calcium variable to generate an eligibility trace for learning, essentially acting as a switch to enable or disable learning depending on the neuron's activity. This minimizes the resources used for the learned inputs. In contrast, the model used here relies on the calcium variable to define the direction of learning (depression or potentiation).

Several VLSI approaches focus on implementing alternative signal processors compared to traditional computing architectures; these systems are useful for neuroscience modelling given that they can accelerate the simulation of complex computational neuroscience models; however, we aim to implement energy efficient and real time systems with biologically realistic time constants (on the order of a few up to hundreds milliseconds) which are more naturally realized with analog circuits in comparison with digital circuits with high frequency clock. Our synapse is bistable on long time scales, therefore equivalent to one bit, nevertheless on short time scales the synaptic weight is fully analog. It has been demonstrated that synapses that have two stable states can dynamically learn with optimal storage efficiency mantaining its memory for an indefinitely long time as palimpsets paradox state (synapse should be very plastic to encode quickly new memories but not too plastic to avoid erasing old memories) [99].

In this chapter we broaden the work presented in [36], where our design of a calcium-based plasticity circuit was originally proposed and compared to a mathematical calcium-based plasticity model using Cadence Spectre simulations. The changes in the model proposed with the purpose of making it more suitable for a compact circuit implementation were described in full details in Sec. 2.3. In Sec. 4.1 we describe the calcium synapse circuit [36], with a focus on its ideal behaviour. Spectre simulations of the circuit are compared with the model in Sec. 4.1.1. Calcium state variable can be expressed as a

voltage or as a current signal, in [36] we presented the first option while here and in [37] we show it as a current value which approximates better with the theoretical model. We also present here an in-depth characterization of the plasticity mechanism and evaluate deviations from the model by analysing the weight change's dependence on initial conditions. The frequency dependence of the learning dynamics is also characterized and discussed in this section. Due to the successful synapse characterization we proceeded with the full chip design by adding I/O interfaces and implementing the full layout circuit which was later sent to fabricate to a foundry. A custom PCB was also designed, produced and populated with discrete devices including voltage regulators, DAC converters and USB interfaces. The PCB had to be integrated in a testing system with measurement equipments such as oscilloscope and a firmware programs implemented in python to control input signal values and spike timings from a PC. Section 4.2 presents the first measured silicon data available for our circuits which was fabricated using a standard AMS $180\ nm$ 1-poly, 6-metal technology, here we replicated the most important simulation results and characterized the response as a function of stimulation frequency. In the last section we discuss our results and future work.

4.1 The Calcium Synapse Circuit

We designed a novel learning circuit that mimics the model described in Section 2.3. This is to our knowledge the first attempt to implement neuromorphic VLSI hardware based on this model. The full circuit comprises 40 transistors and 3 capacitors organized in three functional blocks: *Calcium*, *Synapse Core* and *Bistability* (gm) as shown in Fig. 4.1. The *Calcium* block responds to digital pre- and post-synaptic pulses (Spk_{pre} and Spk_{post}) and produces the current I_{ca} which mimics the calcium concentration described in Eq. 2.15. The output of the *Calcium* block (I_{ca}) is fed to the *Synapse Core*, which implements potentiation and depression depending on the status of the calcium current. The *gm* block is a wide-range OTA in positive feedback, which implements bistability and weight saturation. The *Synapse Core* and *gm* block together define the synaptic weight V_w.

The *Calcium* block, as shown in Fig. 4.2, computes the calcium concentration I_{ca} (see Fig. 4.2b). The calcium waveform is generated by a differential pair integrator (DPI) circuit [87] with two inputs (Spk_{preD} and Spk_{post}). Initially V_{ca} is at its resting potential Vc_{ref} and the output current I_{ca} is null.

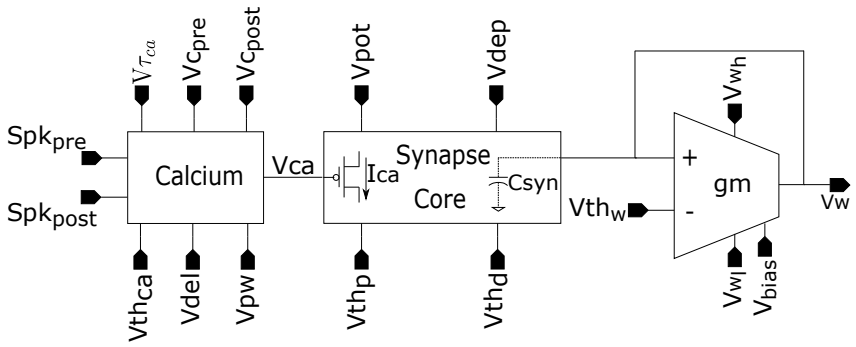


Fig. 4.1: Learning circuit block diagram. The circuit is composed of a calcium block, a synapse core block and a bistability block (implemented with an OTA with positive feedback). The calcium variable is represented by the current I_{ca} produced by the calcium block in response to pre- and post-synaptic spikes and fed to the synapse core. The synapse core produces the appropriate changes in the synaptic weight (represented by voltage V_w). Long term memory is guaranteed by the bistability amplifier which drives the synaptic weight to its stable state in the absence of changes produced by the synapse core block.

Delayed pre-synaptic pulses Spk_{preD} and post-synaptic pulses Spk_{post} turn on transistors Ms_{11} and Ms_{21} respectively and charge can accumulate on the capacitor C_{cal}. The amount of charge accumulated during an input pulse depends on the initial charge, the bias voltages Vth_{ca}, $V\tau_{ca}$, Vc_{pre} and Vc_{post} (for pre- and post-synaptic spikes respectively), and the pulse duration [87]. The capacitor voltage is then converted into a current via the pFET Mp_{x2}. Ideally, Mp_{x2} operates in its subthreshold regime so that linear changes in capacitor voltage cause exponential changes in the current. However, if V_{ca} gets pulled too low, Mp_{x2} may enter strong inversion, and the relationship will become quadratic, adding a nonideality to the dynamics. In the absence of input pulses C_{cal} discharges through the current flowing in transistor Ms_5 at a rate set by the bias voltage $V\tau_{ca}$.

The delayed pre-synaptic pulse is generated by the delay circuit shown in Fig. 4.2a. Upon arrival of a pre-synaptic pulse, the transistor Md_1 is turned on and the capacitor C_{del} is fully charged. When the input pulse ends, the capacitor is discharged linearly over time by the current flowing through transistor Md_2 and set by the bias voltage V_{del}. The triangular pulse on C_{del} is amplified and shifted to the operation range $[0 - vdd]$ by a buffer labelled $B1$ which consists of two inverters in cascade. The resulting output $Vout_1$ is a wider version of Spk_{pre}. The NOR gate generates a short pulse when $Vout_1$ transitions low and the common-source amplifier is charging $\overline{Vout_{1\tau}}$. Simulation results for the delay circuit operation are shown in Fig. 4.4. One

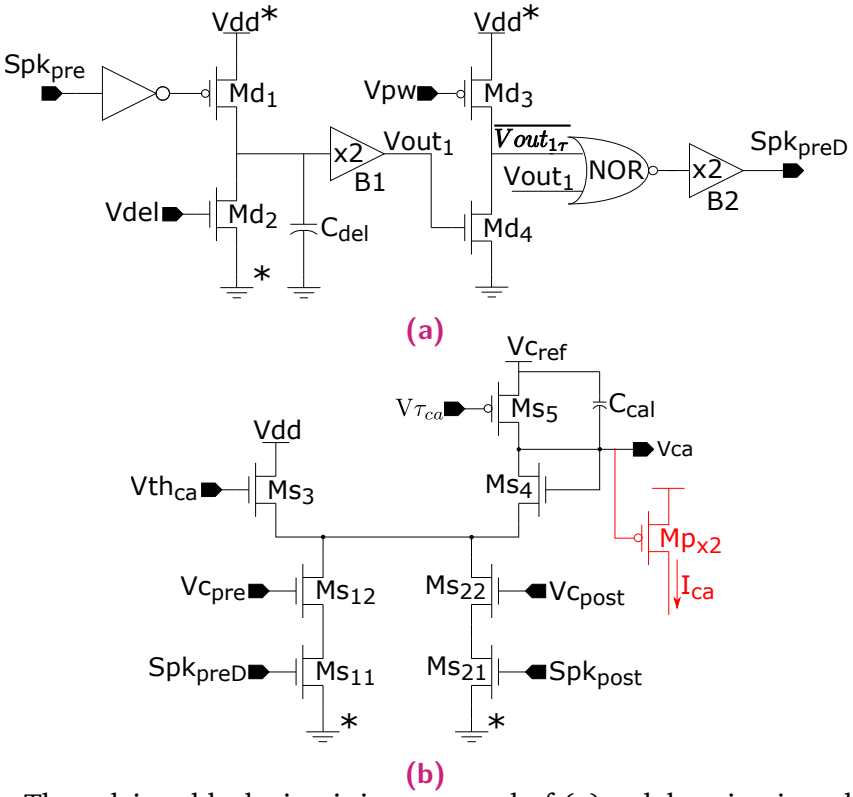


Fig. 4.2: The calcium block circuit is composed of (a) a delay circuit and (b) the calcium core circuit. The delay circuit generates a digital pulse Spk_{preD} in response to an input pulse Spk_{pre}, with a delay controlled by the bias voltage V_{del} and a duration set by the bias voltage V_{pw}. The calcium core circuit generates the calcium dynamics in response to delayed pre-synaptic spikes Spk_{preD} and post-synaptic spikes Spk_{post}. The amplitude of the response is set by Vc_{pre} and Vc_{post} respectively. Vth_{ca} is an independent gain control for the effect of input pulses, and $V\tau_{ca}$ sets the time constant of the calcium current I_{ca}.

disadvantage of this delay circuit is that the gain of the buffer is not very high. Thus the time of $Vout_1$'s transition from high to low varies with the slew rate of Vc_{del}. This dependence causes the output pulse width to be slightly dependent on V_{del}. This circuit could be improved by increasing the gain of the first buffer. In addition, we should note that if the pulse width is too large, V_{ca} will eventually stop decreasing linearly as it approaches Vth_{ca} because of the operation of the differential pair. This saturating effect could lead to model inaccuracies. One way to avoid it is by implementing pulse extenders that fix the spikes' width.

The synapse core implements Eq. 2.13, and its schematic is shown in Fig. 4.3. This circuit is based on the WTA circuit proposed by Lazzaro et al. in the late '80s [90] and is similarly used in other spike-based learning circuits (e.g. [100]). Vw represents the synaptic weight, which is increased or decreased depending on the value of the current I_{ca} representing the calcium

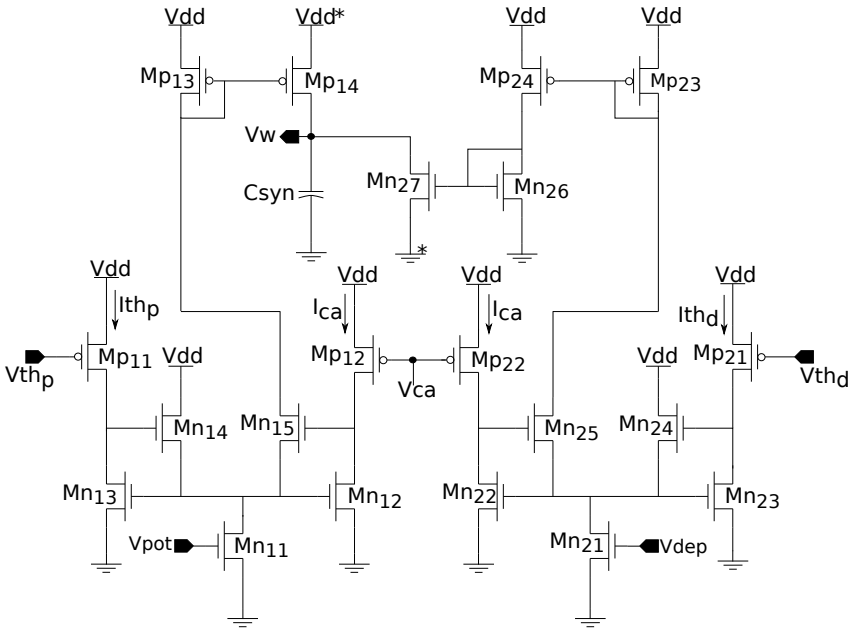


Fig. 4.3: The synapse core circuit generates the synapse dynamics for Vw based on I_{ca}, which represents the calcium concentration. A WTA circuit compares I_{ca} with Ith_p. When $Ica > Ith_p$, a current proportional to $Vpot$ is generated in Mn_{15} and is mirrored through Mp_{13-14} to charge C_{syn}. Similarly, I_{ca} is compared with Ith_d. When $I_{ca} > Ith_d$, a current proportional to V_{dep} is generated in Mn_{25} and then mirrored through Mp_{23-24} and Mn_{26-27} to discharge C_{syn}.

concentration and produced by the *Calcium* block (Fig. 4.2). When I_{ca} is higher than Ith_p, the WTA circuit implemented by transistors Mn_{11-15} and Mp_{11-12} generates an output current which is copied by transistors Mp_{13-14} to charge the capacitor C_{syn}. Similarly, when I_{ca} is higher than Ith_d, the WTA circuit implemented by transistors Mn_{21-25} and Mp_{21-22} generates an output current which is copied by transistors Mp_{23-24} and Mn_{26-27} to discharge the capacitor C_{syn}.

Leakage current is present in Mp_{14} and Mn_{27} even when they do not perform charge or discharge operations in C_{syn}. These leakage currents affect Vw, which therefore cannot retain its value for a long time. In order to reduce this effect, the source voltages of Mp_{14} and Mn_{27} are decreased and increased, respectively. We label the new source voltages Vdd^* and gnd^*. This leads to $Vgs_{Mn27} < 0$ and $Vsg_{Mp14} < 0$, which reduces the leakage current when the transistor operates in the cutoff region. Fig. 4.5 shows this effect, for $Vdd^* = 1.5V$ and $gnd^* = 0.3V$, the leakage current is reduced by more than 3 decades.

When Vdd^* and gnd^* have the same values as the power supply (1.8V and 0V), a considerable leakage current around $15\ pA$ produced by Mp_{14} and Mn_{27}

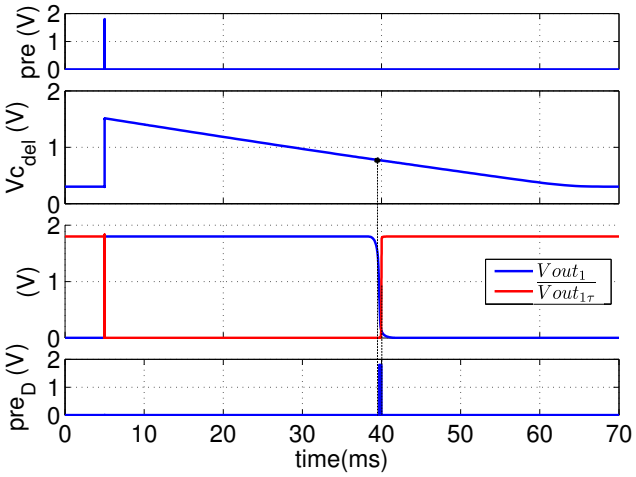


Fig. 4.4: Simulation results for the delay circuit. After C_{del} charges fast when a pre-spike arrives, it discharges slow with a slope set by V_{del}. When the Vc_{del} signal reaches the threshold of the inverter B1, $Vout_1$ switches its level and its inverter version $\overline{Vout_{1\tau}}$ is generated after a delay set by Vpw.

flows in the capacitor C_{syn}. As a result, C_{syn} cannot retain the information for long and discharges in approximately $20ms$. Improvement is observed by reducing Vdd^* and increasing gnd^*, which reduces the leakage current and increases the discharge time. For $Vdd^* = 1.5V$ and $gnd^* = 0.3V$ the leakage current is reduced to less than $10fA$, which leads to an increase in retention time up to $50s$. For values smaller than $Vdd^* = 1.5V$ and greater $gnd^* = 0.3V$ the leakage current reduction is not significant. We also note that shifting these sources reduces the gain of the current mirrors. Alternative techniques to replicate very low currents can be used [101].

The bistability and saturation components of Eq. 2.13 are implemented by an OTA with positive feedback and saturation values V_{wh} and V_{wl} as shown in Fig. 4.1. The wide range transconductance amplifier shown in Fig. 4.6 is based on [102] and is similarly used in [100]. Mp_{22} and Mn_{22} are connected to V_{wh} and V_{wl} with two main purposes: first to reduce the current flow, and second to limit the voltage range of Vw. The OTA injects or leaks current to/from C_{syn} as a function of $(Vw - Vth_w)$:

$$\frac{dV_w}{dt} \propto \frac{I_b}{C_{syn}} tanh\left(\frac{\kappa(Vw - Vth_w)}{2U_T}\right), \tag{4.1}$$

where κ is the capacitive coupling ratio from gate to surface potential and U_T is the thermal voltage. Eq. 4.1 has a proportionality factor because the unequal source voltages at the output current mirrors give them a gain less than one. Since the transconductance amplifier is connected to Vw, it sets the saturation limits of Vw. For values $Vw \approx Vth_w$ Eq. 4.1 is well approximated

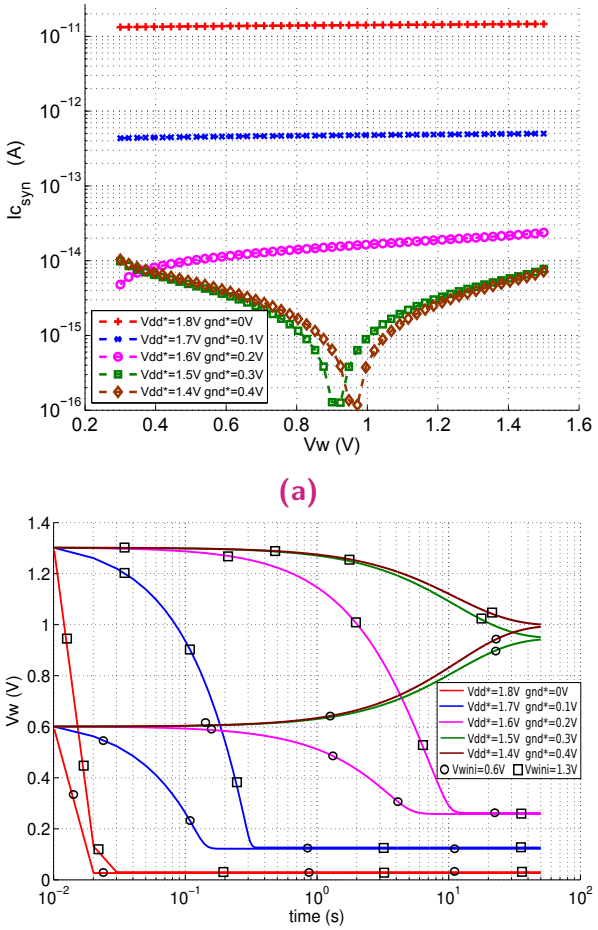


Fig. 4.5: Leakage current simulation results. We estimate the current in C_{syn} when there are no pre- or post-synaptic spikes and bistability is disabled. (a) DC simulation results obtained by replacing C_{syn} with a voltage source and sweeping Vw for different $vddll$ and $gndll$ values. (b) Transient analysis for V_W is shown for two initial Vw_{ini} values (0.6 and $1.3V$) and different $vddll$ and $gndll$. The waveforms are plotted on a logarithmic scale. Results show that for $vddll = 1.5V$ and $gndll = 0.3V$, the capacitor reaches steady state in $50s$.

by a linear function, as is the bistability term in Eq. 2.12; when Vw is considerably higher or lower than Vth_w, a constant current flow is generated; when Vw reaches saturation values the output current is reduced to zero. Simulation results for the OTA shown in Fig. 4.7 reproduce approximately the hyperbolic function of the Eq. 4.1 and leads the input voltage Vw in positive feedback to only two stable values in the extremes voltages of the x-axis (gnd^* and Vdd^*) where the current is zero. Despite the fact that a constant current flow is not reached because of the channel length modulation effect in the CMOS transistors connected to the output terminal, the circuit reproduce faithfully a bistable mechanism.

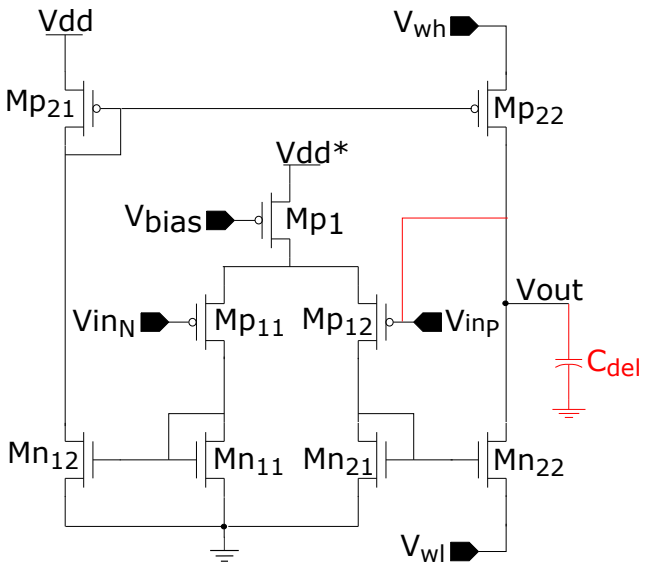


Fig. 4.6: A wide range transconductance amplifier which operates in positive feedback is connected to C_{syn}. The source voltage of Mp_1 is set below Vdd to achieve small current flow. Similarly, we reduce the source voltages of Mp_{22} and Mn_{22} to limit Vw to $[V_{wl}, V_{wh}]$ and to limit their leakage currents.

4.1.1 Simulation Results

We simulated the proposed circuit with Cadence's Spectre simulator to characterize its performance. The simulation data was compared to the simplified theoretical model (Eqs. 2.13,2.15) in Matlab. In order to directly compare these, we must use the same units. The only parameters in [35] which were assigned units were those of time (s) and frequency (Hz). All other quantities were unitless. Here we report those mathematical model parameters in arbitrary units (arb. u.) In the circuit model, the synaptic weight is a voltage (Vw) between V_{wl} and V_{wh}. To compare this voltage to its equivalent quantity in the model, we offset and normalize it, and we refer to this value as "per unit" (pu) as follows:

$$Vw(pu) = \frac{Vw(V) - V_{wl}(V)}{V_{wh}(V) - V_{wl}(V)} \tag{4.2}$$

We set $V_{wl} = 0.3\ V$ and $V_{wh} = 1.5\ V$. To further ease comparisons to the model, the calcium current I_{ca} is plotted as positive.

We first characterized the calcium dynamics in response to isolated pre- and post-synaptic spikes. In Fig. 4.8 the model and circuit implementation of the calcium variable are compared. The dynamics of the waveforms are qualitatively similar. After a delay D following the pre-synaptic pulse, the

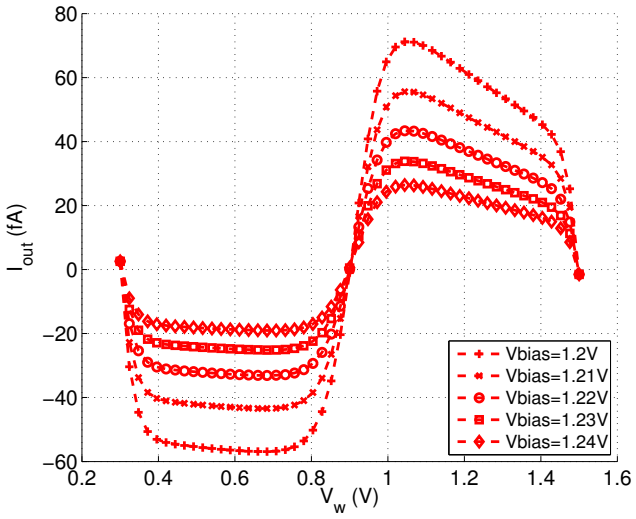


Fig. 4.7: Simulation results for the wide range transconductance amplifier with simple current mirrors and short channel length MOS. The output net Vw is connected to the positive input, in addition a dc voltage source is connected between Vw and gnd to apply a swept signal, Vin_N is set to 0.9V. As observed the output current deviates considerable from an ideal antisymmetric waveform with reference axes Vin_N because of variations in channel length modulation in PMOS and NMOS.

calcium variable quickly rises to a value lower than the depression threshold and then decays exponentially. Immediately after the post-synaptic spike, the calcium variable quickly rises above both thresholds and then decays exponentially. Given the good match of circuit behavior and model data demonstrated above for the calcium dynamics in response to single pulses, we proceeded to more realistic testing with repeated pre-/post-synaptic pulse pairs as shown in Fig. 4.9. As demonstrated in [35], depending on the calcium concentration parameters C_{pre} and C_{post}, threshold values θ_p and θ_d, and potentiation γ_p and depression γ_d coefficients, a plethora of STDP curves can be generated. In these measurements we set the parameters to achieve a classical STDP curve [54] ($C_{pre} < \theta_d < \theta_p < C_{post}$). A complete list of parameters used is provided in Fig. 4.9.

Fig. 4.9a shows the expected LTP of the synaptic weight in response to pairs of pulses with $(t_{post} - t_{pre}) = 12ms$ presented at a frequency of $5\ Hz$. The synaptic weight was initialized to $Vw = 0\ pu$ at the beginning of the experiment. Similarly, LTD is demonstrated by providing input pre-/post-synaptic pulse pairs with $(t_{post} - t_{pre}) = -14ms$ and frequency $5\ Hz$, after initializing the weight to $Vw = 0.75\ pu$.

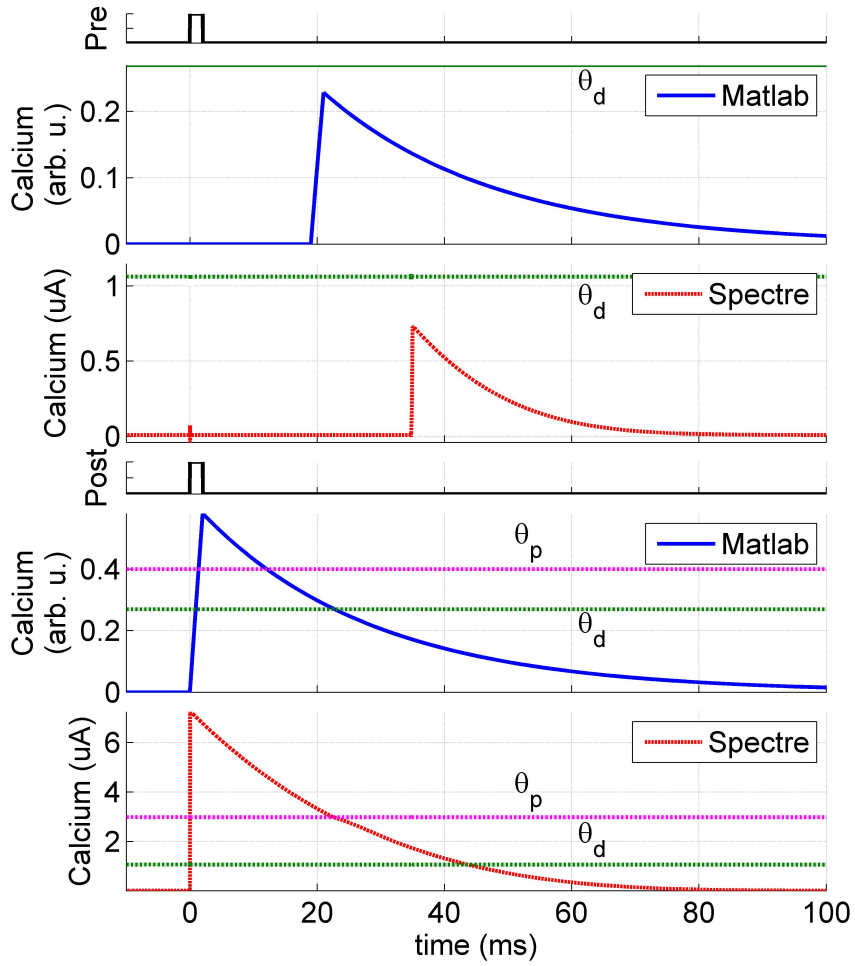


Fig. 4.8: Simulated response of the calcium variable to a pre- (top) or post-synaptic (bottom) pulse for the simplified model (blue) and hardware (red). In order to obtain STDP waveforms the condition $C_{pre} < \theta_d < \theta_p < C_{post}$ is required [35]. See Fig. 4.9 for detailed parameters.

A saturation effect takes place near the upper boundary of the synaptic weight variable which can invert the effect of input spikes. For example, the last spike pairing in Fig. 4.9a results in a small depression, even though the pairing should produce potentiation under normal circumstances. This phenomenon is explained by the calcium dynamics. To undergo LTP, the calcium variable rises above θ_p and increases the weight. When the stimulus ceases, the calcium variable falls below θ_p, but it is still above θ_d for a brief time. This causes a small amount of depression. As long as the calcium is above θ_p for long enough, the synapse will achieve a net potentiation. However, if Vw is close to 1, it cannot increase any further because it's physically limited by Vdd^*. Thus, the small amount of depression will dominate and the weight will decrease slightly as shown in Fig. 4.10b. If Vw is close to 0, a further LTD leads Vw to its minimum value, contrary to the previous case, here the

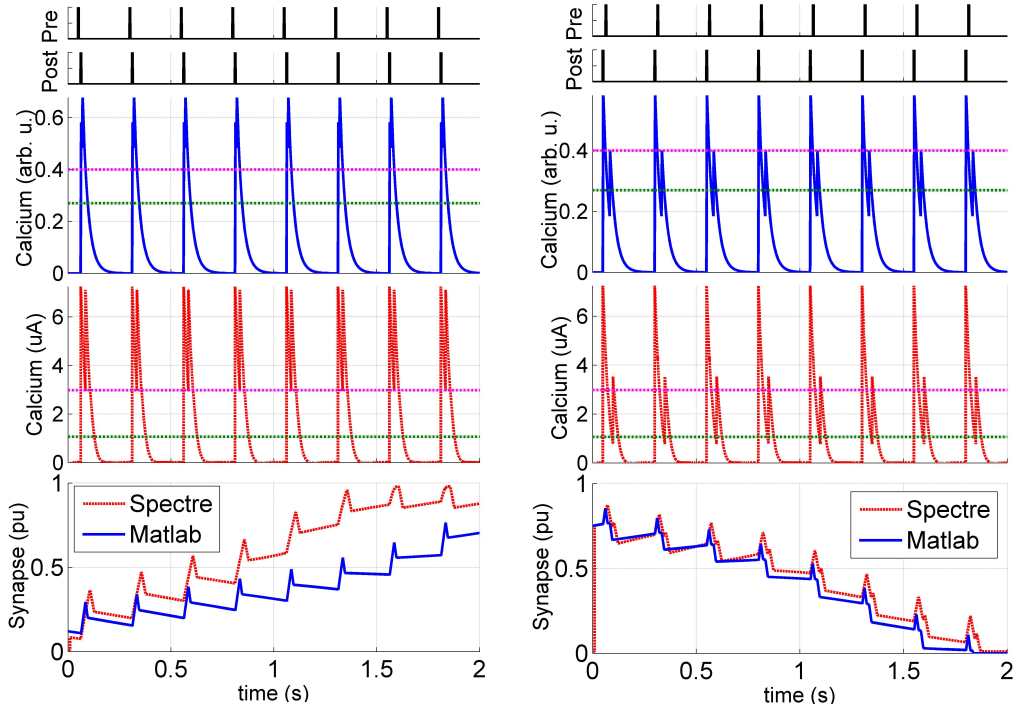
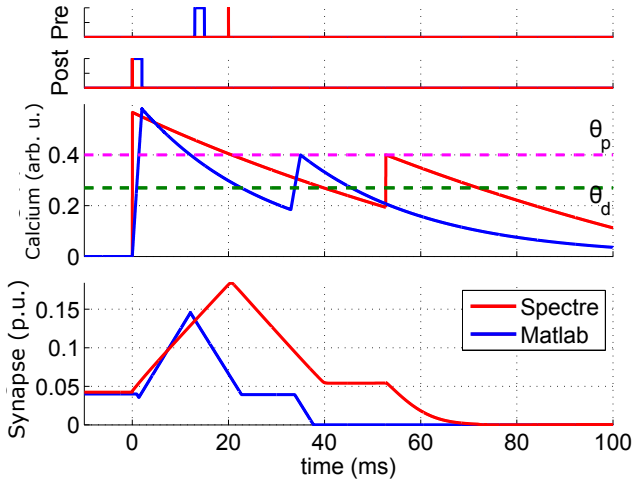


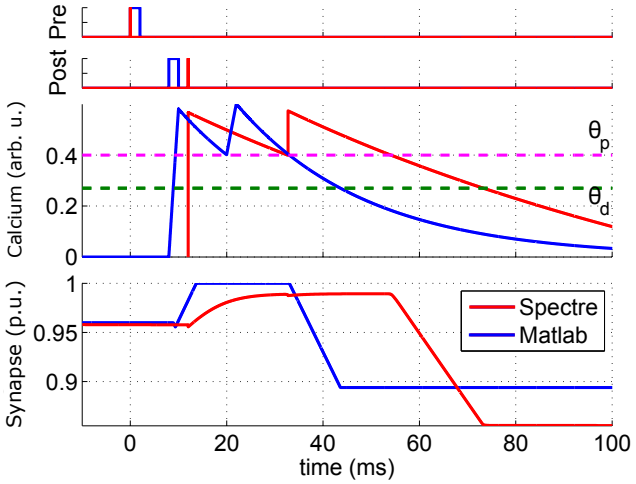
Fig. 4.9: Synapse evolution for spikes with frequency 5Hz (a) potentiation $(t_{post} - t_{pre}) = 12ms$. (b) depression $(t_{post} - t_{pre}) = -14ms$. The comparison between circuit (red) and simplified model (blue) simulations verifies that synapse dynamics depend on the calcium concentration and bistable term. The bistable term generates a slow increase or decrease of the synaptic weight, depending on whether the weight is above or below a threshold value $w_* = 0.5$. The capacitor values are $C_{del} = 100fF$, $C_{ca} = 574fF$ and $C_{syn} = 100fF$, and the transistor parameters are $(W/L)_{Mp_{x3-x4}} = (W/L)_{Mn_{26-27}} = (1/0.54)\mu m$. The parameters for the circuit simulation are spike width $t_{spk} = 20\mu s$, $V_{th_{ca}} = 0.5V$, $V_{\tau_{ca}} = 1.39V$, $V_{C_{ref}} = 1.5V$, $V_{C_{pre}} = 0.544V$, $V_{C_{post}} = 681mV$, $V_{del} = 316mV$, $V_{pw} = 1.65V$, $V_{pot} = 305mV$, $V_{dep} = 269mV$, $V_{th_p} = 1.1V$, $V_{th_d} = 1.23V$, $V_{wh} = 1.5V$, $V_{wl} = 300mV$, $Vth_w = 925mV$, and $V_{bias} = 1.22V$. The parameters for the theoretical simulation are $t_{spk} = 2ms$, $\tau = 2s$, $\tau_{ca} = 27ms$, $k_{bs} = 9$, $C_{pre} = 0.59$, $C_{post} = 1.5$, $\theta_p = 0.4$, $\theta_d = 0.27$, $\gamma_p = 35$, and $\gamma_d = 17.3$.

first spike increases Vw without restriction and the second spike decreases Vw as shown in Fig. 4.10a.

The very good match between model and circuit shown in Fig. 4.9b is not replicated in Fig. 4.9a. The main cause of these deviations is channel length modulation in Mp_{14} and Mn_{27}, which makes $I_{C_{syn}}$ dependent on Vw. This effect is more pronounced when the transistors move into the Ohmic regime. Another cause of nonidealities is that the bistability circuit's OTA does not implement a third-order polynomial, but rather a hyperbolic tangent function. Additionally, this bistability component is asymmetric, which will be discussed in more detail later in this section. Finally, the mathematical equivalence of



(a)



(b)

Fig. 4.10: Synapse saturation effect when (a) depression occurs near 0 and (b) potentiation occurs near 1. For depression, the synapse is reduced until it reaches the bottom limit value. On the other hand, in potentiation, it even decreases the synapse value.

the two systems' calcium dynamics assumes that all transistors operate in the subthreshold regime, which is not always the case. In order to operate the transistors in weak inversion, V_{ca} should swing no lower than $Vdd - Vth$, where Vth for a standard pFET in this process is around $0.37V$. V_{ca}'s maximum value is set by Vc_{ref} in the calcium circuit, which should be lower than Vdd to reduce leakage currents. Therefore, in the calcium circuit V_{ca} swing range is limited to approx. $Vc_{ref} - (V_{dd} - Vth)$ when transistors operate in weak inversion.

The model proposed in [35] provides a mechanistic understanding of how the calcium signal gives rise to the observed multitude of synaptic plasticity forms, and it accounts for plasticity data measured in hippocampus and neocortex in response to stimulation protocols characterized by various spike timing

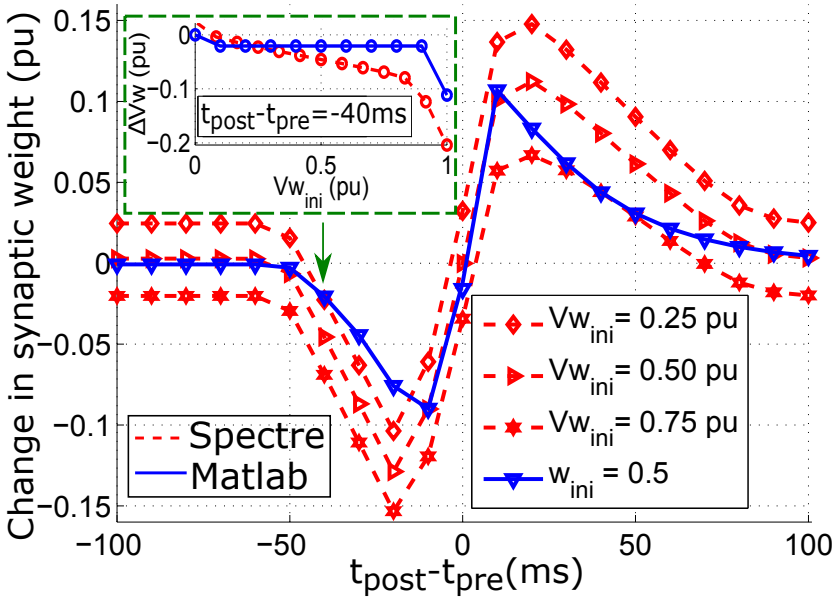


Fig. 4.11: Synapse learning waveforms from the simplified model (blue) and hardware simulation (red). The parameters in this figure are the same as in Fig. 4.9. In the simplified model, the learning waveform is independent of the initial w value. On the other hand, shifts of the learning waveform in the circuit simulation are produced because variations of Vw produce slight variations in its current supplies Mp_{14} and Mn_{27}. The inset figure shows the change in synaptic weight for $t_{post} - t_{pre} = -40ms$ and different initial Vw values. The Matlab model shows a relatively constant change in weight, while the circuit model shows significant variation with Vw_{ini}.

and frequency patterns. The authors thoroughly characterized the parameter space and provided clear definitions of the boundary between different kinds of STDP learning curves. In particular, they showed that classical STDP is achieved in the parameter space region defined by $C_{pre} < \theta_d < \theta_p < C_{post}$. We chose a set of parameters within the classical STDP region and characterized the STDP behaviour as follows. To guarantee that the measured synaptic weight changes are only due to the weight updates triggered by the calcium dynamics, we disabled the bistability circuit, which is equivalent to setting k_{bs} to zero in Eq. 2.13. In the circuit, the bistability is disabled by setting $|Vgs|_{Mp1} \leq 0$. Values smaller than 0 are used to reduce the leakage current. We then applied a classical stimulation protocol [54] to the calcium synapse circuit by providing a single pre- and post-synaptic pulse pair and measured the synaptic weight variation in response to the stimulus. The measured synaptic weight is first normalized as described by Eq. 4.2. The change in normalized weight is plotted in Fig. 4.11 for $t_{post} - t_{pre}$ values between -100 and $+100\ ms$. As a reference, the model response to the same stimulation is plotted using Matlab.

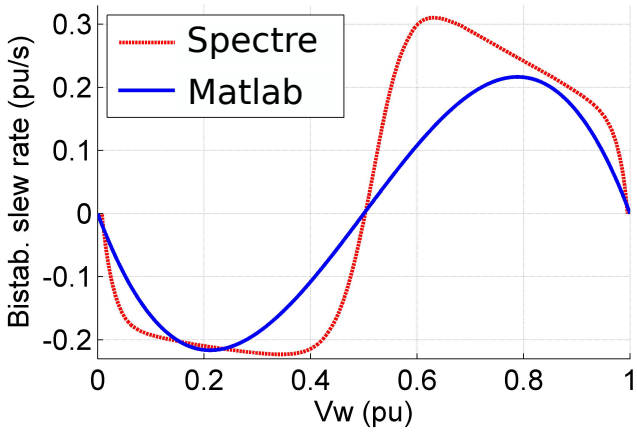


Fig. 4.12: Bistability slew rate (dw/dt) vs. synapse value. In the simplified mathematical model, the slope is a third-degree polynomial. In the circuit model, this slope is created by the current from an OTA in positive feedback with a small bias current, which ideally implements a hyperbolic tangent function. The plot is shown for $k_{bs}/\tau_{ca} = 4.5$ in the mathematical model and for $V_{bias} = 1.22V$ in the circuit model. The non-ideal shape of the circuit's slope is caused by channel-length modulation and saturation of the output transistors at large and small weight values.

The inset of Fig. 4.11 shows that the weight change depends on the value of the weight. This is also seen by the three different STDP curves in the figure: each one was measured for a different initial weight, and it is clear that the curves are shifted vertically depending on the weight. This shift is caused by channel-length modulation: as the weight increases, the current sourced by Mp_{14} decreases and the current sunk by Mn_{27} increases. So if depression and potentiation are perfectly balanced for a particular value of Δ_t at $Vw = 0.5\ pu$, the STDP curve will be biased towards potentiation at $Vw < 0.5\ pu$ and towards depression at $Vw > 0.5\ pu$. At the boundaries (Vw_{ini} near zero and one), the vertical shift is larger, which can be seen in the inset of Fig. 4.11 near $Vw_{ini} = 1pu$. This problem can be improved by using cascode current mirrors in the synapse core circuit and increasing the lengths of the mirror transistors.

Fig. 4.12 shows the bistability term's effect on the synaptic weight. The rate of change of the weight dw/dt is plotted as a function of the weight. In the mathematical model, this function is the polynomial $-k_{bs}\tau^{-1}w(1-w)(w_* - w)$. In the circuit, this function is provided by an OTA, which ideally implements a hyperbolic tangent V-I transfer function when operated in subthreshold, until the amplifier's output transistors move out of saturation, at which point the current reduces to zero. It can be observed that the transconductance amplifier does not maintain a constant current flow in C_{syn} for high values of Vw as expected. This is caused by channel length

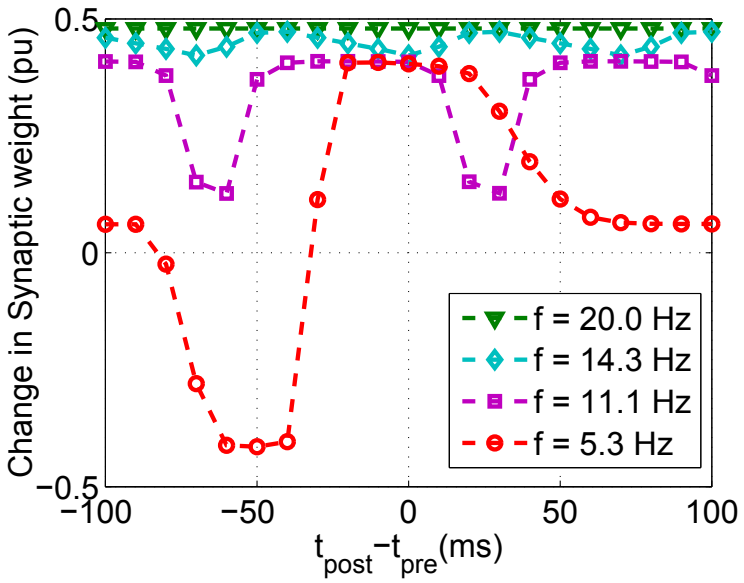


Fig. 4.13: Change in synapse strength for spike trains with different frequencies and pair timings. The synapse weight is measured after $2.1s$. For frequencies around 5Hz the system operates as classic STDP predicts. However, when the frequency increases, the calcium variable is higher than θ_p more often than it is below, and therefore only potentiation is observed. In order to provide high stimulus frequencies, the delay $preD$ was reduced to zero. All other circuit parameters are as in Fig. 4.9.

modulation in Mp_{22} and Mn_{22}, which have short lengths. Because the degree of channel length modulation in the two transistors is different, the bistability curve is asymmetric. This is undesired because the strength of the synaptic weight attractors will differ, which could affect the relative probabilities of LTP and LTD. Cascode current mirrors and an increment of transistor lengths could be employed to greatly reduce the effects of channel length modulation.

The calcium-based model is compatible with the observation that biological synapses are sensitive to both firing rate and timing. Previous work [103] found that in some cells, LTD disappears above $40\ Hz$ and that LTP increases with frequency. We tested this behaviour in the calcium synapse circuit by extending the experiments shown in Fig. 4.11 with repeated pulse stimulation at varying frequency. In each experiment, the synapse converges to a steady-state oscillatory behavior when the jump caused by the calcium and synapse core is exactly offset by the change caused by the bistability circuitry. Lower frequency stimuli take longer to converge than higher-frequency stimuli. We determined the time to convergence for our lowest-frequency stimulus (2.1s in these simulations) and ran all simulations for that amount of time. Fig. 4.13 shows our results. Each data point is the average of the maximum

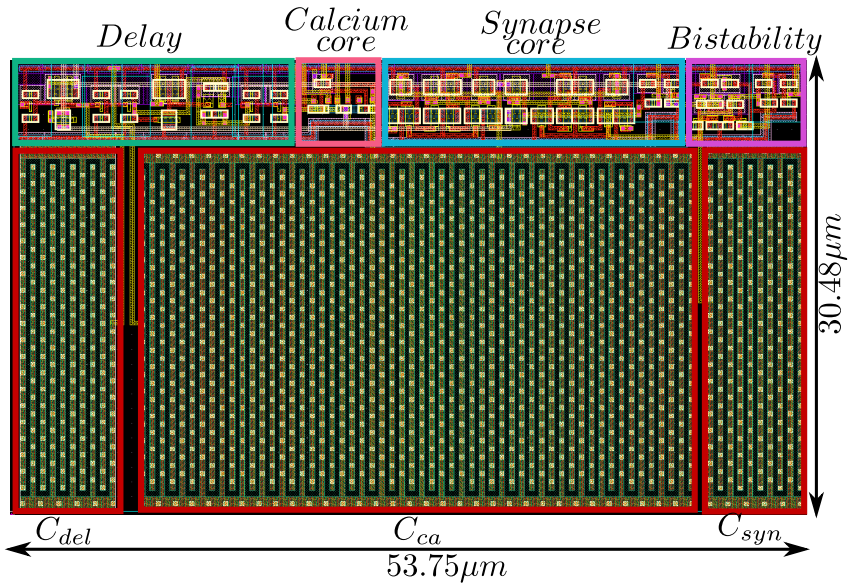


Fig. 4.14: Layout data for the synapse circuit which was fabricated in $180\ nm$ technology. The majority of the area is covered by the capacitors.

and minimum weights in the steady-state oscillation. As expected only potentiation is observed for high frequencies. This effect starts showing up when the inverse of the frequency of stimulation becomes comparable with the time constant of the calcium variable. In this condition and for higher frequencies of stimulation the calcium variable does not have time to return to its resting value before the next pair of pulses is presented. The calcium variable spends more time above the potentiation threshold for intermediate frequencies than for low frequencies, and it is always kept above the potentiation threshold for frequencies higher than 16Hz.

In our circuit the maximum frequency of operation is limited by the delay circuit: if the Interspike Interval (ISI) between two pre-synaptic spikes is smaller than the delay set in the circuit, the delayed version of the second spike is not generated by the circuit. Therefore, we set the delay to zero to perform this experiment. This limitation poses an upper bound on the pre-synaptic spike frequency at $1/D$ (as in Eq. 2.15). If higher frequencies are desired, a digital delay circuit based on registers could be employed. However, this is a limitation common to physical systems, and we expect the biological counterpart to also have an upper bound. Therefore, it would be sufficient to match the upper bound of the circuit implementation to the biological limitations.

The calcium-based learning model can create other learning waveforms besides the classic STDP curve (also known as the Depression-Potentiation

(DP) curve). One such curve is the Depression-Potentiation-Depression (DPD) waveform, where the synapse is potentiated for small Δt and depressed for larger Δt, regardless of spike order. In order to obtain this curve, the parameters should fulfill the conditions c_{pre}, $c_{post} < \theta_d < \theta_p$. In addition, the time constant τ_{ca} is increased to cover the same timing operation range ($200ms$) as in the DP curve. While we were able to obtain a simulated DPD curve, we do not present those results here because of space limitations. Rather, we present a measured DPD curve in Section 4.2.

4.2 Hardware measurement results

We designed and fabricated a testchip which contains all the three components of the calcium plasticity circuit in $180\ nm$ technology. Fig. 4.14 shows the layout of our first prototype circuit. The dimensions are $53.75\mu m \times 30.48\mu m$, and the majority of the area is occupied by the capacitors.

The chip was placed on a custom PCB featuring CPU-controlled DACs. We pinned out the calcium, weight, and delayed pre-synaptic spike voltages because voltage measurement makes data acquisition simpler, requiring only an oscilloscope. Thus the calcium measurements that follow are not the same state variables as the simulations in Section 4.1.1, but rather are proportional to the natural log of the calcium current (assuming subthreshold operation). Nonetheless, these measurements provide useful insight into the behavior of the calcium circuitry. For ease of comparison with the simulations, we plot $V_{C_{ref}} - Vca$. Weight measurements are normalized as described in Eq. 4.2 and reported in pu. As in the simulations, $V_{wl} = 0.3\ V$ and $V_{wh} = 1.5\ V$.

With the aim of replicating simulation results in silicon, we first programmed the CPU-controlled DACs to provide the same bias voltages used in Spectre. Fine tuning was then used to closely match simulation data. We provided precisely timed pre- and post-synaptic spikes using a dual-output function generator. The synaptic weight dynamics for potentiation and depression are shown in Fig. 4.15. Before stimulation, the synaptic weight was initialized to $0\ pu$ for the potentiation experiment (Fig. 4.15a) and to $0.71\ pu$ for the depression experiment (Fig. 4.15b). The calcium variable was always initialized to zero. Synaptic potentiation is shown in Fig. 4.15a for pre/post spike pairings with a time difference of $10\ ms$; depression is shown in Fig. 4.15b for pairings with $\Delta t = -45\ ms$.

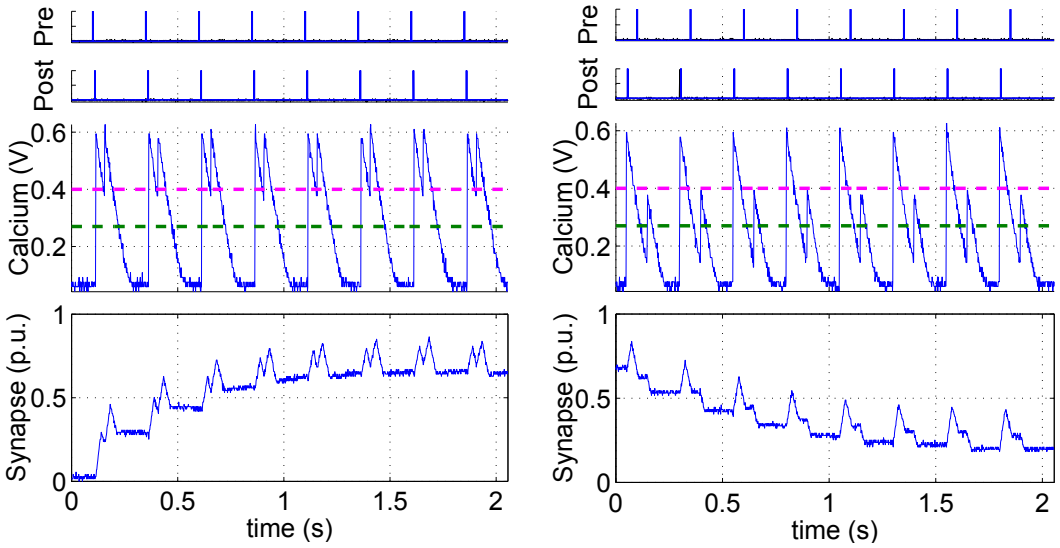


Fig. 4.15: Measured calcium and weight for potentiation and depression experiments. Calcium is plotted as $V_{C_{ref}} - Vca$, and weight is normalized to pu. The waveforms were measured for $(t_{post} - t_{pre})$ = (a) 10ms (b) -45ms. For positive timing potentiation is observed, while negative timing generates depression. The parameters for the measurements are $t_{spk} = 20\mu s$, $V_{th_{ca}} = 0.5V$, $V_{\tau_{ca}} = 1.424V$, $V_{C_{ref}} = 1.5V$, $V_{C_{pre}} = 526mV$, $V_{C_{post}} = 669mV$, $V_{del} = 353mV$, $V_{pw} = 1.7V$, $V_{pot} = 323mV$, $V_{dep} = 321mV$, $V_{th_p} = 1.1V$, $V_{th_d} = 1.23V$, $V_{wh} = 1.5V$, $V_{wl} = 300mV$, $Vth_w = 906mV$, and $V_{bias} = 1.8V$.

Despite a qualitative match of the measured dynamics with the simulated data, minor deviations can be observed. The measurements indicate that the maximum synaptic weight is lower than expected from simulation. The weight cannot reach V_{wh} because the leakage currents are larger than the transconductance amplifier's output current at high output voltages. The circuit can be improved with stacked transistors, cascodes, and a larger C_{syn}.

We characterized the classical STDP learning behaviour as already done in simulation (see Sec. 4.1.1). The bistability block was disabled and single pre-/post-synaptic pulse pairs were presented to the circuit after the initialization procedure described in the following. The calcium voltage V_{ca} was initialized to $V_{C_{ref}}$ and the synaptic weight voltage V_w was set to an initial weight $Vw_{ini} = 0.5\,pu$. This initial voltage was set with a transmission gate connected to the Vw node. The synaptic weight was measured before the first input spike and after the calcium decreased below Vth_d. The change in the weight as a function of the time difference between the pre- and post-synaptic pulses (for a single pulse pair) is shown in the main plot of Fig. 4.16. The inset of Fig. 4.16 shows how the change in weight for a given $t_{post} - t_{pre}$ (indicated by

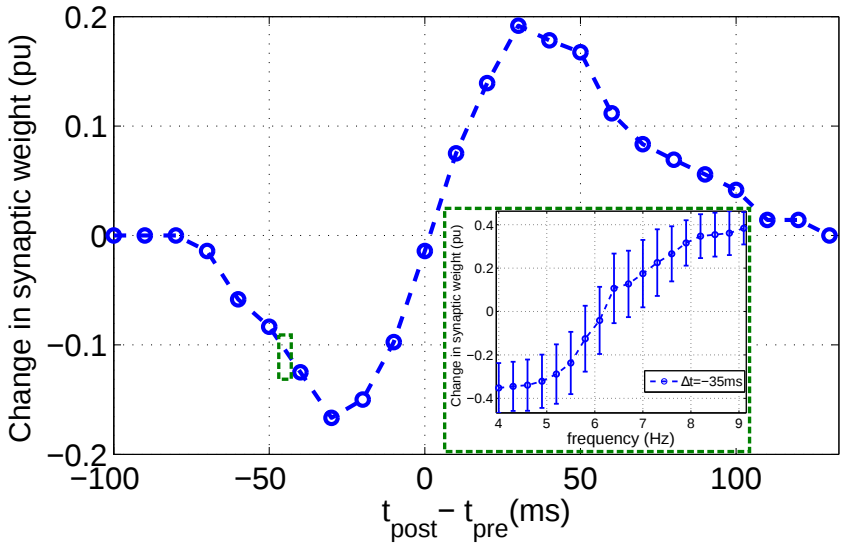


Fig. 4.16: Main plot: the measured STDP learning curve is similar to the simulation results. The bistability is disabled for these measurements. All other parameters are as in Fig. 4.15. Each measurement begins by setting Vw to a nominal value of $Vw_{ini} = 0.5\ pu$. The weight is then measured before and after a single pre/post spike pairing, and the difference between the normalized weights is plotted. This data presents one experiment trial, for which the variation was minor. Inset: the steady-state value of the weight depends on the frequency of stimulus presentation. Bistability is enabled for these measurements ($V_{bias} = 1.268V$). The weight was initialized to $Vw_{ini} = 0.5\ pu$. The relative spike timings were $t_{post} - t_{pre} = -35\ ms$ (corresponding to the green box in the main figure). This stimulus was presented at various frequencies. The weight was recorded after it reached a steady-state oscillatory behavior. Points represent the difference between the average of the maximum and minimum weights measured during steady-state oscillation and Vw_{ini}. The error bars do not represent statistical variation, but rather the difference between the maximum and minimum weights encountered in the steady-state oscillation. Rare outliers encountered during frequency measurements are not included in the dataset because they are infrequent. These outliers could be caused by noise, and focusing on noise reduction techniques could shield the circuit more effectively.

the green box in the main figure) is a function of stimulus frequency. These measurements are comparable to the simulation results and demonstrate that the circuit is able to reproduce the synaptic weight dynamics described by the model.

We next characterized the frequency-dependent behavior of the circuit. We repeated a similar experimental protocol to the one used for Fig. 4.13. We first initialized the weight to $0.5\ pu$. We then provided pairs of spikes with a fixed timing $t_{post} - t_{pre} = -35ms$ and varied the pair repetition frequency using the dual-output function generator. The bistability circuit was enabled

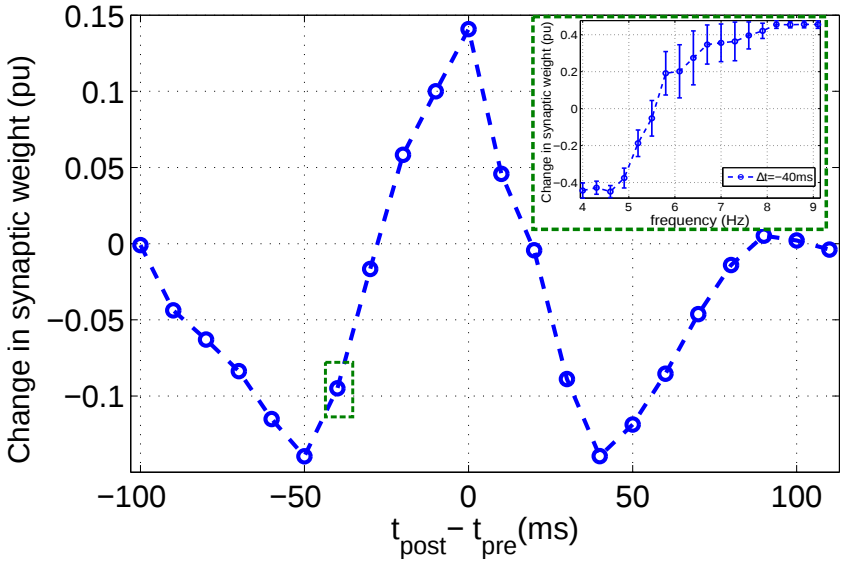


Fig. 4.17: The calcium plasticity circuit can display DPD behavior. Main figure: we applied the same experimental protocol as in Fig. 4.16 but with the following set of parameters: $t_{spk} = 20\mu s$, $V_{th_{ca}} = 0.5V$, $V_{\tau_{ca}} = 1.445V$, $V_{C_{ref}} = 1.5V$, $V_{C_{pre}} = 630mV$, $V_{C_{post}} = 630V$, $V_{del} = 500mV$, $V_{pw} = 1.7V$, $V_{pot} = 323mV$, $V_{dep} = 310mV$, $V_{th_p} = 1.17V$, $V_{th_d} = 1.23V$, $V_{wh} = 1.5V$, $V_{wl} = 300mV$, $Vth_w = 906mV$, and $V_{bias} = 1.8V$. This results in a learning waveform that features potentiation for small time differences and depression for larger differences, regardless of the order of spike presentation. Inset: the DPD curve depends on frequency of presentation. We repeated nearly the same experiment as in the inset of Fig. 4.16 but with the modified parameters, a timing difference of $t_{post} - t_{pre} = -40ms$ (corresponding to the green box in the main figure), and the bistability enabled ($V_{bias} = 1.268V$). The resulting data confirms that an increase of stimulus frequency causes net potentiation.

for this experiment. We then waited for the synapse to reach a steady-state oscillation. The inset of Fig. 4.16 shows our measurements. Each data point is the difference between the average of the maximum and minimum weights encountered in the steady-state oscillation and the initial weight ($Vw_{ini} = 0.5\ pu$). The vertical bars do not represent statistical variation, but rather the difference between the maximum and minimum weights during steady-state oscillation. Thus the bars show the amplitude of the oscillation. Our results are as expected from simulation. When the spike frequency is $4Hz$, the synapse depresses; but when the spike frequency increases, a transition from depression to potentiation occurs. We encountered rare outlying cases of synapses undergoing significantly larger changes than expected. These are not included in the dataset because they are infrequent. They could potentially be eliminated using noise reduction techniques.

In addition to STDP's classic DP learning curve, our circuit is able to display a DPD curve. This was achieved by modifying the circuit's parameters (a full list of parameters is given in the caption). We used the same experimental protocol as in earlier STDP experiments: single pairs of spikes were sent to the synapse, the weight was measured before and after stimulation, and the difference between the two weights was calculated. The results are presented in Fig. 4.17. As expected, potentiation occurs for time differences near zero, while depression occurs for larger differences.

We also characterized the DPD curve's frequency dependence. The experimental protocol is nearly the same as in the earlier STDP experiment: the weight is initialized to $Vw_{ini} = 0.5pu$, a train of pre-/post pairs of spikes was sent to the synapse with a fixed timing of $-40\ ms$, and we waited until the synapse's behavior reached a steady-state oscillation. We then plot the difference between the average of the maximum and minimum voltages seen during oscillation and Vw_{ini}. We use the vertical bars to represent the amplitude of the oscillations. Our results are shown in the inset of Fig. 4.17. The data shows that an increase in the spike frequency causes potentiation for a timing that previously gave depression.

4.3 Discussion

In a theoretical learning, learning process should be slow enough to ensure equal distribution of memory resources (both recent and older experiences should be well remembered). On the contrary, faster learninig could strongly reduce memory lifetime [94]. Small plasticity steps leads to improved memory lifetimes. However, there is a minimum step size below which no further improvement in memory occurs. The minimum limit is also dependent on the signal to noise ratio. The ratio between potentiation and depression should be chosen to find an equilibrium point that generates an equal probability of potentiation and depression [64]. Our circuit bias voltages produce calcium and synapse dynamics comparable to the ones used in theoretical work, which gives rise to biologically plausible transition probabilities. The values are chosen to achieve a STDP window which is $200ms$ wide (τ_{Ca} parameter) and causes no change in the synaptic weight for $t_{post} - t_{pre} = 0$ (delay parameter). k_{bs} and τ are set to get a maximum bistability slope of $0.2pu/s$. γp and γd are related to the amplitude of the change in synaptic weight, so we set to get a maximum rate around 0.1pu. The choice of C_{pre}, C_{post}, θ_p and θ_d should

follow the condition $C'_{pre} < \theta_d < \theta_p < C'_{post}$, where $C'_{pre,post}$ are the maximum peaks of the calcium dynamics (Fig. 4.8b, 4.8e) when the spike ends. We have more flexibility in the choice of these values, but their relationships are important for obtaining the desired symmetry in the STDP characteristic.

The bias voltages were selected so that the circuit exhibits a similar STDP curve to the model (Fig. 7). After fitting this curve, we fine-tuned the parameters like V_{bias} and Vth_w to obtain similar results in the potentiation, depression and bistability experiments (Fig. 6b). We ran our initial chip experiments using the same bias voltages we used in simulation, and we fine-tuned them based on their deviations from the expected behavior. The capacitor values are $C_{del} = 100fF$, $C_{ca} = 574fF$ and $C_{syn} = 100fF$ were chosen such that we could achieve capacitor discharge rates similar to the model with bias voltages no smaller than $100\ mV$. Measured data was obtained using 16-bit Digital to Analog Converters (DACs) with a voltage reference of $3.3V$ which gives us a resolution near $50\mu V$.

L. Abbott et al. in [61] show different learning waveforms found in neocortex-layer 5, 2/3, ELL of electric fish, GABA-ergic neurons in hippocampal culture and neocortex-layer 4 spiny stellates. One may argue that the brain has developed several STDP learning profiles to deal with different learning tasks in different brain areas and animals. The model implemented in our work proposes a general mechanism underlying the various learning profiles which is appealing both as a model for understanding learning in biology and as a substrate for the construction of artificial learning systems.

The idea of using digital inverter and digital gate is commonly used by the neuromorphic community for implementing pulse extenders, some delay circuits are based on static-zero hazard circuits to obtain delays. Our pulse extender circuit after the delay is similarly used in ON-Chip AER Commuication Circuits ([43] and [104]). However, we are working with long time constants on the order of $40ms$, therefore we had to adapt this idea by adding capacitors and low leakage voltage supplies to obtain the desired values. The delay can be tuned by V_{del}, which reduces the amount of discharge current in C_{del}; alternatively we can increase C_{del} to increase the delay time. For our dimensions the maximum delay is reached when we set $V_{del} = gndll$. In this case we have a delay around $79ms$.

Most of the power consumption (as estimated in Spectre simulations) is due to the calcium (11.62μW in potentiation) and synapse (13.45μW in potentiation) blocks. The power consumption in the bistability block (2.4nW) is negligible in comparison. These power consumption numbers are from the experiment of Fig. 4.9a in which Vw rises from a very low level. This power is reduced considerably when Vw_{ini} is set to higher voltages. In the calcium block, the delay circuit consumes almost all of the power, primarily because the voltage on C_{del} decreases slowly, which causes the buffer to draw significant short-circuit power when $V_{C_{del}}$ is approximately midrail. Power consumption could be reduced by detecting when $V_{C_{del}}$ crosses midrail and using positive feedback to force the node low upon this condition. More generally, power consumption reduction techniques will have to be employed in all circuit blocks for the use of our calcium synapse in large scale neural networks.

Capacitor size is a main drawback of this implementation of our circuit. It is large because we used a native metal capacitor. The size of this capacitor could be reduced by around half if it is replaced with a Metal-Insulator-Metal (MIM) capacitor. Alternatively, we can employ techniques to reduce the bias currents in our circuit. We can achieve the same time constants if we scale the bias currents and capacitances down by the same factor (thus achieving a smaller capacitor area). One way to reduce the bias currents is to use high-threshold transistors, since they have lower leakage currents. We can also replace each individual transistor by two transistors in series, exploiting the "stack effect" [105] to reduce their leakage current. Alternative devices to capacitors would be floating gates (FG) and memristors which could store for longer time the synapse values by reducing leakage current, however they would require a redesign (not drop-in replacements for caps), FG's require high voltages and specialized programming circuitry, and memristor programming and integration is in its early stages.

Analog computation is efficient at low precision processing and digital computation at high-precision. Also, physical analog computation is more efficient because it deals with primitives i.e. a wire can be used for adding two input currents [106]. Another disadvantage of synchronous digital compared to analog circuits is that in the former the clock is a power consumption source. Furthermore, we aim at the implementation of real time systems with biologically realistic time constants (on the order of a few up to hundreds milliseconds) which are more naturally realized with analog circuits.

Our synapse is bistable on long time scales, therefore equivalent to one bit. Nevertheless, on short time scales the synaptic weight is fully analog.

We have also discussed a number of nonidealities of the presented implementation. Most importantly, the STDP curve shifts vertically depending on the value of the synaptic weight. This is caused by channel-length modulation in the synapse core, and it can be greatly improved by using cascode current mirrors with large lengths. We also noticed that the weight cannot reach a value of $1\ pu$. This is caused by excess leakage at the weight node. The cascodes discussed earlier help solve this problem because of the stack effect [105]. An increase of the value of the weight capacitor can also improve the circuit performance. Another nonideality of this circuit implementation is that some transistors move out of the subthreshold regime and into strong inversion. We noticed that this causes an asymmetry in the STDP learning window. To compensate for this asymmetry, we increased the delay of the calcium waveform's onset as observed in Fig. 4.8.

A further nonideality not shown here but observed during our experiments is also related to the delay circuit. The width of the output pulse is dependent on the bias voltage V_{del} used to set the delay. This is caused by the low gain of the inverter $B1$ and the low gain from V_{out} to $\overline{V_{out_{1\tau}}}$ (see Fig. 4.2a). This problem does not affect circuit operation; it merely makes setting biases difficult because we must change Vpw to compensate for any changes caused by V_{del}. Nonetheless, we can solve the problem by doubling $B1$'s number of stages, as well as adding an inverter between $\overline{V_{out_{1\tau}}}$ and the NOR gate. Finally, we noted that the bistability characteristic is asymmetric because of channel-length modulation. This can be solved by using cascode current mirrors in the bistability circuitry and increasing the transistors' lengths.

This work offers the opportunity to test the performance of current theories of learning in a realistic environment, overcoming the limitations of traditional digital architectures by imitating computational primitives observed in the brain, leading to advancements in combining models of synaptic plasticity with network-wide activity and the scientific exploration of the role of calcium in large-scale, real-time biologically-plausible neural networks [107].

Future designs will feature large arrays of neurons and synapses, and they will be applied to various tasks such as unsupervised pattern recognition and supervised classification.

5 Second Synapse Circuit Implementation

The development of novel circuits requires several design iterations to reach optimal performances. Our first prototype was extensively tested to identify deviations from the desired behaviour to be addressed in a second design cycle. These issues were improved in a second prototype comprising modifications in all circuit blocks and fabricated in the same $180\ nm$ process used for the first prototype. To further extend our research and investigate learning mechanisms in a small neural network, the calcium synapse was embedded into a neuromorphic array consisting of 8 leaky IF neurons and 120 synapses, 72 of which were calcium-based.

This second chip integrates a 22-bit on-chip bias generator [108] which has been developed over many generations of neuromorphic chips such that it is now quite reliable and precise. It is controlled by AER events sent to the chip. This bias generator is typically shared among all synapses/neurons of the same type, so there is indeed the possibility of mismatch. This variability can be mitigated by carefully sizing the circuits and optimizing their layout. I did not perform such optimizations for this chip but in the next chapter I present simulation results that provide insights for further fabrications.

5.1 The Calcium Circuit

I modified the delay circuit in order to overcome the V_{preD} pulse width dependence on V_{del}. I addressed this problem by increasing the gain of the first buffer by using two buffers $B1$ in cascode instead of only one as in Fig. 4.2a ($x4$ refers to the number of inverters in cascode connection) and moving the second buffer ($x2$) before the NOR gate (see Fig. 5.1a). These changes have the effect of sharpening the two signals $Vout_1$ and $\overline{Vout_{1\tau}}$. The pulse width is a function of the rise times of the signals driven by these inverters, which are in turn a function of how fast their inputs change if the gain is low. Increasing the gain therefore puts a tighter bound on the pulse width. However, the disadvantage of this modification is that an extra

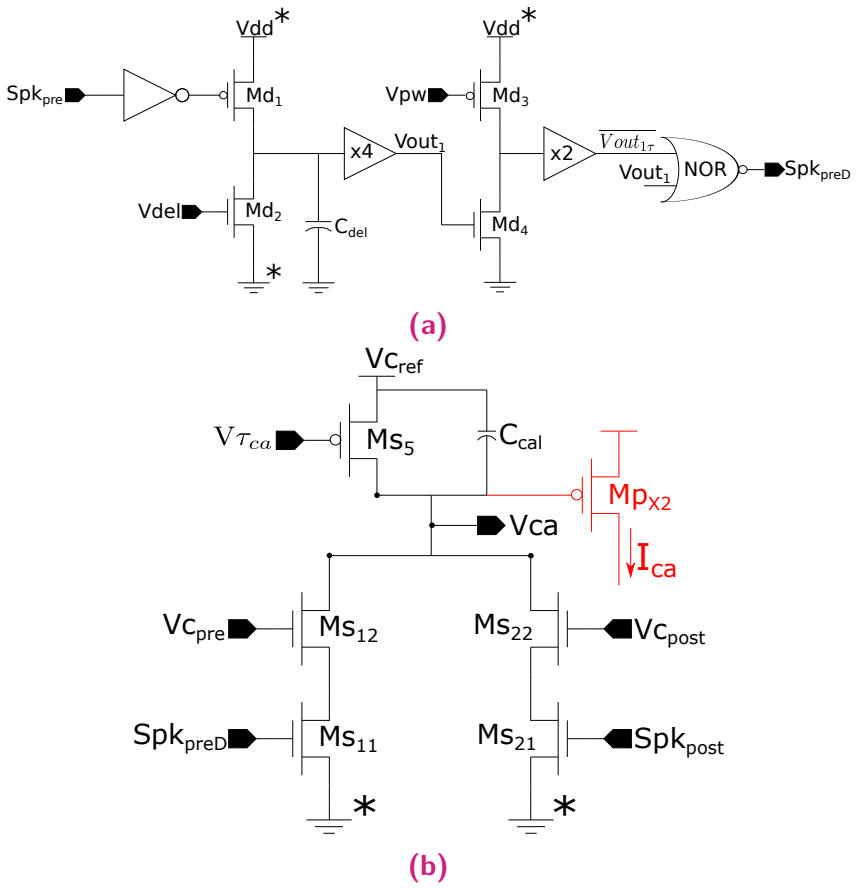


Fig. 5.1: The calcium block circuit is composed of a delay circuit (a) and the Calcium core circuit (b). This delay circuit overcomes the undesired output spike pulse dependence on V_{del} by increasing the gain in $Vout_2$ and $\overline{Vout_{1\tau}}$.

inverter generates additional power consumption; in order to reduce the power consumption an additional feedback circuit to reset Vc_{del} upon it reaches the threshold voltage of the inverter can be added (see Fig. 4.4 Vc_{del} signal).

In our first calcium circuit version we used a DPI circuit for generating the calcium waveform (Fig. 4.2b), this circuit includes an extra bias input to set the calcium dynamics limit; however, when increasing the number of neurons and synapses a big number of required pins can be problematic. By excluding the differential pair Ms_3 and Ms_4 in the DPI Vca limit is defined by the power rail saving an extra pin and silicon area (Fig. 5.1b); this configuration is called Linear Charge-and-Discharge synapse and is similarly used in [109].

Process features	0.18μm, 1 poly, 6 metal
synapse size	56.09 $\times$ 21.04 μm^2
$C_{del} = 102fF$ (MIM) size	12 $\times$ 4.3 μm^2
$C_{ca} = 582fF$ (MIM) size	12 $\times$ 24 μm^2
$C_{syn} = 402fF$ (MIM) size	12 $\times$ 16.6 μm^2
Supply voltage	1.8 V

Table 5.1: Technology and layout features

5.2 The Synapse Core and The Bistability Circuits

In order to address the fact that the weights do not reach the saturation voltage, we employed cascode current mirrors in the weight update circuitry and the bistability circuitry. The stacked transistors reduce the leakage current [105] when no current is being sourced onto the capacitor. We also used a larger capacitor so that we can increase the OTA's tail current and achieve the same bistability slew rate as before. Thus the leakage current is negligible compared to this larger bias current.

The cascode current mirrors in the OTA also address the variation of the bistability current for high and low weight values. They increase the device's output resistance, thus allowing it to source a constant current for a larger range of weight voltages. In addition to adding cascodes, we also increased the devices' lengths to further increase output resistance. The improved circuit for the synapse core and the transconductance amplifier are shown in Figs. 5.2 and 5.3 respectively.

The final layout of the full synapse circuit is shown in Fig. 5.4 and the technology characteristics in Table 5.1. In addition to the modified circuits, I used Metal-Insulator-Metal (MIM) capacitors which have higher capacitance density and use top metals enabling underneath route with lower metal layers; however, the drawback is that contacts to connect this capacitor require considerable distance with the device to satisfy Design Rule Check (DRC).

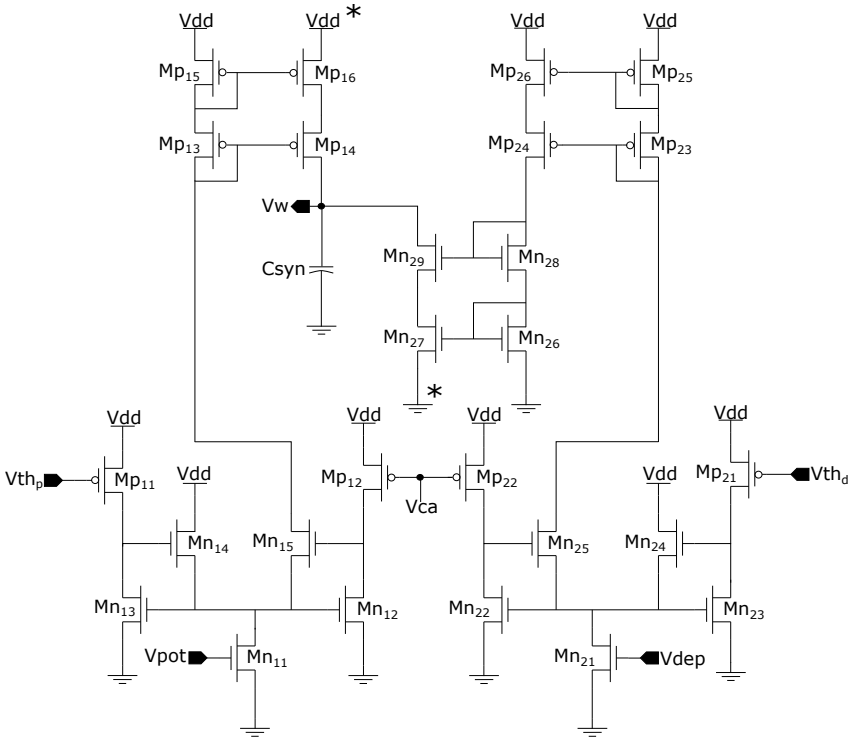


Fig. 5.2: Modification of the synapse core circuit of Fig. 4.3. Cascode current are used instead of the simple one to guide potentiation and depression currents toward C_{syn}. Additionally stack effect is generated to reduce leakage current in the capacitor.

5.3 The Linearizer

The synaptic weight voltage is usually connected to a voltage-current converter block to generate a source current proportional to Vw that will be injected to the membrane potential. One example of this configuration is a DPI [89]. Here the amplitude of the output current after a pre-spike depends exponentially or quadratic of Vw (weak or strong inversion configuration in transistors respectively) and can reach saturation values leading to a fast membrane charge for high values of Vw. In order to reduce strong current supply, an intermediate circuit can provide a linearizer function by generating approximated linear currents as function of Vw. A Source-Follower-Circuit-with-Transconductance-Amplifier (TFS) circuit was proposed in [110] and is shown in Fig. 5.5. The idea here is to bias the transistor Mn_1 in saturation region and vary the output current by taking advantage of the channel length modulation effect where V_{ds} is connected to Vw.

Fig. 5.6 shows simulation results of the output voltage in the linearizer. As observed the swing range is reduced from $< 0.3 - 1.5 >$ to $< 0.4 - 0.8 >$.

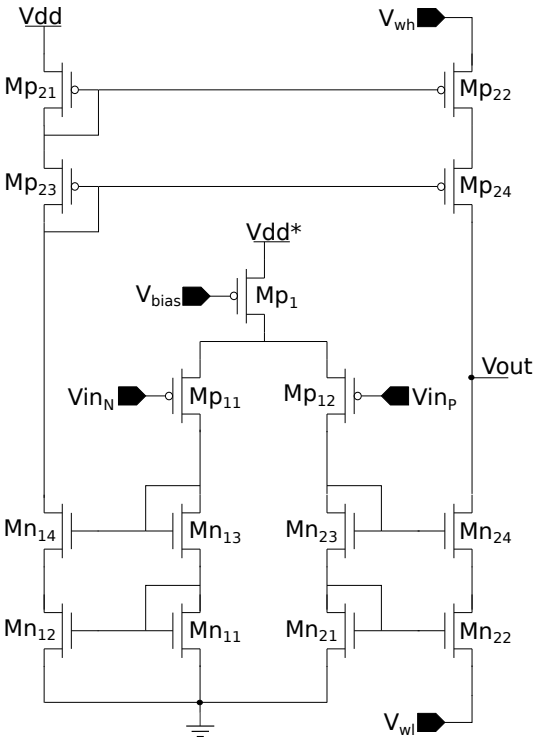


Fig. 5.3: Modification of the wide range transconductance amplifier, cascode current mirrors are used to improve output constant current in saturation region, in addition transistor dimension specially Mp_{2x} and Mn_{2x} were increased to reduce channel length modulation effect and leakage current.

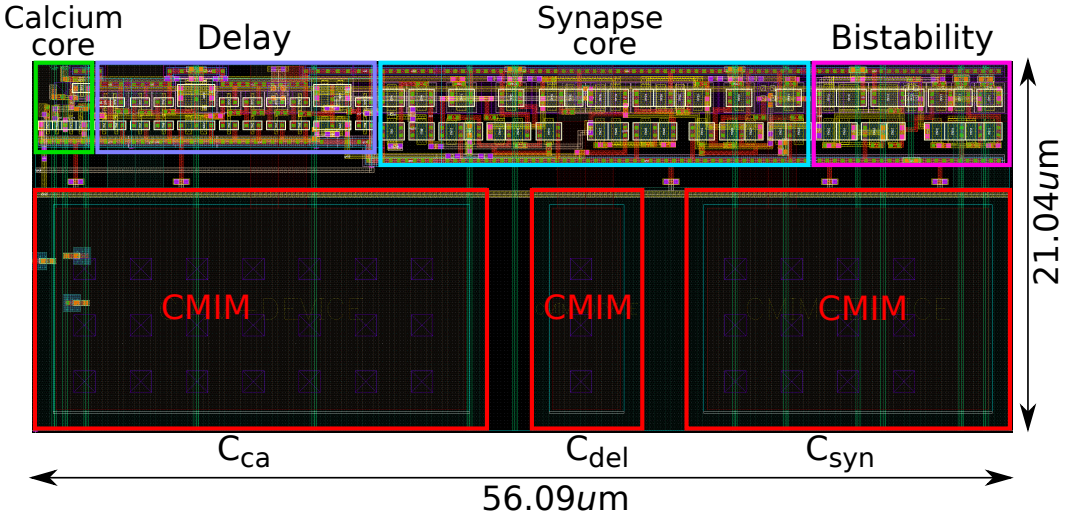


Fig. 5.4: Layout data for the second synapse circuit which was fabricated in AMS $180nm$ technology. Here MIM capacitors are used instead of native ones which has higher density, however still the majority of the area is covered by the three capacitors.

5.4 The Configurable Bias Current Generator

In our firs tape-out we assigned a pin to each reference bias voltage, these pins were connected to Digital-to-Analog Converter (DAC) which were controlled

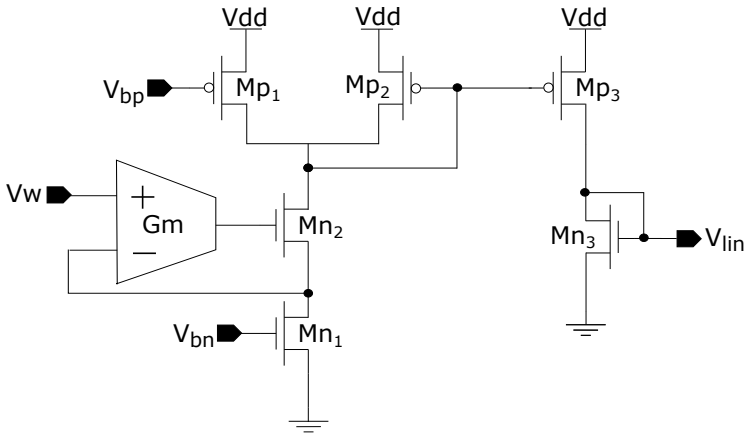


Fig. 5.5: The linearizer circuit originally proposed in [110] consist of a source follower with a transconductance amplifier connected in negative feedback. The transistor Mn_1 takes advantage of the channel length modulation effect and produce currents approximately linear dependent on Vw, V_{bp} bias is used to correct the offset current.

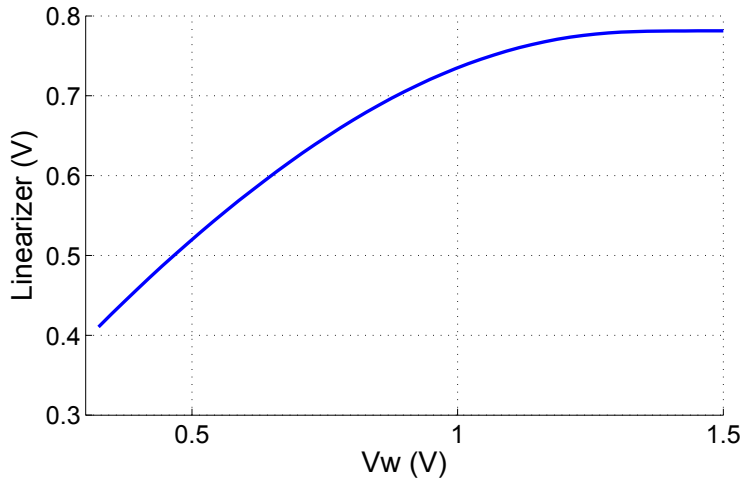


Fig. 5.6: Simulation results of the linearizer circuit. The waveform resembles a linear function respect to Vw however it was not possible to fully remove the voltage offset because of leakage currents, furthermore it is observed that for large gate drain voltages the current tends to move away from a linear function toward a constant one.

by a PC. The disadvantage of this simple configuration is that the bias voltages are sensitive to power supply ripples and also transistors' threshold voltage are dependent on temperature doubling the subthreshold current every 6-8 degrees [43]. In addition, if a chip requires a huge amount of bias voltages we could run out of available pins.

In our second tape-out, we included a configurable bias current generator circuit which is a combination of industry-standard bandgap reference circuits [70, 111] and circuits developed by several research groups [43, 108,

112, 113]. The final layout of this circuit was implemented by the Institute of Neuroinformatics in Zurich.

The configurable bias current generator consists of a master bias (bandgap reference with startup circuit), current splitter, bias buffer, sub-off-current generator and a control system. Here I describe briefly the first three blocks which are also depicted in Fig. 5.7.

The selected bandgap architecture is characterized for implementing a transconductance independent of temperature. The classical idea here is to make I_{out} independent of V_{dd}. For this purpose $M_{p1} - M_{p2}$ current mirror copies I_{bg} to I_{ref} and likewise $M_{c1} - M_{c2} - M_{n1} - M_{n2}$ copies I_{ref} to I_{bg}; so, I_{bg} is bootstrapped [70]; however, because of channel modulation effects on the current mirrors a slight dependence on V_{dd} is expected. The current I_{bg} is obtained by solving the condition in Eq. 5.1.

$$Vgs_{Mn2} = Vgs_{Mn1} + I_{bg}R \tag{5.1}$$

In case all transistor are in strong inversion, we obtain the following expression:

$$I_{bg} = \frac{1}{2\mu_n C_{ox} \left(W/L\right)_N R^2} \tag{5.2}$$

where μ is the electron-effective mobility and C_{ox} is the unit-gate oxide capacitance.

For weak inversion:

$$I_{bg} = ln(4)\frac{U_T}{\kappa R}, \tag{5.3}$$

where U_T is the thermal potential and κ is the body effect coefficient.

Despite this circuit does not implement a temperature-independent reference, it sets a supply-independent transconductance as deducted in Eqs. 5.4 and 5.5 avoiding problems such as noise, small-signal gain and speed [70]. Being more strict a real resistance R varies slightly with the temperature which could lead also to variations in the transistor transconductance.

For strong inversion:

$$gm = \frac{1}{R} \tag{5.4}$$

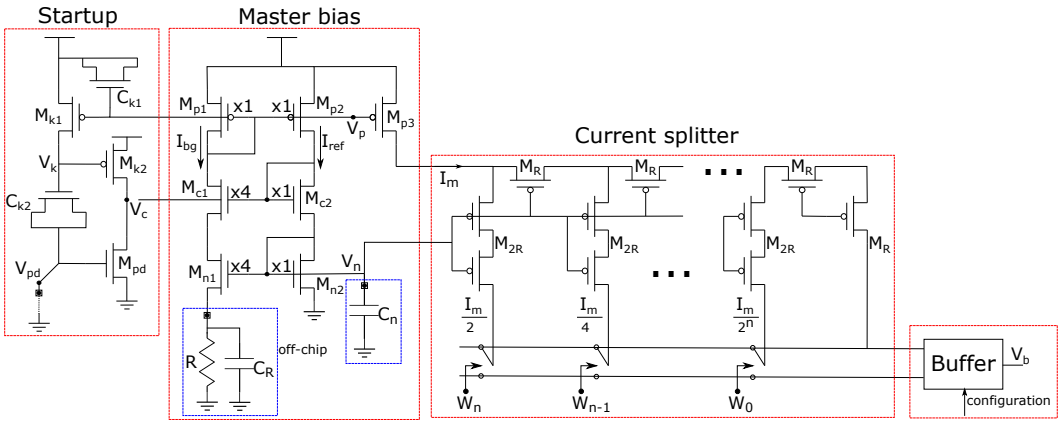


Fig. 5.7: The configurable bias generator circuit proposed in [108, 114] consists of a startup, master bias, current splitter and a configurable buffer. The startup circuit makes sure to activate the Master bias whenever the system is turned on or reset. The master bias circuit uses a bootstrap configuration with dual feedback to generate fixed current values independent of small power supply or temperature variations. The current splitter divides the bias current into geometric ratios. Finally the buffer add all the selected current ratios and generate a respective V_{bias} voltage.

For weak inversion:

$$gm = \frac{\kappa}{R} ln(4) \tag{5.5}$$

The master bias circuit in Fig. 5.7 uses a nMOS cascode current mirror instead of a single one [70] above the resistance terminal to reduce power supply sensitivity; however, this approach is not used in pMOS to preserve voltage swing range (the counter part of cascode configuration is that it increases the headroom voltage).

One critical problem of this circuit is its stability given that the ratio between the nMOS current mirror is larger than 1 (4 in our case). The resistance R is implemented off-chip and therefore includes a several pF capacitance in parallel with it. The circuit can be stabilized by connecting a compensation capacitor C_n several times C_R, or by adding a compensator capacitor parallel to a precreated high impedance node [115].

Generally a startup circuit is added to the bias generator to avoid the undesired solution $I_{bg} = I_{ref} = 0$ in Eq. 5.1 [69]. The function of this circuit is providing an initial current to the current mirror circuits when powering-up. In fig. 5.7 the source current generated in M_{k1} charges the MOS capacitor C_{k2} from an initial value $V_k = 0$ until it reaches V_{dd}; during this transition a

current source is also generated by M_{k2} which likewise activates I_{ref}. The capacitor C_{k1} avoids undesired cut-off operation in I_{bg} when V_{dd} drops suddenly. The power control V_{pd} which in normal operation is connected to ground allows also to shut off the system by setting it to V_{dd}, with this, the activated M_{pd} transistor pulls down V_c voltage which likewise turns off the current mirrors. Setting V_{pd} back to ground returns V_k to low level and therefore restarting the startup circuit.

The current splitter circuit divides the master bias current I_{bg} into geometric proportions of ratio 2^{-k} in each $k-th$ branch. A desired amount of current can be obtained by switching on/off some of the current branches and add them together. The most basic splitter configuration consists of two MOS transistors as shown in Fig. 5.8a which share gate and drain terminals and divide the input current I_{in} into I_1 and I_2 proportional to their device dimensions ratio $\frac{(W/L)_1}{(W/L)_2}$ independently of the transistor region operation [116]. In order to demonstrate this property, a graphical version of the complete all-region model for the CMOS drain current can be used [66], considering the total drain current $I(x) = I_{drift}(x) + I_{diff}(x)$, this can be expressed in terms of the inversion layer charge per unit area at position x as Eq. 5.6 where its implicit component (Eq. 5.7) is plot in Fig. 5.8b.

$$I_d = (W/L) \int_{V_c=V_s}^{V_c=V_d} f(V_g, V_c) dV_c \tag{5.6}$$

where:

$$f(V_g, V_c) = \mu \left(Q_c - \frac{kT}{q} \frac{dQ_c}{dV_c} \right) \tag{5.7}$$

A current injection I_{in} in the splitter shifts the Vm voltage from an initial sate Vm_1 to Vm_2, consequently currents ΔI_1 and ΔI_2 vary proportional to the area under the function $f(V_g, V_c)$ between $Vm_1 - Vm_2$ and given that this area is the same for both currents we can cancel this term when we diving $\frac{\Delta I_1}{\Delta I_2}$ resulting in Eq. 5.8.

$$\frac{\Delta I_1}{\Delta I_2} = \frac{(W/L)_1}{(W/L)_2} \tag{5.8}$$

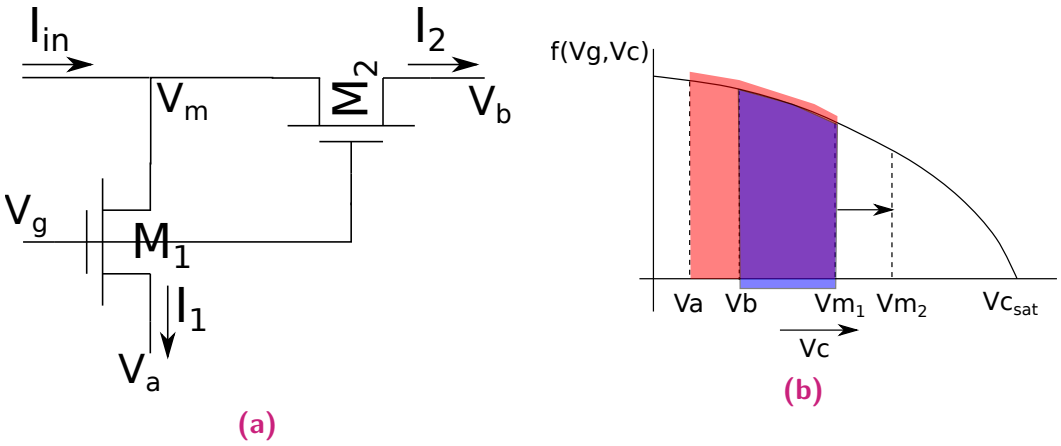


Fig. 5.8: Basic configuration for a current splitter. (a) The circuit consist of two MOS transistors which share at least the same gate and source terminals and can be configure at any regime. (b) A graphical version of the All-Region CMOS model is used to demonstrate its split proportion when a current variation occurs, in this case if an initial $V_c = Vm_1$ voltage shift to V_{m2} the integral of the function $f(V_g, V_c)$ shift in the same value independently of the initial state V_a or V_b and hence current variations depends only of the area between Vm_1 and Vm_2.

Finally, considering V_a and V_b grounded and $(W/L)_1 = (W/L)_2$ we demonstrate that $I_1 = I_2$.

The buffer circuit is based on a current conveyor architecture, which decouples the input current with the output voltage V_{bias}. Switches are included to enable/disable alternative configurations such as cascode output (SW_1), nMOS (I_{nref}) or pMOS (I_{pref}) reference voltages (SW_6), and sub-off current levels (SW_2). The nMOS bias configuration reaches V_{bias} values of $< 0- \sim 1V >$ while the pMOS bias reaches $<\sim 700mV - 1.8V >$.

5.5 The Neural Network Block

This second chip version comprises an array of 8 neurons, each one receiving input from 9 calcium synapses which allows us to implement simple network experiments. Each neuron also receives input from additional 5 synapses (used for independent experiments of our research group) implementing various short-term adaptation mechanisms and filtering properties (e.g. short-term depression, short-term facilitation, band-pass filter and elementary motion detection) and 1 inhibitory synapse. We considered two types of calcium synapses, one that includes the linearizer circuit between the synapse and the DPI and another that contains a comparator (implemented as a

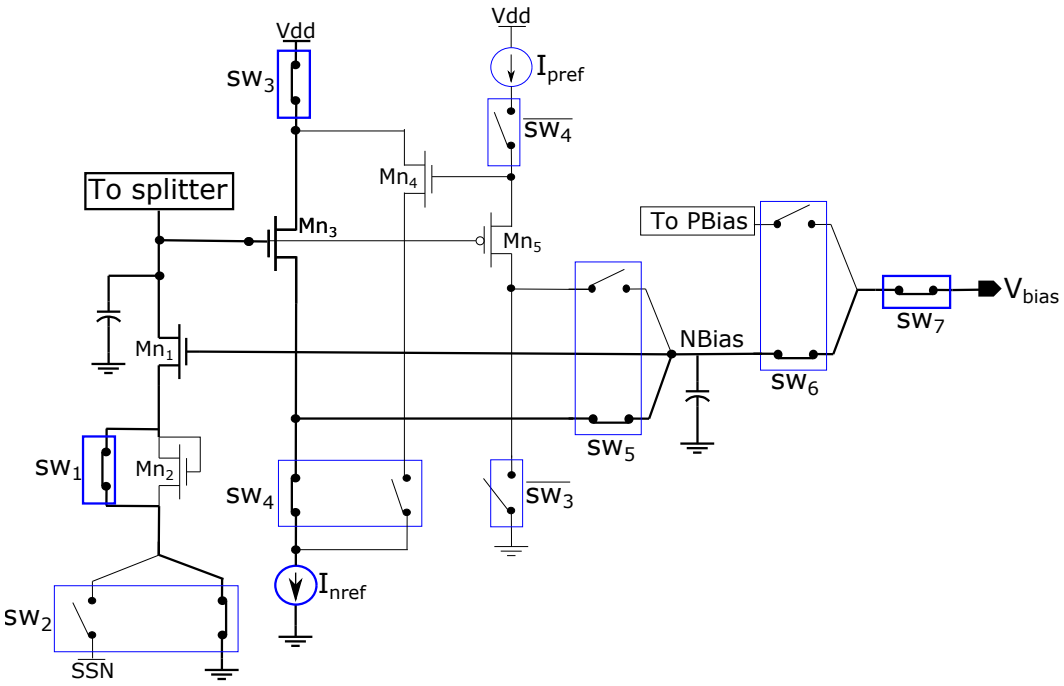


Fig. 5.9: Architecture of the buffer circuit which consist of different configuration options such as NMOS-PMOS output, shifted source bias (which generates saturation currents when $V_{gs} < 0$) and cascode configurations (which increases the output impedance). The default option with NMOS reference current is highlighted in the graphic. The operation is based on a current-controlled conveyor circuit defined by the transistors $Mn_1 - Mn_3$, here the current in Mn_1 sets the voltage in the node NBias independent of its drain voltage and hence the node is clamped to a constant value ensuring that the voltage V_{bias} follows faithfully the the input current from the splitter.

transconductance amplifier with a buffer) circuit instead. Monitor circuit blocks let us measure some test signals such as V_{ca}, V_w, V_{lin} and the output spikes. The system also integrates an asynchronous comunication protocol called AER [117–119]. In AER, a transmitter emits address events of the spikes generated by the neurons; a receiver decodify these bits and can stimulate designated synapses to implement recurrent connectivity patterns. Fig. 5.10 shows the block diagram of the neural network.

5.6 Simulation Results

In the second designed version simulation results demonstrate considerable improvement in potentiation, depression and bistability operations in terms of swing range and symmetry. In order to test the simulation results, I used here the same protocols than in the first chip version.

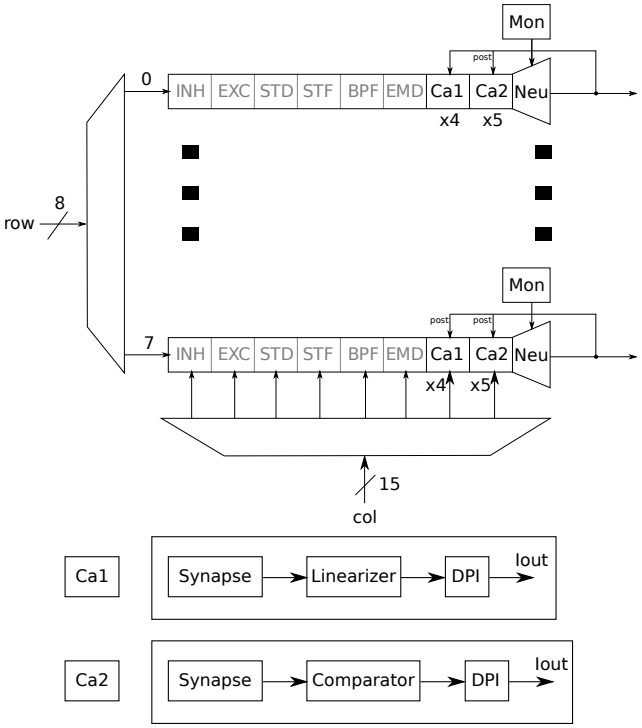


Fig. 5.10: Block diagram of the neural network containing the calcium-based (highlighted) and additional synapses models. A total of 9×8 calcium-based synapse circuits were included together with 8 neurons. Two types of calcium circuits were fabricated, one which includes a linearizer block (Ca_1) and another, a comparator (Ca_2), those are the interface that connect the synapse with the DPI to generate currents which charge the membrane capacitor. Additional peripheral devices such as multiplexor and asynchronous circuits are used to address information to a respective synapse. Furthermore, monitoring circuits are also comprised to trace internal voltages values in the synapse (V_{ca}, V_w and V_{lin}) and neuron (output spikes).

5.6.1 Bistability

Simulation results for the improved OTA are shown in Fig. 5.11. As observed the current curves approach more to an ideal antisymmetric figure, in addition the same bias voltages V_{bias} now produce lower current levels compared to Fig. 4.7; this gives the advantage to have a better bias range for the desired voltage. In learning terms, it reaches a similar probability for getting a high or low memory value.

The bistability circuit implements approximately an hyperbolic function except near the boundaries where it moves to zero from a constant value; on the other hand, the computational model implements a 3rd degree polynomial; this comparison is shown in Fig. 5.12. In order to get the bistability component in the simulation an indirect approach was considered. First in

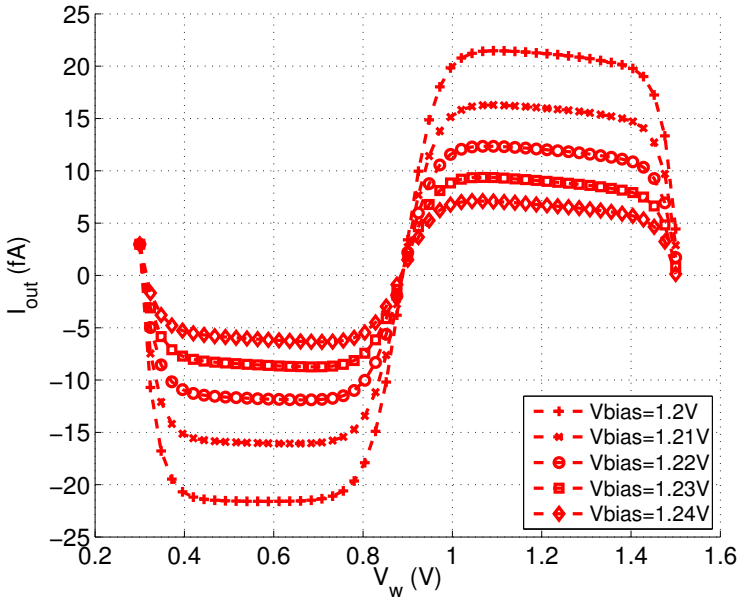


Fig. 5.11: Simulation results of the transconductance amplifier in Fig. 5.3, as observed antisymmetric waveform is improved respect to the simple OTA thanks to the cascode connection which reduces the channel length modulation effect in the transistors connected to the capacitor C_{syn}.

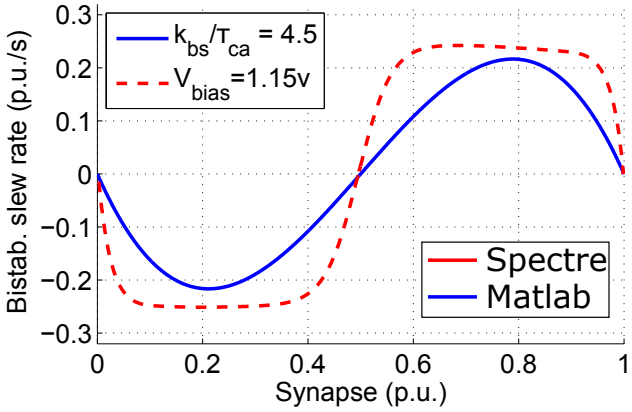


Fig. 5.12: Slope for the improved circuit. Here the positive and negative slope are symmetric. This is achieved by using cascode current mirrors and increasing the channel length of the transistors to reduce channel length modulation effect.

the learning circuit an initial Vw value is set (one slightly above $0.9v$ and other below $0.9v$), then Vw evolution through the time is recorded, later a high degree polynomial regression is used to approximate Vw vs. time and finally Vw (regression) vs. its first derivative is plot.

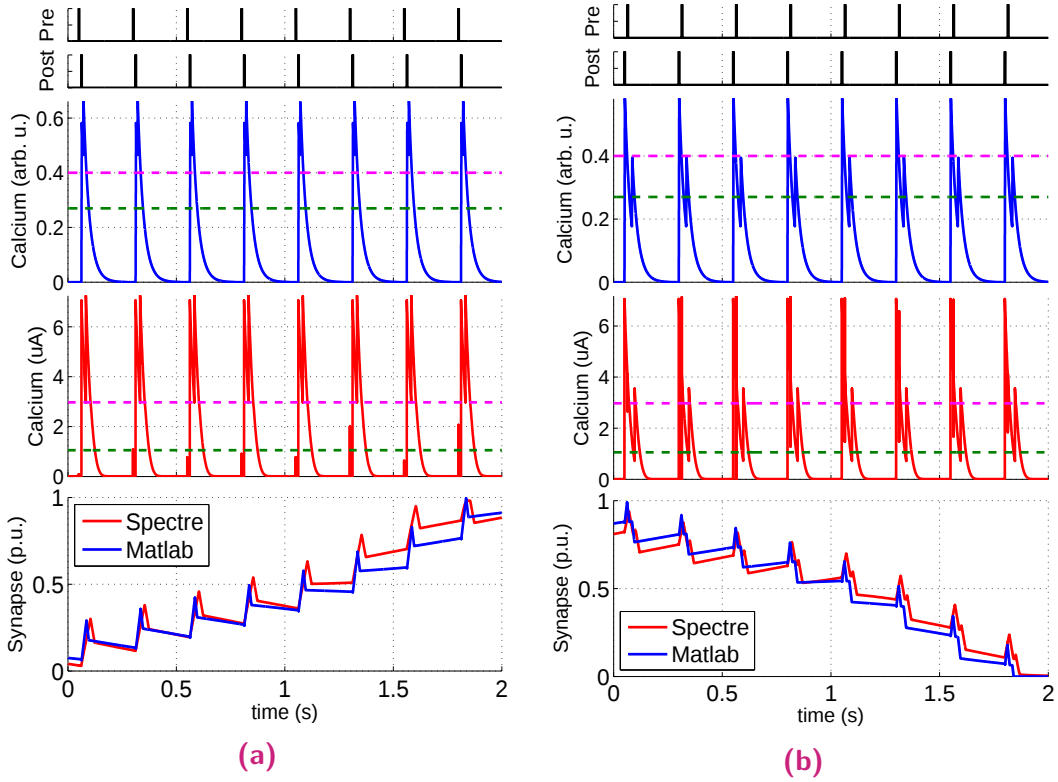


Fig. 5.13: Synapse evolution for (a) potentiation and (b) depression in the improved circuit. A better matching is obtained because of a more symmetric operation in the bistability component and the generated current for potentiation, depression and bistability is less dependent on Vw as in the previous circuit.

5.6.2 Potentiation and Depression

Synapse evolution for potentiation and depression are plot in Fig. 5.13. Compared to Fig. 4.9, the second version resemble more to the computational model. This occurs because potentiation and depression jumps generated from a single spike pair are less dependent from the initial state of Vw compared to the former version and because the bistability component is represented by a more precise antisymmetric waveform.

It was observed however that in the synapse core circuit (Fig. 5.2) Mp_{14} and Mn_{29} convey a leakage current produced in Mn_{15} and Mn_{25} when Vca is greater than the threshold voltages. If V_{pot} and V_{dep} are greater than the transistor threshold voltage ($\approx 0.37V$) the leakage current has considerable value and degradate the synapse performance. The system operation can be improved by adding a feedback mechanism to turn off Mp_{16} and Mn_{27} when Vca is greater than the depression and potentiation thresholds.

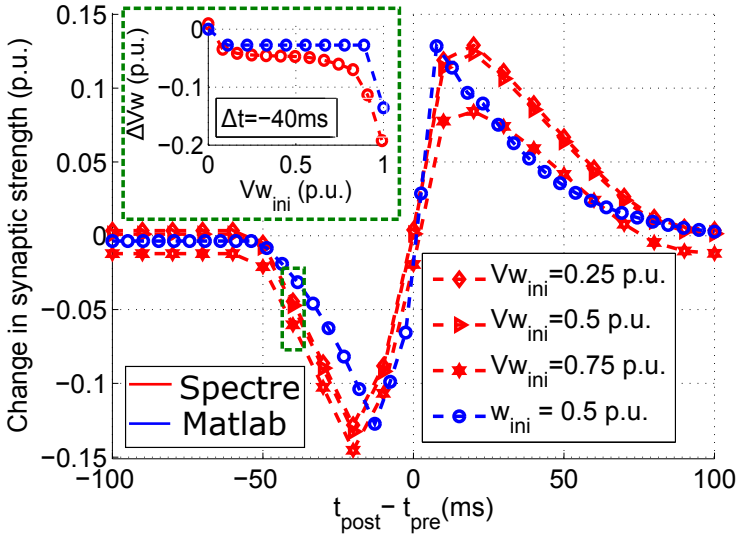


Fig. 5.14: Synapse learning waveforms for Model (blue) and hardware simulation (red) for the improved circuit. Here the learning waveforms in the hardware simulations vary less for different Vw_{ini} values as in Fig. 4.11,this is achieved by improving calcium and synapse core circuit blocks.

5.6.3 STDP Waveform

The STDP waveform is generated by calculating the variation in the synapse potential after a spike pair occurs considering different timing between pre- and post-spikes. The change in normalized weight is plotted in Fig. 5.14 for $t_{post} - t_{pre}$ values between -100 and $+100$ ms. Compared to the first version circuit (Fig. 4.11) the results here are less dependent on Vw_{ini} except for values close to the saturation ones (Vw_h or Vw_l). The inset figure shows the "change in synaptic strength" for timing $t_{post} - t_{pre} = -40ms$, where the Vw values are almost constant for intermediate initial states of Vw_{ini}.

5.6.4 Configurable Bias Circuit

I provide here simulation results for the Master Bias block which resemble a switch-on operation in the power supply considering noise coupling. The noise signal is represented by a sinusoidal waveform with amplitude $150mV$ and frequency 5MHz as shown in Fig. 5.15. The setting time is reached after $\approx 5.94\mu s$ (for a 5% of accuracy), around this time I_{out} and I_{ref} converge to the same value because of both feedbacks. AC Simulation results also state that V_{dd} variations at small frequencies are attenuated in $-65dB$ in V_n, and it reaches a worst attenuation case at 100MHz with $-40.7dB$ (simulation results without load).

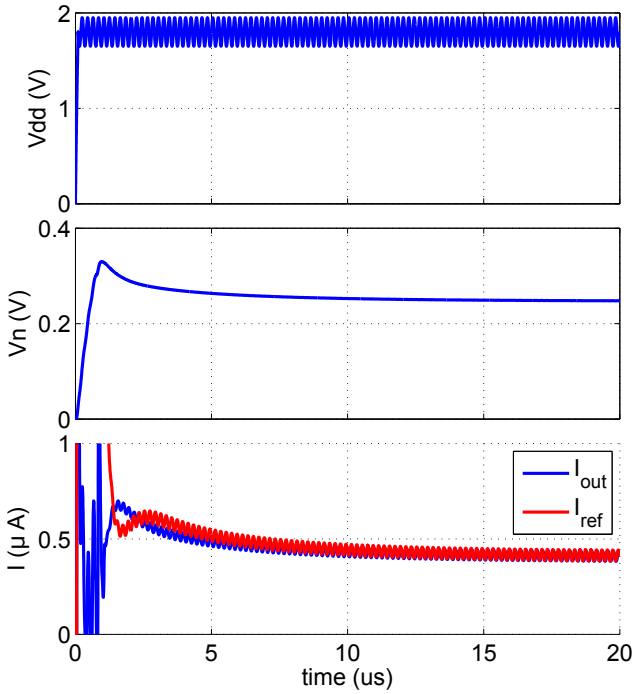


Fig. 5.15: Simulation results of the master bias circuit in Fig. 5.7, a turn on and noise events are mimicked by rising the power supply voltage (vdd) and adding a sinusoidal signal with amplitude 150mV and frequency 5MHz respectively. As observed the reference voltage Vn in the bootstrap circuit stabilizes after 15ms to 0.25V independent of the Vdd value, this time is reached when both currents I_{out} and I_{ref} get the same value as consequence of the dual feedback.

5.7 Hardware measurement results

Given that the second version chip integrates an asynchronous communication protocol and a programmable bias current generator, an automation systems was implemented. We wrote python scripts to generate and send the bytes responsible for configuring the bias values and operation mode in the chip. This includes a learning or recall mode operation, input spikes timing and synapse addressing. I also implemented a control system for the oscilloscope measurements in order to record and plot the data. The python library PyIVI was included for this purpose (http://www.ivifoundation.org/).

5.7.1 STDP Measurement Results

The first step to get the STDP graphics consist of measuring the variation in the synaptic strength ΔVw for different spike timings. In order to facilitate the comparison between the measurements and the computational model

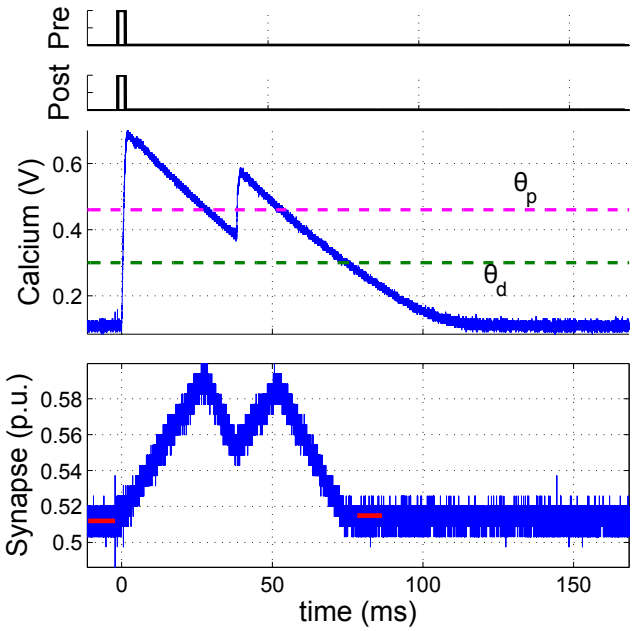


Fig. 5.16: Measurement result of the calcium voltage V_{ca} and synaptic strength Vw when $t_{post} - t_{pre} = 0$. As expected a higher delay time is set in the calcium signal for the pre- spike to keep the synapse invariant at this timing. In order to process the data an initial measurement at $< 10ms - 2ms >$ before the spikes is taken and another one after $< 80ms - 86ms >$ of both spikes as denoted by the red lines, the difference of these values provide the change in the synaptic strength, in addition its standard deviation was calculated.

values, the calcium signal was inverted and shifted ($V'_{ca} = Vc_{ref} - V_{ca}$) and the synaptic weight results (V_w) were normalized with respect to the maximum (V_{wh}) and minimum (V_{wl}) headroom voltages. $V_w(p.u.) = \frac{V_w - V_{wl}}{V_{wh} - V_{wl}}$. An initial measurement is taken before any spike occurs (pre- or post-) and another one after the spike-pair when the synapse stabilizes, then those values are subtracted to get ΔV_w.

Fig. 5.16 shows an example for $t_{post} - t_{pre} = 0ms$ (which is applicable also for $t_{pre} - t_{post} > 0$). From $10ms$ to $4ms$ before t_{post} an average value of Vw_{ini} is calculated, and between $80ms$ to $86ms$ after t_{pre} another average value of Vw_{end} is obtained. Similarly, when $t_{post} - t_{pre} > 0$ an average value Vw_{ini} from $10ms$ to $4ms$ before t_{pre} is calculated, and around $80ms$ to $86ms$ after t_{post} an average value of Vw_{end} is estimated. In both cases, a value of $80ms$ was chosen to ensure that the synapse modification reaches a stable Vw_{end} value for a range of $|t_{post} - t_{pre}| < 100ms$ with $10ms$ of step. Here the mean value μ_{single} and standard deviation σ_{single} are calculated for each timing; this variations are caused mainly by the oscilloscope precision and the signal noise.

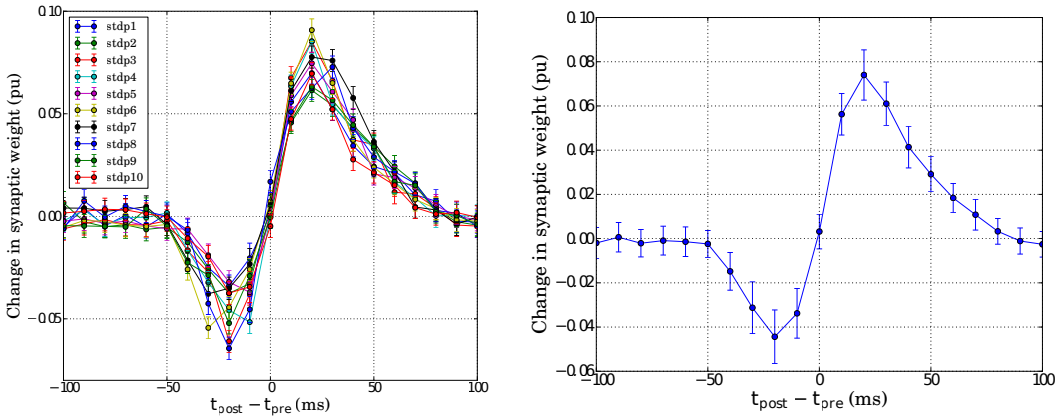


Fig. 5.17: Depression-Potentiation STDP characterization of the measured data. (a) By using the protocol in Fig 5.16, 10 groups of STDP were generated each of them consisting of $t_{post} - t_{pre}$ timings in the range of $< -100ms - 100ms >$, in addition the standard deviation of each point was calculated and shown as vertical bars. (b) The mean value of all the 10 STDP groups is shown together with its standard deviation.

As second step I recorded 10 samples data for each timing. Then the mean value μ_{group} and standard deviation σ_{group} for each timing group are calculated. Variations in each trial are caused by imprecisions in the hardware operation such as latency and pulse widths mismatch. Fig. 5.17a shows the results for all the timings grouped in STDP waveforms. Finally, the total standard deviation for each timing is calculated as $\sigma_{tot} = \sqrt{\sigma^2_{group} + \sigma^2_{single}}$. σ_{tot} reaches higher values when potentiation and depression levels are maximum as observed in Fig. 5.17b. This is because the calcium waveforms generated from pre- and post- spikes are overlapped, consequently a slight variation in one parameter is amplified in the final Vca waveform.

5.7.2 Potentiation and Depression

The synaptic weight dynamics for potentiation and depression are shown in Fig. 5.18 where synapse values are normalized considering the maximum (V_{wh}) and minimum (V_{wl}) power rails. Better symmetry at each synapse modification due to spike pairs is observed here compared to Fig. 4.15; this occurs because the STDP is less dependent on Vw_{ini} and because of a better antisymmetry in the bistability waveform.

Synapse can not reach the maximum value 1 because depression effect turns stronger or equal than potentiation closer to this limit producing therefore

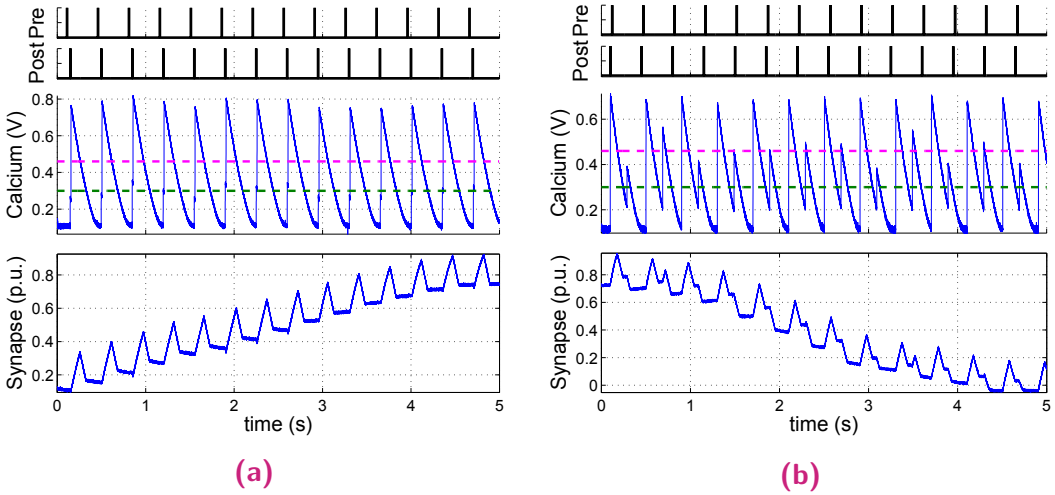


Fig. 5.18: Synapse evolution (normalized) in presence of spike trains with different $t_{post} - t_{pre}$ timings. (a) Potentiation measurements for $t_{post} - t_{pre} = 40ms$, an increase in the synapse after each spike pair is observed given that the calcium signal spend considerable time above θ_p that overcome the $\gamma_d \mathcal{H}\left(c(t) - \theta_d\right)$ factor, in addition a change in the bistability direction is observed when the synapse crosses $W_* = 0.5$. (b) Depression measurement (normalized) for $t_{post} - t_{pre} = -15ms$ timing, mainly the depression effect is caused by each second peak in the calcium signal which generates greater effect in the depression $\gamma_d \mathcal{H}\left(c(t) - \theta_d\right)$ than in the potentiation, in the first peak both effects cancel each other, the bistability sign is positive for values above 0.5 while otherwise is negative.

oscillations near the top value when further spike pairs occur ($\approx 4.5ms$ in Fig. 5.18a). On the other hand, synapse can reach the minimum value 0 when further depression occurs near the lower limit ($\approx 4ms$ in Fig. 5.18b).

5.7.3 Bistability

In order to characterize the operation of the bistability block, this circuit has to be measured in abscence of spike pairs so that only the current generated for the transconductance amplifier in positive feedback with unit gain configuration drives the synapse modification; the slew rate for the bistability is shown in Fig. 5.19. Compared with the first version chip where bistability did not work properly above the threshold value $w_* = 0.5$, here low power design techniques reduce leakage currents therefore improving operation above the threshold w_*, although setting very small current in the bias of the OTA can limit its operation range (I_{ref} and $1.37I_{ref}$ in the figure). I_{ref} is the reference current set in the programmable bias generator current; this referential cur-

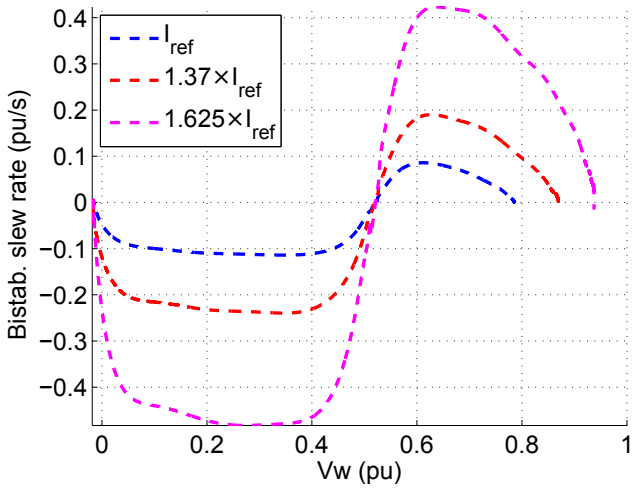


Fig. 5.19: Bistability slew rate measurement results of the circuit in Fig. 5.3 for different V_{bias} values (data normalized in Vw). As observed the change in sign occurs at $W_* = 0.5$. In the case of negative values of the y-axes the circuit resembles a constant and moves toward zero when it approximates to the boundaries 0 and W_*. In the case of positive ones the slew rate can only reach a constant for high enough V_{bias} values; this happens because at sub-off current levels, leakage current in the opposite direction (NMOS transistors) are not neglected.

rent is later reduced by a post-process circuit which includes a current mirror with different transistors' dimensions ratio and source degeneration [70].

5.7.4 Linearizer

The synaptic weight represented by the stored voltage V_w in a capacitor is used to generate a current flow whenever a pre- spike occurs and whose value depends on this voltage. This current charges the neuron membrane potential and eventually generates an action potential. A common circuit for this operation is a DPI which is characterized by a non-linear relationship $(I_{out})_{max}$ vs. Vw (exponential in the case of weak inversion and quadratic for strong inversion), consequently while low Vw voltages generate barely leakage currents, high Vw generates saturation currents. The purpose of the linearizer circuit is to considerably weaken saturation current values for high Vw and strength low current values when Vw is low generating a smooth transfer function [110]. However, the main drawback of the circuit is the complexity to find the precise bias values for controlling the channel length modulation and its offset correction factor given that these characteristics are very sensitive. Measurement results for the linearizer are shown in Fig 5.20; compared to the input voltage range $\langle 0.3 - 1.5 \rangle$ the swing range is limited to $\langle 0.1 - 0.8 \rangle$.

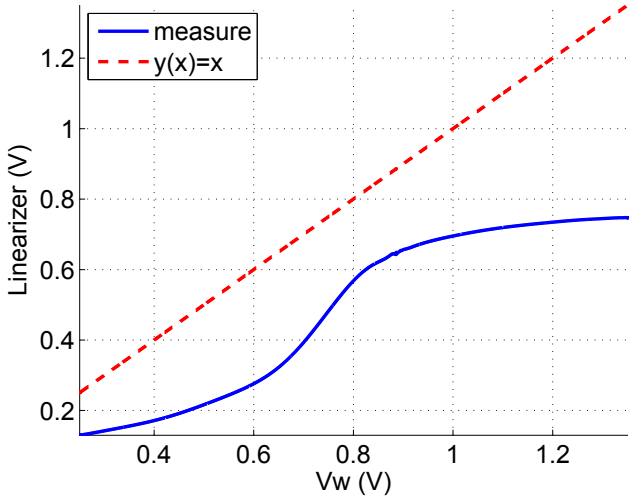


Fig. 5.20: Post- processing result after measuring Vw and V_{lin} evolution through time without spike pulse presence. As observed the linearizer attenuates the Vw value, an inflexion point near $Vw = 0.8$ occurs because the Mn_1 transistor of Fig. 5.5 moves from triode to saturation region in which the function tends to be linear. The linearizer range value was chosen in order to generate significant current values capable to increase the membrane potential overcoming leakage current without generate saturation currents.

5.8 Discussion

In this chapter I presented modifications of the first chip version which came up after identifying inaccurate operations during its measurement. This revised version intended to solve the small operation range of the bistability component by reducing the leakage current in the capacitor that represents the synaptic strength. In addition, a simplified circuit was proposed for the calcium core by removing the threshold parameter Vth_{ca} and thus reducing the number of bias voltages; modifications in the delay circuit were also presented such as adding inverters in cascode configuration to generate more accurate pre_D spikes after this stage, and therefore avoiding interdependence of parameters V_{del} and V_{pw}. Simulation and measurement results show better symmetry in the STDP waveform, potentiation and depression graphics. Furthermore, some extra blocks were added such as a bias generator and an AER which make the circuit more independent of power supply or temperature variations and allow communication among neurons respectively. The linearizer circuit was also presented which works as an interface between the synaptic weight and the current injection in the neuron to avoid saturation currents levels. Concerning the layout design, some modifications included the use of CMIM capacitors which has higher density and thus require less silicon area than native capacitors used in the first version.

Still there is room for improvement with respect to power consumption and mismatch that should be considered in the next chip version. A main source of current consumption is the transistor Md_2 in the calcium core block which sinks current from the capacitor C_{del} during large time ($< 20 - 50 >$ ms), here the drain voltage can be pulled down through a feedback mechanism immediately after it crosses the threshold voltage of the embedded inverter in the buffer $x4$ hence reducing power consumption. Mismatch analysis will be described in a further chapter.

6 Network Operation

In the previous chapters I described the computational models of the neuron and the synapse which are the basic structures of the nervous system. The neuron is a cell that generates electric potentials to transmit information to other cells and the synapse is a structure that connect these neurons. The connection of neurons form a network which is characterized by a topology and connection strengths w_{ij} (from neuron j to neuron i) among pair of neurons. Cognitive functions such as perception and learning motor skills are achieved by specialized networks with many neurons.

In this work we chose the calcium-based model for the synapse, which was also implemented in VLSI. This circuit is bistable in long time scale given that a positive feedback with low slew rate drives slowly the synaptic strength to only two possible values. In the case of short time scale, the synaptic strength is modified by the potentiation and depression effects, and the effect of the bistability is negligible; therefore, infinite states are possible. If a bistable mechanism is not provided, a huge amount of states make convergence without a feedback mechanism unlikely [99].

Neurons in the neocortex region are organized in populations with similar properties where each neuron receives thousands of synaptic inputs. This synaptic connectivity can be modelled as network configurations which are capable to provide solutions to constrained problems [120]. Exploiting the use of the Calcium-based model in this context will lead us to capture some of the canonical principles observed in cortical networks which is therefore useful for understanding its organization, operation and computational potential.

Recurrent neural circuits in the neocortex interconnect neurons within a particular region [121]. Synaptic weights here are set through LTP and LTD mechanisms [122] that take place during learning phase which characterizes for a dominated CA3[1] firing pattern [123]; these weights are later used in the network to retrieve the pattern of activation of the stored memory. In such networks the synaptic weights provide the long-term storage of available

[1] CA3 is one of the subfields that compound “hippocampus proper” which is related to memory and hippocampal learning processes.

memory patterns and the initial network activity of memory retrieval (input data with noise) determines which memory is recalled by choosing the most similar.

Much of the architecture of regions of the brain are made from a proliferation of simple local circuits with well-defined functions; therefore, by understanding the bridge between simple circuits and the complex computational properties of higher nervous systems we can progress in the target of developing new computational capabilities from the collective behavior of large number of simple processing elements [124]. Our basic fabricated neural network consisting of 8 DPI neurons [125, 126] each of them connected to 9 calcium-based synapses provides a platform for this purpose as a necessary step to plan larger and more complex circuits which are essential for cognitive computation.

The most commonly used storage device in computers is Random-Access Memory (RAM) which is characterized by taking the same time to retrieve a word irrespective of its the physical location in the array. However, many applications require searching items in some data structures, such as a data tables in memory; thus, if the data is very large, time-consuming is huge because of two factors: the time used in sending information back and forth to calculate the effective address of the necessary data word (von Neumann bottleneck), and second, the serial nature of the processing, where each piece of information must be handled sequentially. On the other hand, Content-Addressable Memory (CAM) is defined as a collection of storage elements, called associative cells, which are accessed in parallel on the basis of data content rather than by specific address or location, overcoming in this way the problems of RAM [127].

Associative memory, which uses the hardware principles of CAM,retrieves a full item when partial or approximate representation of a stored item is presented. Here, weights are adjusted in the learning phase so that the network has a set of discrete fixed points (energy minima) identical to the patterns of activity that represent the stored memories. Considering few patterns stored, these fixed points can totally or closely retrieve the memory patterns by finding the fixed-point that most closely matches the initial state of the network [39]. The pattern that represents the stored memory can be recalled therefore by reactivating only a small fraction of the stored memory.

This kind of associative network such as in recurrent neural networks has been found in the hippocampal system [128].

The study of learning comprises how synapses are affected by activity during training. The training procedures are classified as supervised, unsupervised and reinforced learning, in this project we focus on the former. In supervised learning, inputs and the corresponding desired outputs are set during training so the network provides the correct answer; this kind of learning requires the availability of labelled data (data that belongs to meaningful class that is desirable to know). Two basic problems addressed in learning are the relationship between the input and output patterns provided during training, and the appropriate outputs for inputs that were not presented during training but are similar [39].

Here I present measurement results for one synapse connected to one neuron as well as two synapses connected to one neuron (simple perceptron). Despite that more than two synapse connections are required for any meaningful classification, the experiments here provide insights of the calcium-based synapse operation in a network and the improvements to consider in the next design version; more complex experiments were not possible in our systems because of mismatch effects as it will be explained in the next chapter.

6.1 Single Synapse Learning

It is believed that the neural code that conveys information processing in the brain is correlated with modulations in firing rate of the neurons in a population. Although the neural spiking is not very reliable and has a lot of variability among neuronal responses, if it is seen as unique block the average population rate is clearly different for different patterns. Therefore, mental functions such as perception and learning motor skills are not accomplished by single neurons alone.

A recognition process can be abstracted as mapping functions which are functions of vectors that restrict the output to a limited set of values. One simple mapping function is a look-up table where all the possible sensory input vectors have corresponding internal representations. Mapping functions are important in many brain processes and have dominated models in cognitive science in the form of multilayer perceptrons [123].

As we described in the previous section 2.1, the membrane potential of a neuron is modelled as a function of the sum of synaptic currents generated by firings of presynaptic cells. This sum of synaptic currents depends on the synaptic strength value for each synapse; therefore, it is important to set a suitable range for the current values depending on the number of synapses that are connected to a neuron, otherwise saturation in the membrane potential can occur. Here I explore two cases were this current flow is high and therefore one input spikes can generate instantly an output spike, likewise I present another case where parameters are set so that neuron firing is obtained only by an accumulation of input spikes.

In order to demonstrate the correct learning operation of a synapse-neuron block following a supervised learning approach, a simple protocol consisting of measuring one output neuron activity before and after learning was assigned. This simple experiment allows us to calibrate the system by finding parameter constraints of the system that provide slow learning rate and potentiation/depression jumps high enough to overcome bistability decay, as initial step I set the same values obtained in the simulations stage previous to the chip fabrication; however, because of fabrication mismatch and parasitic devices this values are slight different from the estimated; therefore, this variation is corrected in the hardware by trial an error approach. This configuration is shown in Fig. 6.1 where the output spikes generated by a teacher are feedback in the synapse as post- spike which together with the input (pre-spike) generates synaptic modification. Here two cases are analysed one when the synaptic strength generates saturation currents and therefore one single pre- spike is enough to rise an action potential, another case happens when the linearizer weakens the synaptic strength effect thus action potential is achieved only after an accumulation of consecutive pre- spikes.

Both experiments are divided into two stages which are called learning and retrieval. The former is where training properly occurs through a teacher signal that fires spikes with different rates depending on the desired stored data; the output spikes generated in the neuron block are forwarded to the input terminal in the synapse which sets the post- spike train. The synaptic weight is driven to high or low level depending on the spike timing and the firing rate. The learning mode is configured by setting in high level the input pin EN_{learn}. In retrieval phase the EN_{learn} signal is set to low level which therefore disables the teacher signal and feedback; in this case the output data is a function of the input spikes and the synaptic weight W that was

obtained at the end of the learning stage. This stage is executed twice, once at the beginning for a low level W an another after W changes its value in the learning stage. The learning rule chosen for this experiment is the same presented in Fig. 5.14 which resembles a STDP for low firing rate, and increases the potentiation probability when the firing rate increases.

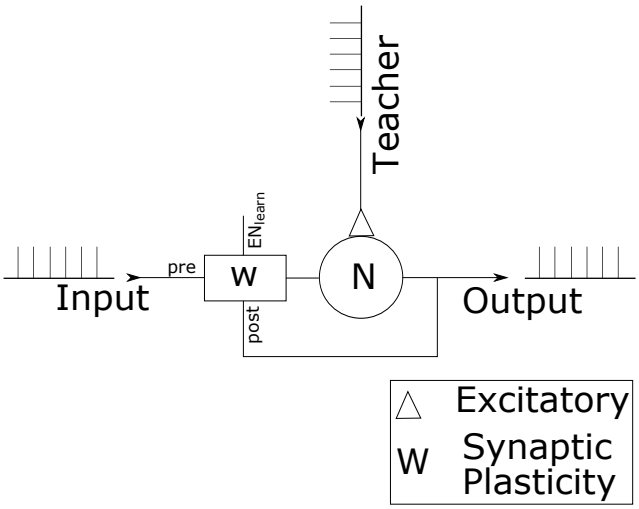


Fig. 6.1: Block diagram of the single synapse learning experiment. At learning stage the teacher signal generates instant output spikes which together with the input ones modify the synaptic strength W. At the retrieval stage the feedback loop is deactivated and the pre-synaptic spike train generates current pulses proportional to the stored value W which charge the membrane capacitance.

The first experiment starts by setting W to low level and disabling learning, so input spikes do not increase the membrane potential nor generate output spikes given that the current that flows through the capacitor that represents the membrane potential is zero. This experiment is shown in (Fig. 6.2a), as observed calcium waveforms are generated following only the input spikes but the synapse remains in zero. As second step learning stage is enabled; therefore, the teacher signal forces the output neuron to fire at its same frequency (current values that charge the membrane potential were set to generate instant post-synaptic spikes), for this experiment the teacher spikes t_{post} were set so that $(t_{post} - t_{pre} > 0)$ hence generating potentiation (Fig. 6.2b). Continuous spike pairs raise the synapse level to its high bistability value remaining there until the system is reset or shutdown, in the figure this occurs after 4 seconds. After the new synapse value converges to high level, a following recalling stage consisting on same previous input spikes than the first step in absence of teaching signal is again presented. Given that the new

synaptic strength value W is high, input spikes this time can generate current flow in the membrane and therefore action potentials (Fig. 6.2c).

In the second experiment, a similar protocol to the first one is carried out consisting on retrieval-learning-retrieval phases. Before, the synaptic strength is initialized to the low level and therefore output spikes are not generated when pre-synaptic spikes appear (Fig. 6.3a). In the learning stage the synaptic strength is driven to high level provided that $(t_{teach} - t_{pre}) > 0$ as shown in Fig.6.3b. Finally, in the new retrieval phase, an output spike is generated after a cumulative number of pre- spikes (Fig.6.3c) given that the synaptic strength for the setting parameters generates smaller currents in the membrane potential compared to the first experiment. As explained in a previous chapter, this currents values are proportional to the linearizer limits; the linearizer is a circuit block configured in cascade after the synapse circuit in order to to map its weight voltage values to a smaller range; this new range is connected to DPI to generate current pulses.

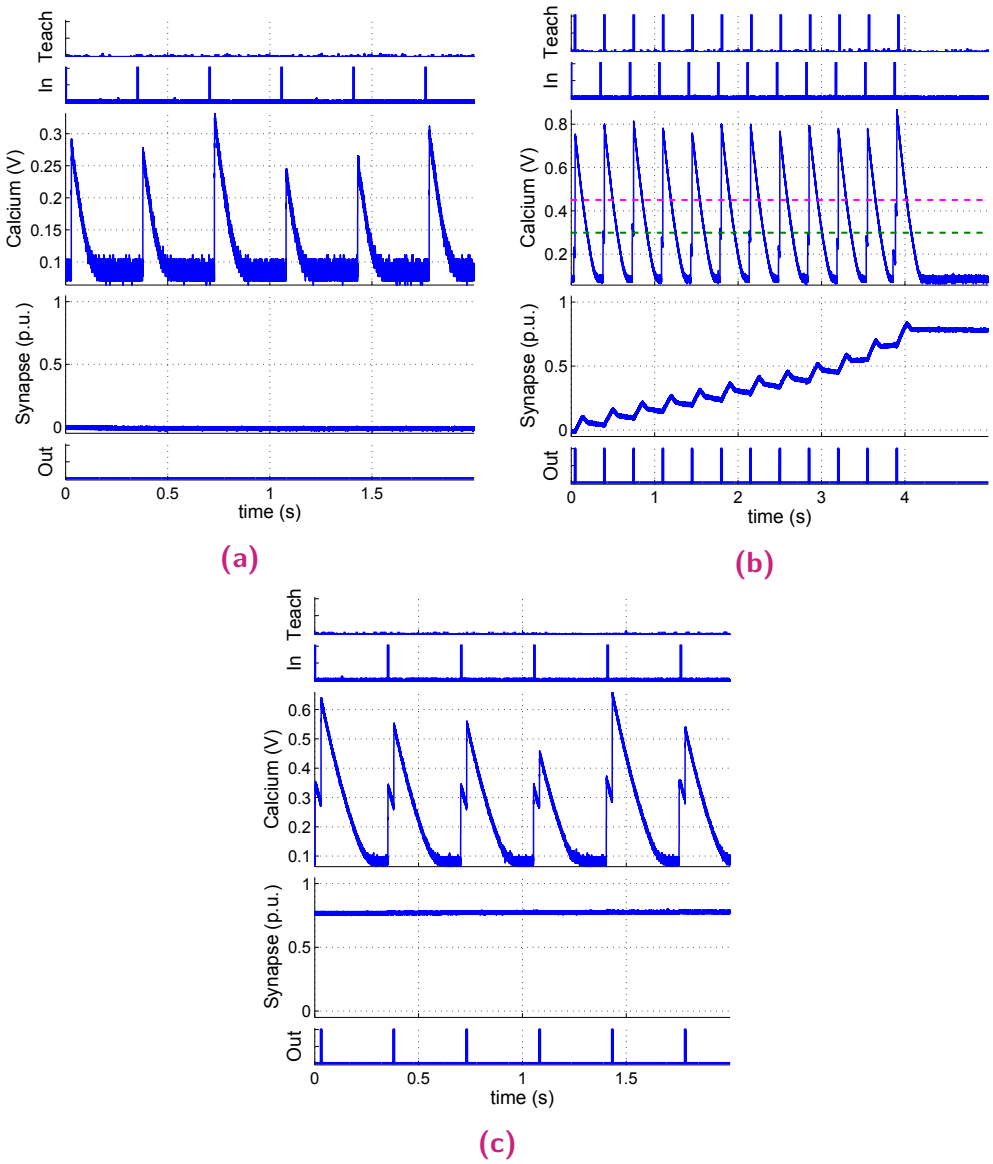


Fig. 6.2: Measurement results for the experiment in Fig. 6.1 configured in such way that synaptic strength generates saturation currents for the membrane potential. (a) In an initial recall stage before learning the synaptic strength has zero value therefore the input spikes do not generate output spikes (b) In the learning stage the synaptic strength is driven to high level considering a positive timing between the teacher (post-synaptic spike) and input spikes (pre-synaptic spike). (c) This time the recalling stage has the synaptic strength at high level therefore pre- spikes generate output spikes.

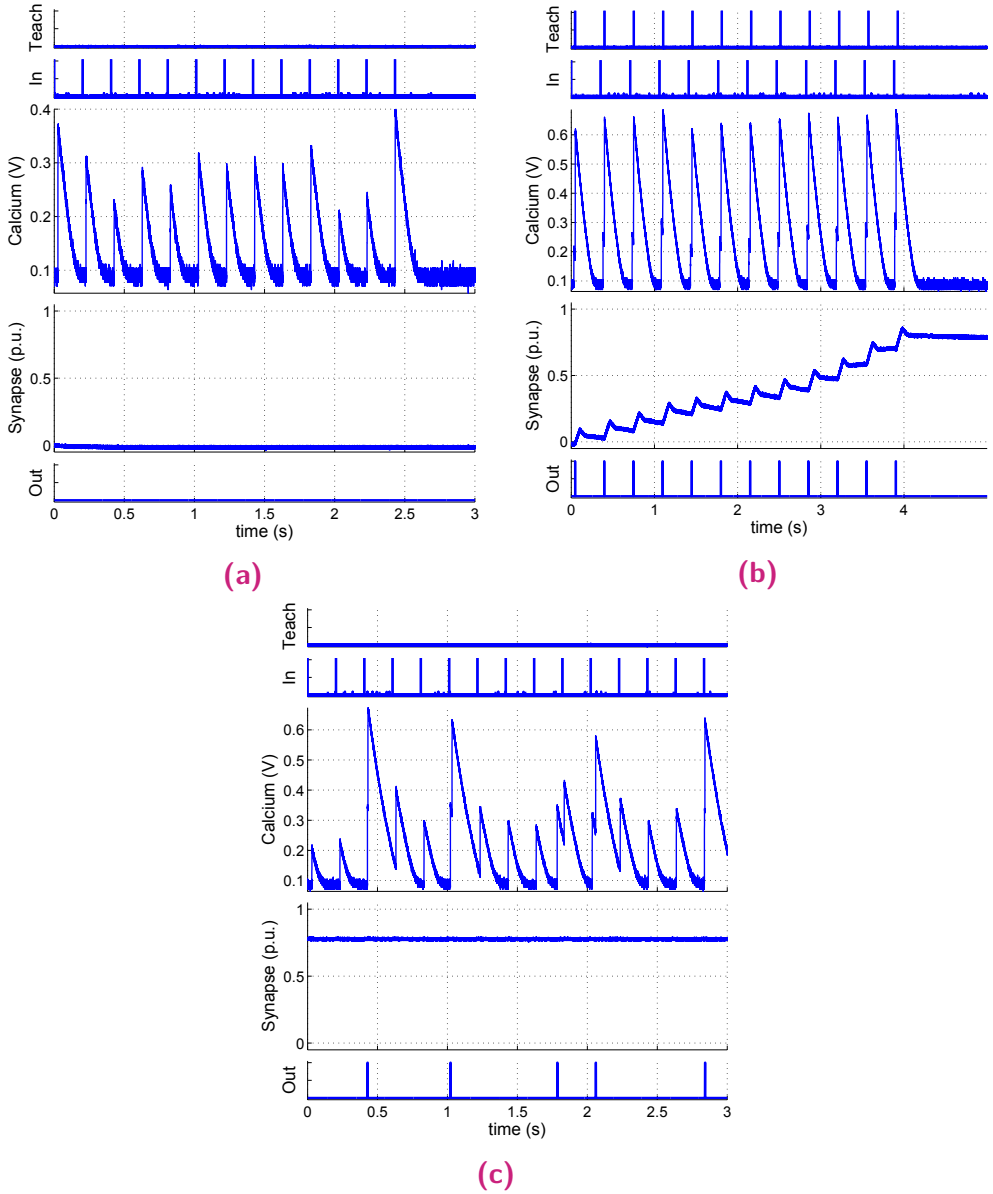


Fig. 6.3: Measurement results for the experiment in Fig. 6.1 configured in such way that synaptic strength generates higher currents that make leakage ones negligible but do not reach saturation levels for the membrane potential. (a) In an initial recall stage the synaptic strength is initialized at zero level, thus input spikes do not generate output ones. (b) In the learning stage where feedback is activated teacher t_{post} and input t_{pre} signals generate potentiation in the synaptic strength. (c) This time in a recalling stage, given that the synaptic strength turned to high level, input spikes generate current pulses that charge the membrane capacitor, the post-synaptic potential in this case is generated by an accumulation of those currents (more than one input spike) instead of one as in the previous case in Fig. 6.2.

Table 6.1: Definition of input spikes in each synapse for $pattern_1$ and $pattern_2$.

•	$pattern_1$	$pattern_2$
input1	2.86Hz	4Hz
input2	4Hz	2.86Hz

6.2 Simple Perceptron

While feedforward networks are not enough to explain cognitive functions alone, they are an important ingredient of brain-style information processing and have contributed greatly to the development of statistical learning theory [123]. The perceptron [129] is a simple neural network model in which a group of input neurons are connected to one output neuron; however, correct classification is restricted to patterns that are linearly separable. Training is provided by a repetitive presentation of patterns which drive the synaptic weights of the network to values that the set an optimal classification of input data. The learning rule here consists of minimizing the mean difference between the output of the feedforward network and the desired state provided by a teacher (error function); this is achieved by changing the weight values along the negative gradient of the error function [123]. Despite the advantageous learning rule for training the network, there is little evidence from biology that a synapse can differenciate between the actual and desired output activity; therefore, the perceptron is categorized as an artificial neural network [48]. The limit number of weight values in a single neuron restricts the complexity of functions that we can represent using a single layer; therefore, an increase in the number of nodes (multi-layer perceptron), and thereby the number of connections with corresponding independent weight values is commonly used [129].

The perceptron implemented in our chip consists of two synapses W_1 and W_2, and a teacher signal all connected to a single neuron N; the output of the neuron is fed back to both synapses as $post$ signal to generate synaptic modification as shown in Fig. 6.4. Input spikes $input_1$ and $input_2$, and teacher spikes are generated from an external PC. The neural network is trained to recognize two different patterns $pattern_1$ and $pattern_2$ which differ in their spike rate. Given that in this experiment I was working with frequency rate instead of timing, parameter values in the chip were configured to produce depression for frequencies lower that 3.5Hz and potentiation for higher frequencies. Table 6.1 summarizes the input frequency for each pattern.

As previous experiments in section 6.1, here a protocol consisting of retrieval-learning-retrieval phases is carried out. In an initial stage synaptic strengths W_1 and W_2 are initialized to low level and the teacher and feedback signals are disabled ($EN_{learn} = 0$); therefore, input spikes in each synapse do not generate action potential in the neuron membrane nor output spikes because of null current flow in the membrane. Measurement results for this stage are shown in Fig. 6.5, here $pattern_1$ (Fig. 6.5a) and $pattern_2$ (Fig. 6.5b) are presented which differ in the firing rate of the input spikes in_1, in_2. As expected the neuron output (out) is zero during all the time and the calcium waveforms are generated for each synapse only when pre-synaptic spikes appear.

During the learning stage the teacher signal force the neuron to fire at its same frequency; these generated post- spikes are fed back into the calcium synapses W_1 and W_2 which together with the pre- spikes pre_1 and pre_2 modify their synaptic strengths. The network is trained with two consecutive patterns, first $pattern_1$ together with the teacher signal firing at $teach_1 = 2.86Hz$ are presented for a certain amount of spikes (15) as shown in Figs. 6.6a and 6.6c. After a short resting time $pattern_2$ together with teacher signal firing at $teach_2 = 4Hz$ are presented with the same amount of spikes (15) as shown in Figs. 6.6b and 6.6d. After the presentation of both patterns the achieved stable synaptic weights remain recorded unless the chip is reset or turned off. For the selected parameters and data frequency W_1 moves to high level and W_2 remains in low level at the end of the learning phase.

In the retrieval phase the teacher and feedback signals are disabled and $pattern_1$ and $pattern_2$ are presented independently. Given that W_1 is in high level, input spikes in pre_1 generate action potentials and therefore output spikes as shown in Fig. 6.7a; on the other hand, W_2 is in low level and input spikes pre_2 do not generate $output$ spikes as shown in Fig. 6.7b. Therefore, for $pattern_1$ the perceptron will generate low frequency output spikes and for $pattern_2$ high frequency.

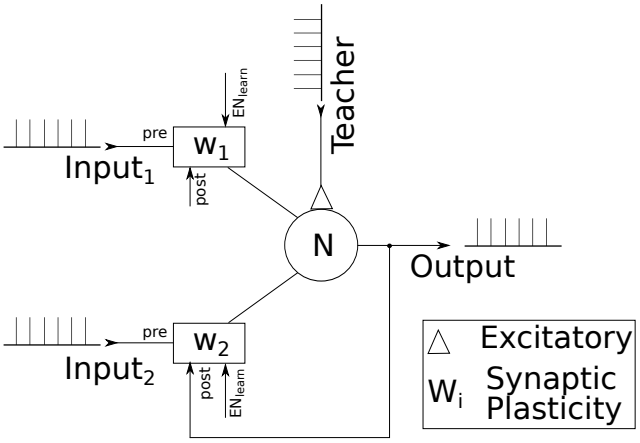


Fig. 6.4: Block diagram of the perceptron experiment using two calcium-based synapses. At the learning stage the feedbacks that connect the output terminal with the post- spikes in each synapse are activated, additionally the teacher signal generates instant spikes in the output terminal, therefore the synaptic strength is modified according to STDP learning rules. In the recalling stage the feedback is deactivated and the input spikes in 1 and 2 generate current pulses in the neuron proportional to W_1 and W_2 values which increase the membrane potential and eventually generate action potentials.

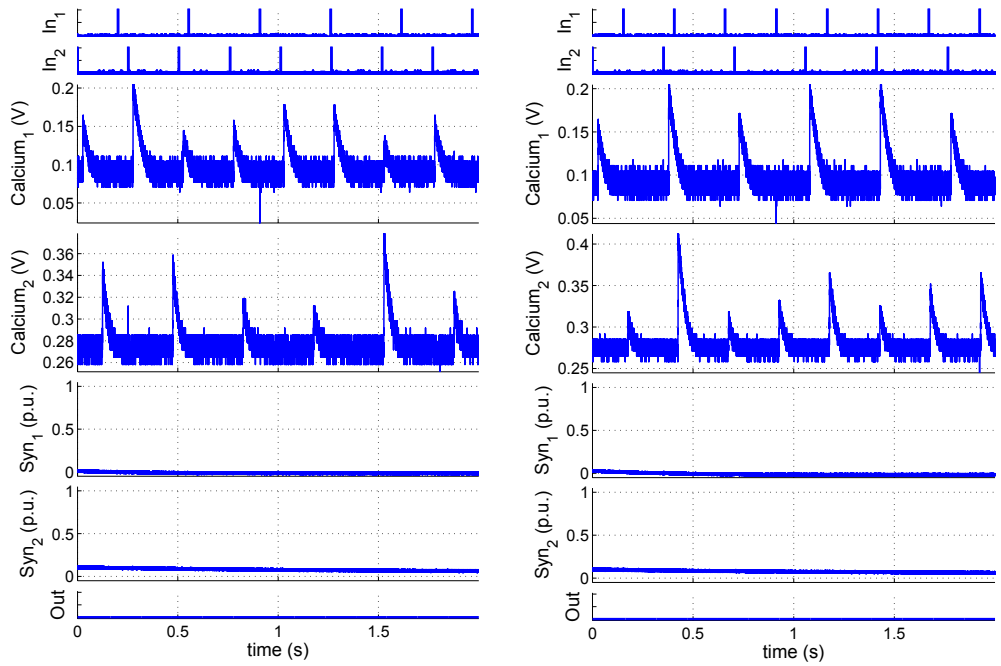


Fig. 6.5: Measurement results of the perceptron in Fig. 6.4 before learning stage in which the synaptic strengths W_1 and W_2 are initialized in zero, two input patterns are presented as stated in Table 6.1. (a) First input pattern consisting of spikes with frequencies $f_1 = 2.86Hz$ and $f_2 = 4Hz$ is presented, however the output is null. (b) Second input pattern consisting of spikes with frequencies $f_1 = 4Hz$ and $f_2 = 2.86Hz$ is presented, similarly here the output is null.

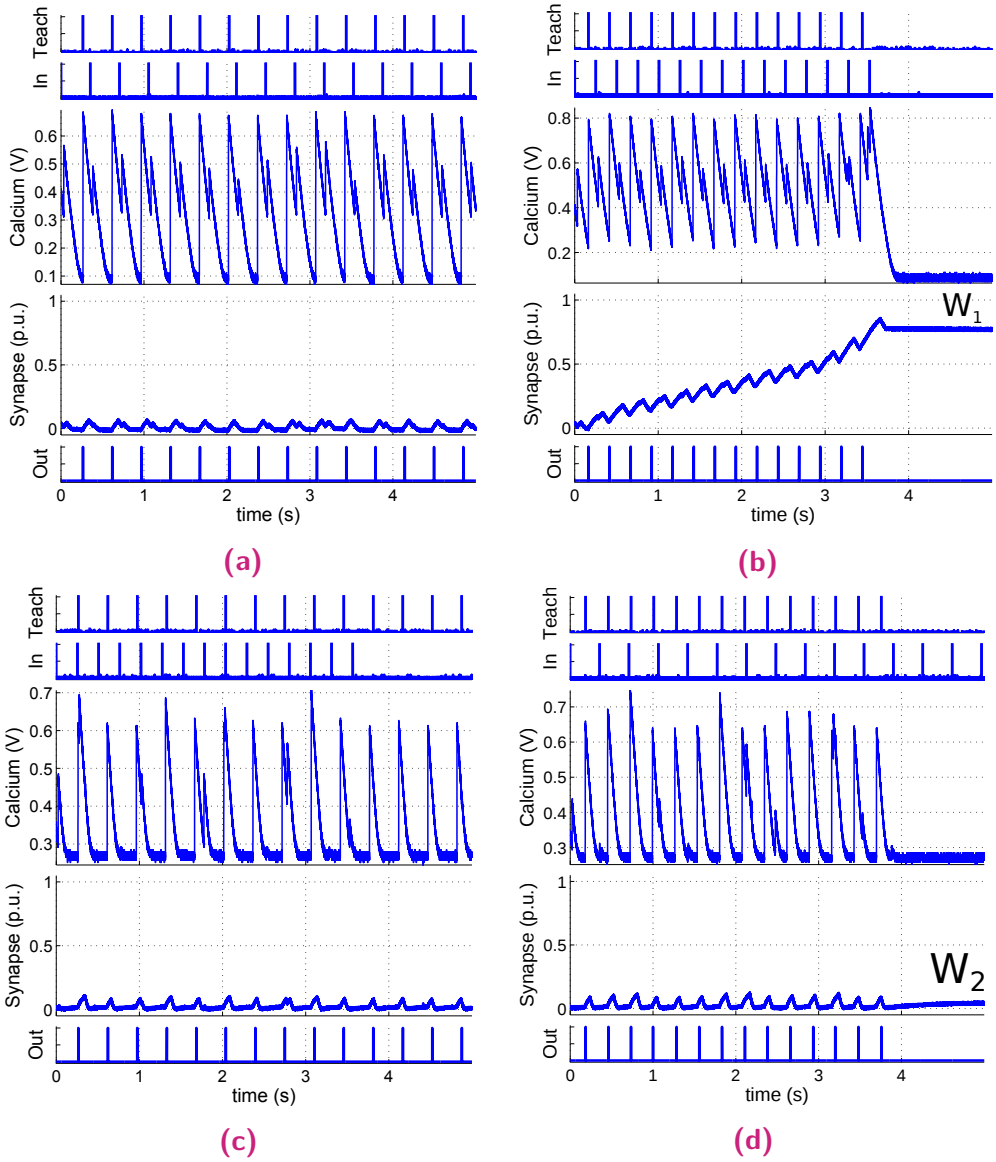


Fig. 6.6: Measurement results of the perceptron in Fig. 6.4 for the learning stage, top and bottom figures show the evolution of W_1 and W_2 respectively. Similarly in the left figures pattern 1 is presented in In_1 and In_2 inputs, while in the right ones, pattern 2 is presented; a resting time between the left and right patters of $2s$ is provided. (a) The pre- (in_1) and post- (teach signal) spike frequencies $f_{pre} = f_{post} = 2.86Hz$ are low therefore depression is generated. (b) The pre_1- (in_1) and post- (teach signal) spike frequencies $f_{pre1} = f_{post} = 4Hz$ are high enough to generate potentiation. (c) The pre- (in_2) and post- (teach signal) spike frequencies $f_{pre2} = 4Hz$, $f_{post} = 2.86Hz$ are not high enough to generate potentiation. (d) The pre- (in_2) and post- (teach signal) spike frequencies $f_{pre2} = 2.86Hz$, $f_{post} = 4Hz$ are not high enough to generate potentiation.

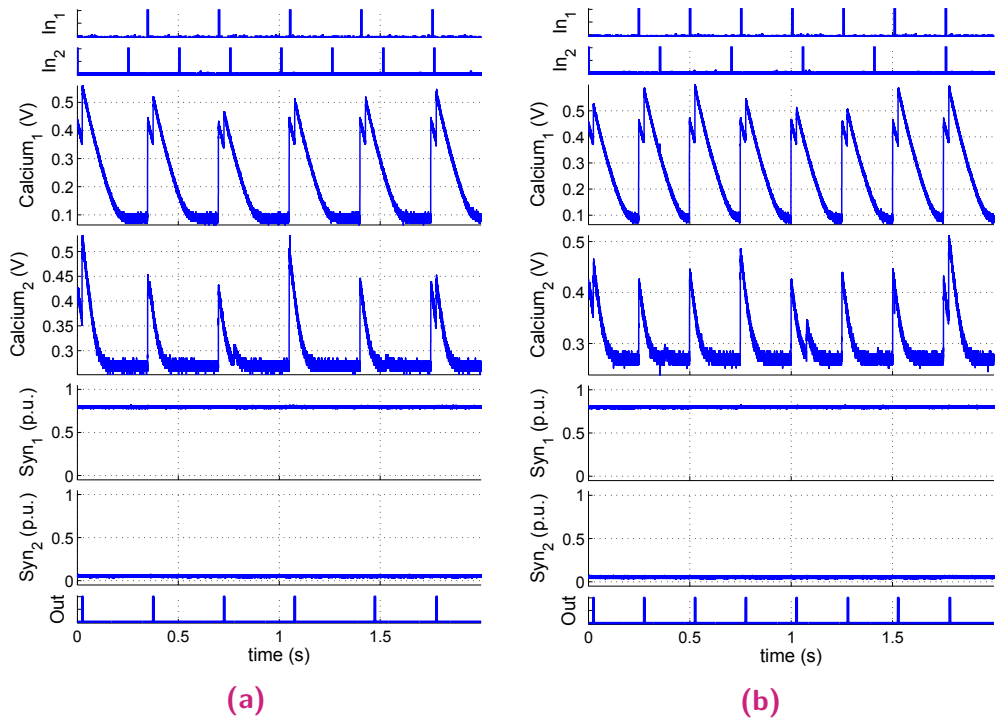


Fig. 6.7: Measurement results of the perceptron in Fig. 6.4 for the recalling stage after learning. As it is observed the synaptic weights are $W_1 \approx 1$ and $W \approx 0$, therefore only input spikes in In_1 generate output spikes (linearizer is not used in this experiment). (a) Output spikes follow the same frequency of In_1 with a slight delay. (b) Similarly, output spikes follow the same frequency of In_1.

6.3 Discussion

In this chapter I presented simple network experiments such as a single synapse connected to a single neuron and two synapses connected to a single neuron (simple perceptron). Although these experiments can not reproduce complex mapping functions, they are important to analyse basic network structures that will be replicated in the next fabricated version. These experiments also provided insight to find suitable parameter values for a correct network operation that resemble look-up tables. I discussed two protocols, one when the input spikes produce instant firing rate at the output and another when cumulative input spikes are required for a single output spike.

The setup for these experiments consisted of a hardware infrastructure (chip and PCB), a software environment which was implemented in Python language, a communication protocol which sends the PC data (bias values, spike

timings and synapse and neuron addresses) to the USB port of the PCB, and measurement instruments (oscilloscope and multimeter). The recorded output neurons response demonstrate a correct integration of the VLSI blocks consisting of the neural network, bias generator and the communication circuit. The measurement results also confirmed that the calcium learning circuit can be setup for implementing STDP and Hebbian learning as firing rate dependency.

A main drawback observed when measuring was that input signals in same unit blocks of the circuit matrix do not reproduce same results despite I used similar parameter values. The main source of this variation was the calcium core circuit in which some transistors used minimum dimensions hence producing considerable mismatch. For relatively distanced synapse blocks, variability in STDP characterization and bistability slew rate were so high that good results in the experiments could not be achieved. Considering even neighbour synapses, variability shifts the same probability of having potentiation and depression for same spikes frequencies. For example in the Figs. 6.6d and 6.6c where only one pre- or post- spike fires at high frequency, the synapse strength should be driven to values slightly greater than 0 but smaller than 0.5 so that it is moved at the end only by the bistability to 0 when spikes are over; this does not occur because mismatch goes in favour of depression reducing the probability of potentiation. The next chapter propose circuit improvements which considerably reduce mismatch effects and estimates their values for the next tape out.

Finally, an improved version hardware should implement more complex neural networks experiments in which each memory is represented by a specific pattern of neural activity (mean firing rates) that is imposed to the network at the time the pattern is memorized.

7 Mismatch Characterization

Measurements of neurons from the visual cortex demonstrate the ISI variability in neural responses [123, 130] especially during spontaneous activity. One reason of this is a considerable fluctuation in the input current to cortical neurons. In the case when external stimulus changes rapidly, neurons in the visual system react more certainly than for constant or slowly moving stimuli given that the majority of spikes follow the changing stimulus instantly although some neurons do not respond with a spike or occur with delay between the stimulus changes [45]. Variability is also found in neuronal parameters such as threshold, membrane time constant, or length of the refractory period which is denominated as slow noise. In order to deal with noise effects, an additional term in the differential equation that describe the synaptic dynamics is added. This single term estimates the effect of all the noisy sources.

On the other hand, variability in electronic circuits consisting of CMOS transistors occurs because different instances of the same block can not produce identical behaviour even when they are biased with same parameter values because of slight variations in their physical dimensions or carriers concentration during fabrication. Despite of the inherent nature of mismatch in the transistors, it is important to reduce their effect otherwise random results not correlated with the external stimulus can occur.

In the previous chapter 3 output waveforms were calculated assuming circuits with perfect symmetry; however, manufacturing variations cause electrical parameter mismatch in CMOS that have identical dimensions, layout and bias conditions. Main CMOS parameters affected by mismatch are the threshold voltage, transconductance and body-effect coefficient which lead to mismatch in gate-source voltage and drain current in different configurations such as current mirrors and differential pairs. In this chapter I analyse local mismatch which results from variations inside a single chip neglecting distance or orientation factors (systematic mismatch).

In the second manufactured chip, experiments involving more than one neuron and synapse were not successful because of considerable mismatch

among blocks, specially those in which transistors' dimensions were smaller such as in the calcium core. Considerable reduction of variability is achieved by increasing the dimensions of the most sensitive transistors in each block. The mismatch contribution is obtained through Montecarlo analysis considering only statistical variation (without process variation). For the result in this chapter I used Virtuosos ADXL with sampling method Latin Hypercube, this is a quasi-random algorithm that requires less samples than the random method, a better approach is the Low-discrepancy sequence sampling method that stop automatically after converging, unfortunately the last method is not available in our current Virtuoso ADXL version.

In the case of a single CMOS current mismatch, variation increases drastically in weak inversion because of the effect of ΔV_{T0}; additionally in absence of body-effect when $V_S = 0$, variation in the slope factor Δn can be neglected ($n = 1/\kappa$), however in the case when $V_S > 0$ the current mismatch is increased by the contribution of Δn, especially in weak inversion. On the other hand, in the case of gate voltage offset, the factors $\Delta\beta$ and Δn increase considerable in strong inversion, thus if $V_S > 0$ the offset is further increased by an amount $V_S \Delta n$ [131].

Considering circuits with more than one transistors, the most common blocks that affect mismatch are the current mirror and the differential pairs given that these blocks convey information in current and voltage respectively. The main target of the current mirror is a precise copy of the current in each branch; however, mismatch between the input and output current occurs as a consequence of the channel length modulation and the output impedance. Given that the drain voltages in both transistors are determine by different bias conditions the drain currents are not the same, a considerable improvement in this case is achieved by using a cascode configuration. In addition, a difference in temperature between two or more transistors inside a die which can be stationary (due to devices at different distance from a heat source) or transient in time (due to a change of ambient temperature that is too fast with respect to the chip thermal time constant) together with variation of process parameters and stress result in gate voltage and current statistical mismatch. A design hint proposed in [132] states "Current mirror mismatch is proportional to g_m/I_{Dsat} ratio therefore the most favourable operation region is strong inversion, if the saturation voltage has to be minimized, the design compromise can be either the limit of strong inversion or moderate inversion operation with an increased transistor area".

In the case of the differential pair, the gate voltage mismatch is the most critical variable because this is the input stage that converts the small input signal into a current. A larger area reduces the voltage mismatch. In addition, given that the voltage mismatch is inversely proportional to the gm/I_{Dsat} ratio, it decreases if the transistor operates in moderate or weak inversion [132].

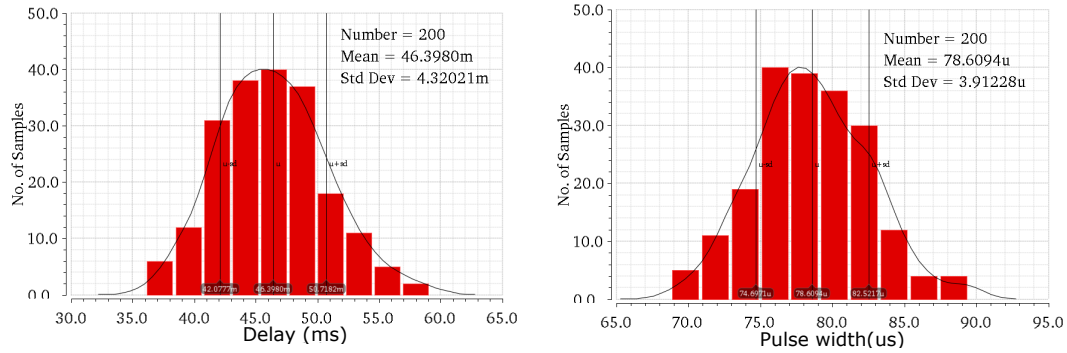
7.1 The Calcium Circuit

The calcium block is the main source of mismatch due to its small transistors' dimensions. These sizes were selected focusing on minimizing silicon area and increasing channel length modulation effect to get a stepper decay in the calcium waveform V_{ca}. In the delay circuit shown in Fig. 5.1a the critical devices are the ones which define the delay threshold (Md_2 and C_{del}) and the pulse width Spk_{preD} (Md_3), thus using at least a dimension ratio of $(W/L) = (2\mu/1\mu)$, ($L \approx 4 \times L_{min}$) improves considerable the mismatch (the remaining transistors can use at least $L \approx 3 \times$ Ł$_{min}$). In the digital circuits like the buffers and nor-gates the mismatch contribution is negligible therefore a size increase is not required. Mismatch simulation results for the delay circuit are shown in Fig. 7.1.

Similarly in the calcium core circuit if a ratio of $W/L = 1$ for at least $L \approx 4 \times$ Ł$_{min}$ is used matching improves considerable; however, a greater delay value needs to be set to compensate a slow slope decay in V_{ca}.

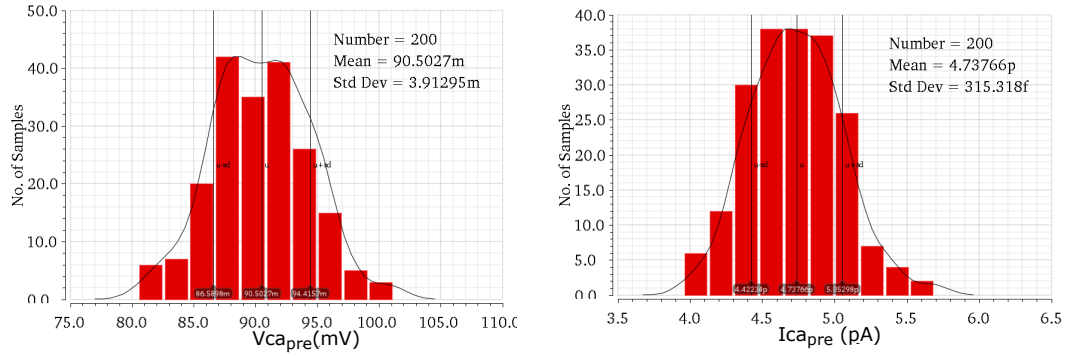
In addition, high threshold voltages transistors were chosen and low leakage voltages (vdd^* and gnd^*) replaced by global power rails to reduce mismatch contribution in the slope factor Δn. Mismatch contribution in the calcium circuit which are associated with delay (D), Ic_{pre}, Ic_{post} and τ_{ca} are shown in Figs. 7.2, 7.3, 7.4.

Simulation results for the new transistors' dimensions predict that local mismatch for most of the parameters are lower than 10%, only Ic_{post} (23.9%) and τ_{ca} (14.1%) are above this value, contrary to the measurement results of the second chip where all the mismatch parameter were greater than 100%. Therefore, similar learning waveforms are highly probably to be obtained in neighbouring neurons for same bias values. The total variability of the synapse model is a complex function of all these mismatch parameters, which makes also difficult to implement a simulation that cover all the possible cases.



(a) (b)

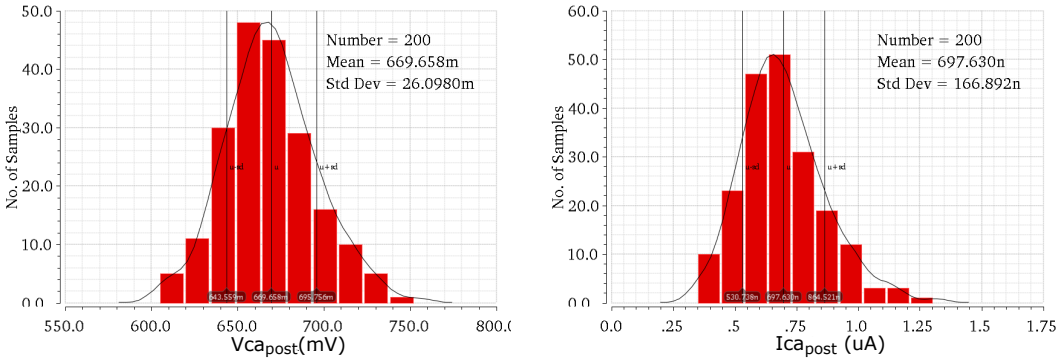
Fig. 7.1: Montecarlo simulation results of the output spike mismatch in the delay circuit shown in Fig. 5.1a for 200 samples. (a) The delay parameter of Eq. 2.14 reaches a coefficient of variation of $\sigma/\mu = 4.32ms/46.4ms \approx 0.093$. (b) The pulse width of the pre_D spike reaches a coefficient of variation of $3.9\mu s/78.6\mu s \approx 0.05$.



(a) (b)

Fig. 7.2: Montecarlo simulation results of the c_{pre} parameter in Eq.2.14 simulated in the calcium core circuit shown in Fig. 5.1b for 200 samples. (a) The output voltages reaches a coefficient of variation of $\sigma/\mu = 3.91mV/90.5mV \approx 0.043$. (b) The previous voltage is converted to current levels when it is injected to the gate of a transistor, in this case the output current reaches a coefficient of variation of $0.32pA/4.73pA \approx 0.067$.

Alternatively, I simulated few specific samples of these parameters together with the ones of the synapse block within the variability range obtaining similar learning waveforms in the case of DP-STDP.



(a) (b)

Fig. 7.3: Montecarlo simulation results of the c_{post} parameter in Eq.2.14 simulated in the calcium core circuit shown in Fig. 5.1b for 200 samples. (a) The output voltages reaches a coefficient of variation of $\sigma/\mu = 26mV/670mV \approx 0.04$. (b) The previous voltage is converted to current levels when it is injected to the gate of a transistor, in this case the output current reaches a coefficient of variation of $167nA/697.6nA \approx 0.24$.

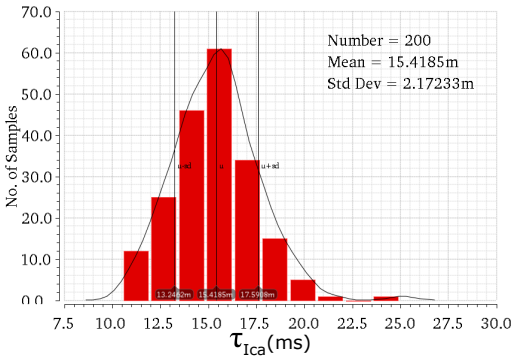
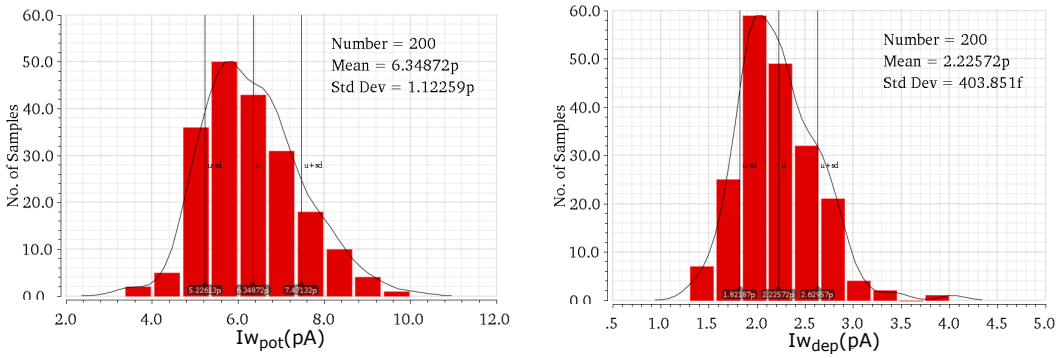


Fig. 7.4: Montecarlo simulation results of the τ_{Ca} parameter in Eq.2.14 simulated in the calcium core circuit shown in Fig. 5.1b for 200 samples. Values here are given as a function of the time that the current takes to return to its zero level after an input spike. The τ_{Ica} value reaches a coefficient of variation of $\sigma/\mu = 2.17ms/15.4ms \approx 0.14$.

7.2 The Synapse Circuit

Mismatch in the synapse core circuit shown in Fig. 5.2 is mainly due to variations in potentiation and depression currents. $Isyn_{pot}$ is defined by the current bias generated in Mn_{11} and the attenuation factor in transistors Mp_{13-16}, therefore increasing the size of Mn_{11} reduces mismatch; given the use of cascode configuration in current mirrors, transistors' area here can be smaller. Nonetheless an important requirement is low leakage design therefore a small ratio (W/L) should be consider as well as high threshold voltage devices. A trade-off condition is found if low leakage sources are replaced by global ones because they could increase the bistability leakage



(a) (b)

Fig. 7.5: Montecarlo simulation results of the γ_p and γ_d parameters in the Eq. 2.13 simulated in the synapse core circuit shown in Fig. 5.2 for 200 samples. (a) The potentiation current $I\gamma_p$ reaches a coefficient of variation of $1.12pA/6.35pA \approx 0.18$. (b) In the case of depression current $I\gamma_d$, it reaches a coefficient of variation of $0.40pA/2.23pA \approx 0.18$.

and set the V_w swing range to the power supply. Similar considerations are given for $Isyn_{dep}$ where transistors Mn_{21}, Mn_{23-26} and Mn_{26-29} are important. The remaining transistors can use at least $L \approx 3 \times L_{min}$.

When measuring mismatch contribution of a single parameter the other ones should be shut down, for instance in the case of potentiation, depression threshold voltage Vth_{dep} is set to zero ($\theta_d = 1$). Similarly in depression, potentiation voltage is set to zero. Synapse parameters variations for I_{pot} and I_{dep} are shown in Figs. 7.5a and 7.5b where values are calculated for $Vw = 0.9V$ (middle point of power rails).

Simulation results for the new transistors' dimensions of the synapse circuit predict that local mismatch for both parameters are close to 18%. The effect of these parameters in the total circuit was tested, as stated in the previous section, together with the ones of the calcium block for few sample data obtaining similar DP-STDP waveforms.

7.3 The Bistability Circuit

Mismatch in the transconductance amplifier shown in Fig. 5.3 affects the output bistable current. The principal transistor that defines this current is Mp_1 therefore larger size there is required (considering that all the other transistor have at least $L > 3 \times L_{min}$); in addition, if its standard model is replaced by a high threshold voltage one, low leakage supply Vdd^* can also be replaced by the global one and thus neglecting mismatch contribution

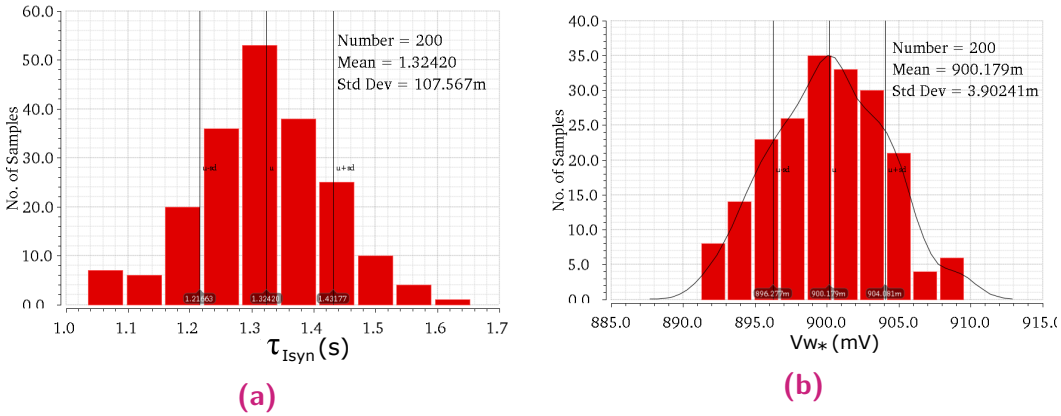


Fig. 7.6: Montecarlo simulation results of the τ_{Isyn} and W_* bistability parameters in the Eq. 2.13 simulated in the full synapse circuit activating only transconductance amplifier of Fig. 5.3 in positive feedback for 200 samples. (a) Mismatch for τ_{Isyn} is calculated as a function of the time that synaptic strength Vw takes to return to its zero level after been initialized in the mid-rail voltage, the coefficient of variation gets the value of $0.11ms/1.32ms \approx 0.08$. (b) Mismatch for Vw_* is calculated as the variations in the negative terminal of the OTA when its value is set through a transmission gate to Vw_*, it reaches a coefficient of variation of $3.9mV/900.18mV \approx 0.004$.

in the slope factor Δn. A main problem in the block is the high headroom voltage of the output stage, mainly in $Mp_{22} - Mp_{24}$ transistors, that's why they should use low ration (W/L). In our case simulation results estimate that a ratio smaller than $1/3$ reduces considerable headroom voltage. Parameter variations such as τ_{syn} and V_* are shown in Fig. 7.6.

Simulation results for the new transistors' dimensions predict that local mismatch is dominated by the τ_{syn} parameter producing around 8% of deviation. Mismatch in this parameter is not related with the learning waveform (as in the two previous sections) but with the long-time convergence in absence of input spikes. The effect of this parameter therefore was tested independently of the other ones giving in all the cases convergence times much greater than $200ms$ (ISI timing for our experiments) which is expected to not interfere with the STDP.

7.4 Discussion

When CMOS transistors operate in weak inversion, mismatch has greater effect compared to strong inversion operation [133]. As analyzed in chapter 3, mismatch can be reduced by increasing the transistor area or by increasing gate-voltage overdrive voltage (in the case of current bias/mirror configura-

tion). It was also demonstrated that by using $L \geq 4L_{min}$ in the most sensitive transistors we can reduce considerably mismatch.

In addition, layout techniques such as common-centroid and dummies should also be kept in mind to reduce this effect. It is also important to have a good floorplanning to avoid excessive routing that can generate parasitic capacitances; similarly, it should be avoided to cross lower metal levels through transistors' gate because signals can be corrupted by crosstalk effect.

Variability of spike timing is a common phenomenon in cortical neurons although the origin of this irregularity in their activity is barely understood. In spiking neuron models such as the one presented in section 2.1, noise is often added explicitly to mimic the unpredictability of neuronal recordings, common ways to introduce this noise is by adding stochastic variables in the threshold parameter, which can let the neuron fire even though the threshold has not been reached, and in the differential equation that represent the membrane potential dynamics (diffusive noise) [45]. Similarly, noise models for synapse can be constructed to represent an average variability of the system and thus predict the accuracy of network results.

In this framework, the simulation results presented here are useful for the next circuit fabrication which will focus on reducing the neurons behaviour mismatch and therefore obtaining similar features in the STDP learning rules, output firing rate in neurons with same bias values and input data. In addition, the statistical values given here provide insightful information for computational models to create noise representations. The normalized variability of the parameters obtained in this chapter can be included in the differential equations that represent the synapse (Eq. 2.13) and the calcium dynamics (Eq. 2.14) to estimate a single noise source of the computational model that equals all the other variabilities ("Noise(t)" term in Eq. 2.12).

8 Conclusions

Through the history of neuroscience it was demonstrated that abstraction from biological observation is a powerful tool to obtain strong technological advance. For example, in the early 1940 simple logic gates, which were abstract representations of neurons, connected in sequential chains could compute effectively any required function [43]. In this sense neuromorphic engineering aims at the development of artificial neural systems in which their architecture and principles are based on biological systems; it provides also a tool to understand the computational strategies used by the brain to overcome constraints such as limited space, wiring and energy. In this framework, the development of this project provides relevant contribution to the memory and learning research fields by exploring biologically inspired technologies for learning systems. While modern learning systems can achieve human-level performance in certain well-defined tasks (e.g. image recognition, attention, scene description, image tracking), there exists no system capable of performing all of these tasks simultaneously, as humans do.

Furthermore, there are no real-time learning systems with power consumption and size comparable to those in nature. The creation of these systems is hindered by the use of conventional (i.e. von Neumann) computing architectures, which are fundamentally different from the biological computation substrate. The von Neumann architecture imposes a physical separation between the computational elements (Central Processing Unit (CPU)) and the information storage elements (memory), which leads to the so called von Neumann bottleneck, which denotes the limited traffic capabilities of the communication channel between the CPU and the memory. Overcoming the limitations of traditional digital architectures by imitating computational primitives observed in the brain, this project offers the opportunity to test the performance of current theories of learning in a realistic environment, leading to advancements in combining models of synaptic plasticity with network-wide activity.

In this thesis I presented my overall progress concerning design, fabrication, and test of two neuromorphic chips which implement a calcium-based synaptic plasticity model together with a neural network. I simplified an

existing model [35] in order to enable its implementation in analog VLSI and presented both simulations and measurements of this circuit's behavior when stimulated with pre- and post-synaptic spikes with different relative timings and presentation frequencies. The measurements showed that the circuit produces the classic STDP behaviour and that an increased stimulus frequency results in increased potentiation, as expected from the mathematical model and observed in biological systems. I showed that learning STDP waveforms can be generated by setting appropriate parameters in the equations describing the calcium and synapse dynamics. I also showed how saturation effects can alter the learning waveforms. In addition, I presented a thorough comparison between the proposed circuital implementation and the simplified theoretical model. Finally I implemented and tested a small array of these synapses connected to aVLSI neurons to characterize their performance in a network and observed that mismatch constraint is the most important problem to deal with in the next chip version. Progress in this field has been made by improving the circuit design with special consideration in increasing transistors' area and quantifying the standard deviation of bias parameters that resemble its computational counterpart; this data will be also useful for computational models of neural networks to predict reliability in population of neurons and thus implement suitable information redundancy blocks. The results of this characterization will guide the next chip design and the exploitation of design techniques for minimizing these variations. The path required to test the synapse behaviour and the simple network experiments include also the development of communication and control hardware infrastructure to interact with the chip and a software interface to set parameter values and input data. In this project we successfully integrated all these components needed for the system setup, which were developed by different research groups throughout years.

A major challenge in my design was the reduction of leakage currents, which are otherwise comparable with the operational currents. More advanced technologies would require stronger reduction of leakage currents and therefore additional circuits. One simple modification is to exploit the "stack effect" by replacing each critical transistor with two transistors in series, thus reducing their current. Additionally, high-threshold devices could be used to reduce leakage current. Finally, the required swing range for V_w has to be guaranteed by an adequate rail voltage.

Most of the chip area in my circuits belong to the capacitors (independent if they are native or MIM capacitors). In order to reduce the silicon area one alternative could be to employ techniques to reduce their bias currents by using high-threshold transistors.

Although my fabricated synapse and neural network circuits are simple representations of their biological counterpart, they reproduce basic principles found in cortical networks which are useful to interpret more complex operations. An important advantage of my system is that it can be scaled up by connecting repeatedly the same core blocks without degrading the overall system performance, likewise scaling down in technology is also feasible given that in my system transistors operate in weak inversion therefore a lower voltage supply wont affect its behaviour.

8.1 Future Work

This work ended with the measurement results of simple neural network experiments and mismatch considerations for further VLSI versions. During the design of the next chip version special focus needs to be put on reducing mismatch so that more complex experiments can be performed. In addition, it would be also important to work in parallel to improve the hardware and software infrastructure, i.e. monitor signals in our systems were restricted to three testing signals which delayed our debug stage schedule; furthermore, some of them even generated undesired leakage currents when connected to V_w reducing bistability performance. I believe that a customized hardware infrastructure instead of using a general one such as in [108] could simplify the circuit analysis and let one deal faster with debug stage.

Given that in the first two chips I intended to come up with a reliable model capable to reproduce the basic characteristics of the proposed calcium plasticity synapse, I didn't focus on reducing power consumption. However, this is also an expected task in the next version; I have proposed some ideas i.e. a feedback mechanism in the delay circuit which was explained in the discus sections. Nonetheless deeper analysis in all the blocks is required.

A long term goal is the implementation of large networks with learning capabilities suitable for autonomous systems. Such real-time compact devices capable of supervised and unsupervised online learning [30, 97] will allow

for robotic systems which can autonomously learn the statistics of their input stimuli (e.g. environmental signals, user input), opening up the possibility for unprecedented adaptive capabilities of technological systems.

Together with the hardware improvement, it is also important to develop computational methodologies that can exploit the advantages of these circuits to achieve complex functionalities. For instance, some inspiration can be taken from recurring canonical microcircuits found in the cortical sheet [134] characterized by excitatory and inhibitory feedback loops shortly called as Soft Winner-Take-All (sWTA). sWTA can be designed in aVLSI to implement basic units of a general purpose neuromorphic processor executed through Finite State Machines (FSMs) [135]. FSMs is a common procedure used in digital design which consists of executing certain actions when the system is in a specific state; a control unit is in charge to organize the order of states to follow by the processor. By embedding state machines in neuromorphic devices we can provide a bridge between digital processors design and neuromorphic engineering. In the long run neuromorphic and conventional computing devices should be merged in hybrid systems able to dynamically devote the most appropriate resources to the current tasks.

Bibliography

[1] R. Borkar, M. Bohr, and S. Jourdan, "Advancing moore's law on 2014, intel," Tech. Rep., Aug 2014.

[2] G. E. Moore, "Cramming more components onto integrated circuits," *Proceedings of the IEEE*, vol. 86, no. 1, pp. 82–85, 1998.

[3] S. Borkar and A. A. Chien, "The future of microprocessors," *Commun. ACM*, vol. 54, no. 5, pp. 67–77, 2011.

[4] R. H. Dennard, "Past progress and future challenges in LSI technology: From dram and scaling to ultra-low-power cmos," *IEEE Solid-State Circuits Magazine*, vol. 7, no. 2, pp. 29–38, 2015.

[5] R. H. Dennard, V. Rideout, E. Bassous, and A. LeBlanc, "Design of ion-implanted mosfet's with very small physical dimensions," *Solid-State Circuits, IEEE Journal of*, vol. 9, no. 5, pp. 256–268, 1974.

[6] Y. Taur and E. J. Nowak, "CMOS devices below 0.1μm: how high will performance go?" in *International Electron Devices Meeting. IEDM Technical Digest*, 1997, pp. 215–218.

[7] M. Bohr, "A 30 year retrospective on Dennard's MOSFET scaling paper," *IEEE Solid-State Circuits Society Newsletter*, vol. 12, no. 1, pp. 11–13, 2007.

[8] M. B. Taylor, "A landscape of the new dark silicon design regime," *Micro, IEEE*, vol. 33, no. 5, pp. 8–19, 2013.

[9] J. Kedzierski, E. Nowak, T. Kanarsky, Y. Zhang, D. Boyd, R. Carruthers, C. Cabral, R. Amos, C. Lavoie, R. Roy, J. Newbury, E. Sullivan, J. Benedict, P. Saunders, K. Wong, D. Canaperi, M. Krishnan, K. L. Lee, B. A. Rainey, D. Fried, P. Cottrell, H. S. P. Wong, M. Ieong, and W. Haensch, "Metal-gate finfet and fully-depleted soi devices using total gate silicidation," in *Digest. International Electron Devices Meeting,*, Dec 2002, pp. 247–250.

[10] D. Hisamoto, W.-C. Lee, J. Kedzierski, H. Takeuchi, K. Asano, C. Kuo, E. Anderson, T.-J. King, J. Bokor, and C. Hu, "Finfet-a self-aligned double-gate mosfet scalable to 20 nm," *IEEE Transactions on Electron Devices*, vol. 47, no. 12, pp. 2320–2325, Dec 2000.

[11] M. M. Waldrop, "The chips are down for Moore's law," February 2016.

[12] H. Esmaeilzadeh, E. Blem, R. S. Amant, K. Sankaralingam, and D. Burger, "Power challenges may end the multicore era," *Communications of the ACM*, vol. 56, no. 2, pp. 93–102, 2013.

[13] D. Patterson, T. Anderson, N. Cardwell, R. Fromm, K. Keeton, C. Kozyrakis, R. Thomas, and K. Yelick, "A case for intelligent RAM," *IEEE Micro*, vol. 17, no. 2, pp. 34–44, 1997.

[14] E. Kandel, J. Schwartz, and T. Jessell, *Principles of Neural Science*. Mc Graw Hill, 2000.

[15] R. J. Douglas and K. A. C. Martin, "Neuronal circuits of the neocortex," *Annual Review of Neuroscience*, vol. 27, pp. 419–451, 2004.

[16] P. J. Sjöström, E. A. Rancz, A. Roth, and M. Häusser, "Dendritic excitability and synaptic plasticity," *Physiological Reviews*, vol. 88, no. 2, pp. 769–840, 2008.

[17] C. Mead, *Analog VLSI and Neural Systems*. Reading, MA: Addison-Wesley, 1989.

[18] W. Harrod, "A journey to exascale computing," pp. 1702–1730, 2012.

[19] A. Bofill-i-Petit and A. Murray, "Synchrony detection and amplification by silicon neurons with STDP synapses," *IEEE Transactions on Neural Networks*, vol. 15, no. 5, pp. 1296–1304, September 2004.

[20] G. Indiveri, E. Chicca, and R. J. Douglas, "A VLSI array of low-power spiking neurons and bistable synapses with spike–timing dependent plasticity," *IEEE Transactions on Neural Networks*, vol. 17, no. 1, pp. 211–221, 2006.

[21] S. Fusi, M. Annunziato, D. Badoni, A. Salamon, and D. J. Amit, "Spike-driven synaptic plasticity: Theory, simulation, VLSI implementation," *Neural Computation*, vol. 12, pp. 2227–2258, 2000.

[22] S. Mitra, S. Fusi, and G. Indiveri, "Real-time classification of complex patterns using spike-based learning in neuromorphic VLSI," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 3, no. 1, pp. 32–42, Feb. 2009.

[23] S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana, "The SpiNNaker project," *Proceedings of the IEEE*, vol. 102, no. 5, pp. 652–665, 2014.

[24] S. B. Furber, D. R. Lester, L. A. Plana, J. D. Garside, E. Painkras, S. Temple, and A. D. Brown, "Overview of the spinnaker system architecture," *IEEE Transactions on Computers*, vol. 62, no. 12, pp. 2454–2467, Dec 2013.

[25] A. Davison, D. Brüderle, J. Eppler, J. Kremkow, E. Muller, D. Pecevski, L. Perrinet, and P. Yger, "PyNN: a common interface for neuronal network simulators," *Frontiers in Neuroinformatics*, vol. 2, pp. 1 – 10, 2008.

[26] D. Fasnacht, A. Whatley, and G. Indiveri, "A serial communication infrastructure for multi-chip address event system," in *International Symposium on Circuits and Systems, ISCAS 2008*. IEEE, May 2008, pp. 648–651.

[27] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. K. Esser, R. Appuswamy, B. Taba, A. Amir, M. D. Flickner,

W. P. Risk, R. Manohar, and D. S. Modha, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, 2014.

[28] A. Martin and M. Nystrom, "Asynchronous techniques for system-on-chip design," *Proceedings of the IEEE*, vol. 94, pp. 1089–1120, 2006.

[29] F. Akopyan, J. Sawada, A. Cassidy, R. Alvarez-Icaza, J. Arthur, P. Merolla, N. Imam, Y. Nakamura, P. Datta, G. J. Nam, B. Taba, M. Beakes, B. Brezzo, J. B. Kuang, R. Manohar, W. P. Risk, B. Jackson, and D. S. Modha, "Truenorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 34, no. 10, pp. 1537–1557, Oct 2015.

[30] N. Qiao, H. Mostafa, F. Corradi, M. Osswald, F. Stefanini, D. Sumislawska, and G. Indiveri, "A re-configurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128k synapses," *Frontiers in Neuroscience*, vol. 9, no. 141, 2015.

[31] 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*, vol. 102, no. 5, pp. 699–716, May 2014.

[32] J. Schemmel, A. Grübl, S. Hartmann, A. Kononov, C. Mayr, K. Meier, S. Millner, J. Partzsch, S. Schiefer, S. Scholze, R. Schüffny, and M. O. Schwartz, "Live demonstration: A scaled-down version of the brainscales wafer-scale neuromorphic system," in *2012 IEEE International Symposium on Circuits and Systems*, May 2012, pp. 702–702.

[33] S. Moradi and G. Indiveri, "An event-based neural network architecture with an asynchronous programmable synaptic memory," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 8, no. 1, pp. 98–107, Feb 2014.

[34] M. Graupner and N. Brunel, "Mechanisms of induction and maintenance of spike-timing dependent plasticity in biophysical synapse

models," *Frontiers in Computational Neuroscience*, vol. 4, no. 136, pp. 1–19, 2010.

[35] M. Graupner and N. Brunel, "Calcium-based plasticity model explains sensitivity of synaptic changes to spike pattern, rate, and dendritic location," *Proceedings of the National Academy of Sciences*, vol. 109, no. 10, pp. 3991–3996, 2012.

[36] F. L. Maldonado Huayaney and E. Chicca, "A VLSI implementation of a calcium-based plasticity learning model," in *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2016, pp. 373–376.

[37] F. L. M. Huayaney, S. Nease, and E. Chicca, "Learning in silicon beyond STDP: A neuromorphic implementation of multi-factor synaptic plasticity with calcium-based dynamics," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 63, no. 12, pp. 2189–2199, 2016.

[38] D. O. Hebb, *The organization of behavior: a neuropsychological theory*. Taylor & Francis, 2012, 1949.

[39] P. Dayan and L. Abbott, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press, 2001.

[40] H. Shouval, M. Bear, and L. Cooper, "A unified model of NMDA receptor-dependent bidirectional synaptic plasticity," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, pp. 10 831–10 836, 2002.

[41] S. Fusi, P. Drew, and L. Abbott, "Cascade models of synaptically stored memories," *Neuron*, vol. 45, pp. 599–611, 2005.

[42] S. Fusi, "Computational models of long term plasticity and memory," in *Oxford Research Encyclopedia of Neuroscience*, 2017.

[43] S.-C. Liu, T. Delbruck, G. Indiveri, A. Whatley, and R. Douglas, *Event-Based Neuromorphic Systems*. John Wiley & Sons, 2014.

[44] A. Hodgkin and A. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve,"

Journal of Physiology, vol. 117, pp. 500–44, 1952.

[45] W. Gerstner and W. Kistler, *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge University Press, 2002.

[46] A. Burkitt, "A review of the integrate-and-fire neuron model: II. inhomogeneous synaptic input and network properties," *Biological Cybernetics*, 2006.

[47] S. J. Gershman, E. J. Horvitz, and J. B. Tenenbaum, "Computational rationality: A converging paradigm for intelligence in brains, minds, and machines," *Science*, vol. 349, no. 6245, pp. 273–278, 2015.

[48] D. Sterratt, B. Graham, A. Gillies, and D. Willshaw, *Principles of Computational Modeling in Neuroscience*. Cambridge University Press, 2011.

[49] E. D. Schutter, *Computational Modeling Methods for Neuroscientists*, 1st ed. The MIT Press, 2010.

[50] T. P. Bliss and G. Collingridge, "A synaptic model of memory: Long term potentiation in the hippocampus," *Nature*, vol. 31, p. 361, 1993.

[51] R. A. Nicoll and R. C. Malenka, "Contrasting properties of two forms of long-term potentiation in the hippocampus," *Nature*, vol. 377, pp. 115–118, 1995.

[52] D. Linden and J. Connor, "Long-term synaptic depression," *Annu Rev Neurosci*, vol. 18, pp. 319–357, 1995.

[53] C. Koch, *Biophysics of Computation: Information Processing in Single Neurons*, M. Stryker, Ed. Oxford University Press, 1999.

[54] 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*, vol. 18, no. 24, pp. 10 464–10 472, 1998.

[55] W. Gerstner, W. Kistler, R. Naud, and L. Paninski, *Neuronal Dynamics. From single neurons to networks and models of cognition*. Cambridge University Press, 2014.

[56] A. Destexhe, Z. Mainen, and T. Sejnowski, "Synthesis of models for excitable membranes, synaptic transmission and neuromodulation using a common kinetic formalism," *Journal of Computational Neuroscience*, vol. 1, pp. 195–230, 1994.

[57] A. Morris, M. Diesmann, and W. Gerstner, "Phenomenological models of synaptic plasticity based on spike timing," *Biological Cybernetics*, vol. 98, no. 6, pp. 459–478, 2008.

[58] J. Lisman, "Three Ca(2+) levels affect plasticity differently: the LTP zone, the LTD zone and no man's land," *The Journal of Physiology*, vol. 532, no. 2, p. 285, 2001.

[59] I. Ismailov, D. Kalikulov, T. Inoue, and M. Friedlander, "The kinetic profile of intracellular calcium predicts long-term potentiation and long-term depression," *The Journal of Neuroscience*, vol. 24, no. 44, pp. 9847–9861, 2004.

[60] J. L. Hellier, D. R. Grosshans, S. J. Coultrap, J. P. Jones, P. Dobelis, M. D. Browning, and K. J. Staley, "Nmda receptor trafficking at recurrent synapses stabilizes the state of the ca3 network," *Journal of Neurophysiology*, vol. 98, no. 5, pp. 2818–2826, 2007.

[61] L. Abbott and S. Nelson, "Synaptic plasticity: taming the beast," *Nature Neuroscience*, vol. 3, pp. 1178–1183, November 2000.

[62] T. Nevian and B. Sakmann, "Spine ca^{2+} signaling in spike-timing-dependent plasticity," *The Journal of Neuroscience*, vol. 26, pp. 11 001–11 013, 2006.

[63] D. H. O'Connor, G. M. Wittenberg, and S. H. Wang, "Graded bidirectional synaptic plasticity is composed of switch-like unitary events," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 27, pp. 9679–9684, 2005.

[64] S. Fusi and L. Abbott, "Limits on the memory storage capacity of bounded synapses," *Nature Neuroscience*, vol. 10, pp. 485–493, 2007.

[65] S. Sze and K. K. Ng, *Physics of Semiconductor Devices*, 3rd ed. New Jersey: Wiley, 2007.

[66] Y. Tsividis and C. McAndrew, *Operation and modeling of the MOS transistor*. Oxford University press, 2010.

[67] M. Mahowald and R. Douglas, "A silicon neuron," *Nature*, vol. 354, pp. 515–518, 1991.

[68] C. Enz, M. A. Chalkiadaki, and A. Mangla, "Low-power analog/RF circuit design based on the inversion coefficient," in *European Solid-State Circuits Conference (ESSCIRC), ESSCIRC 2015 - 41st*, Sept 2015, pp. 202–208.

[69] P. Allen and D. Holberg, *CMOS Analog Circuit Design*, 2nd ed. Oxford University Press, 2002.

[70] B. Razavi, *Design of analog CMOS integrated circuits*. Boston, MA: McGraw-Hill, 2001.

[71] S. Liu, J. Kramer, G. Indiveri, T. Delbruck, and R. Douglas, *Analog VLSI:Circuits and Principles*. MIT Press, 2002.

[72] Y. S. Chauhan, S. Venugopalan, M. A. Karim, S. Khandelwal, N. Paydavosi, P. Thakur, A. M. Niknejad, and C. C. Hu, "Bsim x2014; industry standard compact mosfet models," in *2012 Proceedings of the ESSCIRC (ESSCIRC)*, Sept 2012, pp. 30–33.

[73] Y. S. Chauhan, S. Venugopalan, M. A. Chalkiadaki, M. A. U. Karim, H. Agarwal, S. Khandelwal, N. Paydavosi, J. P. Duarte, C. C. Enz, A. M. Niknejad, and C. Hu, "Bsim6: Analog and rf compact model for bulk mosfet," *IEEE Transactions on Electron Devices*, vol. 61, no. 2, pp. 234–244, Feb 2014.

[74] R. Troutman, "VLSI limitations from drain-induced barrier lowering," *IEEE Transactions on Electron Devices*, vol. ED-26, no. 4, pp. 461–469,

April 1979.

[75] M. Lundstrom and Z. Ren, "Essential physics of carrier transport in nanoscale mosfets," *IEEE Transactions on Electron Devices*, vol. 49, no. 1, pp. 133–141, Jan 2002.

[76] K. Lakshmikumar, R. Hadaway, and M. Copeland, "Characterization and modeling of mismatch in MOS transistors for precision analog design," *IEEE Journal of Solid-State Circuits*, vol. SC-21, no. 6, pp. 1057–1066, December 1986.

[77] M. Pelgrom, A. Duinmaijer, and A. Welbers, "Matching properties of MOS transistors," *IEEE Journal of Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1440, October 1989.

[78] P. R. Kinget, "Device mismatch and tradeoffs in the design of analog circuits," *IEEE Journal of Solid-State Circuits,*, vol. 40, no. 6, pp. 1212–1224, 2005.

[79] M. J. M. Pelgrom, H. P. Tuinhout, and M. Vertregt, "Transistor matching in analog cmos applications," in *Electron Devices Meeting, 1998. IEDM '98. Technical Digest., International*, Dec 1998, pp. 915–918.

[80] P. G. Drennan and C. C. McAndrew, "Understanding mosfet mismatch for analog design," *IEEE Journal of Solid-State Circuits*, vol. 38, no. 3, pp. 450–456, Mar 2003.

[81] D. M. Binkley, C. E. Hopper, S. D. Tucker, B. C. Moss, J. M. Rochelle, and D. P. Foty, "A cad methodology for optimizing transistor current and sizing in analog cmos design," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 22, no. 2, pp. 225–237, Feb 2003.

[82] P. G. Drennan and C. C. McAndrew, "A comprehensive mosfet mismatch model," in *Electron Devices Meeting, 1999. IEDM '99. Technical Digest. International*, Dec 1999, pp. 167–170.

[83] J. Bastos, M. Steyaert, B. Graindourze, and W. Sansen, "Influence of die attachment on mos transistor matching," in *Microelectronic Test*

Structures, 1996. ICMTS 1996. Proceedings. 1996 IEEE International Conference on, Mar 1996, pp. 27–31.

[84] T.-H. Yeh, J. C. H. Lin, S.-C. Wong, H. Huang, and J. Y. C. Sun, "Mis-match characterization of 1.8 v and 3.3 v devices in 0.18 mu;m mixed signal cmos technology," in *Microelectronic Test Structures, 2001. ICMTS 2001. Proceedings of the 2001 International Conference on*, 2001, pp. 77–82.

[85] J. Bastos, M. Steyaert, B. Graindourze, and W. Sansen, "Matching of mos transistors with different layout styles," in *Microelectronic Test Structures, 1996. ICMTS 1996. Proceedings. 1996 IEEE International Conference on*, Mar 1996, pp. 17–18.

[86] P. G. Drennan, M. L. Kniffin, and D. R. Locascio, "Implications of proximity effects for analog design," in *IEEE Custom Integrated Circuits Conference 2006*, Sept 2006, pp. 169–176.

[87] C. Bartolozzi, S. Mitra, and G. Indiveri, "An ultra low power current–mode filter for neuromorphic systems and biomedical signal processing," in *Biomedical Circuits and Systems Conference, BIOCAS 2006*. IEEE, 2006, pp. 130–133.

[88] C. Enz, M. Punzenberger, and D. Python, "Low-voltage log-domain signal processing in cmos and bicmos," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 46, no. 3, pp. 279–289, Mar 1999.

[89] C. Bartolozzi and G. Indiveri, "Synaptic dynamics in analog VLSI," *Neural Computation*, vol. 19, no. 10, pp. 2581–2603, Oct 2007.

[90] J. Lazzaro, S. Ryckebusch, M. Mahowald, and C. Mead, "Winner-take-all networks of $O(n)$ complexity," in *Advances in neural information processing systems*, D. Touretzky, Ed., vol. 2. San Mateo - CA: Morgan Kaufmann, 1989, pp. 703–711.

[91] J. Schemmel, J. Fieres, and K. Meier, "Wafer-scale integration of analog neural networks," in *Proceedings of the IEEE International Joint Conference on Neural Networks*, 2008.

[92] J. Bono and C. Clopath, "Modeling somatic and dendritic spike mediated plasticity at the single neuron and network level," *Nature Communications*, vol. 8, no. 706, 2017.

[93] C. Clopath, L. Büsing, E. Vasilaki, and W. Gerstner, "Connectivity reflects coding: a model of voltage-based STDP with homeostasis," *Nature Neuroscience*, vol. 13, no. 3, pp. 344–352, 2010.

[94] J. Brader, W. Senn, and S. Fusi, "Learning real world stimuli in a neural network with spike-driven synaptic dynamics," *Neural Computation*, vol. 19, pp. 2881–2912, 2007.

[95] P. Häfliger, "Adaptive WTA with an analog VLSI neuromorphic learning chip," *IEEE Transactions on Neural Networks*, vol. 18, no. 2, pp. 551–572, 2007.

[96] E. Chicca, D. Badoni, V. Dante, M. D'Andreagiovanni, G. Salina, L. Carota, S. Fusi, and P. Del Giudice, "A VLSI recurrent network of integrate–and–fire neurons connected by plastic synapses with long–term memory," *IEEE Transactions on Neural Networks*, vol. 14, no. 5, pp. 1297–1307, 2003.

[97] M. Giulioni, F. Corradi, V. Dante, and P. del Giudice, "Real time unsupervised learning of visual stimuli in neuromorphic VLSI systems," *Scientific Reports*, vol. 5, no. 14730, pp. 1–10, 2015.

[98] J. P. Pfister and W. Gerstner, "Triplets of spikes in a model of spike timing-dependent plasticity," *The Journal of Neuroscience*, vol. 26, no. 38, pp. 9673–9682, 2006.

[99] D. J. Amit and S. Fusi, "Learning in neural networks with material synapses," *Neural Computation*, vol. 6, no. 5, pp. 957–982, 1994.

[100] E. Chicca, F. Stefanini, C. Bartolozzi, and G. Indiveri, "Neuromorphic electronic circuits for building autonomous cognitive systems," *Proceedings of the IEEE*, vol. 102, no. 9, pp. 1367–1388, 2014.

[101] B. Linares-Barranco and T. Serrano-Gotarredona, "On the design and characterization of femtoampere current-mode circuits," *IEEE J. Solid-*

State Circuits, vol. 38, no. 8, pp. 1353–1363, August 2003.

[102] R. Sarpeshkar, R. F. Lyon, and C. Mead, "A low-power wide-linear-range transconductance amplifier," *Analog Integrated Circuits and Signal Processing*, vol. 13, no. 1-2, pp. 123–151, 1997.

[103] P. J. Sjöström, G. G. Turrigiano, and S. B. Nelson, "Rate, timing, and cooperativity jointly determine cortical synaptic plasticity," *Neuron*, vol. 32, no. 6, pp. 1149–1164, 2001.

[104] M. Giulioni, P. Camilleri, V. Dante, D. Badoni, G. Indiveri, J. Braun, and P. Del Giudice, "A VLSI network of spiking neurons with plastic fully configurable "stop-learning" synapses," in *International Conference on Electronics, Circuits, and Systems, ICECS 2008*. IEEE, 2008, pp. 678–681.

[105] S. G. Narendra, "Challenges and design choices in nanoscale cmos," *J. Emerg. Technol. Comput. Syst.*, vol. 1, no. 1, pp. 7–49, Mar. 2005.

[106] R. Sarpeshkar, "Analog versus digital: Extrapolating from electronics to neurobiology," *Neural Computation*, vol. 10, no. 7, pp. 1601–1638, October 1998.

[107] L. Abbott and W. Regehr, "Synaptic computation," *Nature*, vol. 431, pp. 796–803, October 2004.

[108] T. Delbruck, R. Berner, P. Lichtsteiner, and C. Dualibe, "32-bit configurable bias current generator with sub-off-current capability," in *International Symposium on Circuits and Systems, ISCAS 2010*. IEEE, 2010, pp. 1647–1650.

[109] J. Arthur and K. Boahen, "Recurrently connected silicon neurons with active dendrites for one-shot learning," in *IEEE International Joint Conference on Neural Networks*, vol. 3, July 2004, pp. 1699–1704.

[110] D. Sumislawska, N. Qiao, M. Pfeiffer, and G. Indiveri, "Wide dynamic range weights and biologically realistic synaptic dynamics for spike-based learning circuits," in *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2016, pp. 2491–2494.

[111] P. G. P. Hurst, S. Lewis, and R. Meyer, *Analysis and design of analog integrated circuits*, 5th ed. New York: Wiley, 2009.

[112] B. Linares-Barranco, T. Serrano-Gotarredona, R. Serrano-Gotarredona, and C. Serrano-Gotarredona, "Current mode techniques for sub-pico-ampere circuit design," *Analog Integrated Circuits and Signal Processing*, vol. 38, no. 2, pp. 103–119, 2004.

[113] G. Rincon-Mora, "Current efficient, low voltage, low dropout regulators," Ph.D. thesis, Georgia Institute of Technology, 1996.

[114] T. Delbruck and A. Van Schaik, "Bias current generators with wide dynamic range," *Analog Integrated Circuits and Signal Processing*, vol. 43, no. 3, pp. 247–268, 2005.

[115] S. Nicolson and K. Phang, "Improvements in biasing and compensation of cmos opamps," *Analog Integrated Circuits and Signal Processing*, vol. 43, no. 3, pp. 237–245, 2005.

[116] K. Bult and G. Geelen, "An inherently linear and compact most-only current-division technique," in *1992 IEEE International Solid-State Circuits Conference Digest of Technical Papers*, Feb 1992, pp. 198–199.

[117] K. Boahen, "A burst-mode word-serial address-event link – I: Transmitter design," *IEEE Transactions on Circuits and Systems I*, vol. 51, no. 7, pp. 1269–80, 2004.

[118] K. Boahen, "A burst-mode word-serial address-event link – II: Receiver design," *IEEE Transactions on Circuits and Systems I*, vol. 51, no. 7, pp. 1281–91, 2004.

[119] K. Boahen, "A burst-mode word-serial address-event link – III: Analysis and test results," *IEEE Transactions on Circuits and Systems I*, vol. 51, no. 7, pp. 1292–300, 2004.

[120] J. Hopfield, "Neurons with graded response have collective computational properties like those of two-state neurons," *Proceedings of the national academy of sciences*, vol. 81, no. 10, pp. 3088–3092, 1984.

[121] R. J. Douglas and K. A. C. Martin, "Recurrent neuronal circuits in the neocortex," *Current Biology*, vol. 17, no. 13, pp. R496–R500, 2007.

[122] M. Mayford, S. A. Siegelbaum, and E. R. Kandel, "Synapses and memory storage," *Cold Spring Harbor Perspective in Biology*, vol. 4, no. 6, 2012.

[123] T. P. Trappenberg, *Fundamentals of Computational Neuroscience*. Oxford Universty Press, 2010.

[124] J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the national academy of sciences*, vol. 79, no. 8, pp. 2554–2558, 1982.

[125] G. Indiveri, B. Linares-Barranco, T. Hamilton, A. van Schaik, R. Etienne-Cummings, T. Delbruck, S. Liu, P. Dudek, P. Häfliger, S. Renaud, J. Schemmel, G. Cauwenberghs, J. Arthur, K. Hynna, F. Folowosele, S. Saighi, T. Serrano-Gotarredona, J. Wijekoon, Y. Wang, and K. Boahen, "Neuromorphic silicon neuron circuits," *Frontiers in Neuroscience*, vol. 5, pp. 1–23, 2011.

[126] G. Indiveri, F. Stefanini, and E. Chicca, "Spike–based learning with a generalized integrate and fire silicon neuron," in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2010, pp. 1951–1954.

[127] T. Jamil, "Ram versus cam," *IEEE Potentials*, vol. 16, no. 2, pp. 26–29, 1997.

[128] P. J. Brasted, T. J. Bussey, E. A. Murray, and S. P. Wise, "Role of the hippocampal system in associative learning beyond the spatial domain," *Brain*, vol. 126, no. 5, pp. 1202–1223, 2003.

[129] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review,*, vol. 65, no. 6, pp. 386–408, Nov 1958.

[130] G. T. Buracas, A. M. Zador, M. R. DeWeese, and T. D. Albrigh, "Efficient discrimination of temporal patterns by motion-sensitive neurons in

primate visual cortex," *Neuron*, vol. 20, no. 5, pp. 959 – 969, 1998.

[131] C. Enz and E. Vittoz, *Charge-Based MOS Transistor Modeling*. John Wiley & Sons, Ltd, 2006.

[132] D. Stefanovic and M. Kayal, *Structured Analog CMOS Design*, 1st ed. Springer Publishing Company, Incorporated, 2010.

[133] F. Forti and M. E. Wright, "Measurement of mos current mismatch in the weak inversion region," *IEEE Journal of Solid-State Circuits*, pp. 138–142, 1994.

[134] R. J. Douglas, K. A. C. Martin, and D. Whitteridge, "A canonical microcircuit for neocortex," *Neural Computation*, vol. 1, no. 4, pp. 480–488, 1989.

[135] U. Rutishauser and R. Douglas, "State-dependent computation using coupled recurrent networks," *Neural Computation*, vol. 21, pp. 478–509, 2009.

Publications

Journal Paper:

1. **Frank L. Maldonado Huayaney**, S. Nease and E. Chicca. "Learning in Silicon Beyond STDP: A Neuromorphic Implementation of Multi-Factor Synaptic Plasticity with Calcium-Based Dynamics," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 63, no. 12, pp. 2189-2199, Dec. 2016.

Refereed Conference Paper:

1. **Frank L. Maldonado Huayaney** and E. Chicca, "A VLSI implementation of a calcium-based plasticity learning model," *IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 373-376, Montreal, QC, May 2016.

2. **Frank L. Maldonado Huayaney**, H. Tanaka, T. Matsuo T. Morie and K. Aihara. "Analysis of Associative Memory Operation in a VLSI Spiking Neural Network," *The 21st Annual Conference of the Japanese Neural Network Society (JNNS 2011)*, pp. 208–209, Okinawa, Japan, Dec. 2011.

3. **Frank L. Maldonado Huayaney**, H. Tanaka, T. Matsuo T. Morie and K. Aihara. "A VLSI Spiking Neural Network with Symmetric STDP and Associative Memory Operation," *Neural Information Processing - ICONIP 2011 Proceedings, Part III*, in book series of *Lecture Notes in Computer Science*, Vol. 7064, pp. 381–388, Springer-Verlag, Shanghai, China, Nov. 2011.