UNIVERSITY OF LIVERPOOL

BIOLOGICALLY MOTIVATED CIRCUITS FOR THIRD GENERATION NEURAL NETWORKS

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy

by

Thomas Dowrick

January 2011

Abstract

As interest in the possibilities of creating systems which can mimic the operation of biological nervous systems grows, small area, low power devices are required which can replicate the important features observed of neural cells. In addition, as advancements in CMOS technology become more challenging and expensive, alternative uses for existing silicon processing technologies and alternative computational paradigms are required, as indicated by the ITRS roadmap.

In this thesis, three separate neural devices, capable of implementing the pertinent features of their biological counterparts, are described. The first is a compact silicon synapse, consisting of either two or three series connected MOSFETs, compatible with spike based communication methods. Plasticity is impleme`nted through an adjustable weight voltage, V_W, which controls the amount of charge in the synapse. Short term depression and refraction are possible through a second control voltage, V_P, which sets the rate at which the synaptic charge is replenished, with recovery times comparable to biology - between 0.5us and 10ms possible. With a transitor count of 3 and circuit area of 2.1µm x 6.2µm, the synapse is, to the author's knowledge, the most compact of any such device reported to date, while offering the same level of functionality.

A neuron circuit, requiring three MOSFETs, is capable of summing excitatory and inhibitory synaptic inputs from the synapse cell, generating biologically plausible post synaptic potentials (PSPs). A MOSFET biased in subthreshold provides a method of adjusting the decay time of the PSP. The addition of a two stage CMOS inverter allows the neuron to generate spike outputs when the triggering voltage of the cell has been reached.

A circuit for the implementation of an axonal delay, requiring only 5 transistors, is also described. Leakage through a subthreshold MOSFET creates a delay path between the output of a presynaptic neuron and the input of a post synaptic neuron, where delay times between 10s of milliseconds and 10s of nanoseconds are possible.

Theoretical analysis, using parameters extracted from test MOS devices, is used to describe the operation of each device. Simulation results and results taken from fabricated chips confirm the validity of the approach.

Methods by which the individual cells can be connected together to create larger scale networks are described, and a number of the issues associated with VLSI neural systems are considered.

Acknowledgments

Firstly, I would like to thank my supervisor, Prof. Steve Hall, for his assistance, guidance and knowledge which have greatly benefited me both in terms of my research and in the preparation of this thesis.

In addition, I would like to thank Dr. O. Buiu, Dr. L. McDaid, Dr. J. Marsland, Dr. Y. Chen, Dr. A. Ghani, Dr. S. Huang and Mr. A. Smith for their collaboration on this work. Their support, criticisms and comments have been greatly appreciated.

I would also like to thank my colleagues at the University of Liverpool who have assisted me in numerous ways throughout the course of my PhD: Dr. D. Donaghy, Dr. L. Tan, Dr. H. Aung, Dr. L. Grossman, Mr. S. Laters, Mr. A. Baleci, Mr. J. Livingstone, Mr. M. Trivedi, Mr. J. Lees, Mr. S. Jones, Mr. M. Round, Mr. S Longmuir, Mr. D. Winstanley and Ms. S. Duffy.

Finally, I would like to thank all of my friends and family who have thoroughly supported me throughout the duration of my PhD studies.

List of Symbols

Symbol	Significance	Unit
A	Area	m^2
C_d	Depletion layer capacitance	F
C_{GS}	MOSFET gate-source capacitance	F
C_{max}	Maximum MOS capacitance	F
C_{min}	Minimum MOS capacitance	F
C_0	Oxide capacitance per unit area	Fm^{-2}
C_P	Neuron parasitic capacitance	F
C_{PSP}	Capacitance associated with V_{PSP} node	F
C_{VIN}	Capacitance associated with V_{IN} node	F
E_C	Conductance band edge	eV
E_F	Fermi energy level	eV
E_g	Energy gap	eV
E_i	Intrinsic Fermi energy level	eV
E_V	Valence band edge	eV
I_D	Drain current	A
ISI	Period between presynaptic inputs	s
k	Boltzman's constant (1.28 x 10^{-23})	JK^{-1}
L	Transistor length	m
m	Gate-channel coupling coefficient	-
N_A	Acceptor doping concentration	m^{-3}
N_f	Total oxide charge density	Cm^{-2}
n_i	Intrinsic doping concentration (1.6 x 10^{16})	m^{-3}
q	Electron charge (1.6 x 10^{-19})	C
Q_d	Depletion region charge	C
Q_f	Fixed oxide charge density	Cm^{-2}
Q_g	MOS gate charge	C
Q_m	Mobile oxide charge density	Cm^{-2}
Q_{sc}	Inversion layer charge	C
Q_t	Trapped oxide charge density	Cm^{-2}
Q_W	Synaptic output charge	C

φ_s	Semiconductor surface potential	V
φ_b	Semiconductor bulk potential	V
S	Subthreshold slope	mV/decade
T	Temperature (300K)	K
t_{ox}	Oxide thickness	m
t_{post}	Post-synaptic neuron firing time	s
t_{pre}	Pre-synaptic neuron firing time	s
t_{rPSP}	PSP rise time	s
t_{rPSP}	PSP fall time	s
V_{DD}	Positive supply voltage	V
V_{DM1}	Synapse virtual drain voltage	V
V_{DS}	Drain-source voltage	V
V_{FB}	Flat Band voltage	V
V_G	Gate voltage	V
V_{GS}	Gate-source voltage	V
V_{IN}	Neuron input node voltage	V
V_{LEAK}	Neuron leakage voltage	V
V_M	Inverter triggering voltage	V
V_{mg}	Mid-gap voltage	V
V_N	Axon circuit capacitor node voltage	V
V_o	Semiconductor oxide voltage drop	V
V_P	Synapse charge recovery control voltage	V
V_{PRES}	Presynaptic voltage pulse	V
V_{PSP}	Neuron postsynaptic potential	V
V_{RR}	Resting potential of V_{IN} node	V
V_{SS}	Negative supply voltage	V
V_{sub}	MOSFET substrate bias	V
V_T	Threshold voltage	V
V_W	Synapse weight voltage	V
V_{Wi}	Inhibitory weight voltage	V
W	MOSFET width	m
W_d	Depletion layer width	m
W_{df}	Equilibrium depletion width	m

W_{do}	Depletion layer width in deep depletion	m
β	MOSFET gain factor	A/V
γ	Body effect factor	$V^{1/2}$
δ	Channel depth	m
ΔT	Pulse width of V_{PRES}	s
ΔV_{PSP}	Change in PSP in response to a single input	V
ΔV_T	Shift in threshold voltage due to substrate bias	V
ε_0	Permittivity of Free Space	Fm^{-1}
ε_{ox}	Relative Permittivity of Silicon Dioxide	-
ε_{si}	Relative Permittivity of Silicon	-
λ	Channel length modulation factor	V^{-1}
μ	mobility	$m^2V^{-1}s^{-1}$
τ_{fPSP}	Fall time of V_{PSP}	s
τ_{r1}	Subthreshold rise time of V_{IN}	s
τ_{r2}	Above threshold rise time of V_{IN}	s
τ_{rPSP}	Rise time of V_{PSP}	s
τ_s	Three terminal synapse charge recovery time	s
Φ_b	Bulk Potential	eV
Φ_m	Metal Work Function	eV
Φ_{ms}	Work function difference	eV
Φ_s	Semiconductor Work Function	eV
χ	Electron Affinity	eV

Table of Contents

Chapter 1: Introduction 1

1.1 Background 1

1.2 Neural networks 2

1.2.1 Biological neural networks 2

1.2.2 Artificial neural networks 5

1.3 Current status of neural networks in VLSI 8

1.3.1 Neurons 10

1.3.2 Synapse circuits 13

1.4 Organisation of the Thesis 17

Chapter 2: Overview of MOS Physics and Device Characterisation 29

2.1 Introduction 29

2.2 MOS Capacitor 29

2.2.1 Extraction of MOS capacitor parameters 33

2.2.2 MOSFET operation 37

2.2.3 Extraction of MOSFET device parameters 40

2.3 Discussion and Conclusions 45

Chapter 3: Silicon Synapse Device and Circuit 47

3.1 Introduction 47

3.2 Two-Terminal Silicon Static Synapse 48

3.2.1 Theoretical Operation 49

3.2.2 Results 60

3.3 The three-terminal dynamic synapse 67

3.3.1 Theory 68

3.3.2 Simulation and Experimental Results 72

3.4 Conclusions and Discussion 77

Chapter 4: Neuron Circuit 79

4.1 Introduction 79

4.2 Theory of operation 79
4.2.1 Derivation of ΔV_{PSP} 80
4.2.2 Rise Time 82
4.2.3 Fall Time 83
4.3 Results 85
4.3.1 Two terminal synapse 85
4.3.2 Three terminal (dynamic) synapse 90
4.4 Triggering circuitry 95
4.5 Inhibitory synapses 98
4.6 Conclusions 99
Chapter 5: Axonal Delay Circuit 100
5.1 Introduction 100
5.2 Axon Circuit and Theory of Operation 101
5.3 Simulation and Experimental results 102
5.4 Series connected neurons 105
5.5 Conclusions and Discussion 106
Chapter 6: VLSI Issues 108
6.1 Introduction 108
6.2 Device variability 109
6.3 Scalability Issues and Standard Neuron Cell 113
6.4 Proposed Circuit for XOR Benchmark Problem 118
6.5 Conclusions 121
Chapter 7: Conclusions and Further Work 123
Appendix – Test Chip Design and Fabrication 126
A1.1 Introduction 126
A1.2 Circuit Layouts 126
A1.3 Output Buffers 130
A1.4 ESD Protection 134

A1.5 Chip Packaging 136

A1.6 Measurement Setup 138

Chapter 1: Introduction

1.1 Background

The proliferation of silicon transistors since the middle of the twentieth century has underpinned a prolonged and continuing increase in the computational power available to humans. The semiconductor industry has successfully kept pace with the prediction made in the 1960s by Gordon Moore that the density of integrated circuits would double every 18 months. At present, the International Technology Roadmap for Semiconductors (IRTS) [1] outlines how the industry will continue to pursue Moore's Law past the current 45nm generation, to 18nm in 2015 and beyond. In addition to this, the roadmap acknowledges that scaling cannot continue indefinitely. The concept of 'More than Moore' was introduced in 2005 to describe devices and systems which could exploit existing semiconductor technologies in unique and diverse ways. Such 'functional diversification' includes System in Package (SiP), Microelectromechanical Systems (MEMS), RF technologies, sensors, actuators and biotechnology.

While the semiconductor industry was undergoing a rapid expansion, improvements in the field of molecular biology, electrophysiology, electron microscopy and computational neuroscience produced substantial improvements in the understanding of the structure and operation of the brain [2]. The molecular components responsible for the growth and modification of neural cells and connections were identified. At the cellular level, advances in electronics allowed for the electrical characteristics of individual neurons to be measured. Higher level brain functions such as vision, hearing, learning and memory were studied under the field of systems neuroscience.

The modern microprocessor can perform mathematical calculations which are orders of magnitude more complex than those which can be computed by the human brain, at a far greater speed. Despite this, there exist several areas in which the operations of the brain appear to be superior to those of a computer. These include visual and auditory processing, language, independent learning, pattern recognition and classification [3]. The superiority of the brain in these situations is thought to be due to the manner in which information is

processed. The massively interconnected network of neurons and adaptive synapses in the brain allows for large volumes of information to be processed in parallel, as opposed to the more serial, algorithmic processing employed by conventional microprocessors. The need for further research into biologically inspired systems is recognised and supported by a number of international programmes, including EPSRC's 'Grand Challenges in Microelectronic Design 4: Building Brains' [4], DARPA's 'SyNAPSE' project [5] and the EU's FP7 FET 'Brain Inspired ICT' call [6].

The appreciation of the processing capabilities of biological systems led to the formation and growth of the field of neural networks. The neural network is a processing paradigm inspired by the way in which biological systems operate. Networks of individual processing elements work in parallel to solve problems. Generally the target problem is one which cannot readily be solved using conventional processing techniques. More recently, the field of neuromorphic engineering has emerged. Electronic circuits are created which attempt to mimic the processes, functions and architectures present in biological systems; in order to create artificial neural systems which closely resemble their biological counterparts.

1.2 Neural networks

In order to properly discuss the topic of artificial neural networks (ANNs), it is first necessary to consider the structure and operation of biological neural networks. An overview of the components and important features of a biological neural network is presented in this section, followed by an introduction to the field of artificial neural networks and their implementations in software and hardware.

1.2.1 Biological neural networks

A typical biological neuron consists of a cell body known as soma, an axon and a network of dendrites; as illustrated in Figure 1.1. A synaptic connection is formed between two neurons when the axon of a presynaptic neuron forms a junction, known as a synaptic cleft, with a dendrite of a postsynaptic neuron. Interneuron communication is achieved through chemical

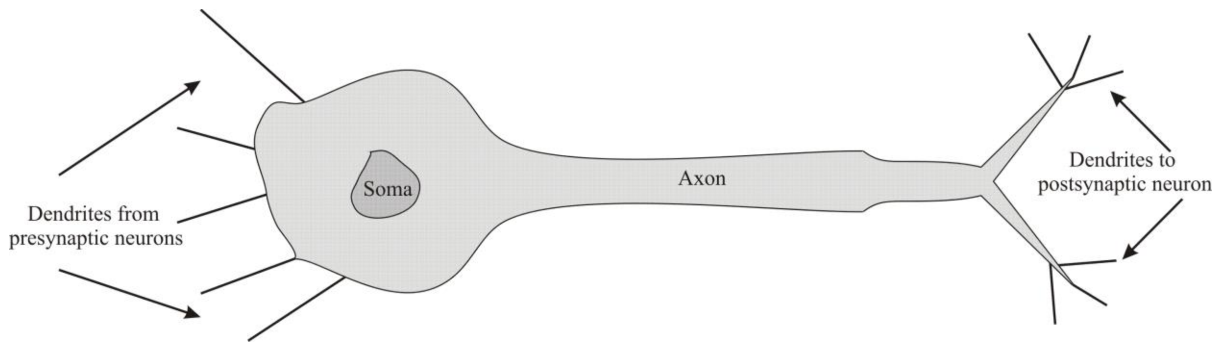

Figure 1.1 - A typical neuron cell.

transmissions across the synaptic cleft. Voltage gated calcium (Ca) channels control the release of neurotransmitters from the presynaptic side into the synaptic cleft. Receptors on the postsynaptic side of the synaptic gap bind with neurotransmitter molecules, opening, or closing, ion channels, which affects the membrane potential of the postsynaptic neuron. Multiple synaptic transmissions are temporally summed in the soma, which has a typical resting potential of -65mV. Synapses can be either excitatory or inhibitory, depending on the types of receptors involved, which increase or decrease the membrane potential. If the combined synaptic inputs increase the membrane potential, commonly referred to as the postsynaptic potential (PSP), above a certain threshold level, the neuron is said to fire and an action potential is generated. The action potential is propagated along the axon, where synaptic connections to neighbouring neurons are activated. Figure 1.2 demonstrates the membrane voltage response to three successive synaptic inputs, with the third initiating the generation of an action potential. Following the generation of an action potential, the neuron enters into a refractory period, typically several milliseconds in length, during which the cell membrane potential does not change in the presence of additional synaptic inputs.

The human brain contains in the order of 10^{11} neurons [7], each with up to 10^3 associated synapses. From a topological standpoint, the number of synapses dominates the neural architecture. The synapse is also responsible for learning and adaption in neural systems, through the modulation of the synaptic strength, also referred to as the weight. The strength of a synaptic connection can be increased (potentiated) or decreased (depressed) to alter the influence it has over an individual neuron. A strong synapse would be able to trigger a neuron's action potential in the absence of other synaptic inputs, whereas a particularly weak synapse may have no net effect on the membrane voltage. One postulate which describes how weight updates can take place is Hebbian Learning [2].

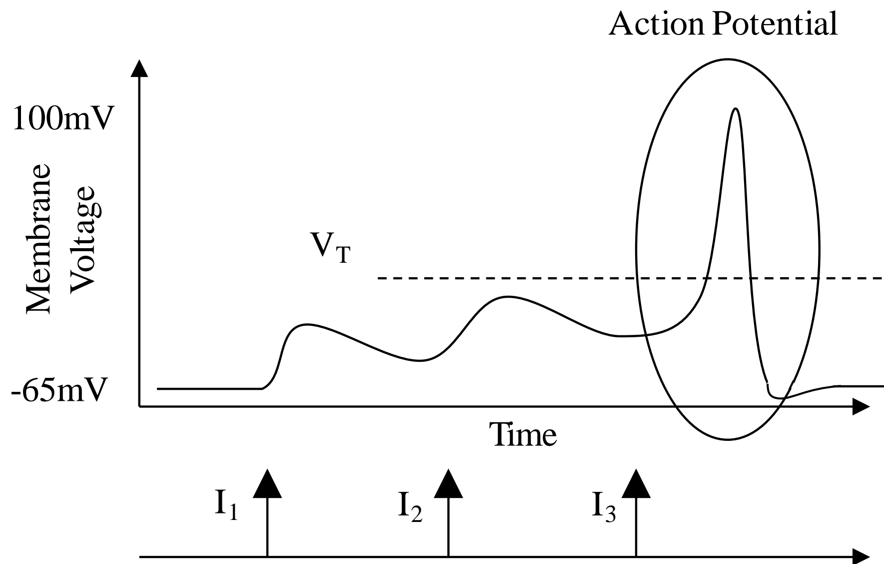

Figure 1.2 - Response of post synaptic potential (PSP) to three synaptic inputs and action potential generation.

Hebb's rule states that the strength of connection between two neurons will be increased if the presynaptic neuron repeatedly contributes to the firing of the postsynaptic neuron.

There are a number of ways in which synaptic weights can be changed. Short term potentiation and depression (STP/STD) temporarily increase or decrease respectively, the synaptic weight following an initial synaptic event. The duration of such short term events is typically between ten milliseconds and several minutes, after which time the synaptic weight returns to its initial value [8]. Long term potentiation and depression (LTP/LTD) induce permanent changes in synaptic weights in response to extended periods of synaptic activity. One of the most widely studied mechanisms for long term weight changes is spike timing dependent plasticity (STDP), which is a more specific form of Hebbian Learning. The relative timing of presynaptic and postsynaptic spikes governs the magnitude and direction of the synaptic weight change [9-11]. For an excitatory synapse, the occurrence of a presynaptic action potential shortly before a postsynaptic action potential results in an increased weight. Depression occurs if postsynaptic firing, at time t_{post}, occurs before presynaptic firing, at time t_{pre}. The magnitude and direction of weight changes are illustrated in Figure 1.3.

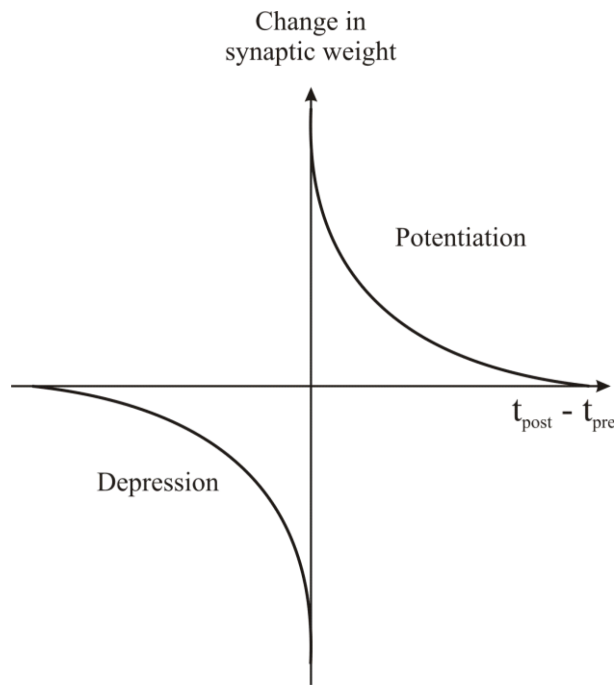

Figure 1.3 - Illustration of STDP for an excitatory synapse.

1.2.2 Artificial neural networks

Artificial neural networks, more commonly referred to simply as neural networks, are computational devices which are inspired by, and attempt to emulate, the operation of biological neural systems. Massively interconnected networks of simple processing elements (neurons) work in parallel to solve a given problem; weighted synaptic connections between neurons store information. The learning process alters these synaptic weights based upon some predefined learning rule, in order to find the optimal solution to the problem(s) being solved. Neural networks are commonly employed in problem areas where an algorithmic solution either does not exist, or is too complex to be found. They are also highly versatile. Properly trained, a single network can perform a variety of tasks including classification, regression, clustering and forecasting. Most commonly implemented in software, current applications include control systems [12], robotics [13], cancer detection [14], pattern recognition [15, 16], image processing [17], forecasting [18] and face detection [19].

A schematic view of a typical neuron is shown in Figure 1.4. The output of the summing junction can be described mathematically:

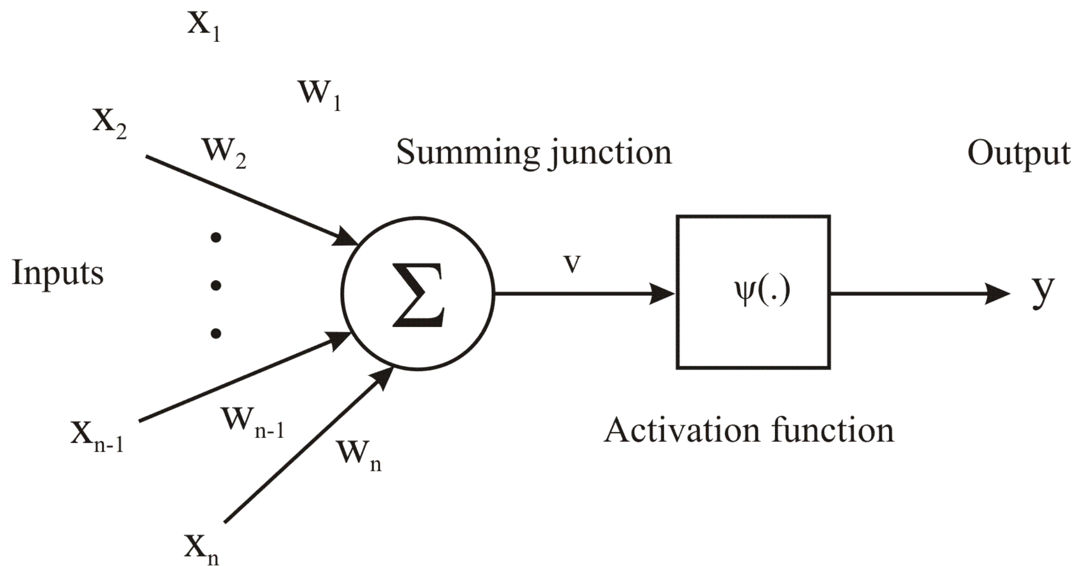

Figure 1.4 - A typical neuron with inputs x_1 - x_n, corresponding synaptic weights w_1 - w_n and output y.

$$v = \sum_{j=1}^{n} x_j \, w_j \tag{1.1}$$

The activation function determines the relationship between the sum of the inputs and the neuronal output. The first generation of neurons, McCulloch-Pitts neurons, employed a simple step function for this purpose, outputting either a 0 or 1. The second generation of neural networks introduced continuous activation functions, which allowed for analog inputs and outputs to be used. Linear activation functions can be used, but the most common type of activation function is the sigmoid function, of which the logistic function is an example:

$$\Psi(x) = \frac{1}{1 + \exp(-ax)} \tag{1.2}$$

where a is the slope parameter which can be adjusted to change the shape of the function.

An example of the type of connectivity employed in a neural network can be seen in Figure 1.5, which shows a fully connected feed-forward network, in which information is only transmitted in the forward direction from inputs to outputs. Networks which include feedback paths are also widely used, as feedback is believed to be an important component of the learning process in real neural networks.

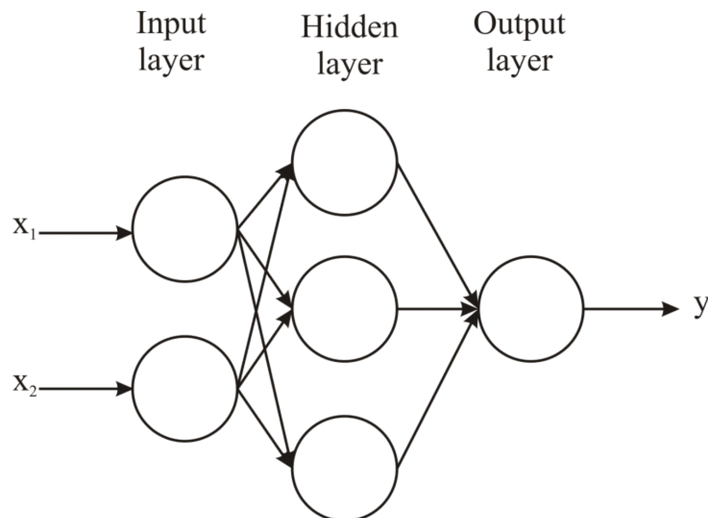

Figure 1.5 - A fully connected feedforward neural network.

The first two generations of neural networks were able to successfully replicate a number of the key features of neural systems; including plasticity, summation and thresholding; to solve a variety of complex problems. They have been demonstrated to out-perform conventional techniques, such as regression analysis, in areas where the input data is highly non-linear, or where robust statistical data is not available [20, 21]. Neurological research suggests that a large part of the computational power of the brain stems from its ability to process large numbers of spiking signals in parallel, where the relative timing of signals plays a crucial part [22]. These temporal dynamics are not taken into account in the neural network models previously outlined. As a result of this, a third generation of neural networks was conceived, where the timing of spikes could be factored into the computations, known as spiking neural networks (SNNs). It has been demonstrated that networks of spiking neurons are computationally more powerful than equivalent networks of static neurons [23].

To date, a significant amount of work has been done in this area, with both software [24-28] and hardware [29-32] based solutions being presented. A software approach to the problem is the easier of the two to implement, but a hardware solution offers a greater degree of realism and potential benefits in terms of speed and parallelism. The computational resources required by software approaches increase significantly as the size of the simulated network grows. This requires either a reduction in operating speed or additional processing power. With hardware implementations, increasing the size of the network increases the complexity of the design process, but the speed of operation need not be affected. The serial nature of modern computer processors means that it is not possible to simulate a truly parallel, real-

time network using a software based approach, whereas parallelism and real-time operating can be engineered into a hardware base approach.

When considering third generation neural hardware, a number of significant problem areas present themselves. The vast scale and massive connectivity of the brain dictates that any neural circuitry created should occupy the minimum possible circuit area and consume the minimum possible power. Secondly, successful neural hardware must be able to correctly emulate a range of neural processes: spatial and temporal summation, thresholding, plasticity, refractory periods and learning; all of which are thought to play a role in the computational power of the brain. A third problem area is that of interconnection. Conventional metal interconnect techniques are unsuited for implementing such a high degree of connectivity. Novel methods are required and several possible solutions have been put forward, including 3-D [33], optical [34] and RF [35] interconnects; Address Event Representation [7, 31], Multiple Valued Logic [36], Network on Chip [37-39] and Pulsed Wave Interconnect [40] systems.

1.3 Current status of neural networks in VLSI

Existing VLSI techniques provide a range of resources for implementing neural systems. The field of neuromorphic engineering aims to replicate the functionality of biological systems using custom silicon chips [32, 41-56]. Analog, digital and mixed-signal implementations of neuromorphic hardware are possible, all of which are constrained by issues such as adaptability, flexibility, scalability and maximisation of speed relative to conventional sequential processors. Neural computing primitives (neurons/synapses) can be realised in a smaller footprint using analog VLSI than in digital. However, efficient communication between individual cells is more effectively handled through a digital implementation. Digital techniques also offer high computational precision, high reliability, and high programmability. While analogue and digital techniques each offer particular advantages, they also have their own drawbacks. Digital implementations can be computationally slower than equivalent analog circuits, requiring larger amounts of silicon and consuming more power. Analog technologies are sensitive to noise and susceptible to interference and process variations. Hybrids of analogue and digital techniques for the implementation of neural networks have shown some potential [46, 48, 51, 57, 58]. Using a hybrid approach to VLSI

neural circuits enables neuromorphic engineers to build dense integrated networks of silicon neurons that run in real time, while capturing the computational power and efficiency of biological neural systems.

Silicon models of specific areas of the nervous system have been built which can perform massively parallel signal processing. Such implementations include retina chips [49, 59, 60], silicon cochlear [61-63], auditory midbrain [64], motion sensing [65-68], and olfaction chips [69]. The silicon retina by Mahowald and Mead [49] detects the contours of a moving stimulus, generating only analogue output. The silicon retina by Zaghloul and Boahen [55, 60], contains a 60×96 array of phototransistors and processing circuits, which is able to generate spiking outputs that mimic the responses of ON-sustained and OFF-sustained retinal ganglion cells. More recently, Koickal et al. [69] presented an analogue VLSI implementation of an adaptive neuromorphic olfaction chip with on-chip chemosensor array and sensor interface. An on-chip spike time dependent learning circuit is used to dynamically adapt weights for odour detection and classification. Rasche [52] described an adjustable and excitable network of spiking units, designed to fit into a multichip neuromorphic system which can perform different visual tasks, such as contour detection, contour propagation, image segmentation and motion detection [70].

To implement higher levels of processing and cognition, a number of multichip approaches and communication protocols between chips have been reported [51, 71-75]. These neuromorphic systems employ a similar design strategy as their biological counterparts - local computations are performed in analogue and the results are communicated by using all-or-none binary events (spikes). A common communication protocol for neuromorphic chips is the address–event representation (AER) system [7, 31, 76-81], which uses time-multiplexing to emulate extensive connectivity between neurons. An address encoder generates a unique binary address for each neuron when it spikes. A digital bus transmits these addresses to the receiving chip where an address decoder selects the corresponding location. The protocol is asynchronous, with the time that the address appears on the bus encoding the spike time directly. Choi et. al. [71] proposed a neuromorphic multichip implementation of orientation hypercolumns in the mammalian primary visual cortex, which consists of a single silicon retina feeding multiple orientation selective image filtering chips [72]. Each chip contains a 2-D array of neurons tuned to the same orientation and spatial frequency, but different retinal locations. All chips operate in continuous time and

communicate with each other using spike-encoded inputs and outputs which are transmitted by the digital asynchronous AER protocol.

A number of large scale projects are being undertaken which aim to simulate the operation of large number of neural cells in as realistic a manner as possible, although each approach does have its particular limitations. The Blue Brain project [82] utilises Blue Gene supercomputers to simulate, down to a molecular level, the neocortical column of a rat. While this is a useful project, in that it aims to further understanding of the architecture and functionality of the brain, it is more focused on the biological side, rather than on potential applications. The SpiNNaker system aims to simulate up to a billion neurons in real time [83-86]. Many thousands of independent multi-core processers can be connected together in a highly parallel, fault-tolerant architecture, to implement a range of neural models. One disadvantage of this approach is that while it is possible to accurately model neurons and synapses when working with a small network, as the size of the system is scaled up, the level of biological plausibility has to be compromised. The FACETS project aims to investigate novel computational paradigms through wafer scale integration of large numbers of high speed analog processing elements [87-91]. The aim is to produce arrays of chips, each containing 384 neurons and 100,000 synapses, on a single silicon wafer, with biological experiments and computer modelling supporting the hardware effort. As with the Blue Brain project, one of the main aims of FACETS is to gain an insight into the computational principles of the brain, rather than to create neural systems that can be used to solve particular problems. However, this does not preclude the possibility of such networks emerging as a by-product of the research process. In addition, each of the above implementations have large overheads in terms of physical space required, cost and power consumption.

1.3.1 Neurons

A number of different approaches to modelling neurons in hardware have been attempted over the years, several of which are shown in Figure 1.6. The circuit diagrams shown are taken from the literature. One approach has been to build highly accurate circuits based upon mathematically described formalisations of neural activity, such as the Hodgkin-Huxley or FitzHugh-Nagumo equations [92-99]. However, the circuit area needed to achieve such a

high level of biological plausibility makes any such implementation unsuitable for a VLSI system.

The majority of work in the area uses the phenomenological approach, whereby circuits are constructed which implement the computationally important features of biological neurons (spiking, plasticity, leakage etc), as efficiently as possible. An axon-hillock circuit proposed by Mead consists of an integrating capacitor connected to two inverters, a feedback capacitor, and a reset transistor driven by the output inverter [100]. An output spike is generated when the voltage across the integrating capacitor reaches the switching threshold of the first inverter. The approach shown in [101] (Figure 1.6a) comprises a neuron circuit made up of 19 MOSFETs which is integrated into a learning system capable of implementing back propagation. Han proposed the circuit in [102] (Figure 1.6i), where an electrically programmable conductance is used to implement the neuron cell. A simplistic integrate and fire neuron, compatible with the AER communication system is presented in [7] (Figure 1.6d). This requires a smaller number of MOSFETs than in previously described work. Excitatory and inhibitory synaptic activity is achieved in [103] (Figure 1.6b), where a spiking neuron cell is described which consumes only 5 transistors and 2 capacitors. Wijekoon et al presented an oscillator circuit in [104] (Figure 1.6f) which, by implementing the Izhikevich equations [105], can simulate the shapes and patterns of a range of cortical cells, requiring 14 MOSFETs and two capacitors. Neural oscillators are also discussed in [106] (Figure 1.6j) and [107] (Figure 1.6e), where several models of neural behaviour are considered.

Several circuits have been reported which concentrate on minimising the power consumption of the neural circuitry. Aunet et al [108] present a three input subthreshold CMOS preceptron circuit using 6 MOSFETs (Figure 1.6c), which has a threshold that can be adjusted in real-time. A floating gate structure is used in [109] (Figure 1.6h) to implement spiking at a rate comparable to biological neurons, with currents as low as 2pA.

A range of more complex integrate and fire neurons circuits have been presented which implement a greater number of useful features - variable threshold voltages, variable refractory periods and output pulse durations; spike frequency adaptation, temporal summation and controllable leakage paths [29-31, 110-112]. (Figure 1.6g, k and l) It is this variety of circuit which shows the most potential for VLSI neural hardware.

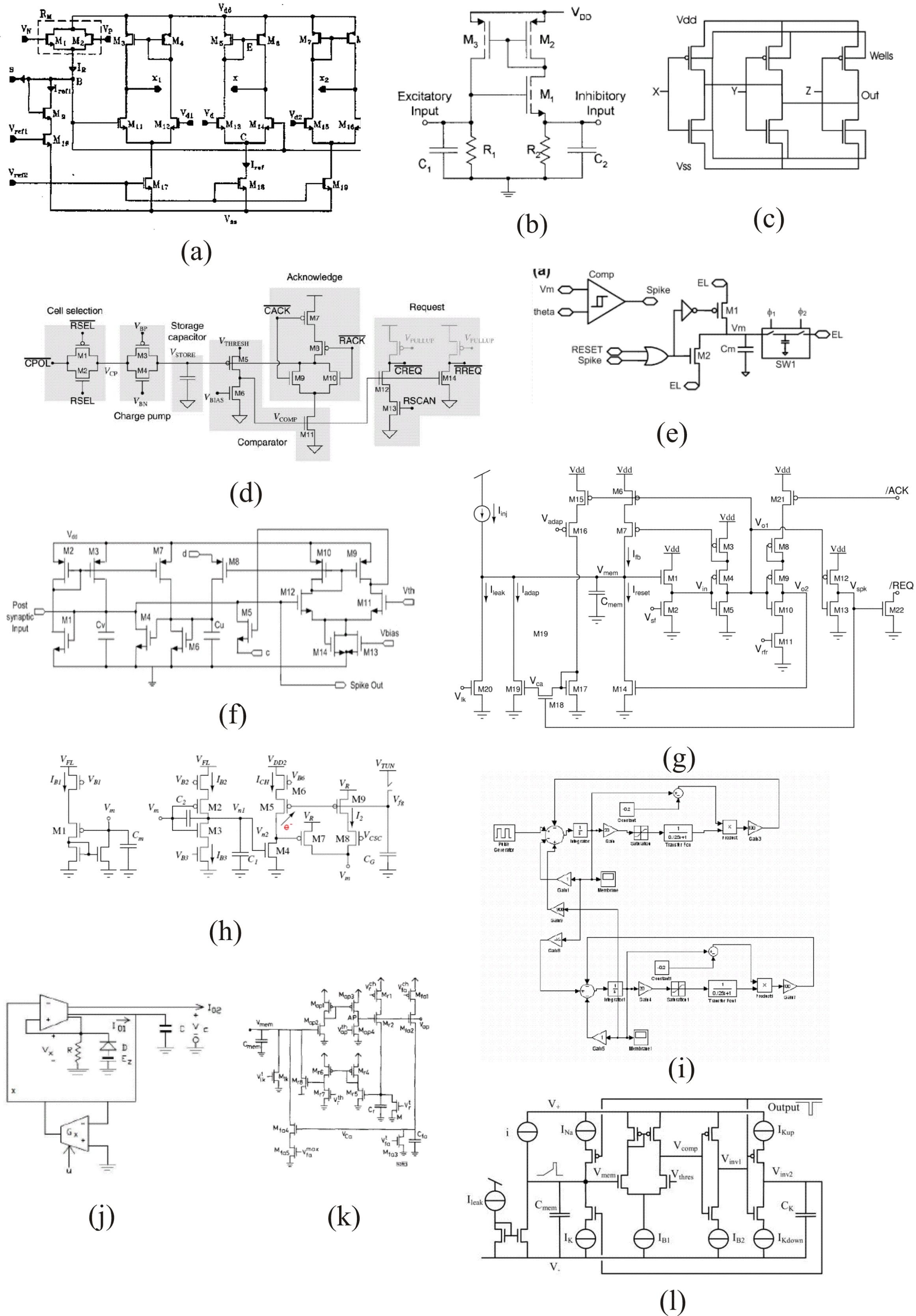

Figure 1.6 - Neuron circuits. (a) [101], (b) [103], (c) [108], (d) [7], (e) [107], (f) [104], (g) [31], (h) [109], (i) [102], (j) [106], (k) [111], (l) [112].

Each of the circuits listed above has some limitations in terms of their ability to create biological plausible neural cells. Some require large amounts of silicon area (g, h, i, k) or bulky control circuitry to operate (a, e,). While not necessarily prohibitively large, further reductions in circuit size are required for progress towards biological scale networks to continue. Other crcuits either fail to implement certain important biological functions (c, d, f, h, i, j, l) or are not able to fully interface with spiking neural systems without some modifications (b, g, k, l).

1.3.2 Synapse circuits

A range of circuits have been presented which attempt to replicate the function of synapses in hardware. These include CMOS circuits [29-31, 80, 110, 113-116], floating gate devices [32, 117-119] and nanoscale devices [120]. The most common approach is to utilise CMOS technology to implement the dynamics and properties associated with a synapse. The complexity and capability of the circuits varies, with each able to implement different synaptic features, including facilitation and depression [29, 113]; plasticity, on both short [121, 122] and long time scales [31, 80]; learning and adaptation. Indiveri et al proposed a highly functional synapse in [31], requiring thirty three transistors and three capacitors. Short term depression of the synaptic weight is achieved by using an adjustable local gain control mechanism, which reduces the weight voltage following each presynaptic input. An STDP circuit compares the relative timings of the pre and postsynaptic inputs and adjusts the weight voltage accordingly. The problem of long-term weight storage is solved by the use of a bistability circuit which drives the synaptic weight to one of two fixed values.

The use of CMOS technology has several advantages in that it is well understood, readily accessible and there are a range of design tools available to aid in development of circuits. The downside is that existing circuits require relatively large numbers of transistors, with a number requiring area consuming capacitors for weight storage. Of the circuits discussed, the smallest [29], has dimensions of approximately 8µm x 6µm in a 0.35um CMOS process.

Another approach has been to utilise floating gate devices to act as artificial synapses. A single floating gate transistor can be engineered to store a non-volatile synaptic weight value,

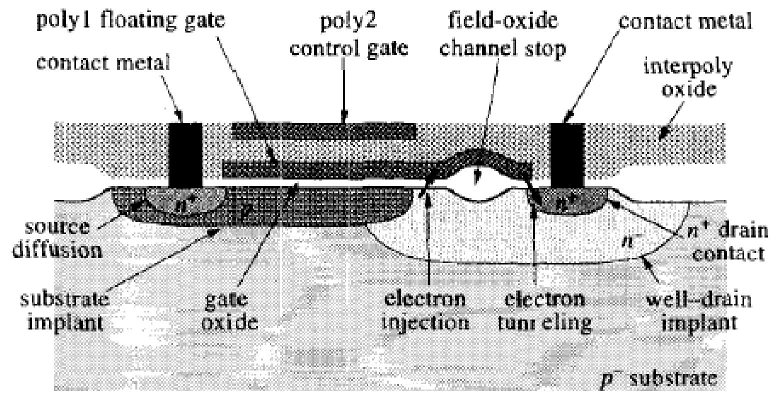

Figure 1.7 - Cross section of floating gate pMOS synapse [119]. Charge is added through electron tunnelling and removed through hot-electron injection.

with bidirectional memory updates achievable through hot electron injection and Fowler-Nordheim tunnelling [119].

The device, shown in Figure 1.7, is examined in more detail in [123]. The weight update rule for a floating gate pFET synapse is considered, where correlation between weight and drain voltages dictates the equilibrium weight voltage.

Unsupervised learning in an array of floating gate MOS synapses has been demonstrated, with weight update rules derived from MOS physics [124]. The increase in packing density afforded by the use of floating gate circuits is substantial. Fewer components are required than in any of the CMOS based approaches previously mentioned. However, this advantage is offset by the need for control circuitry to regulate the processes of adding or removing charge from the floating gate, which consumes a large amount of area. Another limiting factor is the rate at which weight updates can occur. Current technology allows for an array of such synapses to be programmed in under 3 seconds [125]. While the speed of operation is improving, it is still far from the millisecond rate of operation of biological systems. In general, the existing synaptic circuitry either consumes too much area, given that synapses will be far more numerous than neurons in any network implementation, or do not fully implement the required range of features seen in biological synapses.

1.3.3 Axonal Delay

In the field of neuromorphics, the operation of the brain is often simplified so that only the

interactions of synapses and neurons are deemed necessary for successful implementations. However, it has been demonstrated that the delay introduced due to the propagation of signals along an axon plays a role in the computation performed by the brain. One area where this is evident is in the localisation of sound [126]. A number of models have been proposed to explain this phenomenon [127-130]. One conclusion reached is that for computational models of auditory processes to be accurate, a range of axonal delays are required.

Several circuits for the implementation of an axonal delay have been proposed, some circuit diagrams taken from the literature are shown in Figure 1.8. A variable delay set by the rate of current flow through a MOSFET based circuit is presented in [131] (Figure 1.8b), where the length of the delay is influenced by the strength of the synaptic connection. It was shown in [132] (Figure 1.8c) that a programmable dendrite delay can be implemented using a 4-bit bit control bus to ensure the coincidence of spikes during learning. However, there is a requirement for memory capability and additional circuitry to set the logic levels on the bus line. The circuit shown in [133] (Figure 1.8c) ensures coincidence of signals by coupling neural type cells together with a MOSFET based synchronising circuit which injects biasing currents into the neural cells to introduce a delay where necessary. An alternative approach that implements a fixed delay was presented in [134] (Figure 1.8d), whereby a current-mode technique was employed which requires seven MOSFETs. In this approach the associated Miller effect is harnessed to set the delay time constant and multiple time constants can be realised by cascading the current-mode circuits. An interesting approach to inter-neuron signal delay in hardware was presented in [103] (Figure 1.8a) where a series of neuron cells are connected via coupling resistors to form an axon, although the delay introduced is permanently fixed by the choice of resistors. The output from one cell activates the next cell in the chain, and so on, giving rise to a constant propagation velocity where directionality is ensured by the refractory period of the activated cell. Work has also focused on the development of a neuromorphic bidirectional delay line [135] (Figure 1.8f) that uses cascaded CMOS based delay circuits with appropriate coupling conductances. By adjusting the conductances, and other parameters of the delay line, a pulse propagation characteristic similar to what is observed in axon and dendrite trees is possible. As with the neuron and synapse devices, one of the key issues here is that of the area consumed by the devices. In particular, circuits b, c, d and f shown in Figure 1.8 consume large amounts of area and are impractical for use in large scale networks.

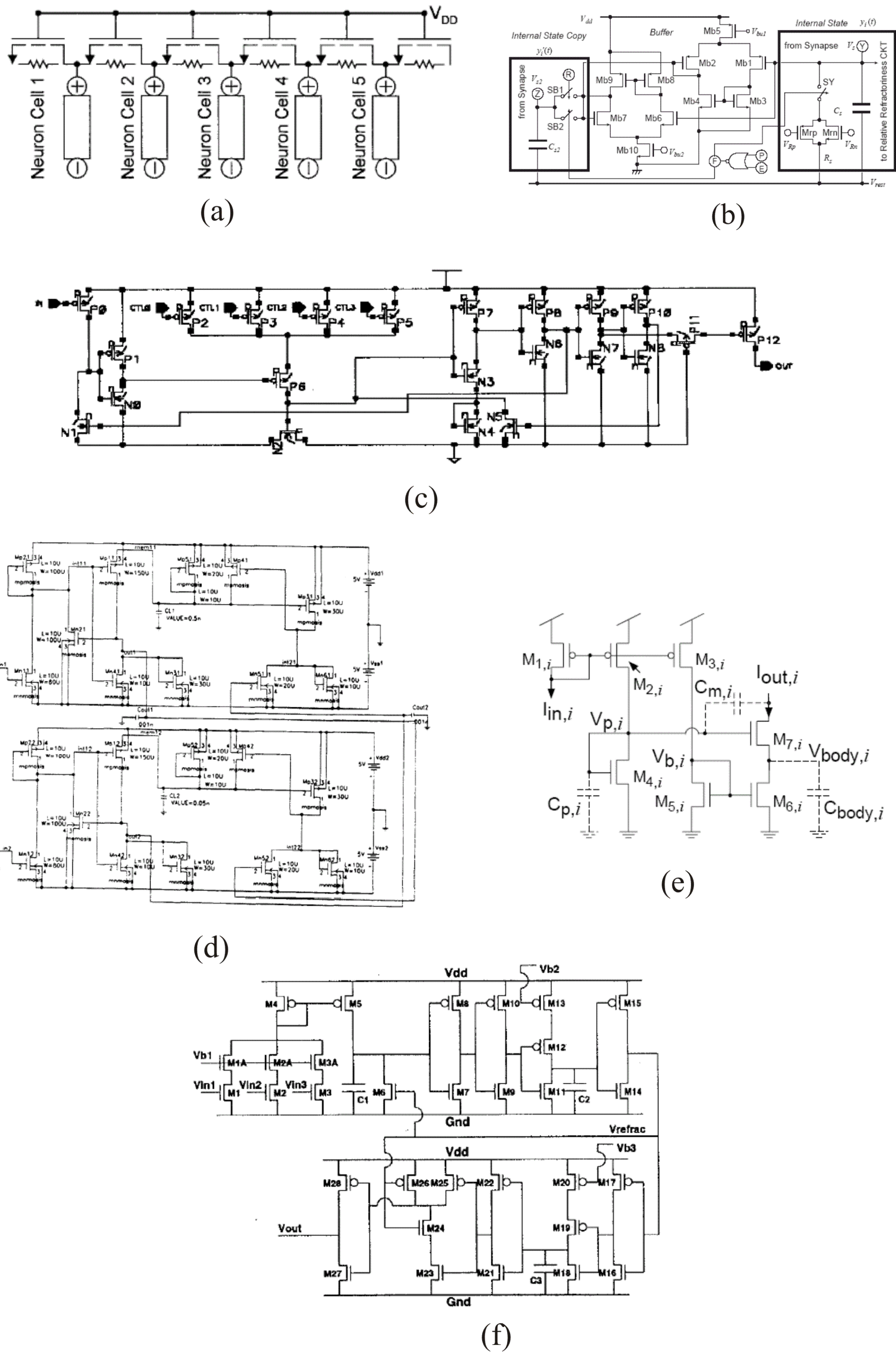

Figure 1.8 - Axonal delay circuits. (a) [103], (b) [131], (c) [132] (d) [133] (e) [134], (f) [135].

1.4 Organisation of the Thesis

Successful, biological scale, neuromorphic hardware must be able to emulate the computationally important processes present in the brain, with a high level of interconnection between elements. In order to effectively implement this, low power, small geometry circuit blocks are essential.

The rest of the thesis is organised as follows. An overview of directly relevant semiconductor physics is given in Chapter 2, where physical parameters of the MOS devices used for this project are extracted. A biologically plausible synapse cell, capable of implementing plasticity, refraction and depression is proposed in Chapter 3. Simulation and experimental results demonstrate the operation of the device, which requires less area than any of the devices previously discussed. A low power compact neuron circuit which incorporates the synapse cell, generating realistic PSPs with an adjustable decay period is presented in Chapter 4. Chapter 5 describes a circuit for the implementation of a variable axonal delay. The possibilities and challenges for VLSI implementations of the neural circuitry are discussed in Chapter 6. A summary of the thesis and proposals for future work are given in Chapter 7.

References

[1] ITRS, "The International Technology Roadmap for Semiconductors," 2009.

[2] D. Purves, Neuroscience, 2 ed.: Sinauer, 2001.

[3] K. Mehrotra, Elements of Artificial Neural Networks: MIT Press, 1997.

[4] "http://www.epsrc.ac.uk/funding/grants/rb/signpost/Pages/ict.aspx."

[5] "http://www.darpa.mil/dso/thrusts/bio/biologically/synapse/."

[6] "ftp://ftp.cordis.europa.eu/pub/fp7/ict/docs/fet-proactive/usef-13_en.pdf."

[7] D. H. Goldberg, G. Cauwenberghs, and A. G. Andreou, "Probabilistic synaptic weighting in a reconfigurable network of VLSI integrate-and-fire neurons," Neural Networks, vol. 14, pp. 781-793, 2001.

[8] W. E. G. Gerald M. Edelman, W. Maxwell Cowan, Synaptic Function: Wiley-Interscience, 1987.

[9] W. B. Levy and O. Steward, "Temporal contiguity requirements for long-term associative potentiation/depression in the hippocampus," Neuroscience, vol. 8, pp. 791-797, 1983.

[10] H. Markram, J. Labke, M. Frotscher, and B. Sakmann, "Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs," Science, vol. 275, pp. 213-215, 1997.

[11] L. F. Abbott and S. B. Nelson, "Synaptic plasticity: Taming the beast," Nature Neuroscience, vol. 3, pp. 1178-1183, 2000.

[12] K. S. Narendra and K. Parthasarathy, "Identification and control of dynamical systems using neural networks," IEEE Transactions on Neural Networks, vol. 1, pp. 4-27, 1990.

[13] F. L. Lewis, A. Yeildirek, and K. Liu, "Multilayer neural-net robot controller with guaranteed tracking performance," IEEE Transactions on Neural Networks, vol. 7, pp. 388-399, 1996.

[14] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," Nature Medicine, vol. 7, pp. 673-679, 2001.

[15] G. A. Carpenter and S. Grossberg, "ART OF ADAPTIVE PATTERN RECOGNITION BY A SELF-ORGANIZING NEURAL NETWORK," Computer, vol. 21, pp. 77-88, 1988.

[16] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, pp. 4-37, 2000.

[17] M. Egmont-Petersen, D. De Ridder, and H. Handels, "Image processing with neural networks- A review," Pattern Recognition, vol. 35, pp. 2279-2301, 2002.

[18] G. Zhang, B. Eddy Patuwo, and M. Y. Hu, "Forecasting with artificial neural networks: The state of the art," International Journal of Forecasting, vol. 14, pp. 35-62, 1998.

[19] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, pp. 23-38, 1998.

[20] J. A. Benediktsson, P. H. Swain, and O. K. Ersoy, "Neural network approaches versus statistical methods in classification of multisource remote sensing data," IEEE

Transactions on Geoscience and Remote Sensing, vol. 28, pp. 540-552, 10 July 1989 through 14 July 1989 1990.

[21] J. V. Tu, "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes," Journal of Clinical Epidemiology, vol. 49, pp. 1225-1231, 1996.

[22] W. Maass and C. M. Bishop, Pulsed Neural Networks. Cambridge, MA: MIT Press, 1999.

[23] W. Maass, "Networks of spiking neurons: The third generation of neural network models," Neural Networks, vol. 10, pp. 1659-1671, 1997.

[24] C. Grassmann and J. K. Anlauf, "Fast digital simulation of spiking neural networks and neuromorphic integration with SPIKELAB," International journal of neural systems, vol. 9, pp. 473-478, 1999.

[25] R. Brette, M. Rudolph, T. Carnevale, M. Hines, D. Beeman, J. M. Bower, M. Diesmann, A. Morrison, P. H. Goodman, F. C. Harris Jr, M. Zirpe, T. NatschlÃ¤ger, D. Pecevski, B. Ermentrout, M. Djurfeldt, A. Lansner, O. Rochel, T. Vieville, E. Muller, A. P. Davison, S. El Boustani, and A. Destexhe, "Simulation of networks of spiking neurons: A review of tools and strategies," Journal of Computational Neuroscience, vol. 23, pp. 349-398, 2007.

[26] M. Migliore, C. Cannia, W. W. Lytton, H. Markram, and M. L. Hines, "Parallel network simulations with NEURON," Journal of Computational Neuroscience, vol. 21, pp. 119-129, 2006.

[27] A. Delorme, J. Gautrais, R. van Rullen, and S. Thorpe, "SpikeNET: A simulator for modeling large networks of integrate and fire neurons," Neurocomputing, vol. 26-27, pp. 989-996, 1999.

[28] M. Schaefer, T. SchÅ“nauer, C. Wolff, G. Hartmann, H. Klar, and U. RÃ¼ckert, "Simulation of spiking neural networks - Architectures and implementations," Neurocomputing, vol. 48, pp. 647-679, 2002.

[29] L. Shih-Chii and R. Douglas, "Temporal coding in a silicon network of integrate-and-fire neurons," Neural Networks, IEEE Transactions on, vol. 15, pp. 1305-1314, 2004.

[30] S.-C. Liu, J. Kramer, G. Indiveri, T. Delbruck, T. Burg, and R. Douglas, "Orientation-selective aVLSI spiking neurons," Neural Networks, vol. 14, pp. 629-643, 2001.

[31] G. Indiveri, E. Chicca, and R. Douglas, "A VLSI array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity," Neural Networks, IEEE Transactions on, vol. 17, pp. 211-221, 2006.

[32] C. Diorio, D. Hsu, and M. Figueroa, "Adaptive CMOS: from biological inspiration to systems-on-a-chip," Proceedings of the IEEE, vol. 90, pp. 345-357, 2002.

[33] J. W. Joyner, P. Zarkesh-Ha, and J. D. Meindl, "Global interconnect design in a three-dimensional system-on-a-chip," Very Large Scale Integration (VLSI) Systems, IEEE Transactions on, vol. 12, pp. 367-372, 2004.

[34] M. S. Bakir, T. K. Gaylord, O. O. Ogunsola, E. N. Glytsis, and J. D. Meindl, "Optical transmission of polymer pillars for chip I/O optical interconnections," Photonics Technology Letters, IEEE, vol. 16, pp. 117-119, 2004.

[35] B. A. Floyd, H. Chih-Ming, and K. K. O, "Intra-chip wireless interconnect for clock distribution implemented with integrated antennas, receivers, and transmitters," Solid-State Circuits, IEEE Journal of, vol. 37, pp. 543-552, 2002.

[36] S. L. Hurst, "MULTIPLE-VALUED LOGIC - ITS STATUS AND ITS FUTURE," IEEE Transactions on Computers, vol. C-33, pp. 1160-1179, 1984.

[37] L. Benini and G. De Micheli, "Networks on chips: A new SoC paradigm," Computer, vol. 35, pp. 70-78, 2002.

[38] W. J. Dally and B. Towles, "Route packets, not wires: On-chip interconnection networks," in Proceedings - Design Automation Conference, Las Vegas, NV, 2001, pp. 684-689.

[39] F. M. Jim Harkin, Liam McDaid, Steve Hall, Brian McGinley, Seamus Cawley, "A Reconfigurable and Biologically Inspired Paradigm for Computation Using Network-On-Chip and Spiking Neural Networks," International Journal of Reconfigurable Computing, 2009.

[40] W. Pingshan, G. Pei, and E. C. C. Kan, "Pulsed wave interconnect," Very Large Scale Integration (VLSI) Systems, IEEE Transactions on, vol. 12, pp. 453-463, 2004.

[41] A. Bofill-i-Petit and A. F. Murray, "Synchrony detection and amplification by silicon neurons with STDP synapses," IEEE Transactions on Neural Networks, vol. 15, pp. 1296-1304, 2004.

[42] 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, pp. 1297-1307, 2003.

[43] D. D. Coon and A. G. U. Perera, "Integrate-and-fire coding and Hodgkin-Huxley circuits employing silicon diodes," Neural Networks, vol. 2, pp. 143-151, 1989.

[44] C. Diorio and R. P. N. Rao, "Neural circuits in silicon," Nature, vol. 405, pp. 891-890, 2000.

[45] R. Douglas, M. Mahowald, and C. Mead, "Neuromorphic analogue VLSI," Annual Review of Neuroscience, vol. 18, pp. 255-281, 1995.

[46] R. H. R. Hahnioser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and H. S. Seung, "Digital selection and analogue amplification coexist in a cortex- inspired silicon circuit," Nature, vol. 405, pp. 947-951, 2000.

[47] G. Indiveri, "A neuromorphic VLSI device for implementing 2-D selective attention systems," IEEE Transactions on Neural Networks, vol. 12, pp. 1455-1463, 2001.

[48] S. C. Liu and R. Douglas, "Temporal coding in a silicon network of integrate-and-fire neurons," IEEE Transactions on Neural Networks, vol. 15, pp. 1305-1314, 2004.

[49] M. A. Mahowald and C. Mead, "The silicon retina," Scientific American, vol. 264, pp. 76-82, 1991.

[50] C. Mead, "Neuromorphic electronic systems," Proceedings of the IEEE, vol. 78, pp. 1629-1636, 1990.

[51] P. A. Merolla, J. V. Arthur, B. E. Shi, and K. A. Boahen, "Expandable networks for neuromorphic chips," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 54, pp. 301-311, 2007.

[52] C. Rasche, "Neuromorphic excitable maps for visual processing," IEEE Transactions on Neural Networks, vol. 18, pp. 520-529, 2007.

[53] J. Schemmel, S. Hohmann, K. Meier, and F. SchÃ¼rmann, "A mixed-mode analog neural network using current-steering synapses," Analog Integrated Circuits and Signal Processing, vol. 38, pp. 233-244, 2004.

[54] C. Song and K. P. Roenker, "Novel heterostructure device for electronic pulse-mode neural circuits," IEEE Transactions on Neural Networks, vol. 5, pp. 663-665, 1994.

[55] K. A. Zaghloul and K. Boahen, "Optic Nerve Signals in a Neuromorphic Chip I: Outer and Inner Retina Models," IEEE Transactions on Biomedical Engineering, vol. 51, pp. 657-666, 2004.

[56] M. Mokhtar, D. M. Halliday, and A. M. Tyrrell, "Hippocampus-inspired spiking neural network on FPGA," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). vol. 5216 LNCS Prague, 2008, pp. 362-371.

[57] M. R. DeYong, R. L. Findley, and C. Fields, "The design, fabrication, and test of a new VLSI hybrid analog-digital neural processing element," IEEE Transactions on Neural Networks, vol. 3, pp. 363-374, 1992.

[58] Y. Horio, K. Aihara, and O. Yamamoto, "Neuron-Synapse IC Chip-Set for Large-Scale Chaotic Neural Networks," IEEE Transactions on Neural Networks, vol. 14, pp. 1393-1404, 2003.

[59] T. Delbrueck, "Silicon retina with correlation-based, velocity-tuned pixels," IEEE Transactions on Neural Networks, vol. 4, pp. 529-541, 1993.

[60] K. A. Zaghloul and K. Boahen, "Optic nerve signals in a neuromorphic chip II: testing and results," Biomedical Engineering, IEEE Transactions on, vol. 51, pp. 667-675, 2004.

[61] E. FragniÃ¨re, A. Van Schaik, and E. A. Vittoz, "Design of an Analogue VLSI Model of an Active Cochlea," Analog Integrated Circuits and Signal Processing, vol. 13, pp. 19-35, 1997.

[62] R. Sarpeshkar, R. F. Lyon, and C. Mead, "A Low-Power Wide-Dynamic-Range Analog VLSI Cochlea," Analog Integrated Circuits and Signal Processing, vol. 16, pp. 245-274, 1998.

[63] B. Wen and K. Boahen, "A silicon cochlea with active coupling," IEEE Transactions on Biomedical Circuits and Systems, vol. 3, pp. 444-455, 2009.

[64] T. Horiuchi and K. M. Hynna, "A VLSI-based model of azimuthal echolocation in the big brown bat," Autonomous Robots, vol. 11, pp. 241-247, 2001.

[65] R. R. Harrison, "A biologically inspired analog IC for visual collision detection," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 52, pp. 2308-2318, 2005.

[66] R. R. Harrison and C. Koch, "Robust analog VLSI motion sensor based on the visual system of the fly," Autonomous Robots, vol. 7, pp. 211-224, 1999.

[67] T. K. Horiuchi and C. Koch, "Analog VLSI-based modeling of the primate oculomotor system," Neural Computation, vol. 11, pp. 243-265, 1999.

[68] R. Sarpeshkar, J. Kramer, G. Indiveri, and C. Koch, "Analog VLSI architectures for motion processing: From fundamental limits to system applications," Proceedings of the IEEE, vol. 84, pp. 969-987, 1996.

[69] T. J. Koickal, A. Hamilton, S. L. Tan, J. A. Covington, J. W. Gardner, and T. C. Pearce, "Analog VLSI circuit implementation of an adaptive neuromorphic olfaction

chip," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 54, pp. 60-73, 2007.

[70] C. Rasche, The Making of a Neuromorphic Visual System: New York: Springer-Verlag, 2005.

[71] T. Y. W. Choi, P. A. Merolla, J. V. Arthur, K. A. Boahen, and B. E. Shi, "Neuromorphic implementation of orientation hypercolumns," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 52, pp. 1049-1060, 2005.

[72] T. Y. W. Choi, B. E. Shi, and K. A. Boahen, "An ON-OFF orientation selective address event representation image transceiver chip," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 51, pp. 342-353, 2004.

[73] G. Indiveri, R. MÃ¼rer, and J. Kramer, "Active vision using an analog VLSI model of selective attention," IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing, vol. 48, pp. 492-500, 2001.

[74] R. J. Vogelstein, U. Mallik, E. Culurciello, G. Cauwenberghs, and R. Etienne-Cummings, "A multichip neuromorphic system for spike-based visual information processing," Neural Computation, vol. 19, pp. 2281-2300, 2007.

[75] R. J. Vogelstein, U. Mallik, J. T. Vogelstein, and G. Cauwenberghs, "Dynamically reconfigurable silicon array of spiking neurons with conductance-based synapses," IEEE Transactions on Neural Networks, vol. 18, pp. 253-265, 2007.

[76] K. A. Boahen, "Point-to-point connectivity between neuromorphic chips using address events," IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing, vol. 47, pp. 416-434, 2000.

[77] J. Fieres, J. Schemmel, and K. Meier, "Realizing biological spiking network models in a configurable wafer-scale hardware system," in Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on, 2008, pp. 969-976.

[78] A. Jimenez-Fernandez, A. Linares-Barranco, R. Paz-Vicente, G. Jimenez-Moreno, and R. Berner, "Spike-based control monitoring and analysis with Address Event Representation," in 2009 IEEE/ACS International Conference on Computer Systems and Applications, AICCSA 2009, 2009, pp. 900-906.

[79] J. Lazzaro, J. Wawrzynek, M. Mahowald, M. Sivilotti, and D. Gillespie, "Silicon auditory processors as computer peripherals," IEEE Transactions on Neural Networks, vol. 4, pp. 523-528, 1993.

[80] S. Mitra, S. Fusi, and G. Indiveri, "A VLSI spike-driven dynamic synapse which learns only when necessary," 2006, p. 4 pp.

[81] R. J. Vogelstein, U. Mallik, and G. Cauwenberghs, "Silicon spike-based synaptic array and address-event transceiver," 2004, pp. V-385-V-388 Vol.5.

[82] H. Markram, "The Blue Brain Project," Nature Reviews Neuroscience, vol. 7, pp. 153-160, 2006.

[83] Part, X. Jin, A. Rast, F. Galluppi, M. Khan, and S. Furber, "Implementing learning on the SpiNNaker universal neural chip multiprocessor," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). vol. 5863 LNCS, 2009, pp. 425-432.

[84] A. D. Rast, Y. Shufan, M. Khan, and S. B. Furber, "Virtual synaptic interconnect using an asynchronous network-on-chip," in Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on, 2008, pp. 2727-2734.

[85] A. D. Rast, M. M. Khan, X. Jin, L. A. Plana, and S. B. Furber, "A universal abstract-time platform for real-time neural networks," in Proceedings of the International Joint Conference on Neural Networks, 2009, pp. 2611-2618.

[86] A. D. Rast, S. Welbourne, X. Jin, and S. B. Furber, "Optimal connectivity in hardware-targetted MLP networks," in Proceedings of the International Joint Conference on Neural Networks, 2009, pp. 2619-2626.

[87] A. Daouzli, S. SaÃ¯ghi, L. Buhry, Y. Bornat, and S. Renaud, "Weights convergence and spikes correlation in an adaptive neural network implemented on vlsi," in BIOSIGNALS 2008 - Proceedings of the 1st International Conference on Bio-inspired Systems and Signal Processing, Funchal, Madeira, 2008, pp. 286-291.

[88] Q. Zou, Y. Bornat, S. Saghi, J. Tomas, S. Renaud, and A. Destexhe, "Analog-digital simulations of full conductance-based networks of spiking neurons with spike timing dependent plasticity," Network: Computation in Neural Systems, vol. 17, pp. 211-233, 2006.

[89] S. Renaud, J. Tomas, Y. Bornat, A. Daouzli, and S. SaÃ¯ghi, "Neuromimetic ICs with analog cores: An alternative for simulating spiking neural networks," in Proceedings - IEEE International Symposium on Circuits and Systems, New Orleans, LA, 2007, pp. 3355-3358.

[90] J. Tomas, Y. Bornat, S. Saghi, T. Lavi, and S. Renaud, "Design of a modular and mixed neuromimetic ASIC," in Proceedings of the IEEE International Conference on Electronics, Circuits, and Systems, Nice, 2006, pp. 946-949.

[91] S. R. N. Lewis, "Spiking neural networks “in silico”: from single neurons to large scale networks," in Fourth International Multi-Conference on Systems, Signals & Devices Hammamet, Tunisia, 2007.

[92] S. Le Masson, A. Laflaquiere, T. Bal, and G. Le Masson, "Analog circuits for modeling biological neural networks: design and applications," Biomedical Engineering, IEEE Transactions on, vol. 46, pp. 638-645, 1999.

[93] A. W. Przybyszewski, P. S. Linsay, P. Gaudiano, and C. M. Wilson, "Basic Difference Between Brain and Computer: Integration of Asynchronous Processes Implemented as Hardware Model of the Retina," Neural Networks, IEEE Transactions on, vol. 18, pp. 70-85, 2007.

[94] C. Rasche and R. Douglas, "Improved silicon neuron," Analog Integrated Circuits and Signal Processing, vol. 23, pp. 227-236, 2000.

[95] M. F. Simoni, G. S. Cymbalyuk, M. E. Sorensen, R. L. Calabrese, and S. P. DeWeerth, "A Multiconductance Silicon Neuron with Biologically Matched Dynamics," IEEE Transactions on Biomedical Engineering, vol. 51, pp. 342-354, 2004.

[96] J. V. Arthur and K. A. Boahen, "Synchrony in silicon: The gamma rhythm," IEEE Transactions on Neural Networks, vol. 18, pp. 1815-1825, 2007.

[97] J. Shin and C. Koch, "Dynamic range and sensitivity adaptation in a silicon spiking neuron," IEEE Transactions on Neural Networks, vol. 10, pp. 1232-1238, 1999.

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

[99] T. Yu and G. Cauwenberghs, "Analog VLSI neuromorphic network with programmable membrane channel kinetics," in Proceedings - IEEE International Symposium on Circuits and Systems, 2009, pp. 349-352.

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

[101] L. Chun, S. Bingxue, and C. Lu, "Hardware implementation of an expandable on-chip learning neural network with 8-neuron and 64-synapse," 2002, pp. 1451-1454 vol.3.

[102] I. S. Han, "Biologically Inspired Hardware Implementation of Neural Networks with Programmable Conductance," in International Joint Conference on Neural Networks, 2007.

[103] Y. Ota and B. M. Wilamowski, "Analog implementation of pulse-coupled neural networks," Neural Networks, IEEE Transactions on, vol. 10, pp. 539-544, 1999.

[104] J. H. B. Wijekoon and P. Dudek, "A Simple Analogue VLSI Circuit of a Cortical Neuron," in EEE International Conference on Electronics, Circuits and Systems, ICECS 2006.

[105] E. M. Izhikevich, "Simple model of spiking neurons," Neural Networks, IEEE Transactions on, vol. 14, pp. 1569-1572, 2003.

[106] P. V. Tymoshchuk and Y. I. Paterega, "Implementation of artificial neural oscillators," in Perspective Technologies and Methods in MEMS Design, 2009. MEMSTECH 2009. 2009 5th International Conference on, 2009, pp. 149-154.

[107] F. Folowosele, A. Harrison, A. Cassidy, A. G. Andreou, R. Etienne-Cummings, S. Mihalas, E. Niebur, and T. J. Hamilton, "A switched capacitor implementation of the generalized linear integrate-and-fire neuron," in Proceedings - IEEE International Symposium on Circuits and Systems, 2009, pp. 2149-2152.

[108] S. Aunet, B. Oelmann, P. A. Norseng, and Y. Berg, "Real-Time Reconfigurable Subthreshold CMOS Perceptron," Neural Networks, IEEE Transactions on, vol. 19, pp. 645-657, 2008.

[109] Y. L. Wong, X. Peng, and P. Abshire, "Ultra-low Spike Rate Silicon Neuron," in Biomedical Circuits and Systems Conference, 2007. BIOCAS 2007. IEEE, 2007, pp. 95-98.

[110] 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," Neural Networks, IEEE Transactions on, vol. 14, pp. 1297-1307, 2003.

[111] S. R. Schultz and M. A. Jabri, "Analogue VLSI 'integrate-and-fire' neuron with frequency adaptation," Electronics Letters, vol. 31, pp. 1357-1358, 1995.

[112] A. Van Schaik, "Building blocks for electronic spiking neural networks," Neural Networks, vol. 14, pp. 617-628, 2001.

[113] E. Chicca, G. Indiveri, and R. Douglas, "An adaptive silicon synapse," 2003, pp. I-81-I-84 vol.1.

[114] J. B. Lont and W. Guggenbuhl, "Analog CMOS implementation of a multilayer perceptron with nonlinear synapses," Neural Networks, IEEE Transactions on, vol. 3, pp. 457-465, 1992.

[115] E. Lazaridis, E. M. Drakakis, and M. Barahona, "A biomimetic CMOS synapse," 2006, p. 4 pp.

[116] H. C. Card, C. R. Schneider, and W. R. Moore, "Hebbian plasticity in MOS synapses," Radar and Signal Processing, IEE Proceedings F, vol. 138, pp. 13-16, 1991.

[117] C. Gordon, A. Preyer, K. Babalola, R. J. Butera, and P. Hasler, "An artificial synapse for interfacing to biological neurons," 2006, p. 4 pp.

[118] C. Diorio, "A p-Channel MOS synapse transistor with self-convergent memory writes," IEEE Transactions on Electron Devices, vol. 47, pp. 464-472, 2000.

[119] C. Diorio, P. Hasler, A. Minch, and C. A. A. M. C. A. Mead, "A single-transistor silicon synapse," Electron Devices, IEEE Transactions on, vol. 43, pp. 1972-1980, 1996.

[120] A. Afifi, A. Ayatollahi, and F. Raissi, "Implementation of biologically plausible spiking neural network models on the memristor crossbar-based CMOS/nano circuits," in ECCTD 2009 - European Conference on Circuit Theory and Design Conference Program, 2009, pp. 563-566.

[121] C. Rasche and R. H. R. Hahnloser, "Silicon synaptic depression," Biological Cybernetics, vol. 84, pp. 57-62, 2001.

[122] M. Boegerhausen, P. Suter, and S. C. Liu, "Modeling short-term syrtaptic depression in silicon," Neural Computation, vol. 15, pp. 331-348, 2003.

[123] P. Hasler and J. Dugger, "Correlation learning rule in floating-gate pFET synapses," Proceedings - IEEE International Symposium on Circuits and Systems, vol. 5, pp. V-387-V-390, 1999.

[124] C. Diorio, "A floating-gate mos learning array with locally computed weight updates," IEEE Transactions on Electron Devices, vol. 44, pp. 2281-2289, 1997.

[125] G. Serrano, P. D. Smith, H. J. Lo, R. Chawla, T. S. Hall, C. M. Twigg, and P. Hasler, "Automatic rapid programming of large arrays of floating-gate elements," 2004, pp. I-373-I-376 Vol.1.

[126] L. A. Jeffress, "A place theory of sound localization," Journal of Comparative and Physiological Psychology, vol. 41, pp. 35-39, 1948.

[127] M. Konishi, "How the owl tracks its prey," American Scientist, vol. 61, pp. 414-424, 1973.

[128] C. E. Carr and M. Konishi, "A circuit for detection of interaural time differences in the brain stem of the barn owl," Journal of Neuroscience, vol. 10, pp. 3227-3246, 1990.

[129] K. Saberi, Y. Takahashi, H. Farahbod, and M. Konishi, "Neural bases of an auditory illusion and its elimination in owls," Nature Neuroscience, vol. 2, pp. 656-659, 1999.

[130] D. McAlpine and B. Grothe, "Sound localization and delay lines - Do mammals fit the model?," Trends in Neurosciences, vol. 26, pp. 347-350, 2003.

[131] Y. Horio, T. Taniguchi, and K. Aihara, "An asynchronous spiking chaotic neuron integrated circuit," Neurocomputing, vol. 64, pp. 447-472, 2005.

[132] R. H. Fujii, G. Sase, Y. Konishi, and H. Amin, "Spike delay controllable neuron," in Midwest Symposium on Circuits and Systems, 2002, pp. II529-II532.

[133] A. Hodge, M. Zaghloul, and R. W. Newcomb, "Synchronization of neural-type cells," in Proceedings - IEEE International Symposium on Circuits and Systems, 1996, pp. 582-585.

[134] M. Ohtani, H. Yamada, K. Nishio, H. Yonezu, and Y. Furukawa, "Analog LSI implementation of biological direction-selective neurons," Japanese Journal of Applied Physics, Part 1: Regular Papers and Short Notes and Review Papers, vol. 41, pp. 1409-1416, 2002.

[135] W. Yang, "Neuromorphic CMOS circuitry for active bidirectional delay lines," in Proceedings - IEEE International Symposium on Circuits and Systems, 1996, pp. 473-476.

Chapter 2: Overview of MOS Physics and Device Characterisation

2.1 Introduction

In this chapter, the fundamentals of semiconductor physics are covered for the MOS capacitor and MOS transistor, which form the basis of the devices used throughout this thesis. The operation of the MOS capacitor is discussed first and the extraction of key device parameters, oxide thickness, substrate doping level and interface quality is described. Test chips fabricated in a 0.35µm process from Austria Microsystems (AMS) are used for parameter extraction, in accordance with the theoretical analysis presented. This is followed by an analysis of the MOS transistor, where device parameters relating to SPICE models are again extracted. Capacitance voltage (CV) measurements were taken with an HP 4192A Impedance Analyser. Transistor characteristics were measured with an HP4155B Semiconductor Parameter Analyser.

2.2 MOS Capacitor

The MOS capacitor is a fundamental building block of semiconductor devices. It consists of a doped semiconductor substrate, an insulating layer of silicon dioxide and a metal/polysilicon gate terminal. Figure 2.1 shows a cross sectional view of the device. The energy band diagram for an ideal MOS capacitor under zero bias is shown in Figure 2.2.µm

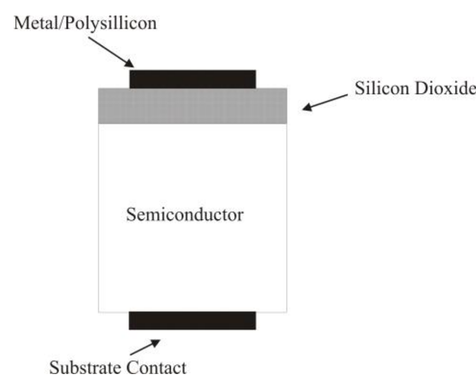

Figure 2.1 - MOS Capacitor

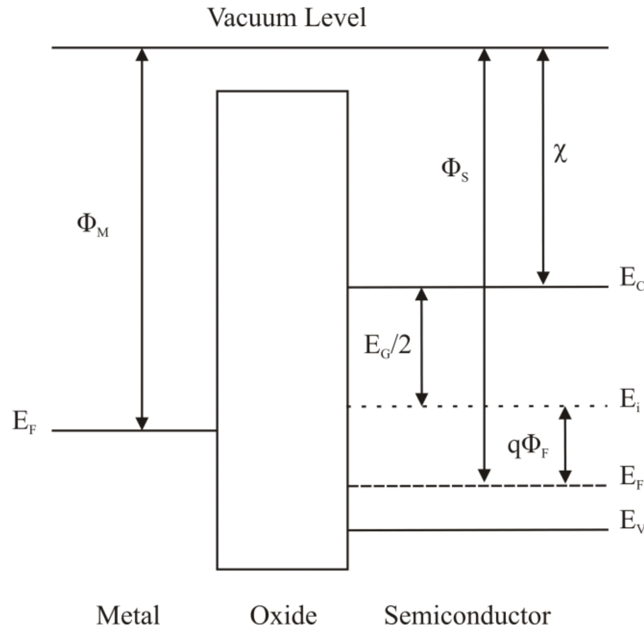

Figure 2.2 - Energy Band Diagram of ideal MOS Capacitor with V_G = 0V. χ is the electron affinity, E_g is the energy gap, φ_b is the Fermi potential/bulk potential.

For an ideal MOS capacitor, the metal work function, Φ_m, is equal to the semiconductor work function, Φ_s and the work function difference, Φ_{ms}, is zero. There are three different modes in which an ideal MOS capacitor can be operated, namely accumulation, depletion and inversion. Figure 2.3 and Figure 2.4 show the charge distribution in the device and the energy band diagrams for each case, assuming a p-type substrate.

A negative gate voltage ($V_G < 0$) causes the device to operate in the accumulation mode. The energy bands near the semiconductor surface bend upwards, as shown in Figure 2.4a. An accumulation layer of holes forms at the interface, shown in Figure 2.3a.The carrier density in the accumulation layer varies exponentially with the energy difference $E_i - E_F$ [1]. The hole density is given by:

$$p = n_i \exp\left[\frac{q(\varphi_s - \varphi_b)}{kT}\right] \tag{2.1}$$

where φ_s is the surface potential at the semiconductor-oxide interface. The upward bending of the energy bands at the surface increases $E_i - E_F$, causing holes from the semiconductor bulk to accumulate at the oxide-semiconductor interface.

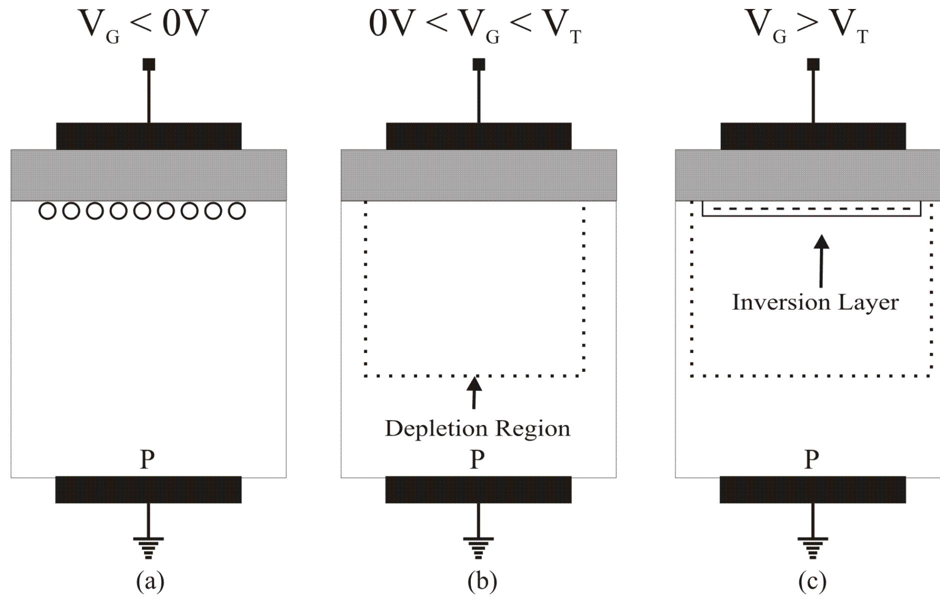

Figure 2.3 – MOS Capacitor in (a) Accumulation (b) Depletion (c) Inversion

A small positive voltage on the gate ($0 < V_G < V_T$) causes the energy bands to bend downwards. Holes in the substrate are repelled from the semiconductor surface and a depletion region is formed, as seen in Figure 2.3b. The total amount of charge in the depletion region is equal to the positive charge on the gate of the capacitor. Assuming the depletion approximation (free charge is negligible), the depletion charge per unit area can be expressed as:

$$Q_d = -qN_A W_d \tag{2.2}$$

where N_A is the substrate acceptor doping concentration and W_d is the width of the depletion region underneath the gate, given by:

$$W_d = \sqrt{\frac{2\epsilon_{si}\epsilon_0\varphi_s}{qN_A}} \tag{2.3}$$

Further increase in gate voltage ($V_G > V_T$) causes even greater band bending; the intrinsic Fermi level at the surface of the semiconductor crosses over the semiconductor Fermi level. The electron concentration is given by:

$$\mathrm{n} = \mathrm{n_i} \exp\left[\frac{q(\varphi_s - \varphi_b)}{kT}\right] \tag{2.4}$$

At this point, shown in Figure 2.4c, $q(\varphi_s - \varphi_b)$ is positive and the number of electrons at the surface is greater than the number of holes. The surface is said to be inverted. This case is illustrated in Figure 2.3c.

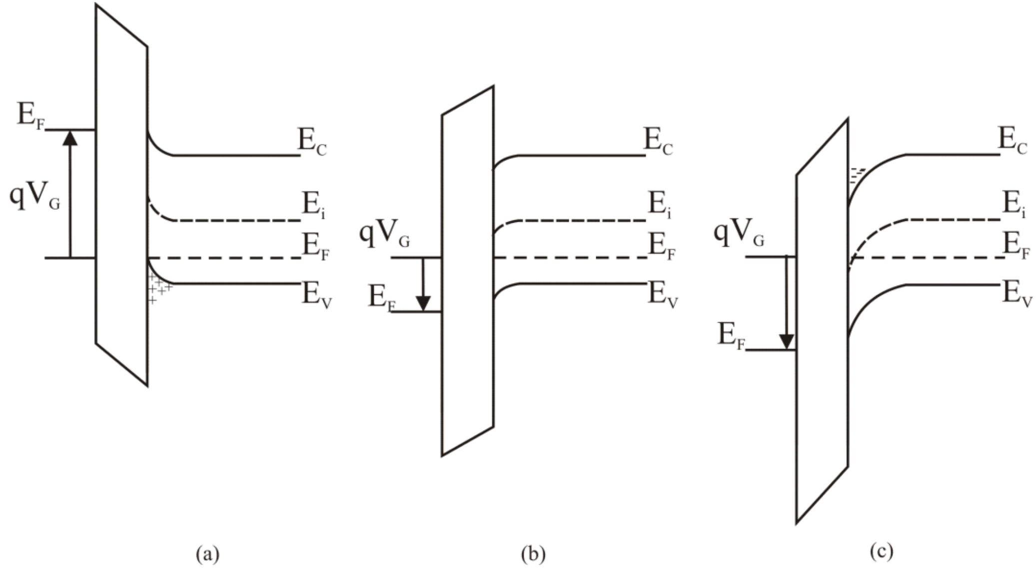

Figure 2.4 - Band Diagrams for MOS capacitor (a) Accumulation (b) Depletion (c) Inversion

The onset of inversion is defined as the point at which the surface potential is equal to twice the bulk potential. Once the inversion case has been reached, the depletion region width reaches a maximum. Further increase in gate voltage is met by increased charge in the inversion layer. The maximum depletion width is generally taken as:

$$W_m = \sqrt{\frac{4\epsilon_{si}\epsilon_0\varphi_b}{qN_A}} \tag{2.5}$$

where the bulk potential, φ_b, is given by:

$$\varphi_b = \frac{kT}{q}\ln\left(\frac{N_A}{n_i}\right) \tag{2.6}$$

In the absence of work function differences between the metal and the semiconductor, the applied gate voltage will fall partially across the oxide and partially across the semiconductor:

$$\mathrm{V_G} = \mathrm{V_o} + \varphi_s \tag{2.7}$$

V_o is the voltage across the oxide, which can be written as:

$$\mathrm{V_o} = \frac{|\mathrm{Q_d}|}{\mathrm{C_o}} \tag{2.8}$$

where C_o is the gate oxide capacitance per unit area. Combining (2.2), (2.5), (2.7) and (2.8) yields an expression for the gate voltage required for the onset of strong inversion, known as the threshold voltage:

$$V_T = \frac{\sqrt{4qN_A\epsilon_{si}\epsilon_0\varphi_b}}{C_o} + 2\varphi_b \tag{2.9}$$

In reality, there are a number of factors which shift the value of the idealised threshold voltage, namely work function differences, oxide and interface charge. The amount by which the threshold voltage is shifted by these effects is referred to as the flat-band voltage, V_{FB}.

The metal work function varies depending on the material used. Aluminium and degenerately doped, n+ polysilicon have work functions of 4.1eV and 3.95eV respectively. Semiconductor work functions vary with the doping concentration [1]. For silicon, the value of Φ_{ms} is always negative. The effect of this is that some downward energy band bending occurs when V_G = 0V causing depletion and accumulation for p and n-type semiconductors respectively. The applied gate voltage required to return to the flat-band condition is equal to the work function difference, Φ_{ms}.

The second component is the presence of charge in the oxide. Ideally, the oxide would be free of charge, but in practice imperfections arise during processing. Oxide charge comprises of fixed charge near to the Si-SiO_2 interface, Q_f, trapped charge throughout the oxide, Q_t, and mobile ionic charge, Q_m. The combined value for the flat-band voltage can be given as:

$$V_{FB} = \phi_{ms} + \frac{Q_f + Q_t + Q_m}{C_o} \tag{2.10}$$

to produce a final expression for the threshold voltage:

$$V_T = \frac{\sqrt{4qN_A\epsilon_{si}\epsilon_0\varphi_b}}{C_o} + 2\varphi_b + \phi_{ms} + \frac{Q_f + Q_t + Q_m}{C_o} \tag{2.11}$$

2.2.1 Extraction of MOS capacitor parameters

A capacitance-voltage (C-V) analysis of a MOS capacitor allows for the extraction of a number of the previously discussed physical parameters. A 100µm x 100µm MOS capacitor was fabricated in the AMS 0.35µm p-well process for two different oxide thicknesses; the standard gate oxide and a thicker ‘mid’ oxide. A high frequency analysis in which a small ac signal is superimposed onto the dc gate voltage was performed. The gate voltage was swept between -4V and 4V at a rate of 0.05V/sec, with an ac component of frequency 1MHz. The results for capacitors with both oxide thicknesses are shown in Figure 2.5.

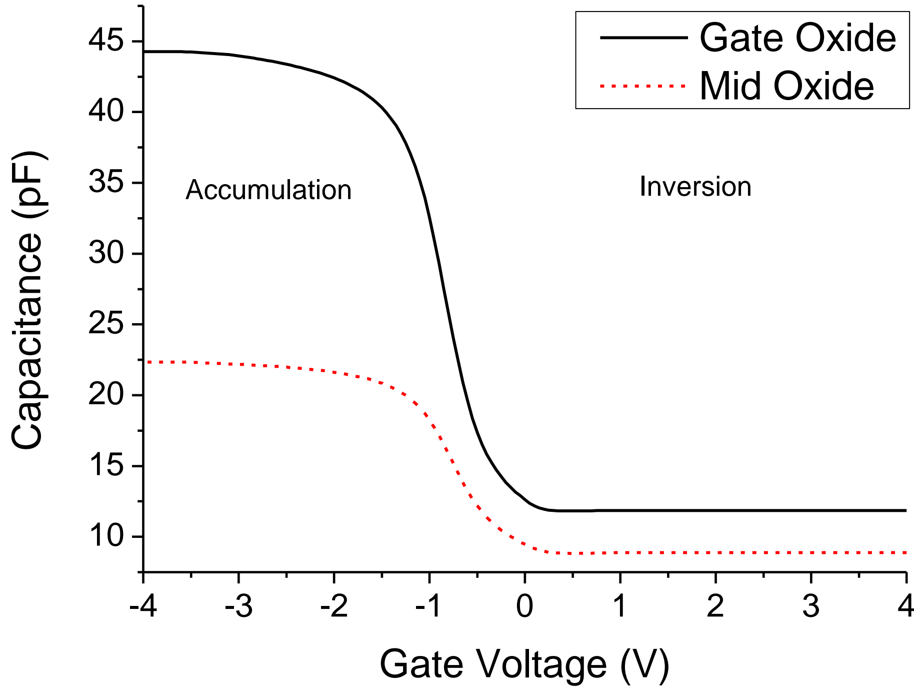

Figure 2.5 - CV Plot for two gate and mid oxide

The maximum and minimum values for the gate oxide and mid oxide are 44.26pF, 11.86pF and 22.33pF, 8.88pF respectively. No hysteresis effects were observed when sweeping the gate voltage in the opposite direction, which indicates that the density of mobile charge carriers in the oxide is low.

Oxide Thickness

Assuming that the accumulation layer capacitance is much greater than that of the oxide when the device is in accumulation, $C_{max} \sim C_{ox}$ and the thickness of the oxide can be extracted from:

$$t_{ox} = \frac{\epsilon_{ox}\epsilon_0 A}{C_{max}} \quad (2.12)$$

where A is the area of the capacitor (10^4 um^2) and C_{max} is the maximum capacitance value taken from the C-V plot. All other values have their usual meanings.

	Measured	**AMS Value**
t_{ox} (Gate oxide)	7.8nm	7.6nm
C_0 (Gate oxide)	4.42 x 10^{-7}Fcm^{-2}	4.54 x 10^{-7}Fcm^{-2}
N_A (Gate oxide)	2.63 x 10^{17}cm^{-3}	2.12 x 10^{17}cm^{-3}
t_{ox} (Mid oxide)	15.5nm	15.0nm
C_0 (Mid oxide)	2.22 x 10^{-7}Fcm^{-2}	2.30 x 10^{-7}Fcm^{-2}
N_A (Mid oxide)	2.15 x 10^{17}cm^{-3}	1.73 x 10^{17}cm^{-3}

Table 2.1 -Values of extracted parameters for capacitors on gate oxide and mid oxide. The equivalent values supplied by AMS are also given.

Doping Density

In the inversion region, the measured capacitance is the series combination of the oxide capacitance and that of the depletion region. The measured capacitance is equal to the minimum capacitance, giving:

$$\frac{1}{C_{min}} = \frac{1}{C_{max}} + \frac{1}{AC_d} \quad (2.13)$$

C_d is the depletion capacitance per unit area:

$$C_d = \frac{\epsilon_{si}\epsilon_0}{W_m} \quad (2.14)$$

where W_m is the maximum depletion width. Combining (2.5), (2.13) and (2.14):

$$\frac{1}{C_{min}} = \frac{1}{C_{max}} + \frac{1}{A}\sqrt{\frac{4\varphi_b}{\epsilon_{si}\epsilon_0 q N_A}} \quad (2.15)$$

which can be solved recursively to yield a value for the substrate doping level, N_A.

The extracted values for oxide thickness, oxide capacitance ($\varepsilon_0\varepsilon_{ox}/t_{ox}$) and doping density are presented in Table 2.1, alongside the nominal values provided by AMS for the 0.35µm process [2]. The measured gate oxide thicknesses are within 3% of the AMS values; the results for N_A are less consistent with the AMS values, the difference between the two results is ~25%.

	Gate oxide	Mid oxide
Experimental mid-gap voltage	-0.35V	-0.15V
Ideal mid-gap voltage	-0.33V	-0.09V
Mid-gap shift	0.02V	0.06V
N_f	5.7×10^{10}cm^{-2}	8.6×10^{10}cm^{-2}

Table 2.2 - Mid-gap voltages and density of oxide charge.

Where device parameters are specified by AMS, typical values extracted during the testing phase are provided. In some cases, minimum and/or maximum values are also given - threshold voltages may vary by +/- 100mV, oxide thicknesses by up to 1nm [2]. Fabricated chips with values outside of these ranges are rejected at the testing phase by the foundry. Other parameters, including doping concentrations, are not used for chip rejection; maximum and minimum values are not provided, only typical ones.

Estimation of oxide charge

As discussed previously, the presence of oxide charge can shift the threshold voltage of a MOS capacitor. This manifests itself on a C-V plot by a shift along the x-axis by an amount ΔV, usually taken at the mid-gap point when the contribution of interface state charge is minimal. ΔV is related to the density of oxide charge by:

$$N_f = \frac{C_{max}\Delta V}{qA} = \frac{C_{max}(V_{mgideal} - V_{mgexperimental})}{qA} \tag{2.16}$$

The ideal mid-gap voltage occurs when the surface potential equals the bulk potential, which can be found using (2.7). The experimental mid-gap voltages can be measured directly from Figure 2.5, where the mid-gap capacitance can be calculated using (2.13). Table 2.2 gives the measured and ideal mid-gap voltages, the mid-gap voltage shift and the density of oxide charge, which is expected to be in the order of 10^{11}cm^{-2} or less for good oxides [1].

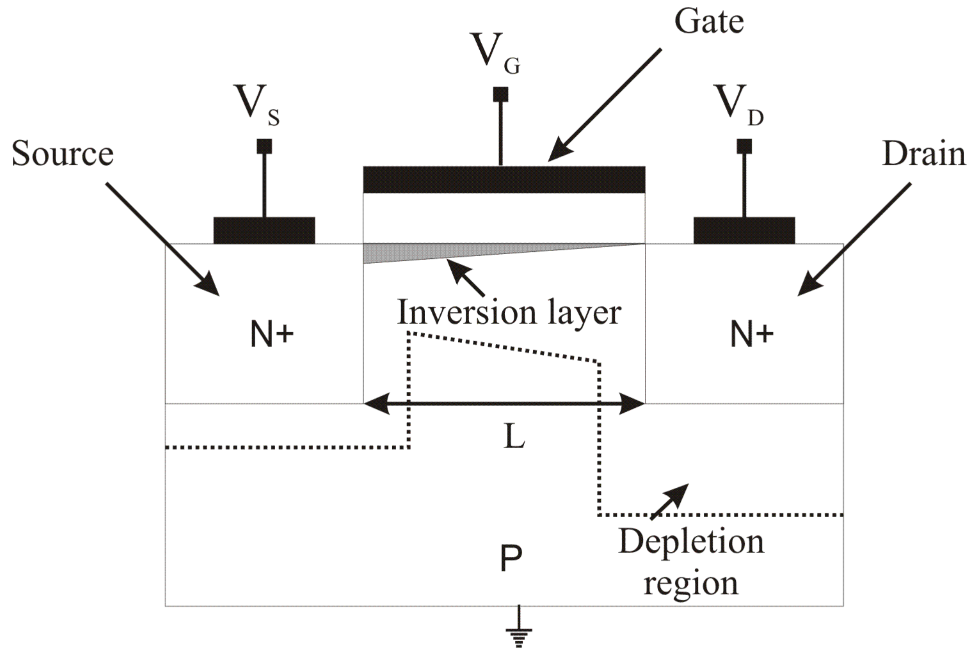

Figure 2.6 - Cross section of n-channel MOS transistor.

2.2.2 MOSFET operation

Before the MOSFET operating characteristics can be derived, a number of assumptions are made: the gate structure is that of an ideal MOS capacitor and there are no interface traps or charges in the oxide; the carrier mobility is constant down the channel and the substrate doping level is uniform; leakage currents in the source drain junctions and through the gate oxide are negligible; drift current is the dominant mechanism of charge transfer; the transverse electric field is much larger than the longitudinal electric field (the 'gradual channel' approximation).

In the inversion regime, the surface potential at a distance y from the source is:

$$\varphi_s = 2\varphi_b + V(y) \tag{2.17}$$

where V(y) is the electron quasi-Fermi potential along the channel with respect to the Fermi potential of the source. The charge in the depletion region is given by:

$$Q_d(V(y)) = -qN_A W_d = -\sqrt{2\epsilon_{si}\epsilon_0 q N_A (2\varphi_b + V(y))} \tag{2.18}$$

The total semiconductor charge is the sum of the depletion charge and the inversion charge; hence an expression for the inversion charge can be found:

$$Q_{inv}(V(y)) = Q_{sc} - Q_d \tag{2.19}$$
$$= -C_0\left(V_{gs} - 2\varphi_b - V(y)\right) + \sqrt{2\epsilon_{si}\epsilon_0 q N_A(2\varphi_b + V(y))}$$

The drain current and the inversion layer charge are related by:

$$\int_0^L I_D\, dy = \frac{W}{L}\mu \int_0^{V_{DS}} -Q_{inv}(V(y))\, dy \tag{2.20}$$

Where W and L are the channel width and length, μ is an average electron mobility in the channel and V_{DS} is the drain voltage relative to the source. Substituting (2.19) into (2.20) and performing the integration gives:

$$I_D = \frac{\mu C_0 W}{L}\left[\left(V_{gs} - 2\varphi_b - \frac{V_{DS}}{2}\right)V_{DS} - 2\frac{\sqrt{2\epsilon_{si}\epsilon_0 q N_A}}{3C_0}\left[(2\varphi_b + V_{DS})^{\frac{3}{2}} - (2\varphi_b)^{\frac{3}{2}}\right]\right] \tag{2.21}$$

Linear operation

For values of $V_{DS} < (V_{GS} - V_T)$, the MOSFET operates in the linear or unsaturated mode and (2.21) simplifies to:

$$I_D = \mu C_0 \frac{W}{L}\left((V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2}\right) \tag{2.22}$$

V_T is the threshold voltage given by (2.11). For $V_{DS} \ll (V_{GS} - V_T)$, the expression can be simplified to:

$$I_D \sim \mu C_0 \frac{W}{L}(V_{GS} - V_T)V_{DS} \tag{2.23}$$

The term $\mu C_0 \frac{W}{L}$ is often referred to as β. It can be seen that the device acts as a voltage controlled resistor.

Saturation operation

The criteria $V_{DS} = (V_{GS} - V_T)$ corresponds to the drain voltage at which the charge in the inversion layer at $y = L_{eff}$ becomes zero, where L_{eff} is the effective or 'electrical' channel

length which defines the so-called pinch-off point. From this point, the MOSFET operates in the saturation regime and the drain current becomes:

$$I_D = \mu C_0 \frac{W}{2L}(V_{GS} - V_T)^2(1 + \lambda V_{DS}) \tag{2.24}$$

where strictly, $L=L_{eff}$ as described below. Ideally, above the pinch-off point, the drain current should reach a saturation value after which further increase in V_{DS} has no effect. In reality, there is a dependence on V_{DS} particularly for short channel devices due to the dependence on V_{DS} of L_{eff}. The depletion region associated with the drain n+ implant expands into the channel as V_{DS} increases, reducing the effective channel length, where the length at the pinch-off point is taken as L_{eff}. Reductions in L_{eff} above this point are modelled by the inclusion of the right hand term, where λ is the channel-length modulation factor. The parameter λ, is usually taken as a constant value, but it has some dependence on the values of both V_{GS} and V_{DS}.

Subthreshold operation

When $V_{GS} < V_T$, the MOSFET operates in the subthreshold region, where diffusion current dominates and the charge density in the channel is exponentially dependent on the gate voltage. The surface potential for a given value of V_{GS} can be obtained by equating the dielectric displacement across the gate oxide and depletion region in the substrate, hence the charge injected from the source into the channel can be found. Assuming that this charge diffuses to the drain, with no recombination along the channel and considering the diffusion equation, it can easily be shown that the current depends on the gate voltage as:

$$I_D = I_0 \exp\left(\frac{qV_{GS}}{mkT}\right)\left[1 - exp\left(-\frac{qV_{DS}}{kT}\right)\right] \tag{2.25}$$

where I_0 is the MOSFET off-current and m is the gate-channel coupling coefficient:

$$m = 1 + \frac{C_d}{C_0} \tag{2.26}$$

For $V_{DS} > 3kT/q$, (2.25) simplifies to:

$$I_D = I_0 \exp\left(\frac{qV_{GS}}{mkT}\right) \tag{2.27}$$

2.2.3 Extraction of MOSFET device parameters

Nominal values of V_T, I_0, m and N_{SS}, the density of surface states, were extracted from I-V plots taken from fabricated MOSFETs, using the 0.35µm AMS process test chip. NMOS transistors were fabricated on both the gate oxide and the mid oxide; PMOS transistors were fabricated only on the gate oxide. All transistors are of dimensions 100µm x 100µm.

Figure 2.7 shows the I_D – V_{GS} characteristics of the three MOSFETs used throughout this thesis on log-lin scales, for V_{DS} = 3.0V. The value of I_0 is equal to the extrapolated intercept on the y-axis and m is related to the subthreshold slope by:

$$S = m\frac{kT}{q}\ln(10) \tag{2.28}$$

where S is the slope of the plot, expressed in mV/decade. N_{SS} can be estimated, as the presence of surface states increases the value of m according to:

$$m = 1 + \frac{C_d}{C_o} + \frac{C_{SS}}{C_o} \tag{2.29}$$

$$N_{SS} = \frac{C_{SS}}{q} \tag{2.30}$$

Typical values for N_{SS} are low 10^{10}cm^{-2} [1]. The inset of Figure 2.7 shows the drain induced barrier lowering (DIBL) effect; there is a shift of approximately 1.5mV between the plots for V_{DS} = 0.1V and V_{DS} = 3.0V. The threshold voltage can be measured by plotting I_D against V_{GS} for a value of V_{DS} such that the device is operating in the linear region. This is plotted in Figure 2.8 for each of the devices, where V_{DS} = 0.1V. Linearly extrapolating the graph and finding the x-intercept yields the threshold voltage of the MOSFET. A value for β can also be extracted from the slope of the graph. As $\beta = \mu C_0 W/L$, a value for the low field mobility can also be calculated. The mobility at high field values can be calculated [3] according to:

$$\mu_{eff} = \frac{\mu_0}{1 + U_a\left(\frac{V_{GS}+V_T}{t_{ox}}\right) + U_b\left(\frac{V_{GS}+V_T}{t_{ox}}\right)^2} \tag{2.31}$$

where U_a = 4.7 x 10^{-10}m/V, U_b = 1.47 x 10^{-18}m/V^2. The minimum mobility value, corresponding to V_{GS} = 3V is 0.83μ_0.

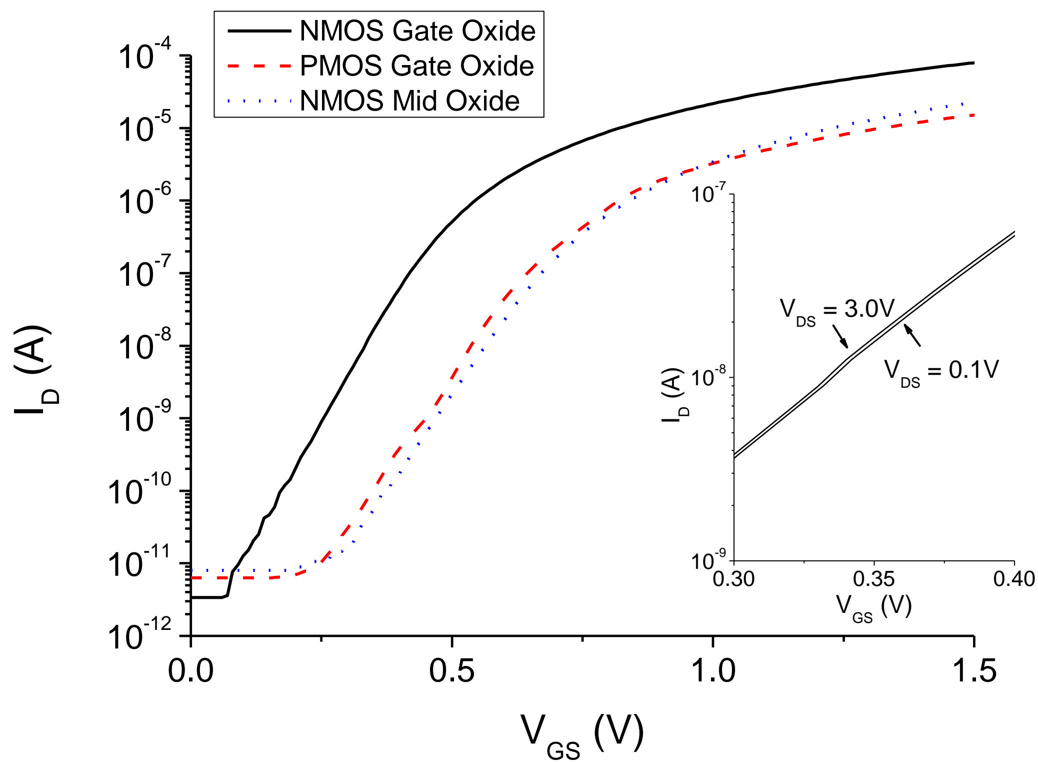

Figure 2.7 - I_D vs. V_{GS} plot for three different MOSFETs. V_{DS} = 3V. W = 100µm, L = 100µm. Inset shows the DIBL effect for the NMOS on gate oxide. Shift is ~1.5mV

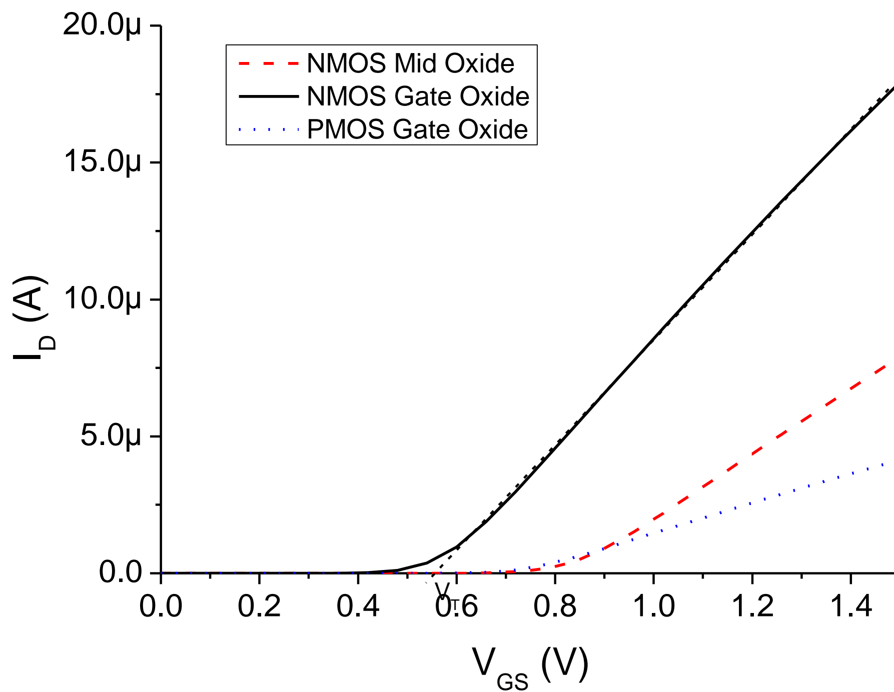

Figure 2.8 - I_{DS} vs. V_{GS}. V_{DS} = 0.1V. Linear extrapolation of the curves allows for an estimation of the threshold voltage. W = 100µm, L = 100µm.

Figure 2.9 plots the output characteristics of the three MOSFETs. As the test MOSFETs are long channel devices, the effect of λ in (2.24) is negligible and the characteristics are essentially flat when the devices are operating in the saturation region.

For shorter channel devices, it is possible to estimate the value of λ by measuring the slope of the graph in the saturation region. An estimate of the value of λ was found by simulating the output characteristics of both short ($L = L_{MIN}$) and long (L=3.5μm) channel devices.

A value for the source-drain resistances, $R_S + R_D$, of the MOSFETs can be found from the inverse of the slope of the characteristics in Figure 2.9. The total source drain resistance, between the contacts, R_{DS} consists of the source and drain resistances, and the resistance of the channel:

$$R_{DS} = R_S + R_D + R_{CH} \tag{2.32}$$

R_{CH} can be found from (2.23) as:

$$R_{CH} = \frac{1}{\beta(V_{GS} - V_T)} \tag{2.33}$$

$$R_S + R_D = R_{DS} - R_{CH} \tag{2.34}$$

Figure 2.10 shows a typical plot of R_{DS} against 1 / ($V_{GS} - V_T$) for the n-channel MOSFET on gate oxide. The intercept on the y-axis gives $R_S + R_D$, equal to 2.3kΩ. Assuming a symmetrical transistor, $R_S = R_D$ = 1.15kΩ. This is approximately double the value extracted from simulation, 0.56kΩ. With maximum operational currents typically in the 1-10uA range, the worst case source and drain voltage drop due to the additional resistance is approximately 5.9mV. The voltage drops due to source-drain resistance for the NMOST and PMOST on mid oxide are 2.8mV and 5.1mV respectively. The slight biasing of the source as a result of voltage drop across R induces a small shift of V_T also; estimated to be a few mV maximum.

The measured values of subthreshold slope, β, λ, I_0, m, μ, R_S, N_{SS} and V_T are presented in Table 2.3. Nominal values supplied by AMS are also given. It is worth noting that the interface state density for the pMOST is very high; N_{SS} =2.0 x 10^{12}cm^{-2}, but is consistent with the increased subthreshold slope of the pMOST.

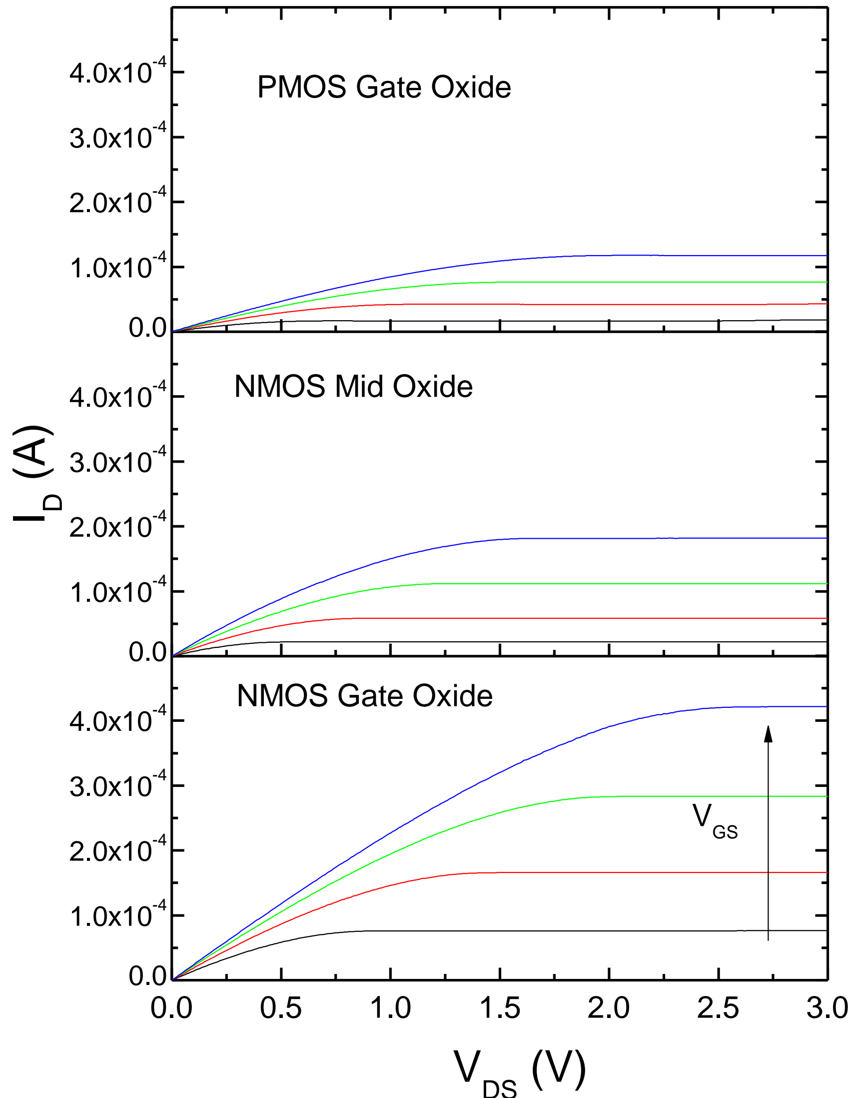

Figure 2.9 - MOSFET output characteristics, for V_{GS} = 1.5V, 2.0V, 2.5V, 3.0V. W = 100µm, L = 100µm.

The 16.4mV increase in S is equivalent to an N_{SS} of 1.8 x $10^{12}cm^{-2}$. Source drain resistances are also about twice those quoted by AMS. For comparison, the measured and simulated characteristics of an n-channel MOST on the mid-oxide is shown in Figure 2.11. Plot (a) shows the original data obtained from simulation; for (b) the simulated data was shifted, within the accepted ranges of variation for V_T, to create a match to the experimental data. Theoretical values are also shown, but are discontinuous around the threshold voltage as (2.24) and (2.25) do not model the transition from subthreshold to above threshold operation.

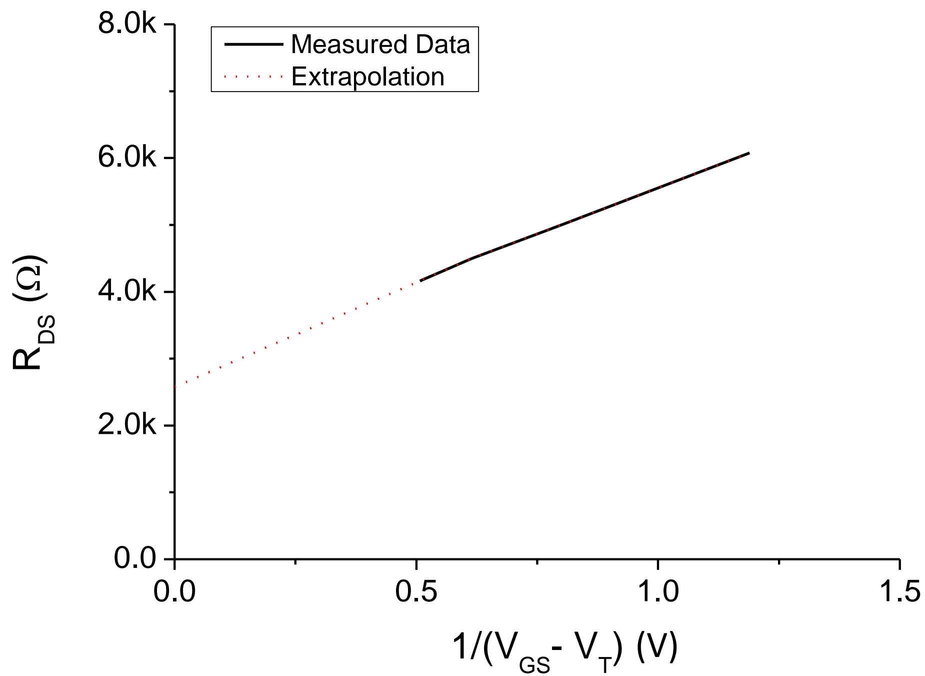

Figure 2.10 - R_{DS} vs $1/(V_{GS} - V_T)$ for nMOST(gate oxide). Intercept from extrapolated values is 2.3kΩ.

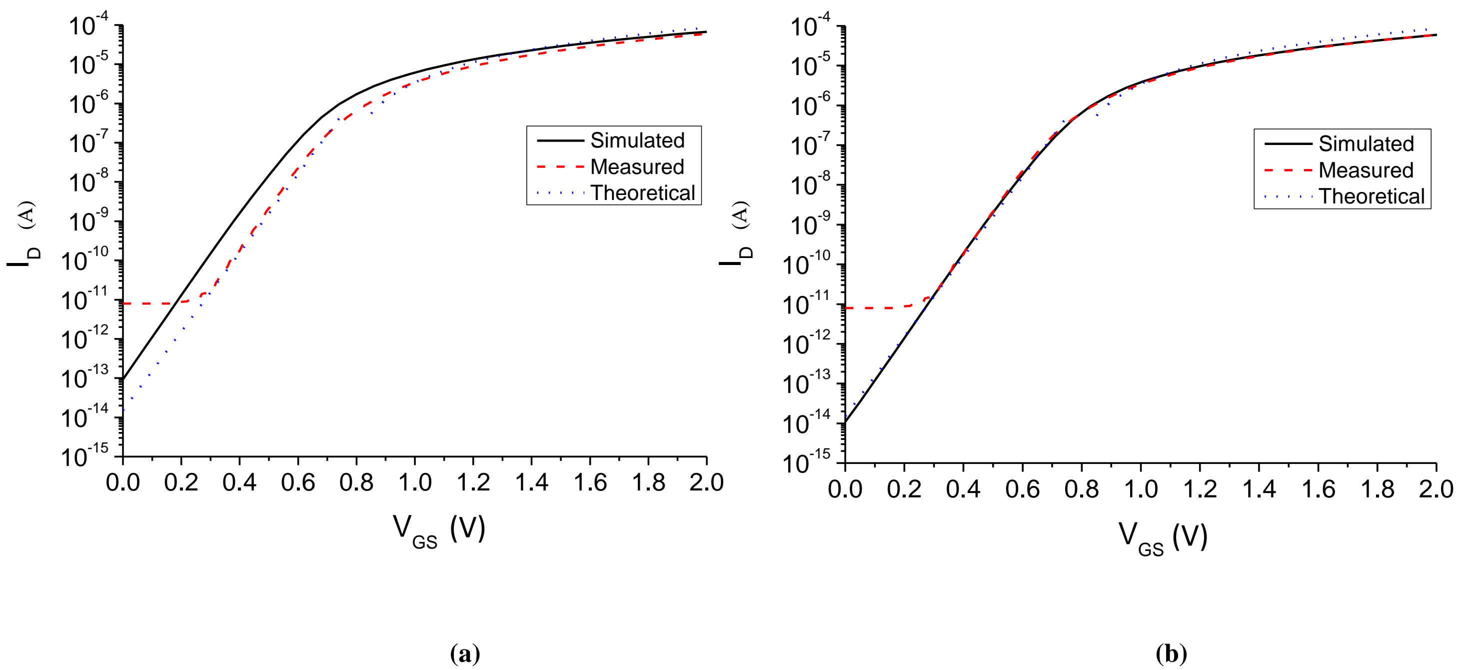

Figure 2.11 - Measured, simulated and theoretical I_D vs V_{GS} for 100µm x 100µm nMOST on mid oxide. Original simulation data is shown in (a), the simulated data is shifted in (b) to match the threshold voltage of the experimental data.

	NMOS gate oxide	NMOS mid oxide	PMOS gate oxide
I_0	640fA	11fA	9fA
Subthreshold slope	82.3mV/decade (80.1mV)	99.6mV/decade (94.7mV)	98.4mV/decade (82.0mV)
m	1.43	1.73	1.71
V_T	0.55V (0.46V)	0.8V (0.7V)	0.7V (0.68V)
β	191uA/V^2 (170uA/V^2)	110uA/V^2 (100uA/V^2)	51uA/V^2 (58uA/V^2)
μ	432cm^{2}V^{-1}s^{-1}	495cm^{2}V^{-1}s^{-1}	115cm^{2}V^{-1}s^{-1}
λ ($L = L_{MIN}$)	23.0mV^{-1}	9.8mV^{-1}	90.0mV^{-1}
λ (L = 3.5μm)	9.9mV^{-1}	0.4mV^{-1}	11.2mV^{-1}
R_S, R_D	1.15kΩ (0.56 kΩ)	0.52kΩ (0.24k)	0.94kΩ (0.43k)
N_{SS}	1.2 x 10^{11}cm^{-2}	9.7 x 10^{10}cm^{-2}	2.0 x 10^{12}cm^{-2}

Table 2.3 – Extracted MOSFET characteristics. Bracketed terms indicate typical values, where provided, from AMS, or extracted from simulations in the case of the subthreshold slope and R_S,R_D; the β and μ values corresponds to low field.

2.3 Discussion and Conclusions

An overview of MOS physics has been provided in this chapter. The MOS capacitor and transistor are the main devices used throughout this thesis; their basic operation has been considered and relevant device parameters have been extracted from test structures fabricated in the 0.35µm CMOS process. The differences observed between measured values and the typical values provided by AMS are within the bounds specified for the fabrication process.

It should be noted that while all of the devices tested in this chapter were fabricated on the same test chip, it was not possible to use the same chip for all of the measurements taken throughout the work reported in this thesis. Even on a single chip, uniformity between different devices will be present. As such, where measurements are taken from fabricated chips in the following chapters, the device parameters of those chips may vary from the values presented in this chapter. AMS tolerances are taken into account when comparing experimental, simulated and theoretical results, throughout the thesis.

References

[1] S. M. Sze, Semiconductor Devices: Physics and Technology, 2nd ed.: John Wiley & Sons, 2001.

[2] AMS, "0.35µm CMOS C35 Process Parameters," July 2007.

[3] M. C. Yuhua Cheng, Kelvin Hui, Min-chie Jeng,, J. H. Zhihong Liu, Kai Chen, James Chen, Robert Tu,, and C. H. Ping K. Ko, BSIM3v3 Manual: University of California, Berkeley, 1996.

Chapter 3: Silicon Synapse Device and Circuit

3.1 Introduction

Synapses are the dominant processing elements in the brain, forming an electrochemical junction between the axon of a presynaptic neuron and a dendrite of a postsynaptic neuron. Information processing and storage are possible because the strength of synaptic connections between neurons can be changed. Of all of the features of a synapse, this phenomenon, known as synaptic plasticity, is the most widely studied and implemented in neural networks.

In this chapter, two implementations of a compact spiking silicon synapse are described. The first device implements synaptic plasticity by means of a weight voltage on the gate of an n-channel MOSFET and constitutes a static synapse. A second MOSFET acts as a transfer terminal, controlling the transfer of the weight charge to an output terminal, which produces an output current/voltage spike. The second synapse device adds an additional control voltage on the gate of a third MOSFET, allowing for the implementation of synaptic depression and thus corresponds to a dynamic synapse. Design equations for both devices are derived from MOS physics and a theoretical model is presented. Simulation and experimental results validate the operation of the device. Where required, the MOSFET parameters extracted in Chapter 2 are used for calculations. For convenience, these values are given in Table 3.1.

The rest of the chapter is organised as follows, the two-terminal synapse is described in Section 3.2. A theoretical model for the operation of the device is developed; simulation and experimental results confirm the operation of the device. The analysis is repeated for the three gate synapse in Section 3.3. Conclusions and discussion are presented in Section 3.4.

	NMOS gate oxide	NMOS mid oxide	PMOS gate oxide
I_0	640fA	11fA	9fA
Subthreshold slope	82.3mV/decade	99.6mV/decade	98.4mV/decade
m	1.43	1.73	1.71
V_T	0.55V	0.8V	0.7V
β	191uA/V^2	110uA/V^2	51uA/V^2
μ (low field)	432cm^{2}V^{-1}s^{-1}	495cm^{2}V^{-1}s^{-1}	115cm^{2}V^{-1}s^{-1}
N_A	2.63 x 10^{17}cm^{-3}	2.2 x 10^{17}cm^{-3}	2.63 x 10^{17}cm^{-3}
C_0	4.42 x 10^{-7}Fcm^{-2}	2.22 x 10^{-7}Fcm^{-2}	4.42 x 10^{-7}Fcm^{-2}

Table 3.1- MOSFET Device parameters

3.2 Two-Terminal Silicon Static Synapse

A cross-sectional view of the two-terminal synapse cell is shown in Figure 3.1. The layout is that of two series-connected MOSFETs with a common source/drain connection. Both MOSFETs are fabricated using the mid-oxide process option described in Chapter 2, with a gate oxide thickness of ~15nm. Using the thicker gate oxide reduces the magnitude of the synaptic output current for a given weight voltage. This allows the neuron circuit, discussed in Chapter 4 to operate over a wider voltage range. A schematic view of the device is shown in Figure 3.2. The output n+ region of the synapse is connected to a saturated load MOSFET, M3, which forms a current mirror with M4 and acts as the input branch of the neuron circuit, which is discussed in more detail in Chapter 4.

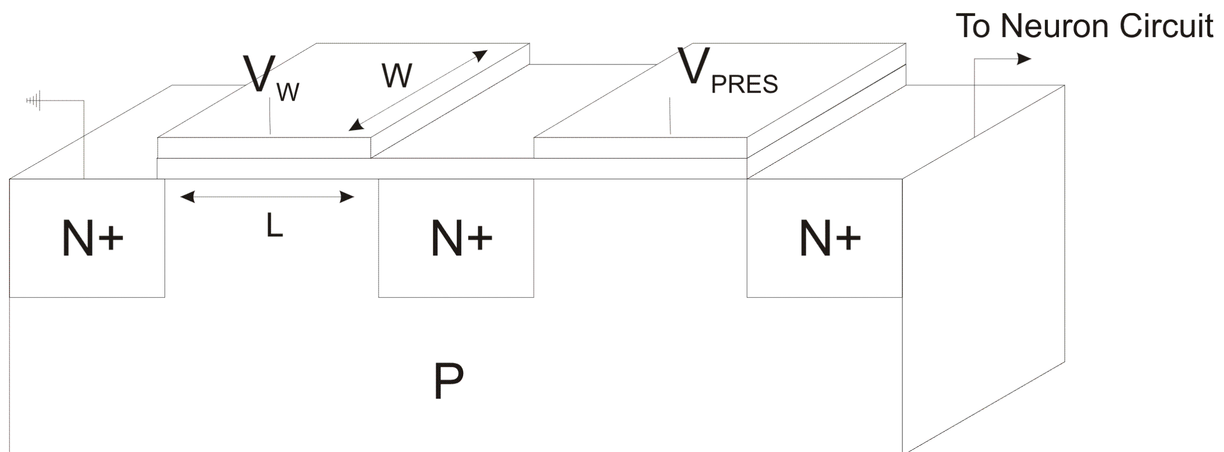

Figure 3.1 - Silicon Synapse

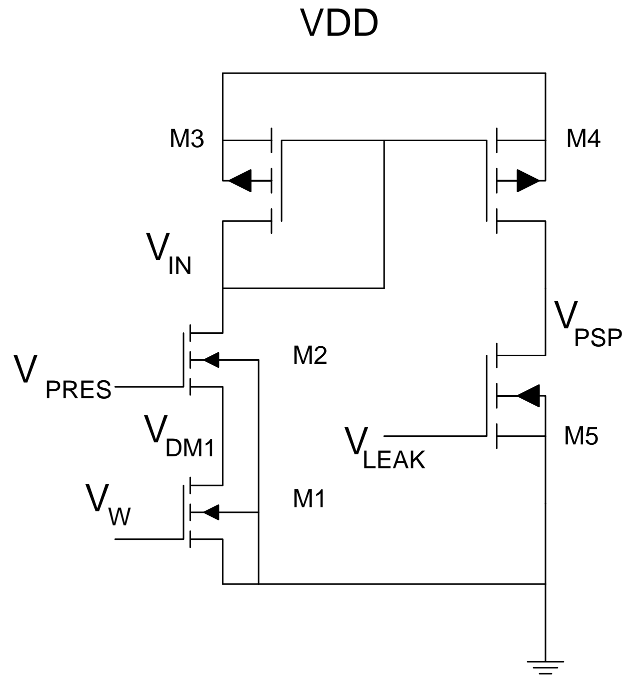

Figure 3.2 - Schematic view of synapse cell (M1,M2). Synapse output is connected to the input arm of neuron circuit, described in Chapter 4.

Synaptic plasticity is implemented through the weight voltage, V_W, which controls the amount of charge present underneath the gate of M1. Under quiescent operation conditions, with V_{PRES} = 0V, the voltage at the output node of the synapse, V_{IN} will sit at a value just below V_{DD}, due to leakage. The application of a voltage pulse to the V_{PRES} terminal initiates the transfer of the weight charge from M1 to the output node, V_{IN}. As will be shown later, the transfer of charge to the output node will typically be manifested as a current spike at the synapse output for the duration of the input pulse and a proportional reduction in the value of V_{IN}. When V_{PRES} returns to 0V, the voltage at V_{IN} will relax back to its resting value via current flow through M3.

3.2.1 Theoretical Operation

In a resting state, with V_{PRES} = 0V, the voltage at V_{IN} will sit at a value ~0.25V less than V_{DD}, as M3 must supply leakage current to the MOSFET chain. The weight voltage V_W induces a channel of electrons under the gate of M1 and the drain/source voltage of M1/M2, V_{DM1}, will be ~0V as there is no significant current flow in the branch.

The total synaptic output charge, Q_W, produced in response to an input pulse on the V_{PRES} terminal represents the weight of the synapse. It will be shown in Chapter 4 how this charge is stored and communicated to the point neuron, via M4. The derivation of Q_W differs significantly depending on whether M1 is operating above ($V_W > V_{Tn}$) or below threshold ($V_W < V_{Tn}$).

The case of $V_W > V_{Tn}$ is considered first. The application of a voltage pulse of magnitude V_{DD}, pulse width ΔT and rise time τ_{rp} to the gate of M2 will initiate a transient flow of electrons from M1 to the V_{IN} node, effectively discharging its capacitance. Thus V_{IN} will be reduced by an amount ΔV_{IN} and V_{DM1} will increase by an amount ΔV_{DM1} to ensure current continuity. The final values of V_{IN} and V_{DM1} will be dependent on the value of V_W and will be such that:

$$I_{DM1}(V_W) = I_{DM2} = I_{DM3} \tag{3.1}$$

These currents will remain constant for the duration of the input pulse. During the rising portion of the voltage pulse, the output charge will be equal to:

$$Q_{wr} = \int_0^{\tau_{rp}} I_{DM2}(t)dt \tag{3.2}$$

The total output charge of the synapse, Q_W, is:

$$Q_w(V_W) = \int_0^{\tau_{rp}} I_{DM2}(t)dt + I_{DM1}(V_w)\Delta T \tag{3.3}$$

If $\Delta T >> \tau_{rp}$, then the integral term can be ignored and the total charge becomes:

$$Q_w(V_W) = I_{DM1}(V_w)\Delta T \tag{3.4}$$

For a fixed pulse width, the charge will scale according to the current through M1. Initially, V_{IN} falls about equally across M1 and M2 and they will be saturated:

$$Q_w(V_W) = \frac{\beta_n}{2}(V_W - V_{Tn})^2\Delta T \tag{3.5}$$

Where $\beta_n = \frac{u_n C_{On} W_n}{L_n}$. In order to satisfy the current requirements of M1 and M2, V_{IN} must be such that M3 is operating above threshold. The value of V_{IN} for a given value of V_W can be found by equating the currents in M1 and M3:

$$\frac{\beta_n}{2}(V_W - V_{Tn})^2 = \frac{\beta_p}{2}\left(V_{DD} - V_{in} - V_{Tp}\right)^2 \tag{3.6}$$

Rearranging for V_{IN} gives:

$$V_{IN} = V_{DD} - V_{Tp} - \sqrt{\frac{\beta_n}{\beta_p}}(V_W - V_{Tn}) \tag{3.7}$$

Substituting V_{DD} = 3V, W_n =0.8µm, L_n = 0.6µm, W_p = 0.35µm, L_p = 0.35µm and the values extracted for V_{Tn}, V_{Tp}, C_{0n}, C_{0p}, μ_n and μ_p in Chapter 2, gives:

$$V_{IN} = 3.51 - 1.51V_W \tag{3.8}$$

Despite variations in mobility for higher values of V_{GS}, as described in Section 2.3.2, the square root of the ratio β_n/β_p remains approximately constant, varying by less than 2% and (3.8) remains valid.

Given that I_{DM1} = I_{DM2} and both transistors are saturated; the V_{GS} of both transistors must be approximately equal. Hence, an expression can be derived for the value of V_{DM1} by equating gate-source overdrive voltages of M1 and M2. The output impedance given by λ = 9.8mV^{-1}, corresponds to an output impedance, R_{out} ≈ 750kΩ which is taken as infinite. Assuming also a constant mobility V_{DM1} can be found from:

$$V_W - V_{Tn} = V_{GSM2} - V_{Tn} = V_{PRES} - V_{DM1} - V_{Tn} \tag{3.9}$$

However, there will be a shift in the threshold voltage of M2 due to substrate bias caused by V_{DM1}. The effect of substrate bias on threshold voltage can be expressed as:

$$V_T = V_{Tn0} + \gamma\left(\sqrt{V_{sub} + 2\varphi_b} - \sqrt{2\varphi_b}\right) \tag{3.10}$$

where V_{sub} is the substrate to source bias, in this case equal to V_{DM1}, V_{Tn0} is the threshold voltage with no substrate bias and γ is the body effect factor:

$$\gamma = \frac{\sqrt{\epsilon_{si}\epsilon_0 q N_A}}{C_0} \tag{3.11}$$

which is calculated as 0.85V^{-1} (The value calculated using AMS values is 0.75V^{-1}). It should be noted that (3.10) assumes a uniform substrate doping and a long channel device. Substituting (3.10) into (3.9) gives:

$$V_W - V_{Tn} = V_{PRES} - V_{DM1} - V_{Tn} - \gamma\left(\sqrt{V_{DM1} + 2\varphi_b} - \sqrt{2\varphi_b}\right) \tag{3.12}$$

Substituting V_{PRES} = 3V, V_{Tn} = 0.8V, φ_b = 0.41V and rearranging gives:

$$V_W - 3.78 + V_{DM1} = -\gamma\left(\sqrt{V_{DM1} + 0.82}\right) \tag{3.13}$$

Squaring both sides produces a quadratic equation, which can be solved to give V_{DM1} as a function of V_W:

$$V_{DM1}^2 + (2V_W - 8.41)V_{DM1} + ((V_W - 3.78)^2 - 0.78) = 0 \quad (3.14)$$

If the variation in mobility is included and the analysis repeated, the expression becomes:

$$V_{DM1}^2 + \left(2\sqrt{\frac{u_1}{u_2}}(V_W - 0.8) - 2.98\right)V_{DM1} + \left(\sqrt{\frac{u_1}{u_2}}((V_W - 0.8) - 2.98)^2 - 1.2\right) = 0 \quad (3.15)$$

Where μ_1 and μ_2 are functions of V_W and V_{PRES} respectively. With $V_{PRES} = 3V$, the maximum difference between (3.14) and (3.15) will occur when $V_W = 0.8V$. Substituting and solving both equations for V_{DM1} gives a difference between the two values of only 2mV, validating the earlier assumption that the mobility can be considered constant.

(3.14) gives V_{DM1} as a function of V_W while M2 is saturated. An expression for V_{DM1} when M2 is unsaturated can be found by equating the currents in M1 and M2 and solving. A quadratic solution can be obtained if the effect of substrate bias on the threshold voltage of M2 is ignored:

$$V_{DM1}^2 - 4.4V_{DM1} + 1.28V_W^2 - 9.2V_W + 3.05 = 0 \quad (3.16)$$

The full analysis, including the effect of substrate bias, can not readily be solved analytically, but can be solved computationally for a given value of V_W. Where values for V_{DM1} are required in further analysis, the computationally obtained values are used. The saturation points of M1 and M2 can be expressed, respectively, as:

$$V_W - V_T - V_{DM1} = 0 \quad (3.17)$$

$$V_{PRES} - V_T(V_{DM1}) - V_{IN} = 0 \quad (3.18)$$

Figure 3.3 plots (3.17) and (3.18) as a function of V_W, giving the saturation voltages as V_W = 1.26V for M2 and V_W = 1.67V for M1, at which point V_{DM1} = 0.62V. For values of V_W > 1.67V, the total synaptic charge becomes:

$$Q_w(V_W) = \beta_n\left((V_W - V_{Tn})V_{DM1} - \frac{V_{DM1}^2}{2}\right)\Delta T \quad (3.19)$$

When M1 is unsaturated, the use of (3.8) and (3.14) to predict the values of V_{IN} and V_{DM1} is no longer valid as they were derived for the case of M1 being saturated. In order to simplify, a reasonable assumption to make is that the value of V_{DM1} remains approximately constant at 0.62V when M1 is unsaturated.

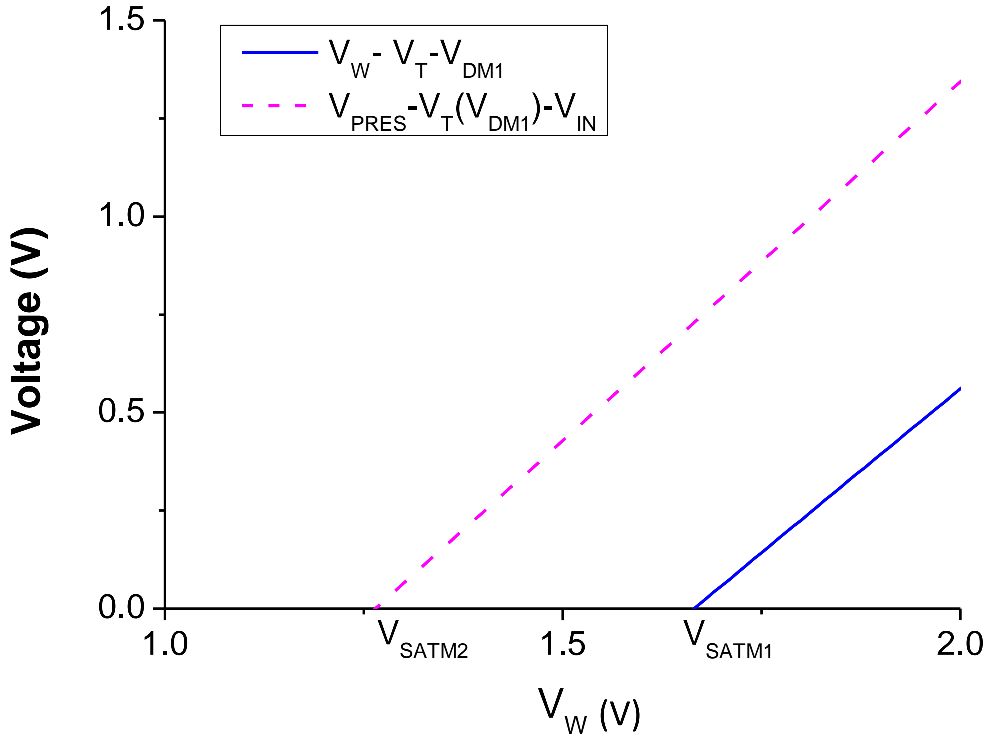

Figure 3.3 - Graphical representation of the saturation points of M1 and M2.

This approximation is validated later through comparison with simulation results. In reality, V_{DM1} will decrease slightly for increased values of V_W, to maintain the condition described in (3.1). Plotting (3.5) and (3.19) against V_W produces the graph shown in Figure 3.4. The synaptic output charge and electron density (Q_W/q) are plotted on a log scale for ΔT = 1ns, 10ns, 100ns and 1000ns. In practice, the range of V_W values used is limited by the neuron circuit. Typically, the maximum required value of V_W is 1.3V or less.

3.2.1.1 Synapse operation when $V_W < V_T$

When V_W < 0.8V, M1 will operate in the subthreshold region. In order for (3.1) to be satisfied, M2 must also operate subthreshold, where the current is given by:

$$I_D = I_0 \exp\left(\frac{qV_{GS}}{m_n kT}\right) \tag{3.20}$$

assuming that V_{DS} of both MOSTs is greater than 3kT/q.

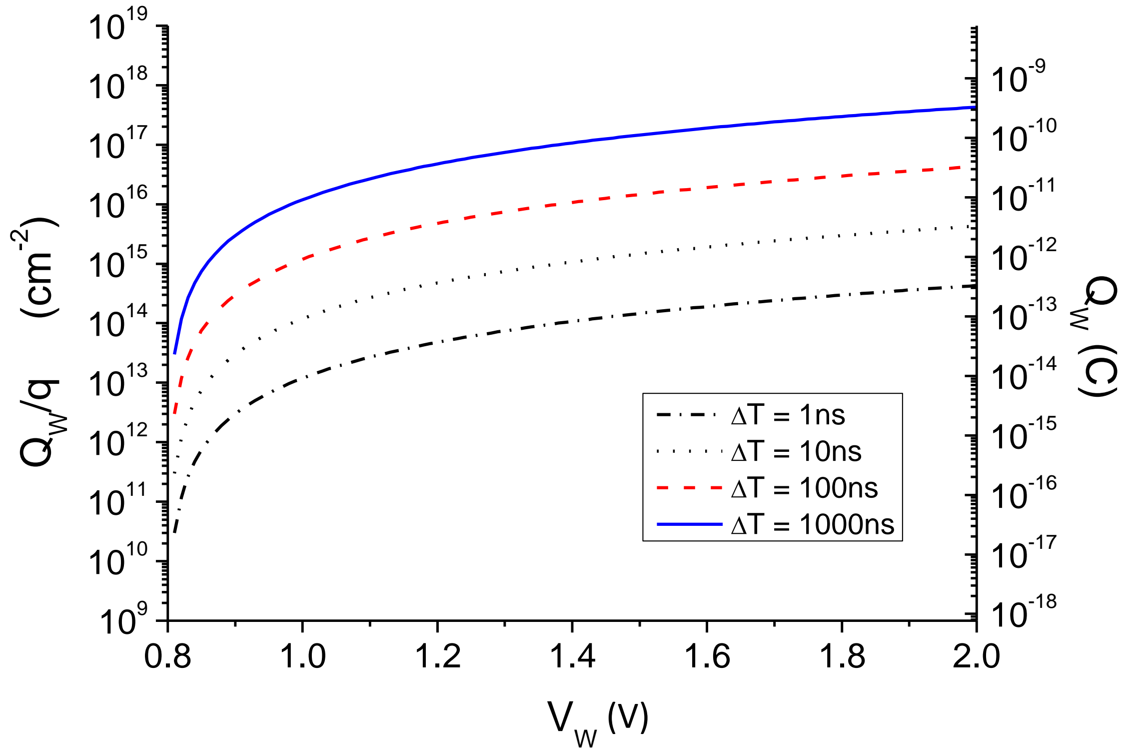

Figure 3.4 - Synaptic output charge and electron density against V_W for different pulse widths.

Setting $V_W = V_T$ in (3.8) and (3.14) allows the drain source voltages of M1 and M2 to be calculated as 1.47V and 0.83V respectively, confirming that both transistors are saturated when operating subthreshold. Given that both transistors must have the same V_{GS}, equating the currents in M1 and M2 yields an expression for V_{DM1} as a function of V_W:

$$I_{0n} \exp\left(\frac{qV_W}{m_n kT}\right) = I_{0n} \exp\left(\frac{q(V_{PRES} - V_{DM1} - \Delta V_T)}{m_n kT}\right) \tag{3.21}$$

where the ΔV_T term accounts for the increased threshold voltage of M2, due to the effective substrate bias, V_{DM1}, at the source of M2. Equating the gate voltages of the two transistors gives:

$$V_W = V_{PRES} - V_{DM1} - \gamma\left(\sqrt{V_{DM1} + 2\varphi_b} - \sqrt{2\varphi_b}\right) \tag{3.22}$$

which is equivalent to the expression in (3.12). The subsequent analysis is identical, which means that (3.14) can also be used to estimate V_{DM1} when M1 is operating below threshold.

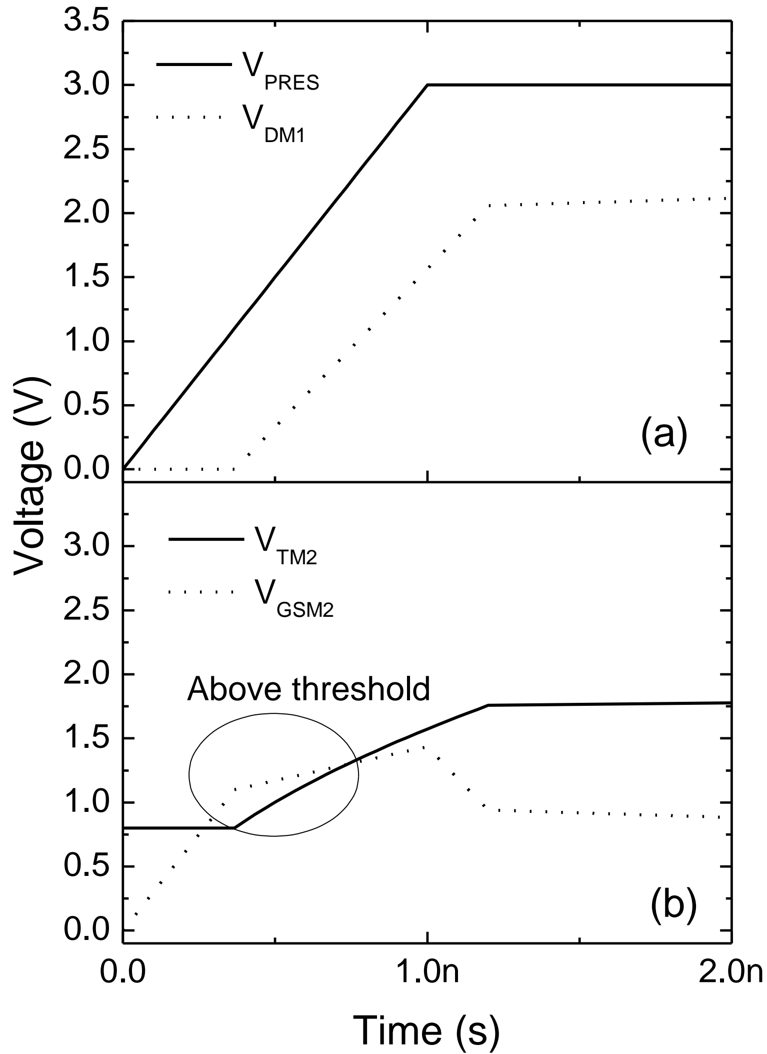

Figure 3.5 - Voltages during rising edge of input pulse for the case of V_W = 0.7V. (a) shows V_{PRES}, V_{DM1}. (b) shows the gate-source and threshold voltages of M2. The circled region indicates where M2 is operating above threshold.

Equation (3.20) corresponds to the synaptic output current that will flow for the duration of the input pulse, once the values of V_{IN} and V_{DM1} have settled at their final values. However, as indicated in (3.3), there is also a transient current component during the rising edge of the input pulse. In fact, for values of $V_W < 0.8$V, this current becomes much greater than the final steady state current. This will be demonstrated in the following analysis. During the rising edge of the V_{PRES} pulse, V_{DM1} will charge up to its final value according to:

$$I = I_{DM2} - I_{DM1} = C\frac{dV_{DM1}}{dt} \tag{3.23}$$

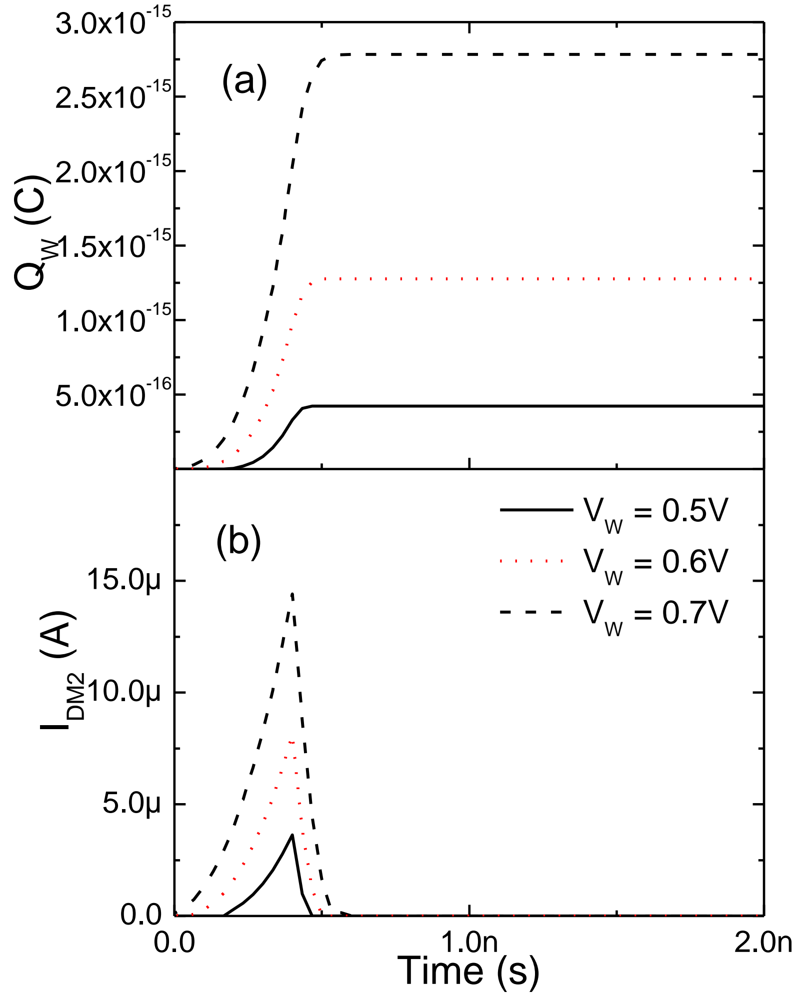

Figure 3.6 - Synapse output charge (a) and current (b) when $V_W < V_T$. The output charge saturates once M2 returns to the subthreshold region of operation.

where C is the capacitance associated with the V_{DM1} node. I_{DM1} is a constant for a given value of V_W and I_{DM2} is a function of both V_{PRES} and V_{DM1}. Substituting values for I_{DM1} and I_{DM2} gives:

$$I_0 \exp\left(\frac{q(V_{PRES}(t) - V_{DM1})}{m_n kT}\right) - I_0 \exp\left(\frac{qV_W}{m_n kT}\right) = C\frac{dV_{DM1}}{dt} \tag{3.24}$$

Rearranging, and integrating both sides gives:

$$t = C\int_0^{V_{DM1final}} \frac{dV_{DM1}}{I_0\left(\exp\left(\frac{q(V_{PRES}(t) - V_{DM1})}{m_n kT}\right) - \exp\left(\frac{qV_W}{m_n kT}\right)\right)} \tag{3.25}$$

As V_W becomes smaller, the charging current at the V_{DM1} node ($I_{DM2} – I_{DM1}$) becomes larger and the slew rate of V_{DM1} increases. To fully solve the integral it is necessary to formulate an expression for V_{DM1} as a function of V_{PRES}. However, a sufficient approximation to the experimental and simulated results is that V_{DM1} charges in a linear manner from 0V to its final value, given by (3.14), over a period of time equal to the rise time of the input pulse.

Figure 3.5 illustrates the charging of the drain of M1 up to its final value of V_{DM1}, and its effect on the V_{GS} and V_T of M2, for V_W = 0.7V. Figure 3.5a shows the rise of V_{PRES} and V_{DM1}, where the rise time of V_{PRES} is set at 1ns. The gate-source voltage and threshold voltage of M2 are plotted in Figure 3.5b. The circled region indicates the region where M2 is operating above threshold. After this, M2 returns to the subthreshold region, where it remains for the duration of the input pulse. The data provided in Figure 3.5b is used to plot the synapse output charge and current, shown in Figure 3.6 for V_W = 0.7V, 0.6V and 0.5V. It can be seen that for each case, the output current during the period where M2 is operating above threshold is much greater than the subthreshold current. Figure 3.6a shows that the synaptic charge, Q_W, undergoes a rapid initial increase, after which there is little change in the total output charge. As such, it can be concluded that for the case of $V_W < V_T$, the synaptic output charge Q_W is independent of the pulse width ΔT and varies only with the weight voltage V_W. The implications of this are considered in further detail when the operation of the three terminal synapse is discussed.

3.2.1.2 Dynamics of the V_{IN} node

Once the voltage pulse applied to the V_{PRES} terminal returns to 0V, the current through M3 will charge the node V_{IN} back up to its original value, according to:

$$I_{DM3} = C_{VIN} \frac{dV_{IN}}{dt} \tag{3.26}$$

where C_{VIN} is the total capacitance at the V_{IN} node. It is comprised of the output capacitance of the synapse(s), C_{syn}, which is that of a pn junction; the gate capacitances of M3 and M4; and the capacitance of the metal interconnects in the circuit.

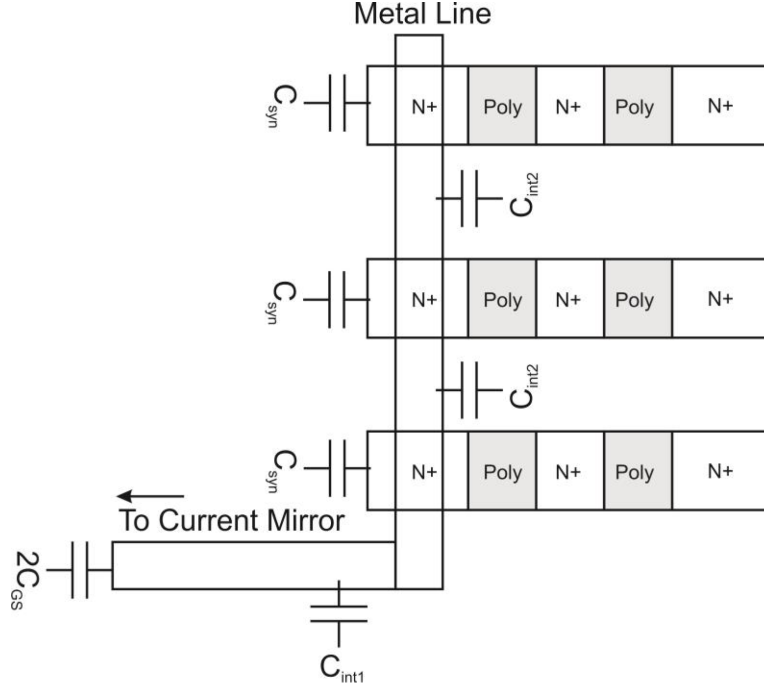

Figure 3.7- Capacitances at the V_{IN} node.

The interconnect capacitance has two components. There is a fixed capacitance between the output of the synapses and the input to the current mirror, C_{int1}, and an interconnect capacitance for each additional synapse, C_{int2}. Figure 3.7 illustrates the various components of capacitance. The total capacitance can be expressed as:

$$C_{VIN} = n\sqrt{\frac{qN_A\epsilon_{si}\epsilon_0}{2(\varphi_b + V_{DD} - V_{IN})}}A + 2C_{GS} + C_{int1} + C_{int2}(n-1) \tag{3.27}$$

where n is the number of parallel synapses connected to the V_{IN} (the fan-in) and A is the area of the synapse output region. The gate-source capacitance, C_{GS} can be approximated as:

$$C_{GS} = \frac{2}{3}C_0WL + C_0 0.1L \tag{3.28}$$

where the 0.1L term accounts for the capacitance due to the gate-source oxide overlap which is assumed to be 10% of L.

During the rise time of V_{IN}, M3 can operate either in the subthreshold region or above threshold and saturated. Considering first the subthreshold case:

$$I_{0p}\exp\left(\frac{qV_{GS}}{m_pkT}\right) = C_{VIN}\frac{dV_{IN}}{dt} \tag{3.29}$$

Since $V_{DD} - V_{IN} = V_{GS}$, the variable of the integration can be changed as $dV_{IN}/dt = -dV_{GS}/dt$. When the rise time is at its maximum, this corresponds to V_{GS} going from V_{Tp} to its resting value of V_{RR}, which is ~0.25V. Separating variable gives:

$$\frac{I_{0p}}{C_{VIN}}\int_0^{\tau r1} dt \ = \int_{V_{Tp}}^{V_{RR}} \exp\left(-\frac{qV_{GS}}{m_p kT}\right) dV_{GS} \tag{3.30}$$

where $\tau_{r1}$is the rise time of V_{IN}. Integrating both sides gives:

$$\frac{I_{0p}}{C_{VIN}}\tau_{r1} \ = \frac{m_p kT}{q}\left[\exp\left(-\frac{qV_{GS}}{m_p kT}\right)\right]_{V_{Tp}}^{V_{RR}} \tag{3.31}$$

Expanding the brackets and rearranging gives:

$$\tau_{r1} = C_{VIN}\frac{m_p kT}{I_{0p}q}\left[\exp\left(-\frac{qV_{RR}}{m_p kT}\right) - \exp\left(-\frac{qV_{Tp}}{m_p kT}\right)\right] \tag{3.32}$$

When M3 is operating above threshold the V_{IN} node charges according to:

$$\frac{\mu C_{0p}W}{2L}\ \left(V_{GS} - \ V_{Tp}\right)^2 = C_{VIN}\frac{dV_{IN}}{dt} \tag{3.33}$$

Changing the variable of integration to $-dV_{GS}/dt$ and separating variables gives:

$$\frac{\mu C_{0p}W}{2LC_{VIN}}\int_0^{\tau r2} dt \ = -\int_{V_{DD}}^{V_{Tp}} \frac{1}{\left(V_{GS} - \ V_{Tp}\right)^2} dV_{GS} \tag{3.34}$$

where τ_{r2} is the rise time while M3 is above threshold. Here, the maximum rise time is for V_{GS} between V_{DD} and V_{Tp+}, a value slightly above threshold. Integrating gives:

$$\frac{\mu C_{0p}W}{2LC_{VIN}}\ \tau_{r2} \ = \left[\frac{1}{V_{GS} - \ V_{Tp}}\right]_{V_{DD}}^{V_{Tp+}} \tag{3.35}$$

which becomes:

$$\tau_{r2} \ = \frac{2LC_{VIN}}{\mu C_{0p}W}\left[\frac{1}{V_{Tp+} - \ V_{Tp}} - \frac{1}{V_{DD} - \ V_{Tp}}\right] \tag{3.36}$$

Figure **3.8** shows plots of τ_{r1} and τ_{r2} against the fan-in, n. The width and length of M3 and M4 are 0.4μm and 0.35μm; C_{int1} and C_{int2} in (3.27) are 2.77fF and 0.2fF. The voltage dependence of C_{VIN} is omitted for this analysis. Since $\tau_{r1} >> \tau_{r2}$, the total relaxation time is almost entirely dependent on the subthreshold element, τ_{r1}, and will be in the order of milliseconds.

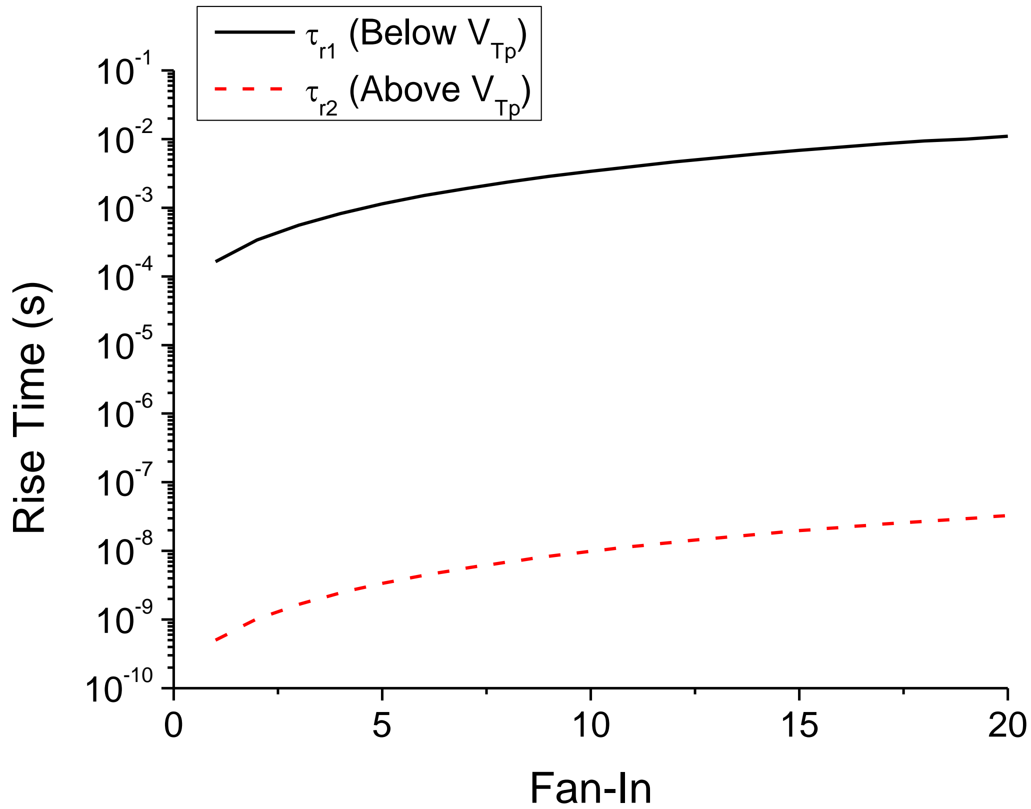

Figure 3.8 - The two components of the rise time, τ_{r1} and τ_{r2}, as a function of the fan-in.

However, the current flow onto the output branch of the current mirror, considered in Chapter 4, is greatest when M4 is operating above threshold. The exact effect of τ_{r1} and τ_{r2} on the operation of the neuron circuit depends on the values of V_W, V_{LEAK} and ΔT. This is considered in Section 4.2.2.

3.2.2 Results

In this section, simulation and experimental results are presented which confirm that the operation of the synapse is consistent with the theory presented in the previous section. The Cadence software suite was used to obtain simulation results, by implementing the SPICE general-purpose analogue circuit simulator. Cadence can also be used to extract parasitic capacitances and resistances from circuit layouts and annotate them to a schematic view for inclusion in the SPICE simulations. The inclusion of parasitics makes the results more accurate, so this is done for all simulation results presented.

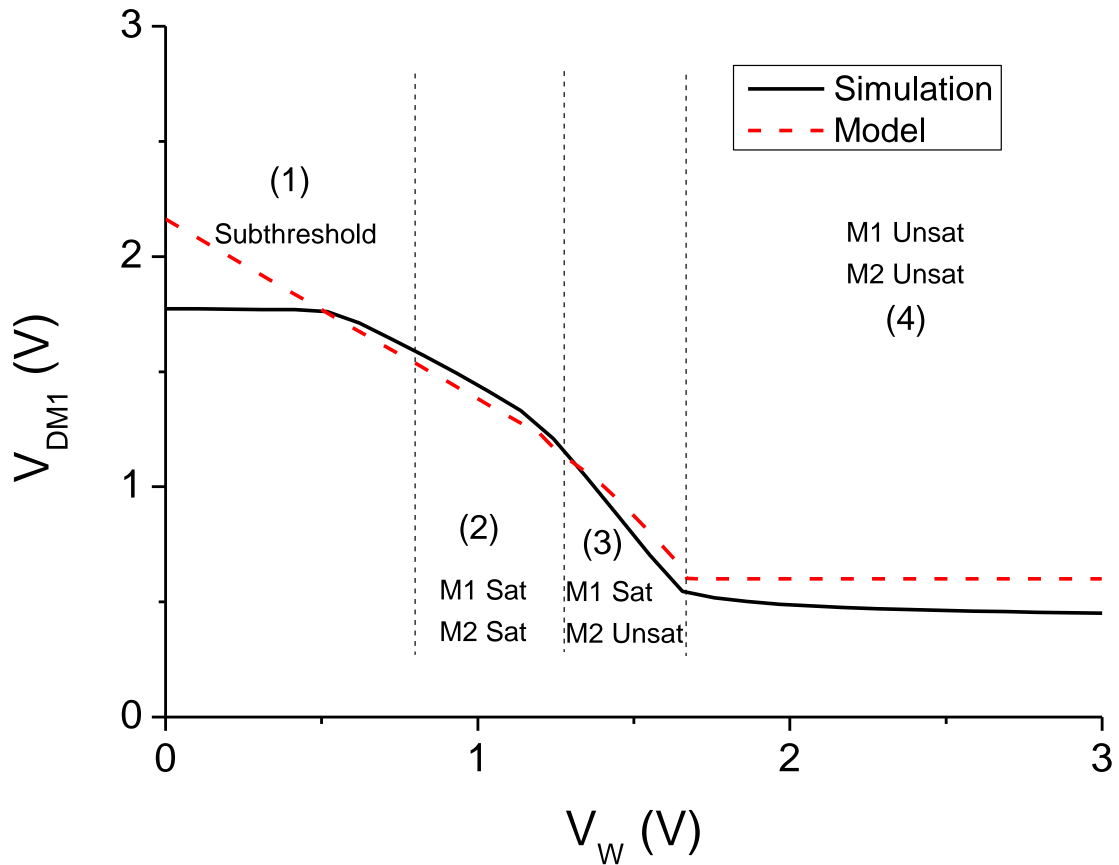

Figure 3.9 - Simulated and modelled values of V_{DM1}. The different regions of operation are indicated and numbered.

Experimental results are obtained through measurements taken from fabricated test chips, as described in Appendix 1. Unless otherwise stated, V_{DD} is set to 3V and transistor dimensions are: M1,M2 (0.8μm x 0.6μm), M3,M4,M5 (0.35μm x 0.35μm). These are the minimum dimensions allowed for transistors on the mid oxide and gate oxide respectively.

First, the model derived for the value of V_{DM1} in the previous section can be compared against simulation results. Figure 3.9 plots the two sets of results, the different regions of operation are indicated on the plot. There is reasonable agreement in regions 2 and 3, and for $V_W > 0.4V$ in region 1. The discrepancy between the modelled and simulated results is greatest in region 4, where both M1 and M2 are unsaturated, and also when $V_W < 0.4V$.

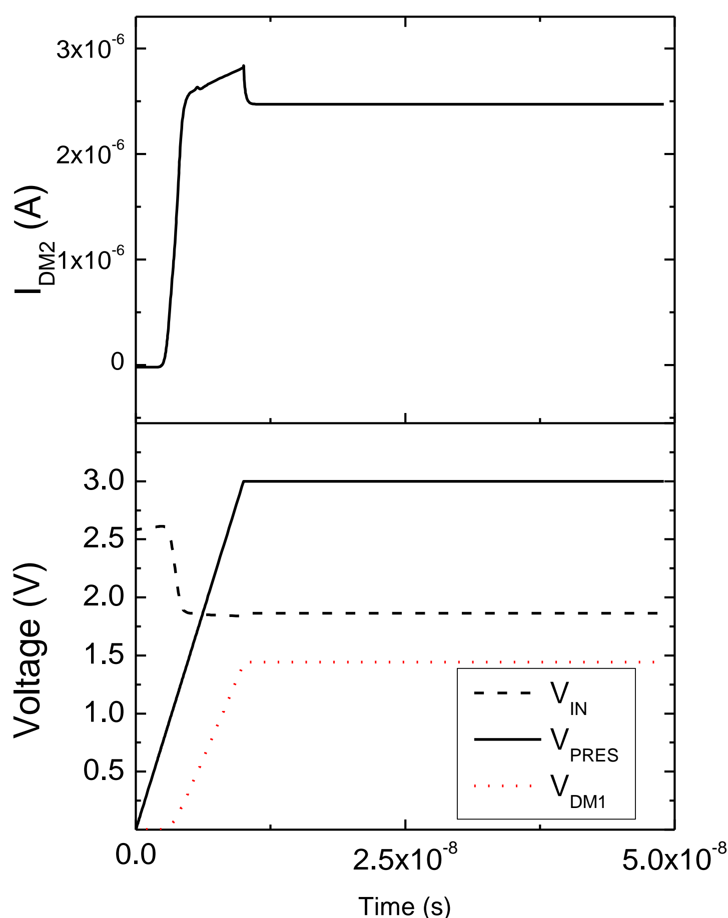

Figure 3.10 - Synapse node voltages and output current when V_W = 1.0V.

Simulation data can now be used to verify the claim that when M1 is above threshold, the total charge is a function of V_W and the pulse width ΔT. Figure 3.10 shows simulation plots of V_{PRES}, V_{IN}, V_{DM1} and I_{DM2} following the application of an input pulse to the V_{PRES} terminal, with V_W set to 1V (that is $V_W > V_T$). The initial negative undershoot of I_{DM2} is due to the capacitive coupling between the gate and source of M2. As V_{PRES} rises from 0V to 3V the output current increases and settles to 2.4uA, where it remains for the duration of the input pulse. The corresponding theoretical current for V_W = 1V is 2.67uA. The final values reached for V_{DM1} and V_{IN} are 1.86V and 1.44V respectively. The values for V_{IN} and V_{DM1} predicted by (3.8) and (3.14) at this point are 2.0V and 1.38V. Given the approximations of constant mobility and infinite output impedance have already been considered, the quality of the fit could be improved by the use of a more accurate expression for the substrate bias dependence of V_T. As previously stated, (3.10) does not take into account short channel effects, , or variations in the substrate doping profile. As a result, (3.10) overestimates the

dependence of V_T on V_{sub} and excludes a uniform positive shift and slight V_{DS} dependence caused by short channel effects [1]. Thus V_{DM1} is under-estimated for low values of V_W and overestimated for high values.

The case where M1 is operating below threshold is now considered. The theoretical analysis presented earlier predicted that the synapse output would consist of an initial current spike, followed a much smaller constant current for the duration of the input pulse. Simulation results for this case are shown Figure 3.11a, where V_W = 0.5V. As predicted, there is a current spike over the duration of the rise time (10ns), with a peak value of 3.2uA. The total amount of charge transferred does not vary significantly with the rise time. If the rise time is increased, the magnitude of the I_{DM2} current spike is proportionally reduced in magnitude and the overall charge remains the same. The results obtained from the theoretical model are presented in Figure 3.11b for comparison. The magnitude of the current spike is comparable, 3.6uA, but the duration is shortened, due to the simplified modelling of V_{DM1} with respect to time. As such, the theoretical model will underestimate the total amount of output charge from the synapse. Knowing the value of V_{DM1}, it is possible to extract the V_{GS} and V_T of M2 from the simulated data. These are plotted in Figure 3.12. It can be seen that over the period of the current spike, M2 is operating above threshold. After this, M2 operates ~300mV below the threshold voltage and the output current is less than 1nA.

The fabricated chips include a number of large area synapses, where the output n+ region is connected to a pad (parasitic capacitance of ~100fF) from which the synapse output current can be measured directly. The large synapse dimensions are 100µm x 0.6µm, 75µm x 0.6µm and 50µm x 0.6µm. Figure 3.13 shows plots of the synapse output current as a function of V_W for the three synapses sizes. The relationship between output current and weight voltage is approximately linear in the range $V_T < V_W < 1.8V$. Figure 3.14 gives a comparison of simulated and experimental transient results taken from the large synapses. The responses of the 100µm x 0.6µm synapse to an input pulse of width 10us and rise time 1us, is plotted for two values of V_W.

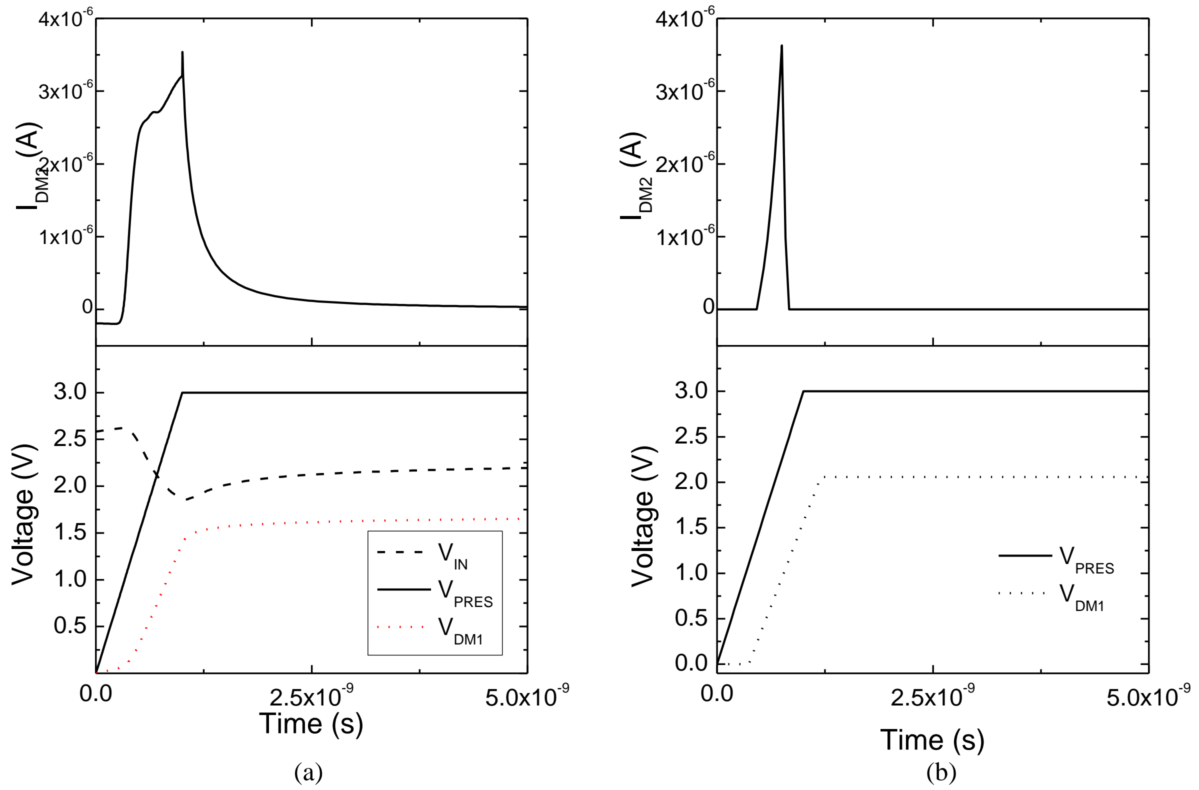

Figure 3.11 - Synapse node voltages and output current when V_W = 0.5V. A single current spike is seen at the output, after which I_{DM2} returns to ~1nA. Simulation results shown in (a), theoretical in (b).

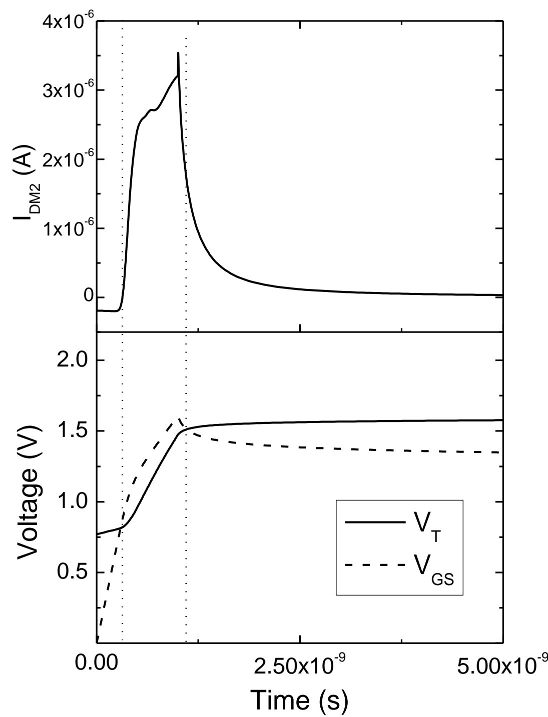

Figure 3.12 - V_{GS} and V_T of M2. The area between the dotted lines indicates the range over which a significant current flows.

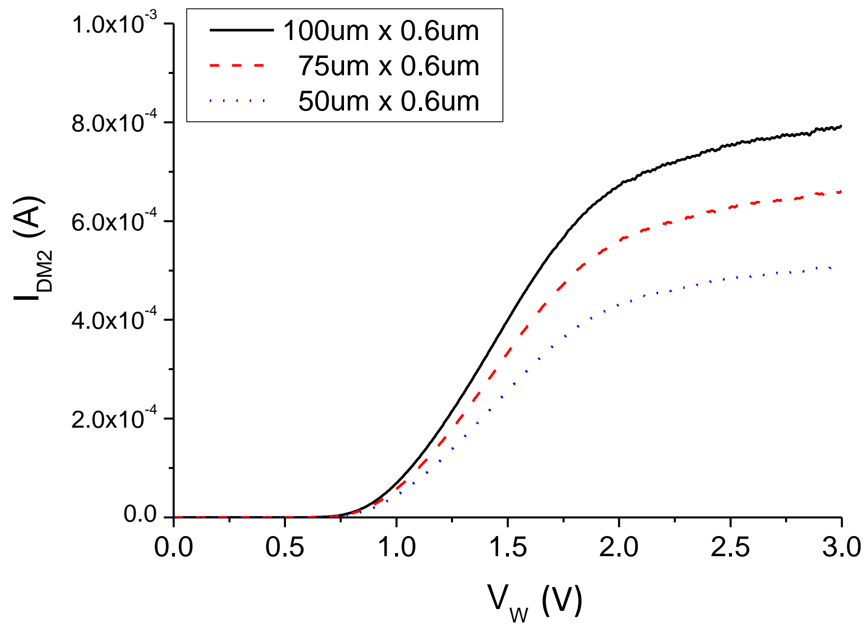

Figure 3.13 - Synaptic output current against Vw for large area synapses.

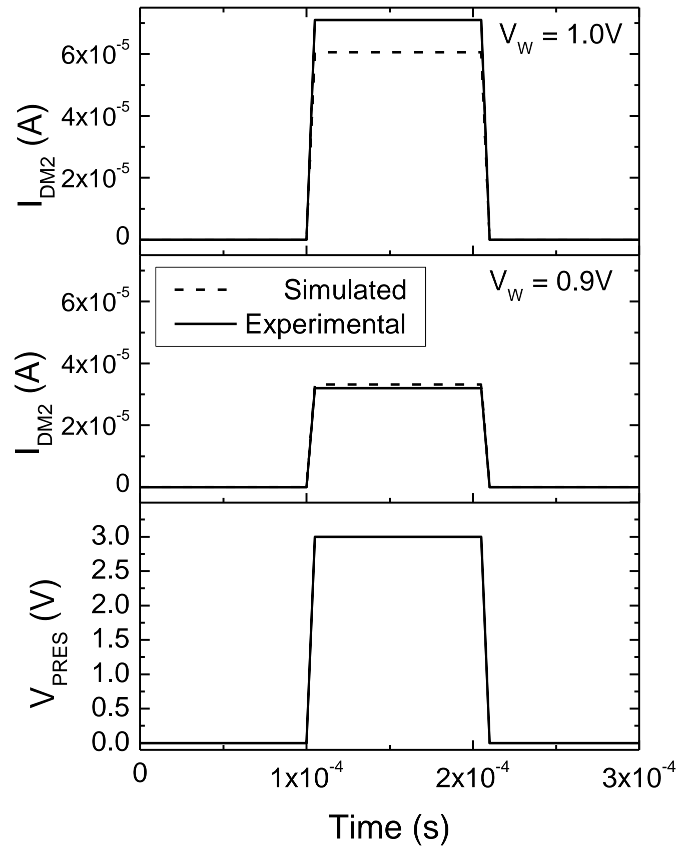

Figure 3.14 – Simulated and measured synapse output in response to a 10us input pulse with 1us rise/fall times. Measurements taken from 100µm x 0.6µm synapse.

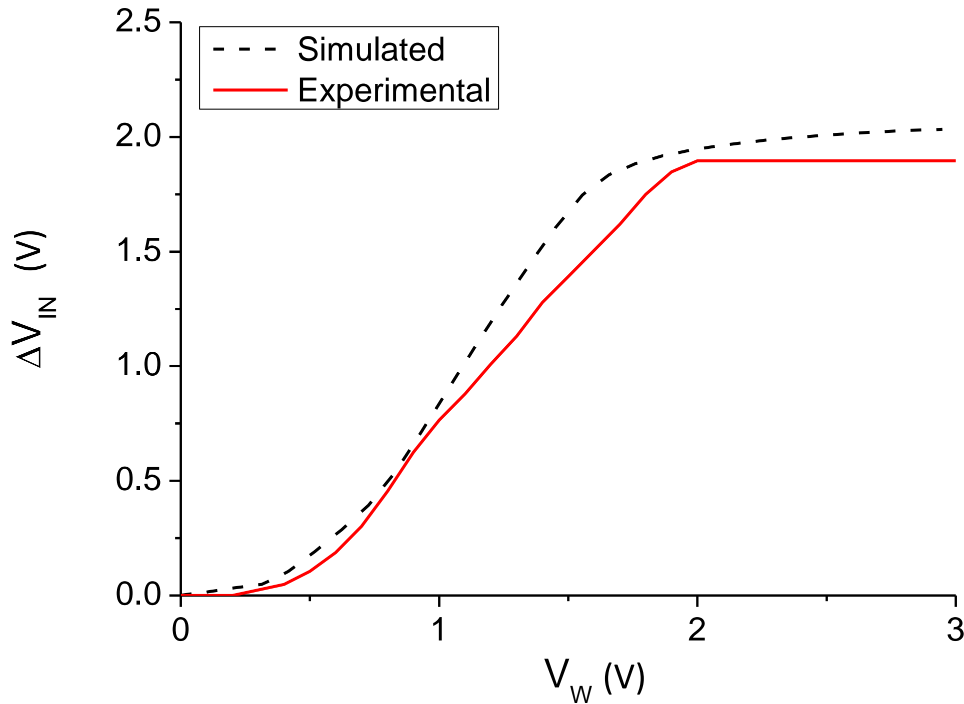

Figure 3.15 - Simulated and measured values of V_{IN}.

The combined synapse/neuron circuit of Figure 3.2 was included on the test chip, with an output pad for measuring V_{IN}. Figure 3.15 shows the simulated and measured values of ΔV_{IN}, which is the change in V_{IN} following the application of an input pulse to the V_{PRES} terminal, as a function of V_W. The levelling-off of the characteristics above ~2V is due to the fact that the V_{DS} of M2 ($V_{IN} - V_{DM1}$) decreases for increasing V_W. Since M2 operates in saturation for $V_W > 1.26V$, a point will eventually be reached where increases in V_W are offset by the decreased value of V_{DS}. In the linear portion of the graph, the slope of the experimental data is less than the slope of the simulated data. This indicates a difference in β values in either or both the n channel and p channel transistors The value of V_{IN} varies with the square root of β_n/β_p as given in (3.7).

Finally, the simulated and measured relaxation of the V_{IN} node is shown in Figure 3.16. A 3V voltage pulse with rise/fall times of 1ns and pulse width 100ns is applied to the V_{PRES} terminal at t = 0.2ms. The 10% - 90% rise/fall times of the simulated results are 368μs and 19μs; for the experimental results they are 604μs and 32μs. This agrees with the theoretical analysis which predicted that the rise time of V_{IN} will be of the order of hundreds of microseconds for a fan-in of 1.

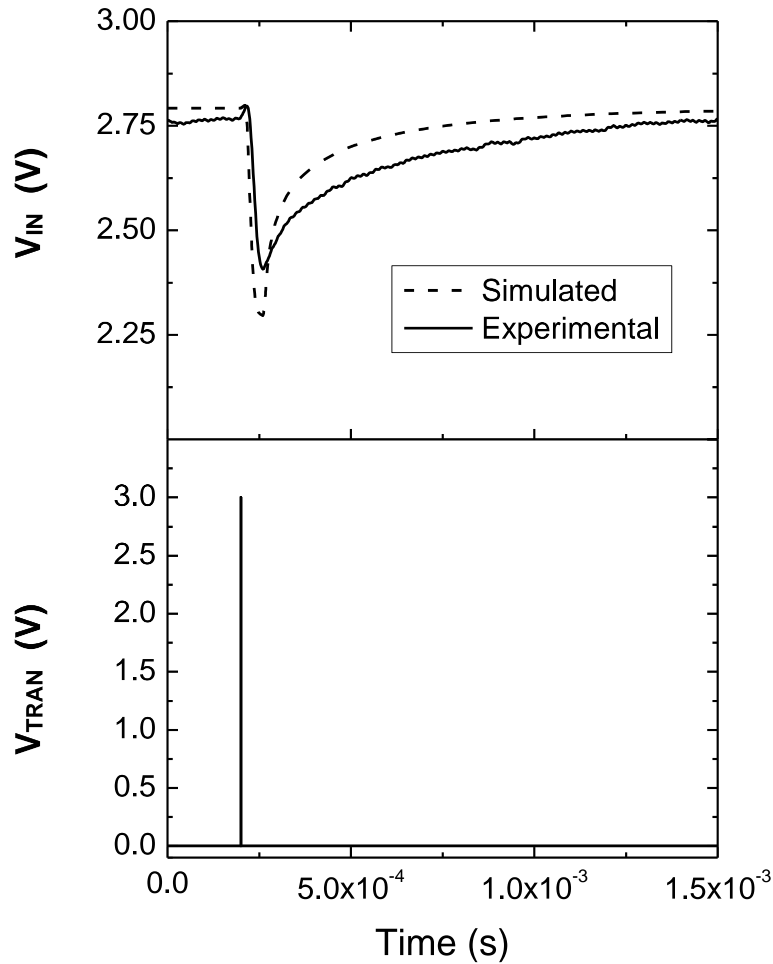

Figure 3.16 - Relaxation of the V_{IN} node following an input pulse.

The difference between the two rise times, ~40%, is equivalent to an additional parasitic capacitance on the fabricated chip of approximately 1.2fF, based upon the estimated node capacitance in (3.27).

3.3 The three-terminal dynamic synapse

A more advanced synapse design is discussed in this section, where an additional terminal is added to increase the functionality of the device. Figure 3.17 shows the device, with a schematic view shown in Figure 3.18. An additional control gate is added, the purpose of which is to set the rate at which the charge under the weighted gate is replenished following the application of an input pulse to the V_{PRES} terminal.

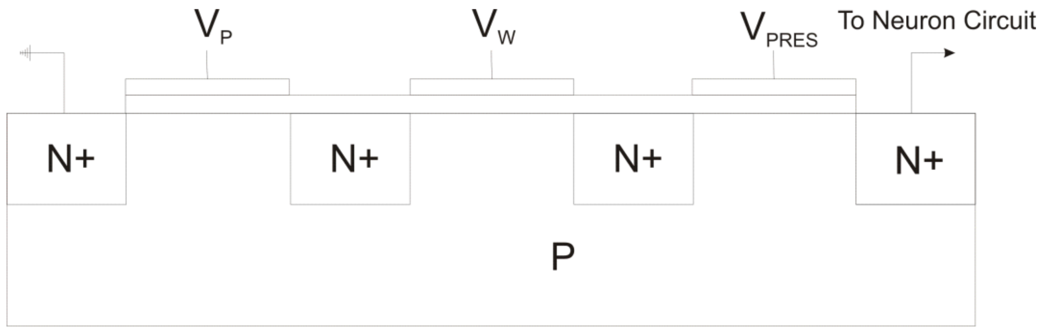

Figure 3.17 - Three gate synapse

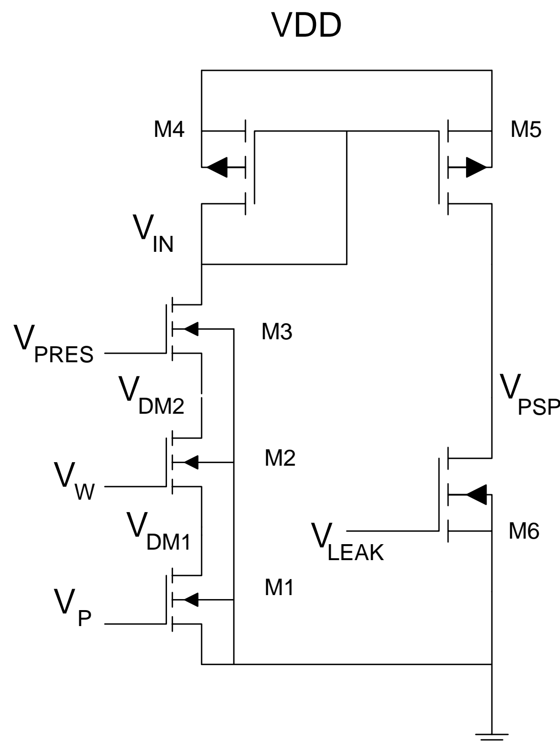

Figure 3.18 - Schematic of three gate synapse and neuron circuit.

This effect implements synaptic depression. The level of depression can be controlled by adjusting either the voltage V_P or the inter spike interval (ISI) between consecutive input pulses.

3.3.1 Theory

The theoretical analysis provided for the two gate synapse can be expanded to cover the three gate synapse, as the basic operation of the device is essentially the same. Following a V_{PRES} pulse, the voltages V_{DM1}, V_{DM2} and V_{IN} will settle to values such that:

$$I_{DM1}(V_P) = I_{DM2}(V\text{-}_W) = \ I_{DM3} = I_{DM4} \tag{3.37}$$

If $V_P > 0.8V$, M1 will operate above threshold and the operation of the device is essentially the same as for the two terminal synapse. When $V_W > 0.8V$, the output charge is given as $I_{DM3}\Delta T$. When $V_W < 0.8V$, the output charge becomes independent of the pulse width as the substrate bias effect causes M3 to operate subthreshold after V_{PRES} reaches its final value.

The operation of M1 in subthreshold will have the same effect as operating M2 in subthreshold. To satisfy (3.37) the steady state node voltages will be such that M2, M3 and M4 will also operate in subthreshold. The consequence of this is that, when $V_P < 0.8V$, the output charge, Q_W is independent of the pulse width for all values of V_W. From the scaling point of view, this is advantageous, as it increases the consistency between different synapses, despite possible variations in input pulse width.

The depressing action of the synapse can now be considered. The application of an input pulse transfers the charge from underneath the gate of M2 to the output node. In order for charge neutrality to be maintained in M2, the depletion region expands further into the bulk of the device, a condition known as deep-depletion. The weight charge is replenished through M1, at a rate set by V_P and this is illustrated in Figure 3.19. The time required for the weight charge to be completely replenished is labelled τ_s. If the ISI is greater than τ_s, then consecutive synaptic outputs will be of the same magnitude. If the ISI is less than τ_s, then the magnitude of the output will reduce following the first input pulse, as illustrated in Figure 3.20. An expression for τ_s can be found by equating the charge components in M2:

$$Q_g + Q_d + Q_{inv} = 0 \tag{3.38}$$

Q_g, Q_d and Q_{inv} are the gate, depletion and inversion layer charges respectively of M2. Substituting values gives:

$$C_0(V_W - \varphi_s) + qN_AW_d + Q_{inv} = 0 \tag{3.39}$$

Differentiating (3.39) with respect to time gives:

$$-C_0\frac{d\varphi_s}{dt} + qN_A\frac{dW_d}{dt} + \frac{dQ_{inv}}{dt} = 0 \tag{3.40}$$

φ_s can be expressed as:

$$\varphi_s = \frac{qN_AW_d^2}{2\epsilon_{si}\epsilon_0} \tag{3.41}$$

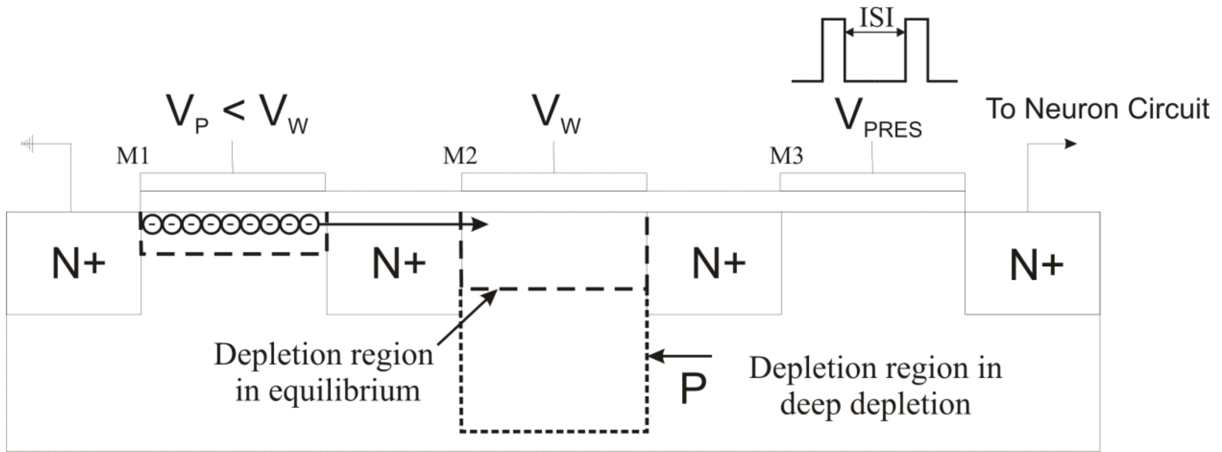

Figure 3.19 - Between input pulses, the weight charge is replenished through M1.

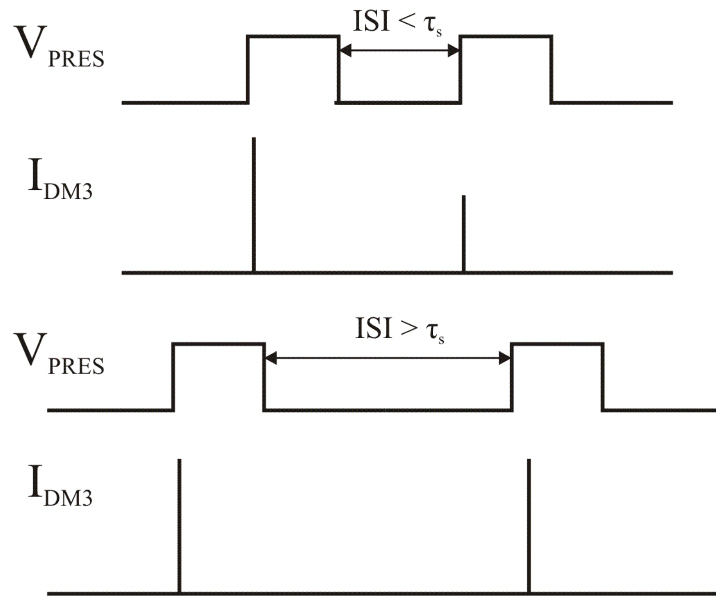

Figure 3.20 - Illustration of the depressing effect. When the ISI < τ_s, the charge recovery is incomplete.

Differentiating both sides gives:

$$\frac{\mathrm{d}\varphi_s}{\mathrm{dt}} = \frac{qN_A W_d}{\epsilon_{si}\epsilon_0}\frac{dW_d}{dt} \tag{3.42}$$

Also:

$$\frac{dQ_{inv}}{dt} = \frac{I_{DM1}}{W\delta} \tag{3.43}$$

Where W is the width of the gate of M1 and δ is the channel depth, assumed to be 5nm [2]. Rewriting (3.40):

$$-C_0 \frac{qN_A W_d}{\epsilon_{si}\epsilon_0}\frac{dW_d}{dt} + qN_A \frac{dW_d}{dt} + \frac{I_{DM1}}{W\delta} = 0 \tag{3.44}$$

Rearranging and separating variables gives:

$$\left[1 - C_0 \frac{W_d}{\epsilon_{si}\epsilon_0}\right] dW_d = -\frac{I_{DM1}}{W\delta}\frac{1}{qN_A} dt \tag{3.45}$$

Initially, W_d will be equal to the deep depletion width, W_{do}, which is given by:

$$W_{do} = \sqrt{\frac{2\epsilon_0 \epsilon_{si} \varphi_{do}}{qN_A}} \tag{3.46}$$

The surface potential in deep depletion, φ_{do}, can be found from the quadratic equation relating gate voltage to surface potential:

$$V_G = \frac{\sqrt{2qN_A \epsilon_{si}\epsilon_0 \varphi_s}}{C_o} + \varphi_s \tag{3.47}$$

With V_G = 3V and substituting values of other parameters, listed in Table 3.1, a value of φ_{do} = 1.52V is obtained by solving the quadratic equation, (3.47). At equilibrium, the depletion width, W_{df}, is approximately equal to the depletion width at inversion:

$$W_{df} = \sqrt{\frac{4\epsilon_0 \epsilon_{si} \varphi_b}{qN_A}} \tag{3.48}$$

Integrating (3.45) between W_{do} and W_{df} gives:

$$\left[W_d - C_0 \frac{W_d^2}{2\epsilon_{si}\epsilon_0}\right]_{W_{do}}^{W_{df}} = -\frac{I_{DM1}}{W\delta}\frac{1}{qN_A}\tau_s \tag{3.49}$$

Substituting the boundary conditions and rearranging for τ_s gives:

$$\tau_s = -\frac{W\delta qN_A}{I_{DM1}}\left[W_{df} - W_{do} + \frac{C_0}{2\epsilon_{si}\epsilon_0}(W_{do}^2 - W_{df}^2)\right] \tag{3.50}$$

When M1 is operating in subthreshold, τ_s can be expressed as a function of V_P:

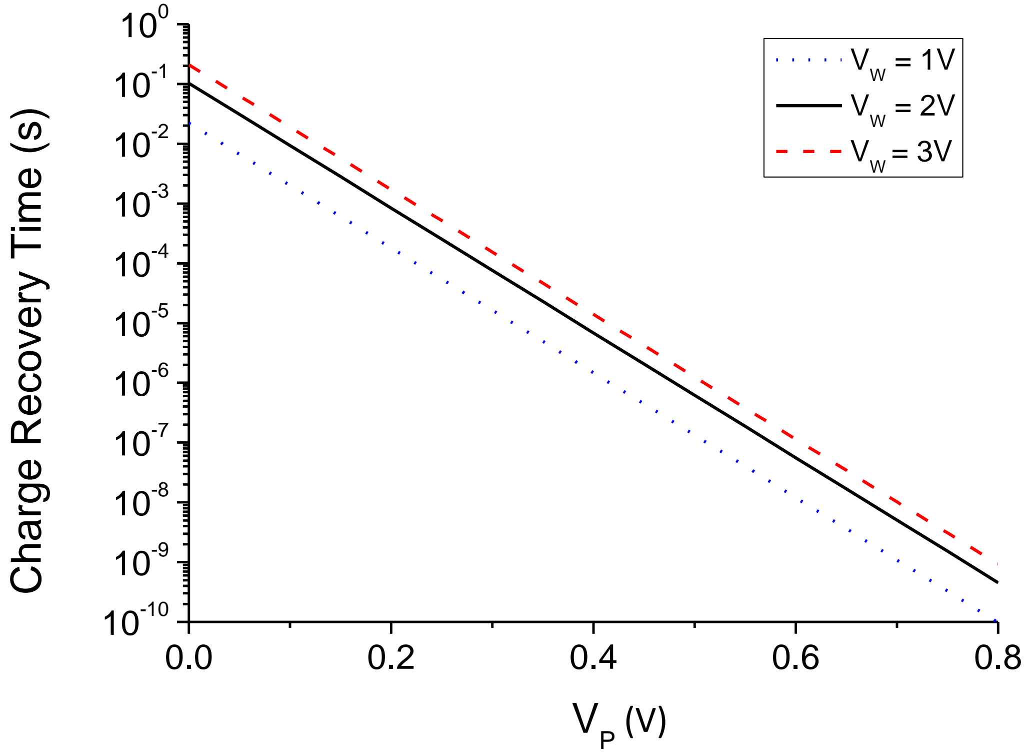

Figure 3.21 - Charge recovery time as a function of V_P, for V_W = 1V, 2V and 3V.

$$\tau_s = -\frac{W\delta q N_A}{I_{0n}\exp\left(\frac{qV_P}{m_n kT}\right)}\left[W_{df} - W_{do} + \frac{C_0}{2\epsilon_{si}\epsilon_0}(W_{do}^2 - W_{df}^2)\right] \tag{3.51}$$

Figure 3.21 shows plots of τ_s as a function of V_P for V_W = 1V, 2V and 3V.

3.3.2 Simulation and Experimental Results

In this section, results are presented which demonstrate that the updated synapse circuit can implement synaptic depression. Simulations results are again obtained using Cadence. The synapse/neuron circuit of Figure 3.18 was fabricated to provide experimental results. An output pad was again used to measure the value of V_{IN}. For the following results, transistor dimensions are M1,M2, M3 (0.8µm x 0.6µm), M4, M5, M6 (0.35µm x 0.35µm). The simulation results of Figure 3.22 show the depressing effect. With V_P set to 0.4V, input pulses are applied every 0.2ms. In this case, the ISI is insufficient to fully replenish the weight charge.

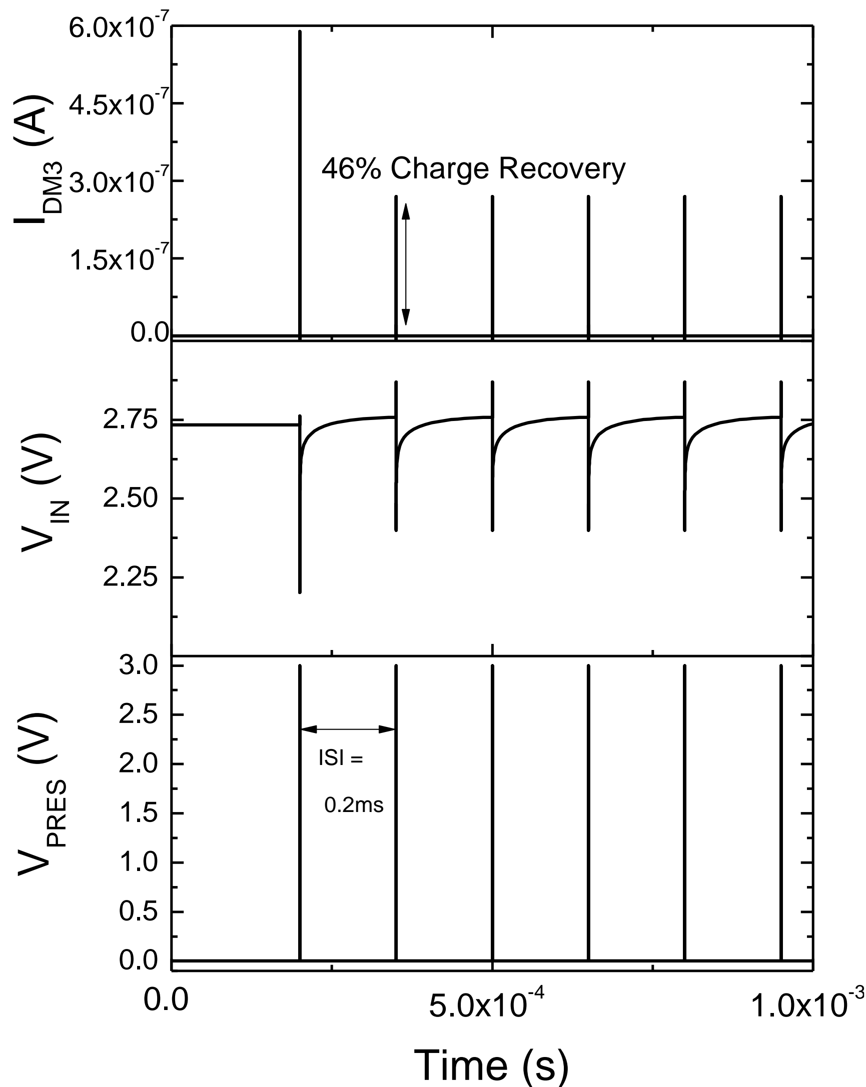

Figure 3.22 - Simulation results showing effect of too short an ISI on the synapse output. V_P = 0.4V.

The first output pulse has a magnitude of 0.59uA, each subsequent pulse has a peak output current of 0.27uA. The depressing action also affects ΔV_{IN}. For the first pulse, ΔV_{IN} = 0.56V, for each subsequent pulse, ΔV_{IN} = 0.35V. As the ISI is increased, the amount of charge replenished increases until total charge recovery is achieved. Figure 3.23 shows V_{IN} and I_{DM3} for an ISI of 1.5ms, which is greater than the minimum required value. It can clearly be seen that for both input pulses, ΔV_{IN} and I_{DM3} take the same values.

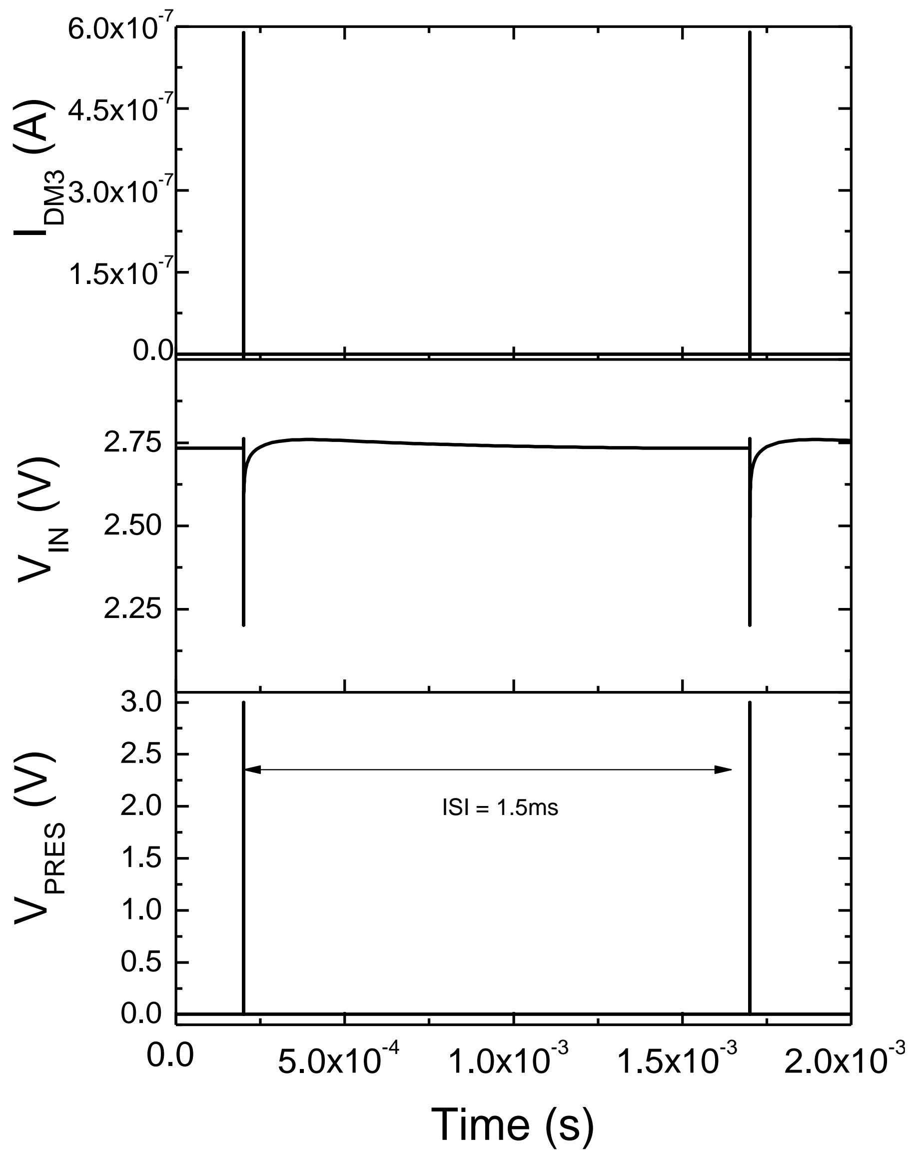

Figure 3.23 - Simulation showing full charge recovery for V_P = 0.4V.

Using the fabricated test circuit, it is only possible to measure the value of V_{IN}. The current through the synapse cannot be directly measured, but the charge recovery time can be measured by observing ΔV_{IN}. Experimental results are plotted in Figure 3.24. Input pulses with ISIs of 1.5ms and 3ms were applied to the circuit and V_{IN} was recorded. In this case V_P is set to 0.2V. As was observed for the simulated results, a sufficiently small ISI results in a reduced ΔV_{IN} for subsequent inputs. A 3ms ISI is required for full charge recovery.

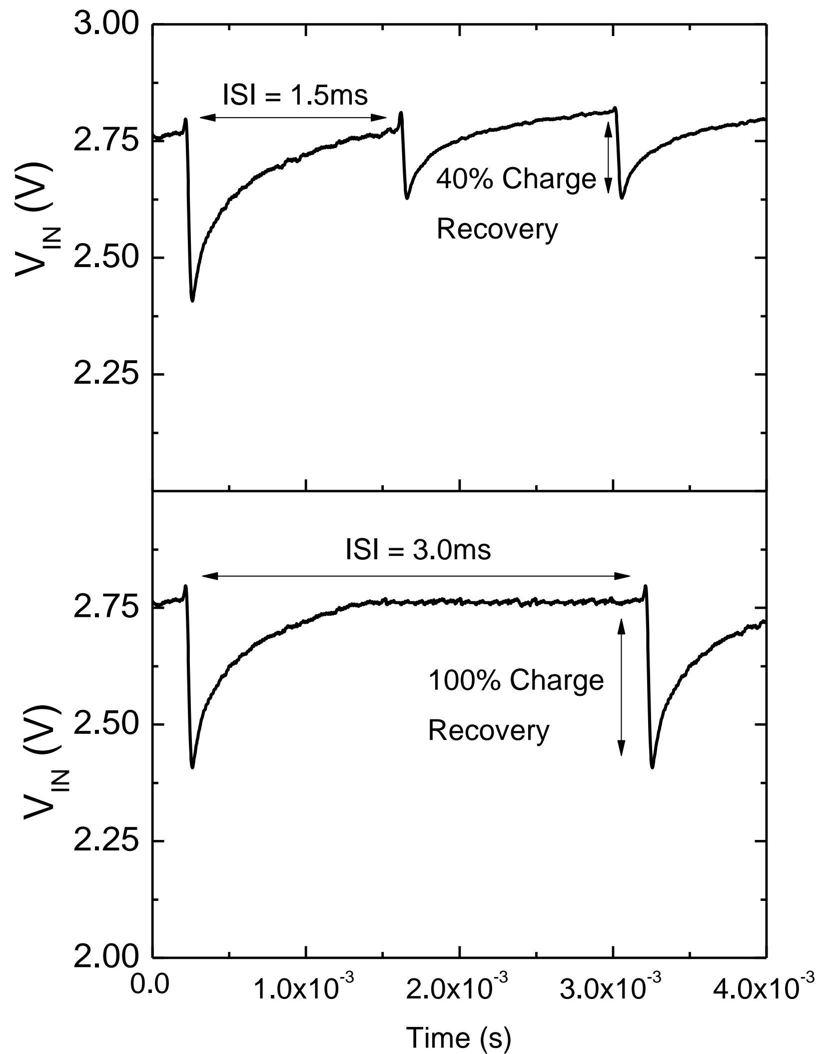

Figure 3.24 - Experimental results showing partial and full charge recovery . V_P = 0.2V

Repeating the measurements described above for different values of V_P yields the plot of Figure 3.25a, where the charge recovery time is plotted against V_P. The results show that the synapse is able to implement a depressing action with recovery times spanning more than four orders of magnitude, from tens of milliseconds to fractions of a microsecond. This effect also provides the synapse with an adjustable refractory period, over which the synapse output is greatly reduced.

There is some discrepancy between the experimental, theoretical and simulation results. The gradients of the simulated and experimental sets of results are approximately equal, but there is a linear shift along the x-axis between them. This indicates a threshold voltage difference between the fabricated chips and the SPICE model used for simulation of ~80mV.

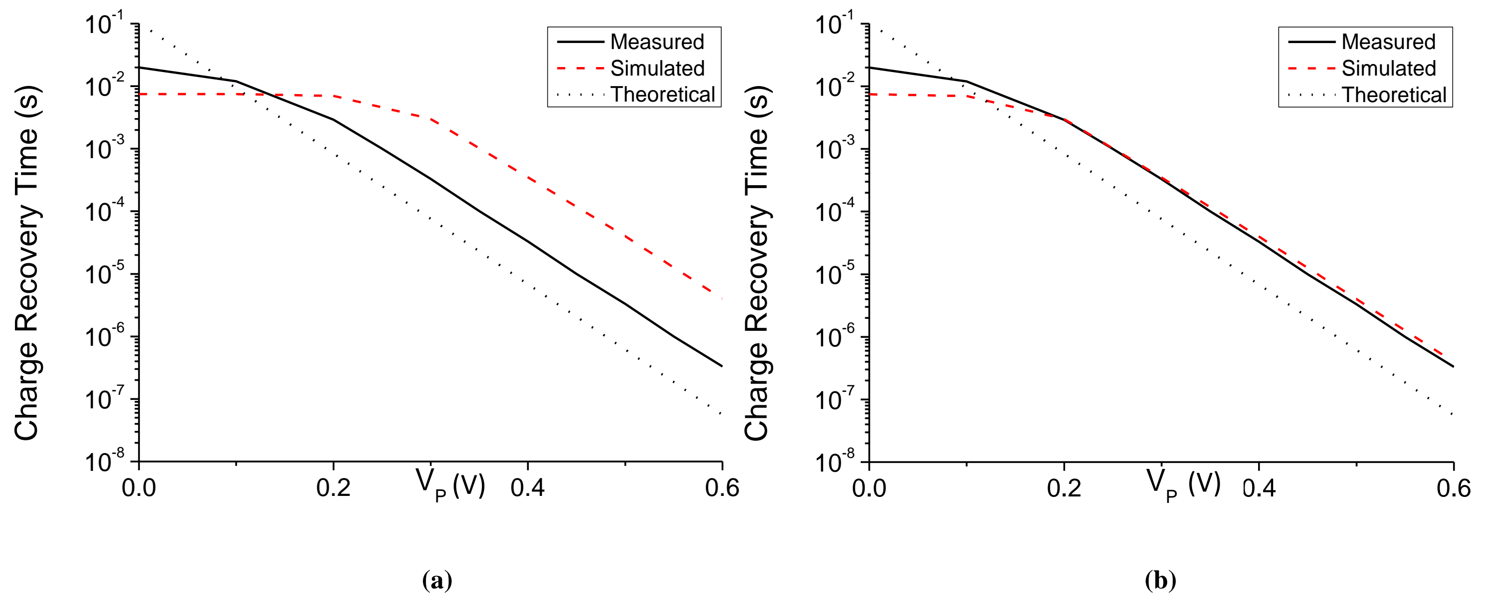

Figure 3.25 - Charge recovery time against V_P, V_W = 2V. Simulated, theoretical and experimentally measured results are plotted. Original simulation results are plotted in (a), results shifted by 80mV are shown in (b).

As explained in Chapter 2, MOSFET threshold voltages can vary from the typical values by up to 100mV, so such variations are within the expected limits. Figure 3.25b plots the results where the simulation data has been shifted along the x-axis to account for the perceived difference in threshold voltages. In both cases, the theoretical values are lower than the measured and simulated values, by roughly three quarters of one order of magnitude.

For a constant value of V_P, the charge recovery can be plotted against time by measuring the amount of charge recovered for increasing values of ISI. This is plotted in Figure 3.26, for V_P = 0.1V. The simulated and experimental recovery times are 7.5ms and 12ms, consistent with the values plotted in Figure 3.25a. Table 3.2 compares the depressing action of the silicon synapse with that of a biological synapse. The silicon device is able to introduce a greater degree of charge reduction, while maintaining the variability in recovery time.

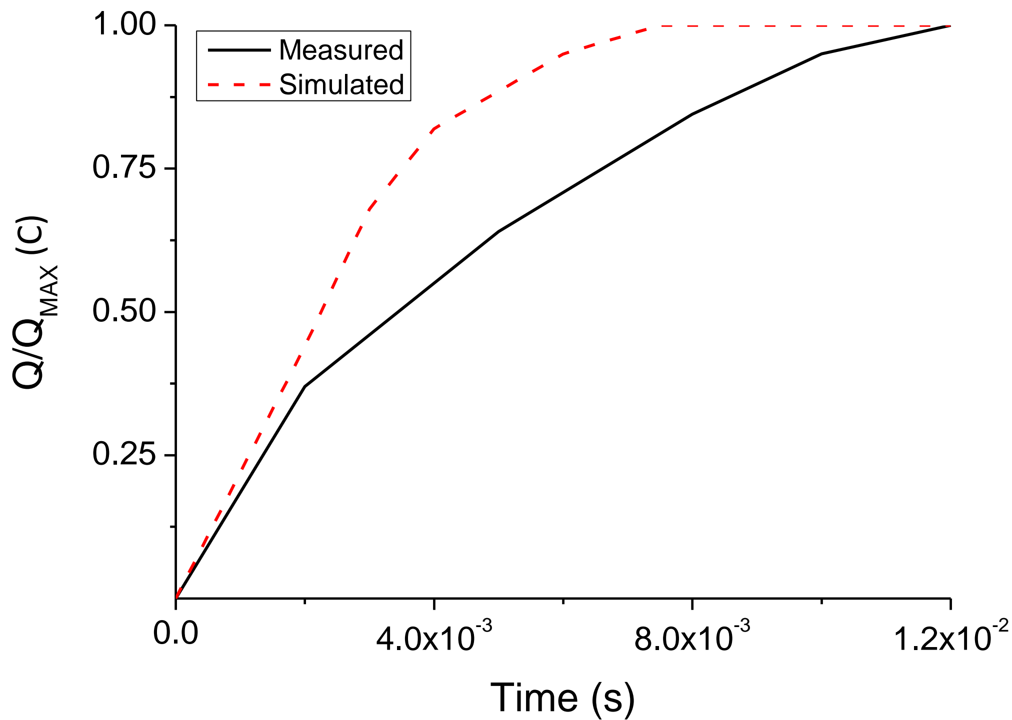

Figure 3.26 – Normalised charge recovery against time. V_P = 0.1V V_W = 2V.

	Silicon Synapse	**Biological Synapse [3]**
Magnitude of charge reduction	0 -100%	0 – 15%
Recovery time	0.5us – 10ms (~4 orders of magnitude)	100ms – 10 minutes (~4 order of magnitude)

Table 3.2 - Comparison of depression between silicon and biological synapses.

3.4 Conclusions and Discussion

In this chapter, implementations of static and dynamic silicon synapses have been described. Both types of synapseare compact, with layout dimensions of 2.1µm x 5.0µm (static) and 2.1µm x 6.2µm (dynamic) when fabricated in a 0.35µm process. When compared to the synapse devices considered in Chapter 1, Section 1.3.2, the dynamic synapse requires only a quarter of the area of the smallest equivalent device reported in the literature. The power

consumption of an individual synapse can be considered as low, with a V_{DD} of 3V, transient operating currents of several uA and resting currents in the nA range.

Synaptic plasticity is implemented through a weight voltage, V_W, which controls the amount of charge delivered by the synapse. For the two terminal static synapse, the total weight charge can also be controlled by varying the pulse width of input signals. A potential downside to this is that for large scale implementations, it may be necessary to include additional circuitry to regulate the pulse width, as unwanted variations could adversely affect the overall performance of the network.

The more advanced, three-terminal dynamic synapse is not affected by variations in pulse width. The inclusion of an additional control gate, biased in subthreshold renders the weight charge independent of pulse width. The main function of the additional gate is to implement synaptic depression by controlling the rate of recovery of the weight charge. Recovery times spanning four orders of magnitude are achievable, through adjustments to the value of the control voltage V_P. This range of values is comparable to that seen in biological systems. This depressing mechanism can also be viewed as an implementation of a refractory period. After the synapse has fired, there is a period of time, dependent on V_P, over which the synapse output is less than 1% of its original value.

These two devices provide an effective, compact implementation of the most ubiquitous element in biological neural systems – the synapse. Several of the key processes of biological synapses are successfully implemented. In the following chapter, it is shown how the synapse integrates with the neuron circuit to create a compact, functional neural block, capable of producing biologically plausible PSPs.

References

[1] M. C. Yuhua Cheng, Kelvin Hui, Min-chie Jeng,, J. H. Zhihong Liu, Kai Chen, James Chen, Robert Tu,, and C. H. Ping K. Ko, BSIM3v3 Manual: University of California, Berkeley, 1996.

[2] S. M. Sze, Semiconductor Devices: Physics and Technology, 2nd ed.: John Wiley & Sons, 2001.

[3] W. E. G. Gerald M. Edelman, W. Maxwell Cowan, Synaptic Function: Wiley-Interscience, 1987.

Chapter 4: Neuron Circuit

4.1 Introduction

In this chapter, the circuit used to implement hardware neuron cells are presented. The neuron can take inputs from the synapse devices described in the previous chapter, and replicate the functionality of biological neurons, based upon the Leaky Integrate and Fire (LIF) model. The operation of the neuron with both static (two terminal) and dynamic (three terminal) synapses is considered. A single neuron receives signals from *n* presynaptic neurons via a series of *n* synapses ($S_1 - S_n$). The firing of a presynaptic neuron generates a weighted, time decaying voltage change in the cell body of the neuron, which is referred to as a post synaptic potential (PSP). The shape of a PSP is typically characterised by a fall time which is much greater than the rise time. As multiple synaptic signals are summed in the cell body, the magnitude of the PSP may exceed the neuronal threshold voltage. When this happens, a signal known as an action potential is generated and propagated to postsynaptic neurons along axons. While the shape of individual PSPs may vary between neurons, the action potential is thought to be invariant, as such it can be considered to be a 'binary' event [1].

The neuron circuit and the generation of PSPs are described in Section 4.2; experimental and simulated results for static and dynamic synapses are presented in Section 4.3. Triggering circuitry is considered in Section 4.4 and a possible method of connecting inhibitory synapses to the neuron is discussed in Section 4.5. Discussions and conclusions are given in Section 4.6.

4.2 Theory of operation

The neuron circuit, with connected static synapse, is reprinted for convenience in Figure 4.1. The circuit consists of a current mirror (transistors M4/M5) and a leakage transistor, M6, which regulates the fall time of the PSP. Synaptic currents in the drain of M4 are mirrored through M5 to the V_{PSP} node, where they induce a change in voltage, ΔV_{PSP}. In order to fully describe the PSP, it is necessary to know the magnitude of ΔV_{PSP} and the rise/fall times.

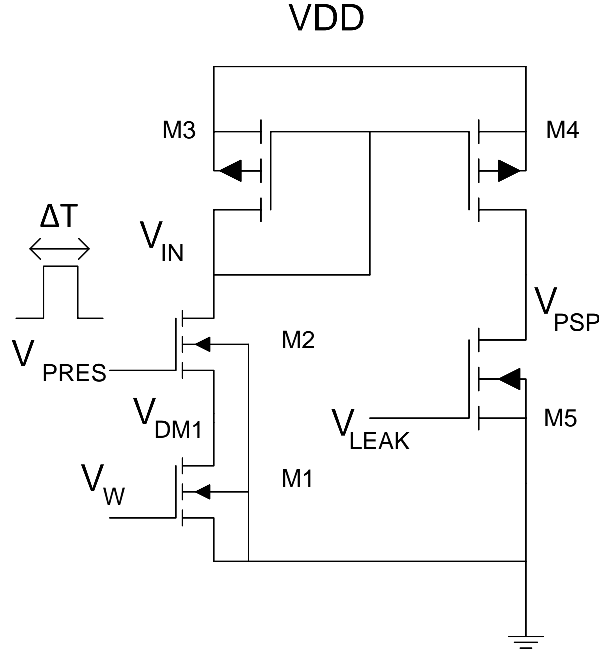

Figure 4.1 - Neuron circuit with two terminal (static) synapse.

4.2.1 Derivation of ΔV_{PSP}

The change in voltage of V_{PSP} in response to a synaptic input can be expressed as:

$$\Delta V_{PSP} = \frac{Q_W}{C_{PSP}} \tag{4.1}$$

where C_{PSP} is the voltage dependent capacitance associated with the V_{PSP} node and Q_W is the synaptic output charge, derived in Chapter 2:

$$Q_w(V_W) = \frac{\beta_n}{2}(V_W - V_{Tn})^2 \Delta T \tag{4.2}$$

ΔT is the width of the V_{PRES} pulse and V_W is the weight voltage. C_{PSP} consists of the capacitance associated with the drain of M5 and the parasitic capacitances associated with the node:

$$C_{PSP} = \sqrt{\frac{qN_A\epsilon_{si}\epsilon_0}{2(\varphi_b + V_{PSP})}}A + C_P \tag{4.3}$$

With V_{PSP} initially 0V and C_P estimated to be 2fF, from a back-annotated simulation. C_{PSP} = 3fF.

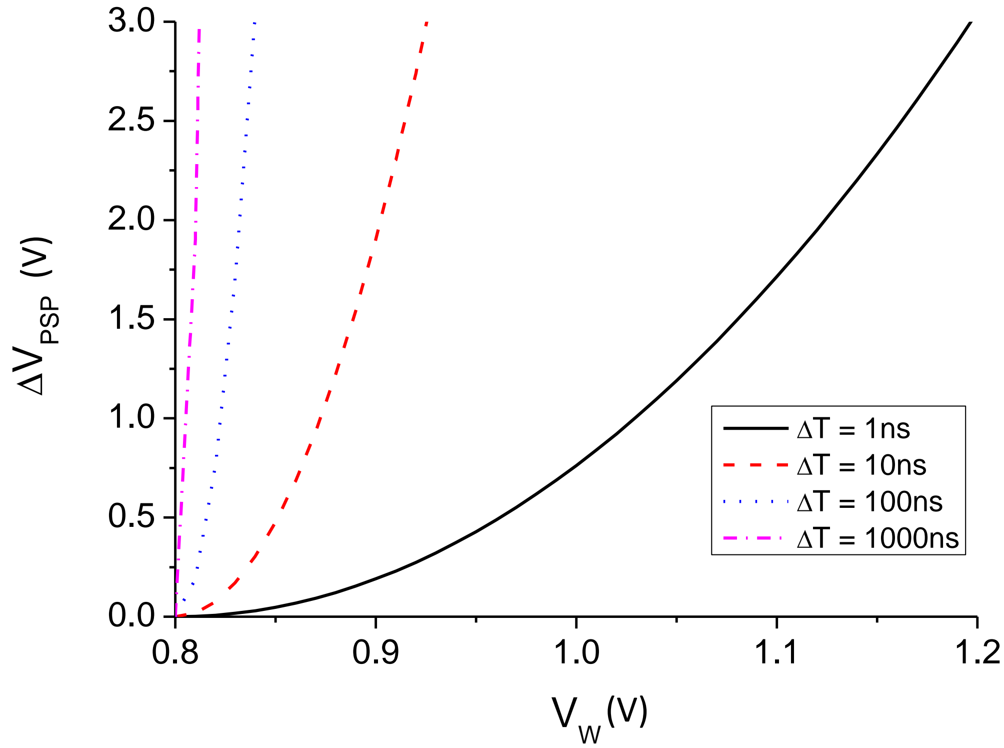

Figure 4.2 - V_{PSP} against V_W. Input pulse widths are indicated. V_{DD} = 3V.

It is assumed that the mirrored current, $I_{M4} = I_{M3}$. In practice the mirroring action is affected by differences in the drain-source voltages of M3 and M4 and the λ parameter, extracted in Chapter 2. The fabricated chips used for taking measurements have current mirror transistors with short channel lengths, 0.35µm, and a λ value of $90mV^{-1}$. For further iterations, devices with longer channels would form a more effective current mirror, due to their flatter output characteristics and lower value of λ, $11.2mV^{-1}$ when L = 3.5µm. The maximum difference between I_{M3} and I_{M4} will occur when there is the greatest difference between the drain source voltages of the two devices, and can be calculated using the expression for the drain current of a MOSFET in saturation:

$$I_D = \mu C_0 \frac{W}{2L}(V_{GS} - V_T)^2(1 + \lambda V_{DS}) \tag{4.4}$$

This corresponds to V_{PSP} = 0V and V_{IN} = 2.3V ($V_{GS} = V_T$). For short channel devices, where $\lambda = 90mV^{-1}$, this gives $I_{DM4} = 1.20I_{DM3}$. The long channel value of $\lambda = 11.2mV^{-1}$ gives $I_{DM4} = 1.03I_{DM3}$. To maintain the correct current levels, it would be necessary to also increase the widths of M3 and M4 to match the length. In reality, 100% accuracy may not be required, as discussed further in Chapter 6.

Another assumption made for the purposes of the analysis is that the leakage transistor M5 will have no effect on the initial value of ΔV_{PSP}. The total charge Q_W will in fact be reduced by a small amount due to a constant leakage current through M5. This introduces a slight dependence of ΔV_{PSP} on the value of V_{LEAK}, which will shift the ΔV_{PSP} vs. V_W curve along the V_W-axis by a fixed amount in the positive direction. If V_{LEAK} is < 0.4V, then the effect is negligible. The shift at V_W = 0.55V (V_T) is 21mV.

Figure 4.2 plots ΔV_{PSP} for ΔT = 1ns, 10ns, 100ns and 1000ns. The maximum value V_{PSP} can take is V_{DD}, which is set to 3V. As ΔT increases, it can be seen that the operational voltage range decreases, making ΔV_{PSP} more sensitive to changes in V_W. If necessary, it would be possible to introduce an additional capacitive loading onto the node, which would allow the dynamic range of the device to be increased. This is considered in more detail in Chapter 6.

4.2.2 Rise Time

The rising portion of V_{PSP} is controlled by the current ($I_{M4} - I_{M5}$). The initial value of V_{IN} following the application of an input pulse was found in Chapter 3, (3.1) to (3.8), to be given by:

$$V_{IN} = 3.51 - 1.51V_W \tag{4.5}$$

V_{IN} remains at this value for the duration of the input pulse, ΔT, after which it will return to its resting value in the manner described in Section 3.2.1.2. The value of V_{PSP} will increase while $I_{M5} > I_{M4}$. For a given value of V_{LEAK}, the rise time can be estimated by finding the voltage at which $I_{M4} = I_{M5}$, V_R:

$$V_R = m_p \frac{kT}{q} ln\left(\frac{I_{0n}}{I_{0p}}\right) + \frac{\mathrm{m}_p}{m_n} \mathrm{V}_{LEAK} \tag{4.6}$$

The rise time of V_{PSP} will be equal to the time taken for the voltage at the V_{IN} node to reach V_R. Having previously developed expression for the rise time of the V_{IN} node, in Chapter 3, section 3.2.1.2 the rise time of V_{PSP} is given by:

$$t_{rPSP} = \Delta T + \tau_{r1} + \tau_{r2} \tag{4.7}$$

$$\tau_{r1} = C_{VIN}\frac{m_p kT}{I_{0p}q}\left[\exp\left(-\frac{qV_R}{m_p kT}\right) - \exp\left(-\frac{qV_{Tp}}{m_p kT}\right)\right] \tag{4.8}$$

$$\tau_{r2} = \frac{2LC_{VIN}}{3\mu C_{0p}W}\left(V_{IN} - V_{Tp}\right)^3 \tag{4.9}$$

It is assumed that V_R will be such that M5 is operating in subthreshold. For values of $V_{LEAK} > 0.45V$, the corresponding value of V_R is above threshold and τ_{r1} can be disregarded. Values obtained from (4.7) are plotted in the results section. Once V_{PSP} has reached its final value, the voltage at the V_{IN} node continues to rise until it reaches its resting potential of ~0.25V.

4.2.3 Fall Time

During the falling portion of the V_{PSP} waveform, $I_{M5} > I_{M4}$. In order for the fall time to be dependent only on the value of V_{LEAK}, it is necessary to set V_{LEAK} to a minimum value such that $I_{M5} >> I_{M4}$. To satisfy this, $I_{M5} = 10I_{M4}$ is taken as a minimum requirement. Taking into account the difference in subthreshold parameters between the nMOST and pMOST devices, setting $V_{LEAK} > 0.3V$ satisfies the criteria. Given this, the rate at which charge leaks away from the V_{PSP} node can be expressed solely as a function of V_{LEAK}.

$$I_{M5} = I_{0n}\exp\left(\frac{qV_{LEAK}}{m_n kT}\right) \tag{4.10}$$

The fall time can be calculated according to:

$$t_{fPSP} = \frac{Q_W}{I_{M5}} \tag{4.11}$$

The fall time is more conveniently expressed as a function of ΔV_{PSP}:

$$t_{fPSP} = \frac{\Delta V_{PSP} C_{PSP}}{I_{M5}} \tag{4.12}$$

Figure 4.3 plots (4.12) for V_{LEAK} = 0.35V, 0.40V and 0.45V. Figure 4.4 plots the fall time against V_{LEAK} for a constant ΔV_{PSP} of 1V. It is estimated that the fall time can be set between several milliseconds and tens of nanoseconds. Biological neurons are observed to have fall times in the order of milliseconds.

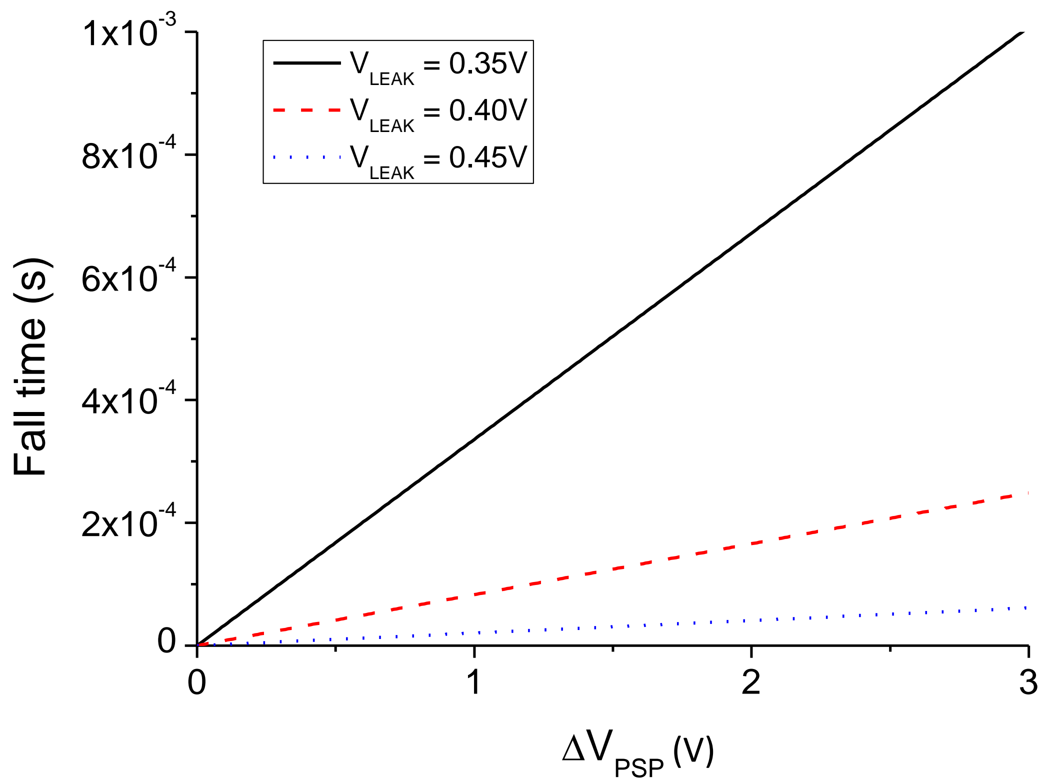

Figure 4.3 - Fall time against ΔV_{PSP} for different values of V_{LEAK}.

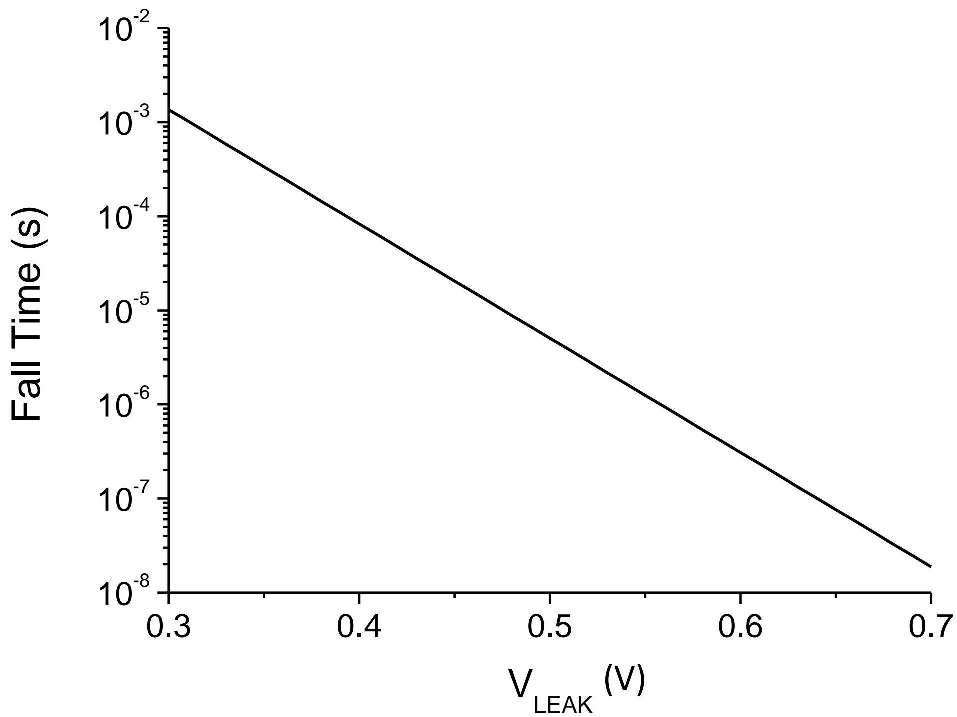

Figure 4.4 - Fall time against V_{LEAK} for $\Delta V_{PSP} = 1V$.

4.3 Results

In this section, the characteristics of PSPs obtained from simulations and measured from fabricated chips are presented and compared with the theoretical predictions. Individual PSPs are examined to confirm the biological plausibility of the neuron circuit.

4.3.1 Two terminal synapse

The results obtained in the previous section describing the operation of the neuron cell can be compared against simulated and experimentally obtained values. Figure 4.5 plots the value of ΔV_{PSP} against V_W for each of the different methods. For comparison, the dynamic range, midpoints (V_W at $\Delta V_{PSP} = 1.5V$) and percentage deviation from the measured values of each set of data are shown in Table 4.1. Simulations of the circuit were also conducted including the additional MOSFET drain/source resistance measured in Chapter 2. However, due to the small currents involved, only a 0.6mV shift along the V_W-axis seen between the two sets of results.

The agreement between the simulated and measured results is good, with at most an 8% difference. There is poorer agreement between the theoretical and experimental results, with a maximum of 13% difference. Given that the potential for variation in the experimental values of β_n, V_{Tn} and C_{PSP} is up to 20% under AMS specifications, 13% is deemed to be an acceptable mismatch.

	ΔT = 10ns		**ΔT = 1ns**	
	Range	**Midpoint**	**Range**	**Midpoint**
Theoretical	123mV (11%)	0.89V (11%)	332mV (13%)	1.07V (4%)
Simulated	145mV (5%)	0.93V (7%)	321mV (8%)	1.05V (6%)
Measured	138mV	1.00V	295mV	1.11V

Table 4.1 - Ranges and midpoints of data in Figure 4.5. Bracketed terms indicate percentage difference from experimentally measured values.

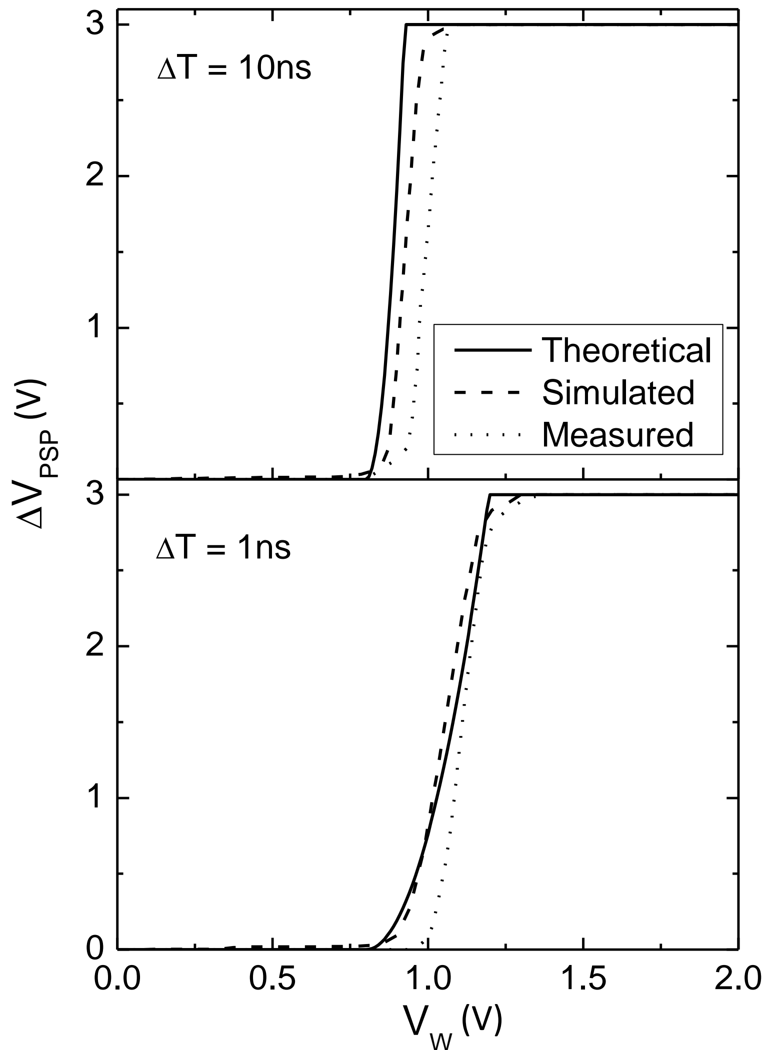

Figure 4.5 - ΔV_{PSP} against V_W for ΔT = 1ns and 10ns.

The rise-time of the PSP, predicted by (4.7), is compared against simulated 0%-100% rise-times in Figure 4.6. Only the section corresponding to the dynamic range of the neuron is shown for each case. The theoretical model generally underestimates the value of the rise time, by up to 3ns for lower values of V_W. Experimental results are omitted from this plot, as measurement noise makes it difficult to discern exact values, given that the variations can be less than 0.1ns in magnitude.

The final aspect of the theoretical prediction to be evaluated is the fall-time. The three sets of results, for ΔV_{PSP} = 1V, are plotted in Figure 4.7a. While there is reasonably good agreement between the theoretical and measured results, there is a shift of ~60mV between the simulated and measured results.

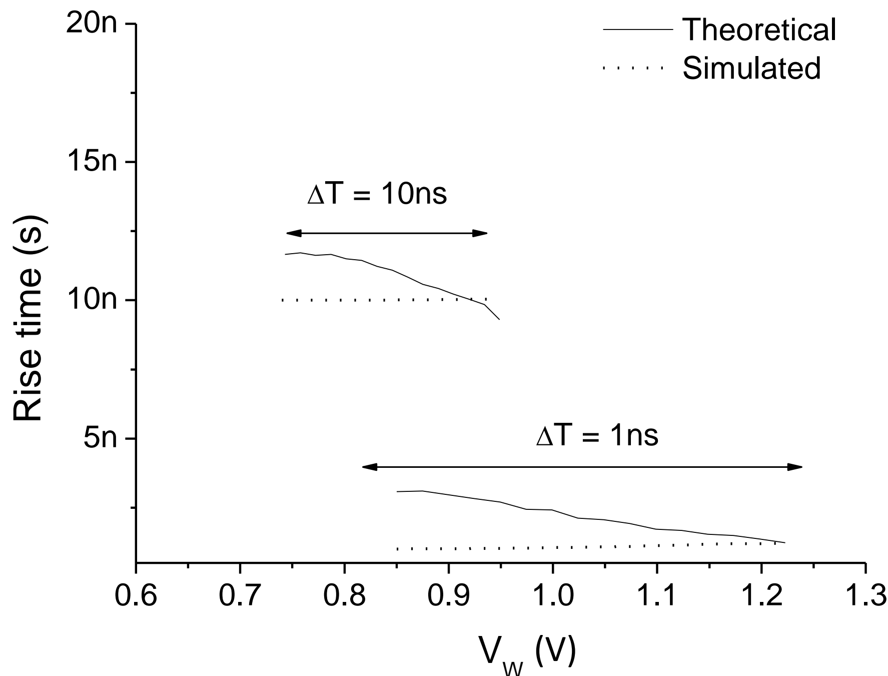

Figure 4.6 - Simulated and theoretical rise times over the appropriate dynamic range. V_{LEAK} = 0.4V

Referring back to Table 4.1, it can be seen that the voltage difference between the midpoints of the simulated and measured results is also ~60mV. This suggests that the threshold voltage on the chip used for measurements is ~60mV higher than the typical value used for simulations. Again, this falls within the range of variability specified by the chip manufacturer. Figure 4.7b plots the data with the simulated values shifted along the x-axis by 60mV.

Examples of PSPs obtained from the neuron circuit are shown in Figure 4.8 and Figure 4.9, which demonstrate the temporal summation of multiple synaptic inputs. Simulation and experimental results are shown. The values of V_W and V_{LEAK} have been adjusted in the simulation such that waveforms similar to the experimental results are obtained. In both cases, the rise/fall time of the V_{PRES} pulse is 0.1ns and the pulse width is 10ns. Figure 4.8 shows a PSP generated following three input spikes applied to the V_{PRES} terminal. For Figure 4.9, twenty inputs were applied, with lower values of V_W and V_{LEAK} used.

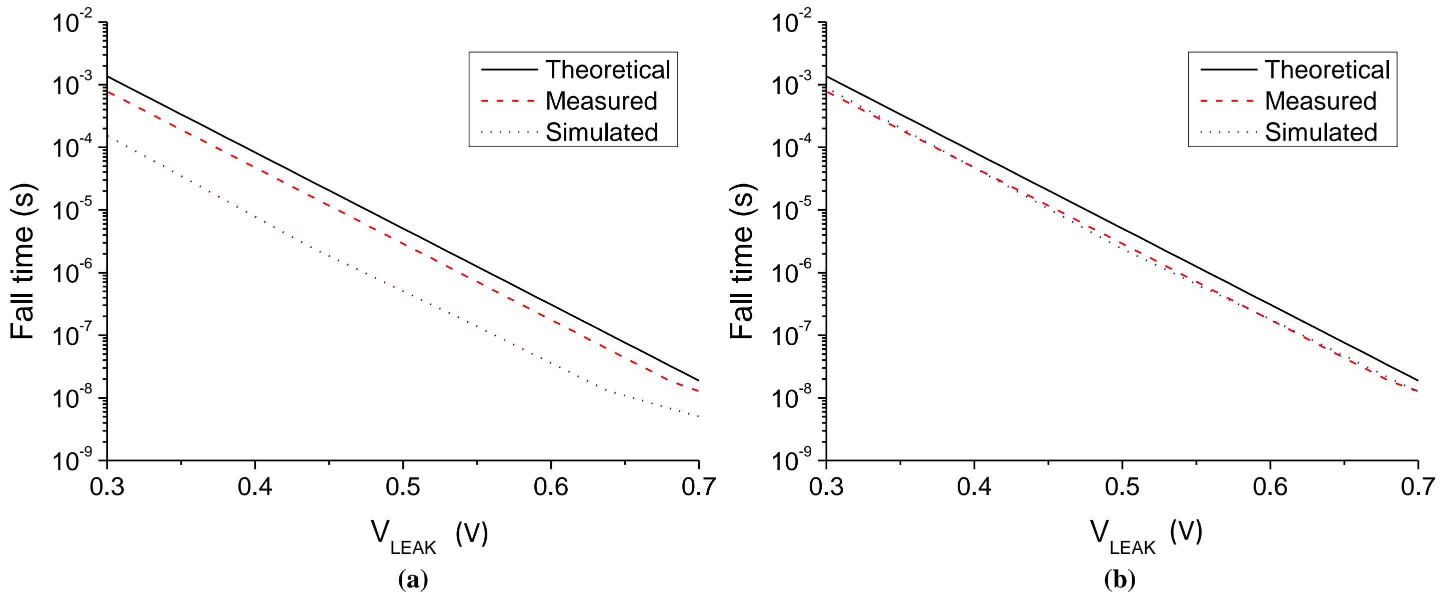

Figure 4.7 - Fall time against V_{LEAK}. ΔV_{PSP} = 1V, (a) original results, (b) simulation results shifted 60mV along x-axis.

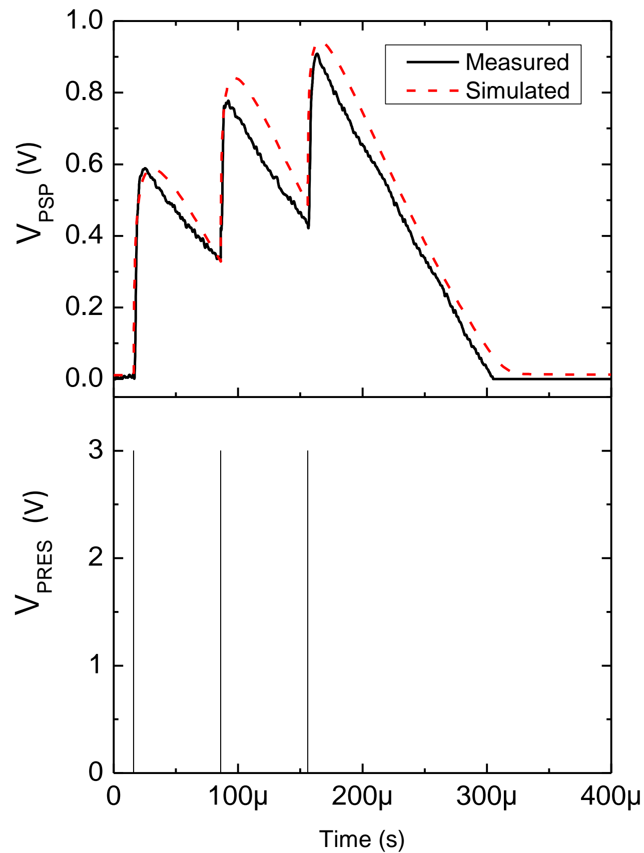

Figure 4.8 - PSP in response to 3 inputs. Experimental values: V_W = 0.95V, V_{LEAK} = 0.36V. Simulation values: V_W = 0.89V, V_{LEAK} = 0.3V. V_{PRES} rise/fall time 0.1ns, pulse width 10ns. V_{PSP} Fall time = 140us.

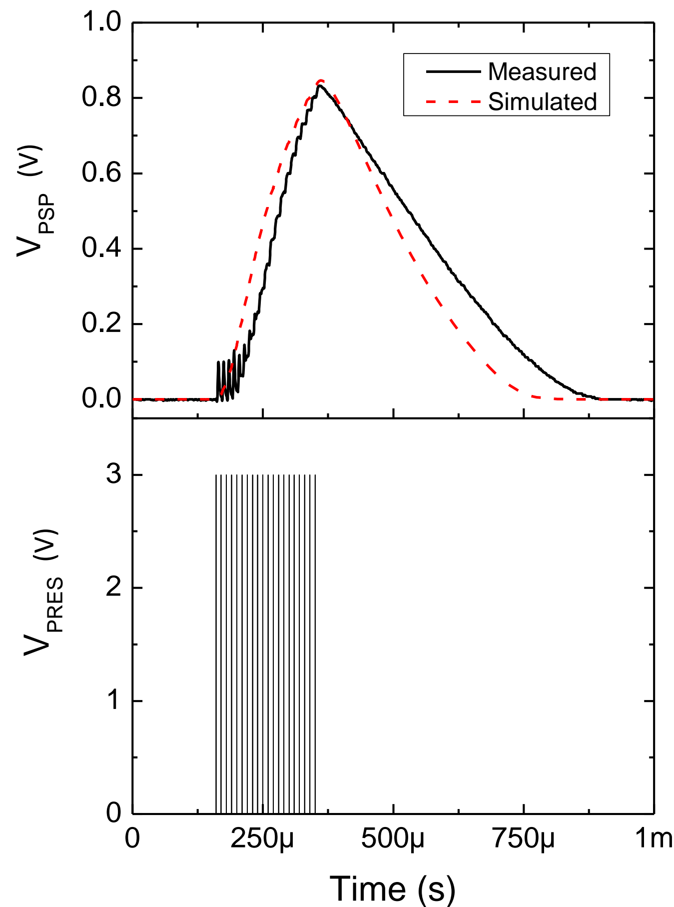

Figure 4.9 - PSP in response to 20 inputs. Experimental values: V_W = 0.9V, V_{LEAK} = 0.3V. Simulation values: V_W = 0.84V, V_{LEAK} = 0.26V. V_{PRES} rise/fall time 0.1ns, pulse width 10ns. V_{PSP} Fall time = 500us.

Given that the values of V_W remain unchanged, it would be expected that the change in V_{PSP} following each synaptic input will be the same. Taking values from Figure 4.8, it can be seen that this is not the case. The change in voltage following the first two synaptic pulses are 0.58V and 0.44V respectively. As V_{PSP} increases, the capacitance associate with the V_{PSP} node also increases. Since $\Delta V_{PSP} = Q_W / C_{PSP}$, this reduces the voltage change for subsequent input spikes. The implications of this are discussed further in Chapter 6.

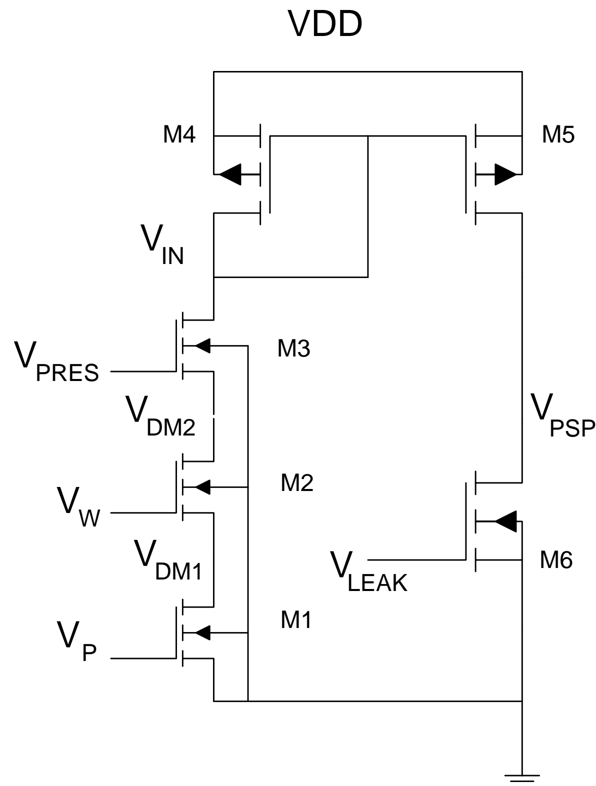

Figure 4.10 - Three terminal synapse and neuron circuit.

4.3.2 Three terminal (dynamic) synapse

The neuron circuit, connected to the three gate synapse is shown in Figure 4.10. In this section, results are presented which demonstrate the ability of this configuration to produce depressing PSPs.

The ΔV_{PSP} against V_W characteristic is shown in Figure 4.11. The smaller output charge of the three terminal device results in a much larger dynamic range, approximately 2V, than observed for the two gate synapse.

Generated PSPs are shown in Figure 4.12, Figure 4.13 and Figure 4.14, with different combinations of V_P and inter spike interval (ISI). As for the previous set of PSPs generated, the simulation and experimental control voltages used have been adjusted so that the two sets of results match. The exact voltages used are indicated in the figure captions. The effect of changing V_P on the magnitude of the PSP is demonstrated in Figure 4.12.

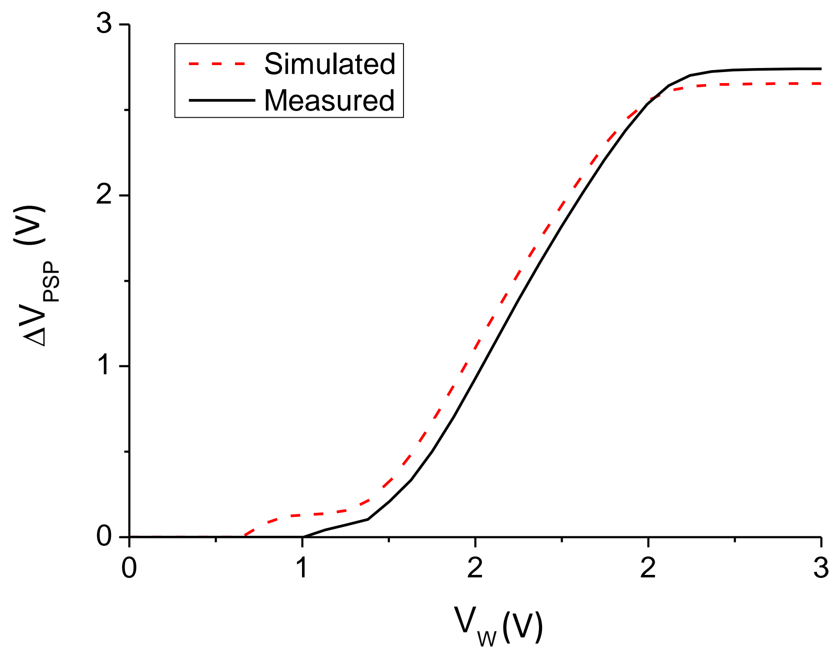

Figure 4.11 - PSP magnitude against V_W. The dynamic range is ~2V.

For a fixed ISI, the magnitude of the PSP decreases after the first synaptic input as the value of V_P is decreased. A similar effect is seen in Figure 4.13, where V_P is kept constant, but the length of the ISI adjusted. Finally, Figure 4.14 shows PSPs generated when multiple depressing synaptic inputs are temporally summed. The result is a depressing PSP, the exact shape of which depends on the ISI and/or the value of V_P.

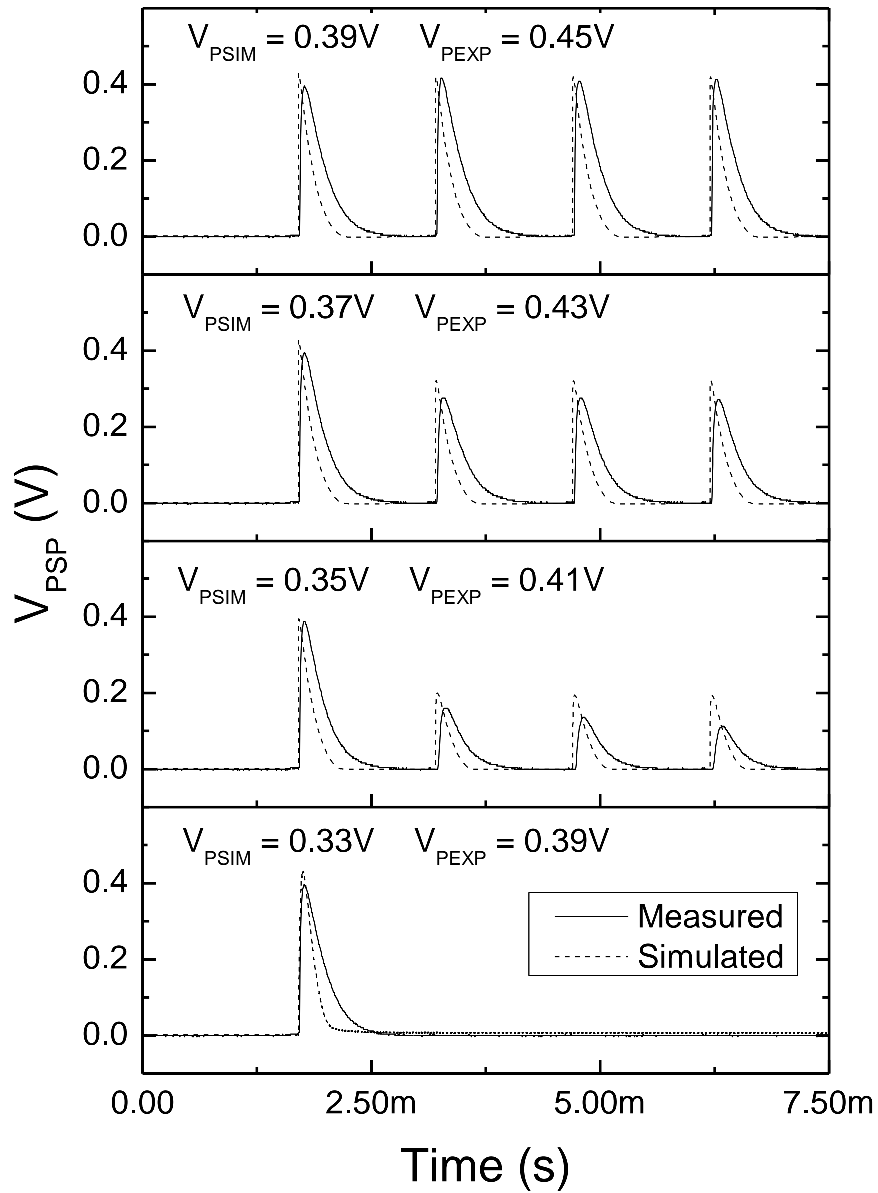

Figure 4.12 - Effect of changes in V_P on the PSP. ISI is fixed at 1.5ms. Experimental values: $V_W = 1.25V$, $V_{LEAK} = 0.44V$. Simulation values: $V_W = 1.19V$, $V_{LEAK} = 0.38V$.

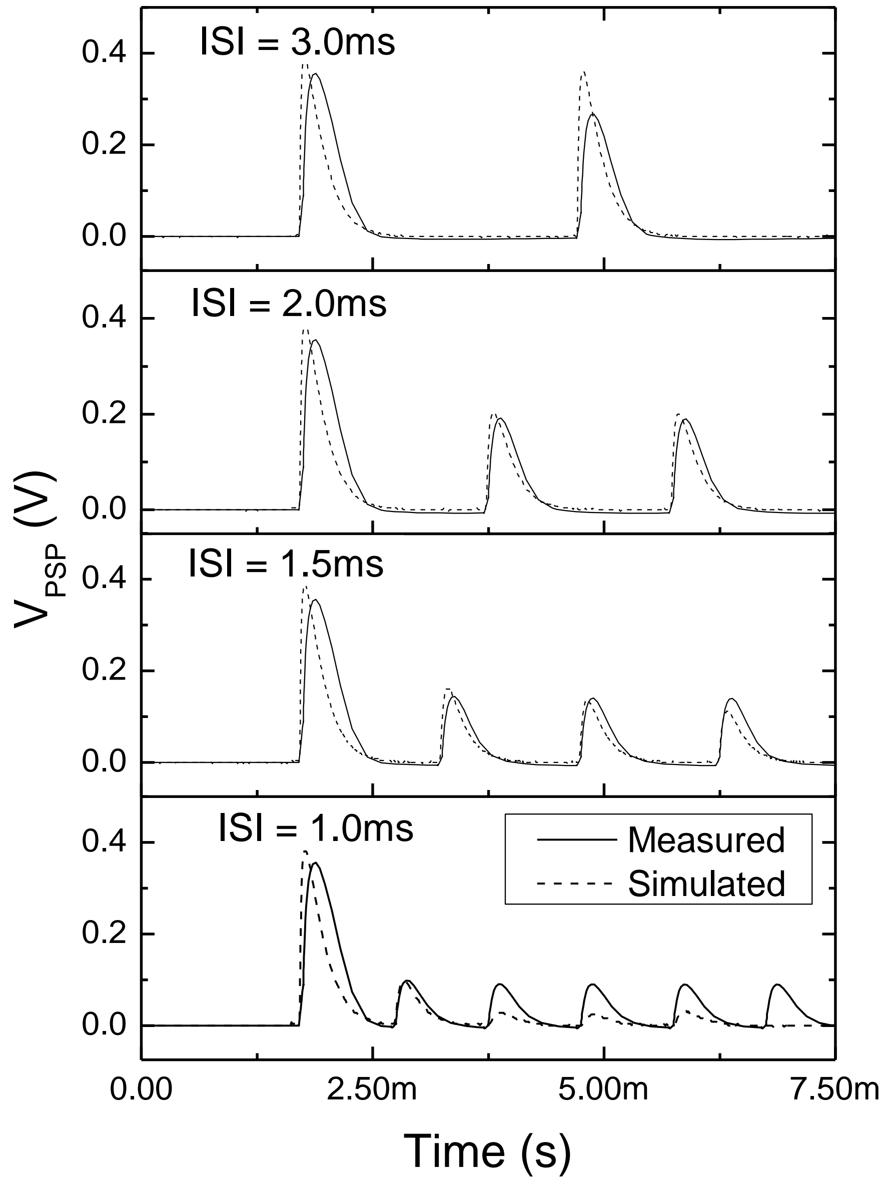

Figure 4.13 - Effect of changing ISI with a fixed V_P. Experimental values: V_W = 1.25V, V_{LEAK} = 0.44V, V_P = 0.41V. Simulation values: V_W = 1.19V, V_{LEAK} = 0.38V, V_P = 0.35V.

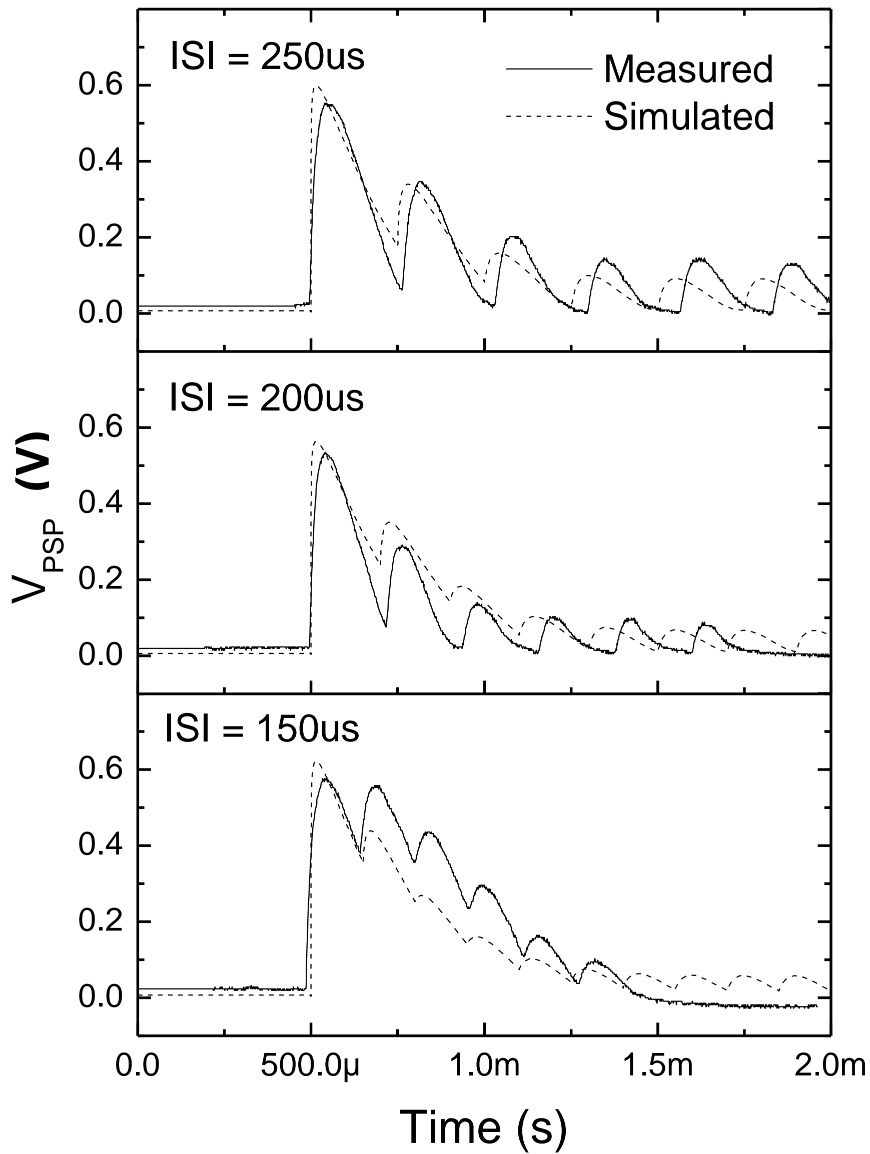

Figure 4.14 - An appropriate combination of synaptic inputs can create a depressing PSP. Experimental values: V_W = 1.3V, V_{LEAK} = 0.48V, V_P = 0.41V. Simulation values: V_W = 1.24V, V_{LEAK} = 0.42V, V_P = 0.35V.

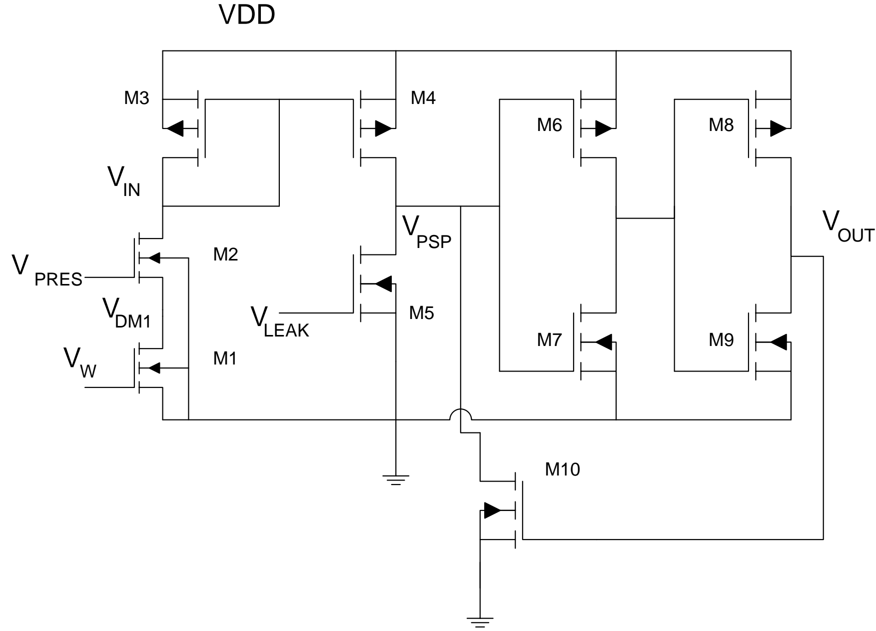

Figure 4.15 - Neuron with triggering and reset circuitry. Transistor dimensions: M6/M8 – 3.5µm x 0.35µm, M7/M9 – 0.4µm x 3.5µm, M10 – 0.4µm x 0.35µm.

4.4 Triggering circuitry

The neuron circuit is capable of integrating multiple synaptic inputs. In order to fully replicate the functionality of a biological neuron, triggering circuitry is required. A possible implementation is shown in Figure 4.15. The neuron is connected to a dual inverter chain (M6/M7 and M8/M9) and a reset transistor M10. In a large scale network implementation, V_{OUT} would serve as a presynaptic input to a separate synapse/neuron cell.

Initially, with V_{PSP} = 0V, V_{OUT} = 0V. When V_{PSP} exceeds the triggering voltage of the first inverter, its output undergoes a high-low transition. V_{OUT} goes high; M10 turns on and discharges the V_{PSP} node back to 0V. The exact triggering point, V_M, can be set by appropriate transistor sizing, where V_M varies according to:

$$\sqrt{\frac{\beta_n}{\beta_p}} = \frac{V_{DD} - V_M - V_{Tp}}{V_M - V_{Tn}} \tag{4.13}$$

That is:

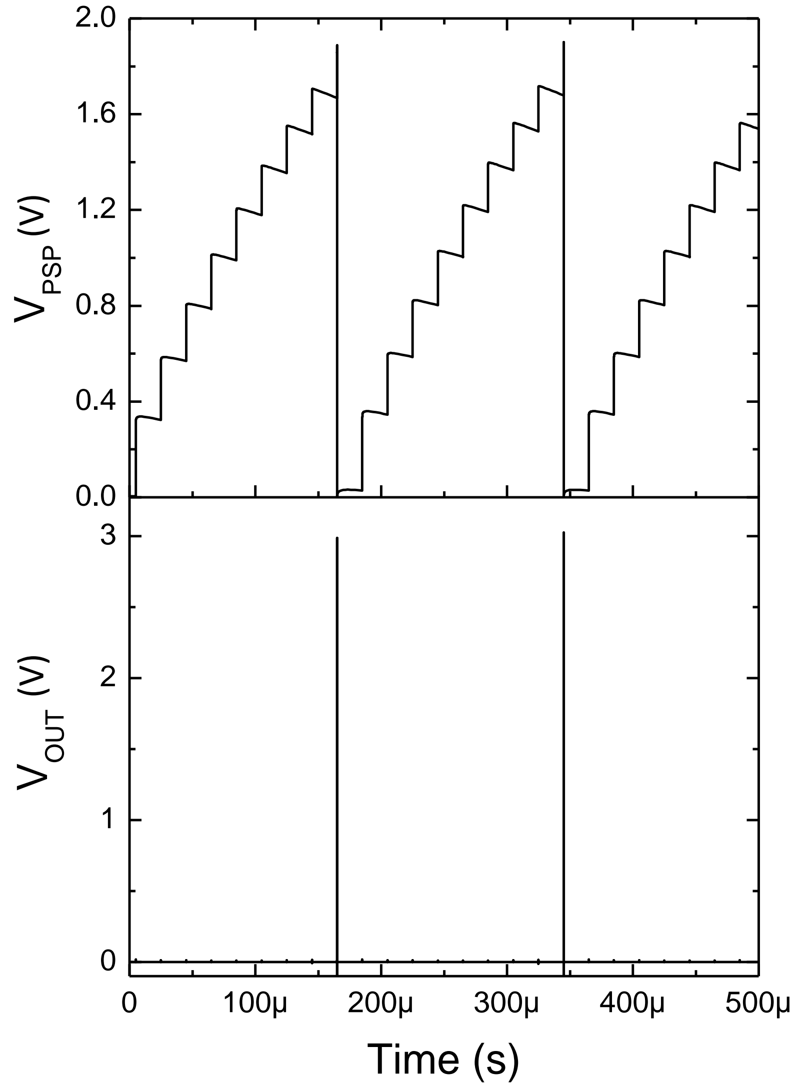

Figure 4.16 - V_{PSP} and V_{OUT}. When V_{PSP} reaches the triggering level, V_{OUT} undergoes a low-high transition and the device is reset. Synaptic inputs are applied every 20us. V_W = 0.7V, V_{LEAK} = 0.3V.

$$V_M = \frac{V_{DD} - V_{Tp} + \sqrt{\frac{\beta_n}{\beta_p}} V_{Tn}}{1 + \sqrt{\frac{\beta_n}{\beta_p}}} \tag{4.14}$$

By setting W_n/L_n to 10, W_P/L_P to 1 and vice-versa, values of V_M between 0.87V and 1.89V are obtainable. Choice of the value for V_M will depend on the network architecture. Higher triggering voltages would increase the range of usable weight voltages, producing neurons which require larger number of inputs before firing; lower triggering voltages would create

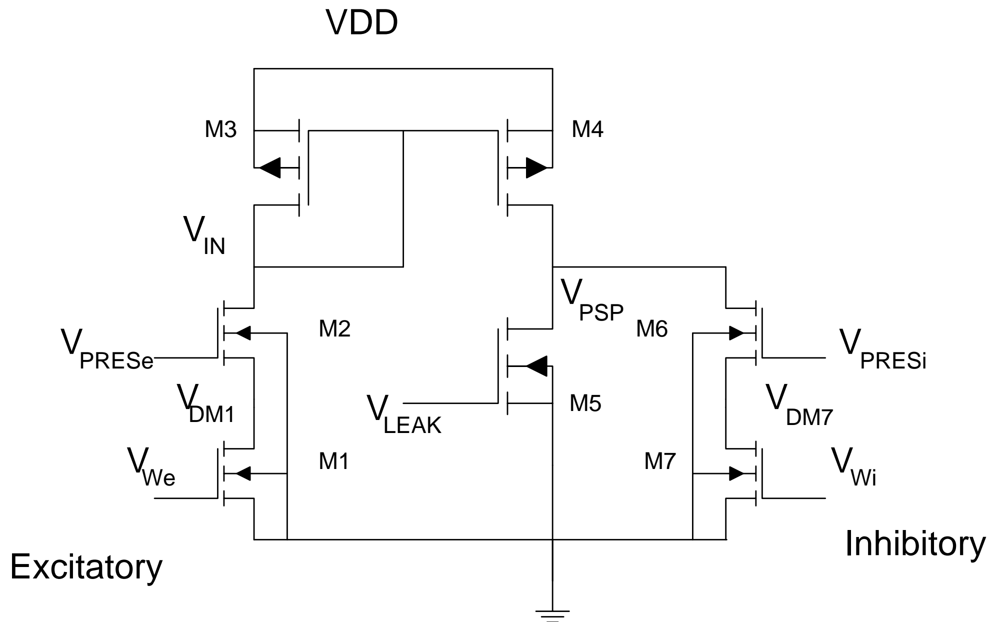

Figure 4.17 - Neuron circuit with excitatory and inhibitory synapses.

neurons which are more sensitive to synaptic inputs, which would subsequently fire more frequently. Further tuning of the switching voltage is possible through the application of substrate bias to the p-channel devices. All n-channel devices share a common substrate connection, so the application of bias to these is not practical. A bias of 0.5V applied to M6 will increase the triggering voltage by up to 300mV, depending on the sizing of the MOSFETs. The necessity of a variable triggering voltage is dependent on the overall network architecture and training methodologies chosen. In many cases the feature may not be required.

Simulation results confirming the operation of the triggering circuitry are shown in Figure 4.16. Presynaptic pulses are repeatedly applied to the V_{PRES} terminal, causing V_{PSP} to increase. When the triggering voltage, 1.65V, is reached, V_{OUT} can be seen to go high, resetting V_{PSP} to 0V. The width of the output pulse can be increased/decreased by altering the dimensions of M10 to increase/decrease the rate at which the V_{PSP} node is discharged. In this case, the pulse width is ~30ns.

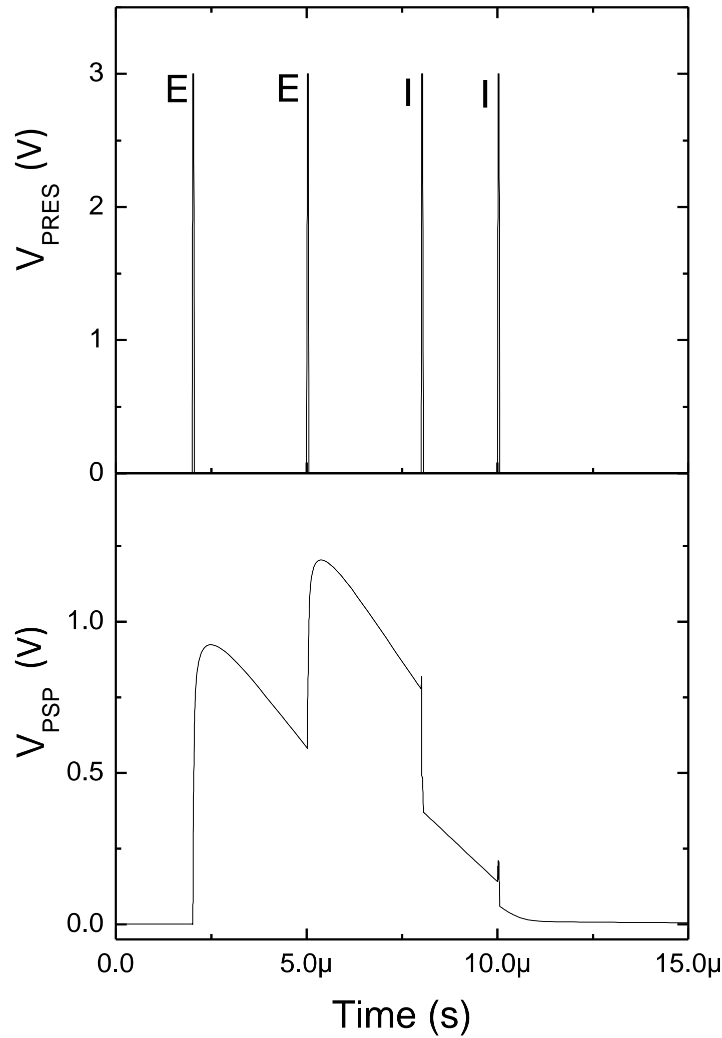

Figure 4.18 - V_{PRES} and V_{PSP}. E/I labels indicate whether the synapse firing is excitatory or inhibitory. V_{We} = 0.55V, V_{Wi} = 0.6V, V_{LEAK} = 0.4V.

4.5 Inhibitory synapses

All of the synaptic activity considered so far has been excitatory in nature, increasing the value of V_{PSP}. Many neural network algorithms require the use of inhibitory synapses which provide for negative weights, which is equivalent to a reduction in the value of V_{PSP}. Figure 4.17 shows the proposed implementation. The excitatory synapse is connected as normal, between the drain of M3 and ground. If a synapse is also connected between the drain of M4 and ground, then it will function in an inhibitory fashion.

The application of a pulse to the V_{PRESi} terminal initiates the transfer of charge through the synapse as normal. However, in this case the charge is removed from the V_{PSP} node, reducing the potential towards 0V.

Inhibitory synapses only have an effect if excitatory synapses have recently fired, as the potential of the neuron cannot be moved below its resting potential, 0V in this case. Figure 4.18 shows the results of a simulation where two excitatory inputs are followed by two inhibitory inputs. After increasing in response to the excitatory inputs, V_{PSP} can be seen to be reduced by the inhibitory inputs.

4.6 Conclusions

In this chapter, the structure of the neuron circuit has been described and results have been presented which validate its operation. In combination with the synapse cell, the neuron can produce biologically plausible PSPs. Plasticity, temporal summation, adjustable fall times, refractory periods and depression have all been demonstrated through measurements taken from fabricated devices. In addition, potential implementation of inhibitory synapses and a triggering/reset circuit have been described. While further work is required on these two areas, the simulation results presented indicate that they are viable prospects which will increase the functionality of the neuron cell even further.

References

[1] R. H. S. Carpenter, Neurophysiology, 4th ed.: Arnold, 2003.

Chapter 5: Axonal Delay Circuit

5.1 Introduction

The combined synapse/neuron circuit of the previous chapter provides a small area building block for hardware neural networks, implementing key features seen in biological neural networks. The functionality of the cell can be extended further, through the inclusion of additional circuitry to replicate the effect of an axonal delay.

In biological systems, interneuron communication is achieved through the transmission of action potentials along axons. Human axons are typically hundreds of micrometres to several millimetres in length, but can extend up to a metre. Conductance velocities are between 0.5m/s and 120m/s [1, 2]. This produces a typical range of delays between one microsecond and several milliseconds. The delay introduced by axons has been shown to play a computational role in the brain, which is particularly evident in the localisation of sound [3]. In this chapter, a circuit is presented which can implement an axon delay line between the output of a presynaptic neuron and the associated synapse. The circuit comprises a subthreshold MOSFET in series with a CMOS inverter chain consisting of two inverters. If an input is received from a presynaptic neuron, charge leaks through the transistor onto a capacitor which is in parallel with the inverter chain, and the input voltage to the inverter increases. The delay time is defined as the time between the presynaptic input being received and the output of the second inverter turning high. The associated rate of change of the inverter input voltage will be a function of the capacitance value and the MOSFET gate voltage. By adjusting the gate voltage of the MOSFET, it is possible to introduce a delay with a duration ranging from hundreds of milliseconds to tens of microseconds.

The chapter is organised as follows. In Section 5.2, the axon circuit is described and design equations are produced. Section 5.3 shows results achieved through SPICE simulations and through measurements taken from fabricated test chips. A complete neural cell consisting of the axon integrated with the synapse and neuron circuits described in previous chapters is shown in Section 5.4. Discussion and conclusions are given in Section 5.5.

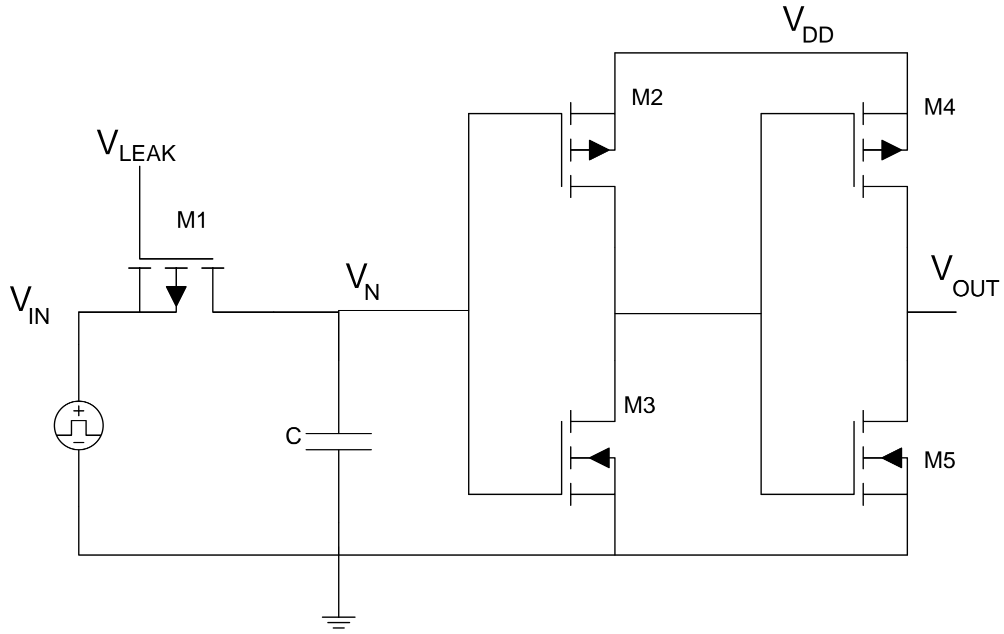

Figure 5.1 - Circuit for the implementation of an axonal delay.

5.2 Axon Circuit and Theory of Operation

The axonal delay circuit is shown in Figure 5.1. It consists of a p-channel leakage transistor M1, a capacitor C, and two inverters, M2, M3 and M4, M5. The p-MOSFET sits an n-well, thus allowing the substrate to be tied to source, to prevent back-bias effects.

The input to the circuit is to be taken as the output from a presynaptic neuron, while the output feeds into a postsynaptic neuron. To achieve the desired range of delays and to reduce the overall power consumption of the cell, the value of V_{LEAK} is set to bias M1 in the subthreshold region, where the threshold voltage of M1 is 0.7V.

Initially, with V_{IN} = 0V no current will flow through M1. The voltage across the capacitor, V_N, and the output voltage, V_{OUT}, will be 0V. Consider the arrival of a voltage pulse of magnitude V_{DD} at time t = 0 at the V_{IN} terminal. V_N will jump to a voltage set by the capacitive division between M1 and C:

$$V_{N0} = V_{DD} \frac{C_{M1}}{C_{M1} + C} \tag{5.1}$$

AMS provides a value for the drain/source capacitance associated with M1 of 1.36fFμm^{-2}, which gives a value for C_{M1} of 0.4fF, using values provided by AMS. V_N will increase according to:

$$I_{M1} = C\frac{dV_N}{dt} \tag{5.2}$$

In the subthreshold region, the current through M1 can be assumed to be constant for a given value of V_{GS} if $V_{DS} > 3V_{th}$:

$$I_{M1} = I_0 \exp\left(\frac{V_{GS}}{mV_{th}}\right) \tag{5.3}$$

An expression for the total delay time can be found by combining (5.2) and (2.25), rearranging and integrating between $V_N = V_{N0}$ and $V_N = V_{TI}$, the triggering voltage of the first inverter:

$$I_0 \exp\left(\frac{V_{GS}}{mV_{th}}\right)\int_0^t dt = C\int_{V_{N0}}^{V_{TI}} dV_N \tag{5.4}$$

$$t = \frac{C}{I_0 \exp\left(\frac{V_{GS}}{mV_{th}}\right)}(V_{TI} - V_{N0}) \tag{5.5}$$

I_O and m can be assigned nominal values extracted from the AMS 0.35μm CMOS process, 9fA and 1.71 respectively. M1 is 0.4μm x 0.35μm, the minimum dimensions allowed. The value of V_{N0} was calculated using (5.1) for each value of C. Figure 5.2 shows a plot of (5.5) as a function of the V_{GS} of M1, in the subthreshold region, for C = 1fF, 10fF and 100fF. The threshold voltage of the pMOST is taken to be 0.7V, as measured in Chapter 2.

5.3 Simulation and Experimental results

The circuit of Figure 5.1 was fabricated in the 0.35μm CMOS process, as described in Appendix 1. The widths and lengths of all transistors were set to 0.40μm and 0.35μm respectively. A two polysilicon layer capacitor was used, with nominal value for the capacitor, C = 100fF. The delay time was measured both experimentally and through SPICE simulations, using default parameters provided by AMS. With V_{DD} set to 3V, a 0V-3V voltage step was applied to the V_{IN} terminal. Assuming a finite pulse width, the input signal can be considered a voltage spike, which is typical of the type of signal that would be encountered in systems communicating through neuronal output spikes.

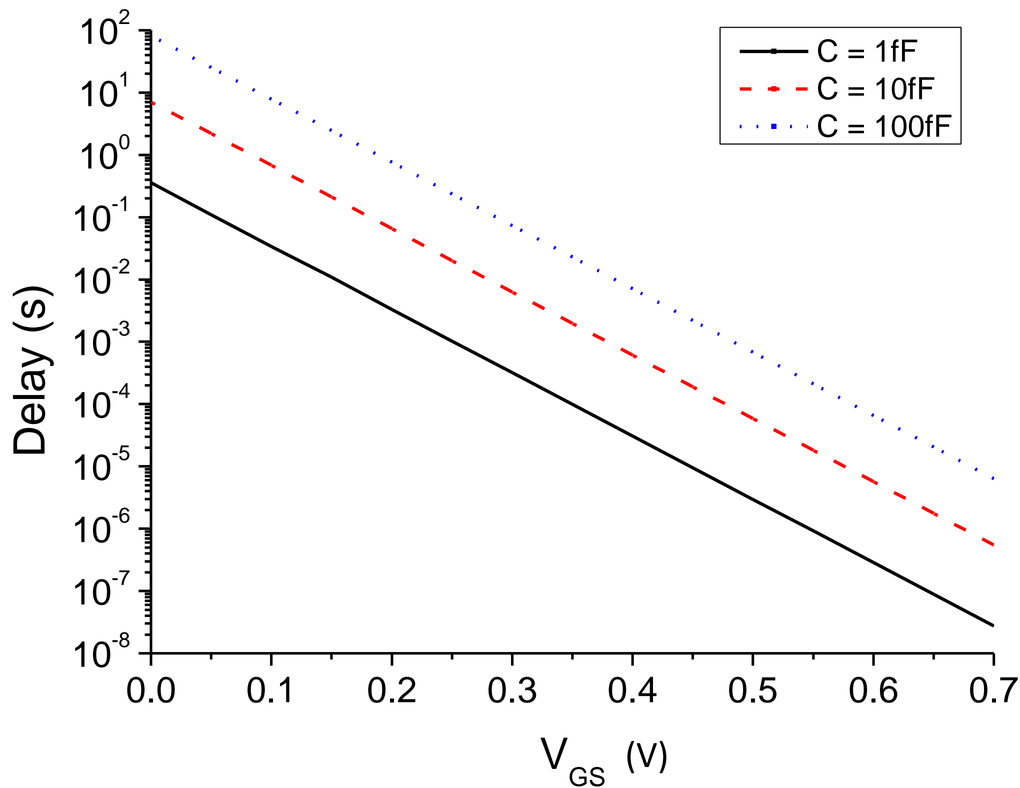

Figure 5.2 – Delay time as predicted by (5.5) for three values of C. The threshold voltage of M1 is 0.7V.

The ability of the circuit to work with such signals is demonstrated in Section 4.4. The delay time is measured as the time difference between V_{IN} going high and V_{OUT} going high, as shown in Figure 5.3. The experimental and simulated delay times are shown alongside theoretical values, predicted by (5.5), in Figure 5.4a.

For 0V < V_{GS} < 0.3V, the simulated and experimental results are constant at 30ms and 53ms respectively, indicating a constant leakage current flowing through M1. When V_{GS} > 0.3V, there is an exponential fall off of the delay time. There is a linear shift between the two sets of results, equivalent to a threshold voltage difference of approximately 60mV ($V_{Texp} = V_{Tsim}$ - 60mV), which is within the +/- 100mV tolerance of the AMS process. The theoretical predictions show good agreement with the other results for V_{GS} > 0.3V. It is possible to improve the quality of fit of the theoretical data by modelling the delay time for V_{GS} < 0.3V as a constant value:

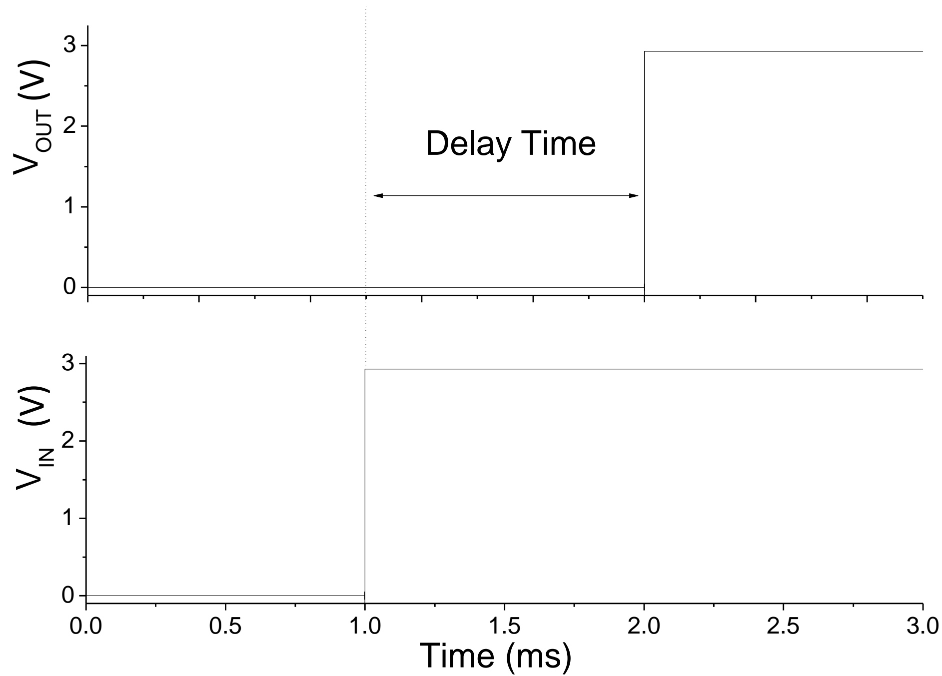

Figure 5.3 - Measurement of delay time as observed on oscilloscope. V_{GS} = 0.45V.

$$t = \frac{C}{I_0 \exp\left(\frac{0.25}{mV_{th}}\right)}(V_{TI} - V_{N0}) \tag{5.6}$$

With a subthreshold slope of 98.4mV/decade, a threshold voltage shift of -60mV will increase the value of I_0 by a factor of $10^{60/98.4} = 4$. Solving (5.6) under this condition gives a value for the constant delay of 15ms. The results produced after making these adjustments are plotted in Figure 5.4b, where the simulation results are also modified to take into account the 60mV threshold voltage shift, assuming that it arises from a higher than expected fixed oxide charge. Reasonably good agreement between the three sets of values is shown across the entire voltage range. The V_T could also be shifted due to a variation in subthreshold slope, S, for the different p-MOSTs but this is difficult to incorporate into simulations due to the complexity of the model employed in SPICE. However, the p-MOSTs were seen, in Chapter 2, to exhibit a very high level of interface states, $N_{SS} = 2.0 \times 10^{12}cm^{-2}$. This is consistent with the increased subthreshold slope of 16.4mV, which corresponds to an N_{SS} of $1.8 \times 10^{12}cm^{-2}$.

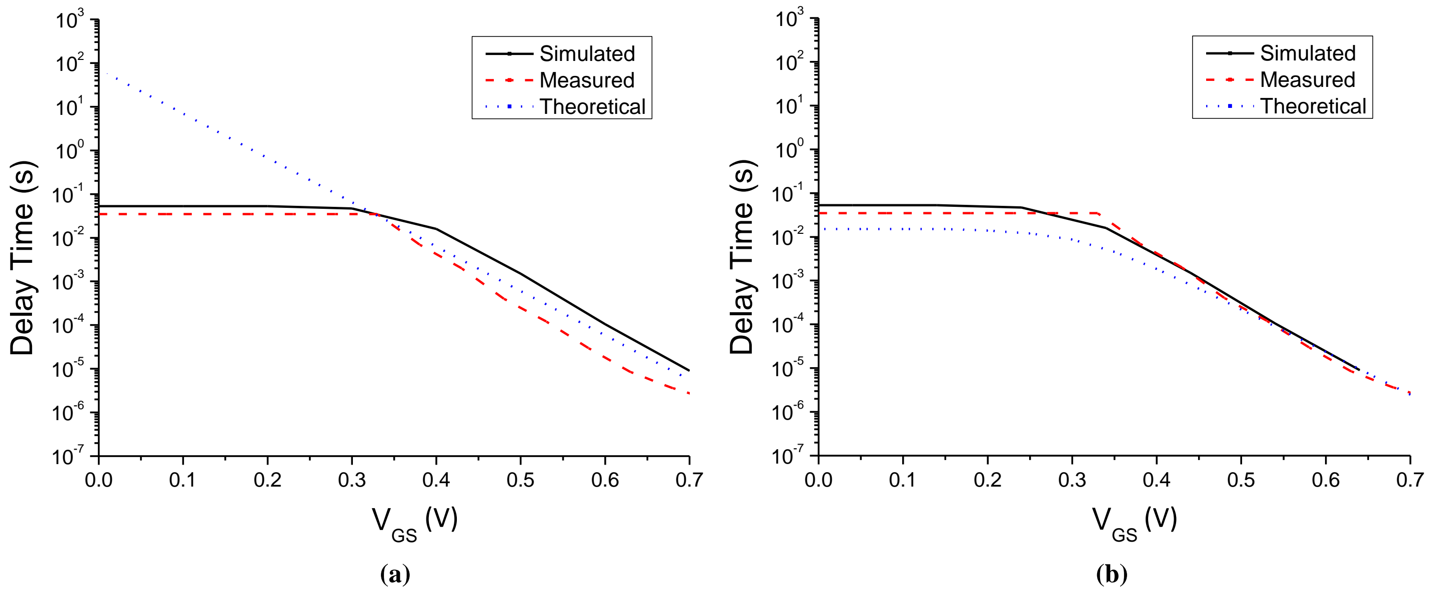

Figure 5.4 - Simulated, theoretical and experimentally measured delay times. (a) Original simulation data and theoretical values calculated using (5.5); (b) Shifted simulation data and theoretical values calculated using (5.6).

5.4 Series connected neurons

In the schematic of Figure 5.5, the axon circuit is used to introduce a delay between the output of Neuron 1 and the input of Neuron 2. The output of Neuron 1, V_{PSP1}, feeds the triggering circuitry, as described in Chapter 4. The axon circuit is used to delay the output of the triggering circuitry before it is fed into the presynaptic input terminal of Neuron 2. The reset transistor, M20, has been connected to the axon output rather than the output of the triggering circuitry, to ensure that the circuit is not reset until neuron 2 has fired.

Simulation results for the circuit are shown in Figure 5.6, which shows V_{PSP1}, V_{TRIG}, V_{PRES2} and V_{PSP2}. Two scenarios are shown, one where the delay is 1us and another with a 7ms delay, values consistent with the range of biological values described in the introduction to this chapter. Synaptic inputs are applied to Neuron 1 until the output of the triggering circuit goes high. When V_{PRES2} goes high The synapse M15/M16 is activated, dumping charge onto the V_{PSP2} and the value of V_{PSP1} is reset to 0V. Where the longer delay is generated in Figure 5.6b, it is necessary for V_{PSP1} to remain above the triggering voltage for the duration of the delay period.

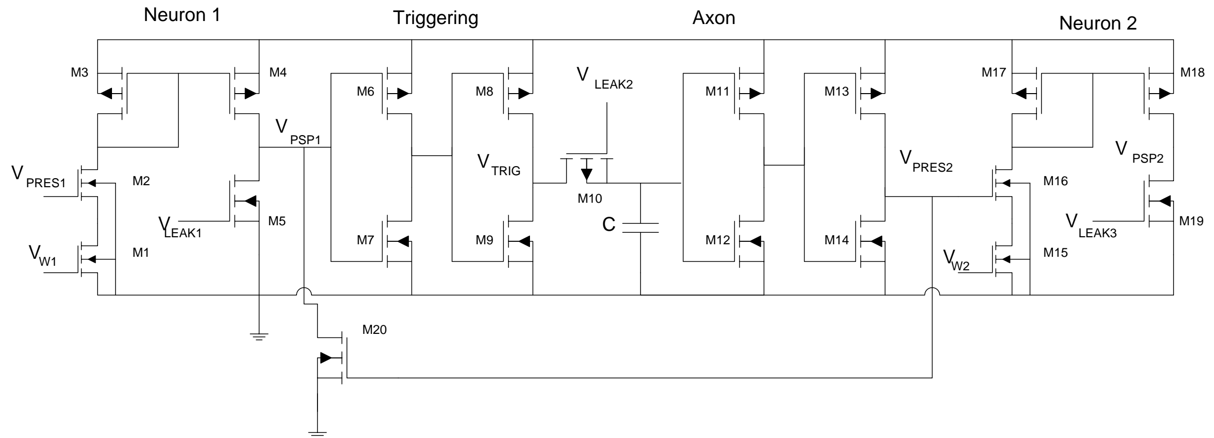

Figure 5.5 - Schematic of two neurons connected by an axon. The presynaptic input to Neuron 2 is the output of Neuron 1, delayed by the axon circuit.

In practice, as hardware neural networks are often operated faster than biology, delays in the order of milliseconds may not be necessary.

5.5 Conclusions and Discussion

A circuit has been presented which can introduce a delay into a neural pathway by utilizing the leakage through a sub threshold MOSFET. The approach requires only five MOSFETs and a capacitor to produce the delay, which can be engineered to last between several microseconds and tens of milliseconds. Compared to the circuits considered in Chapter 1, Section 1.3.3, this implementation allows a higher degree of control over the length of the delay and has a lower transistor count. A two-stage CMOS inverter chain serves to propagate the input signal to subsequent circuit elements once the delay period has elapsed. The theory and experimental results presented show clearly the feasibility and scope of the approach, which can be integrated with existing spiking neural circuits to produce biological scale delays.

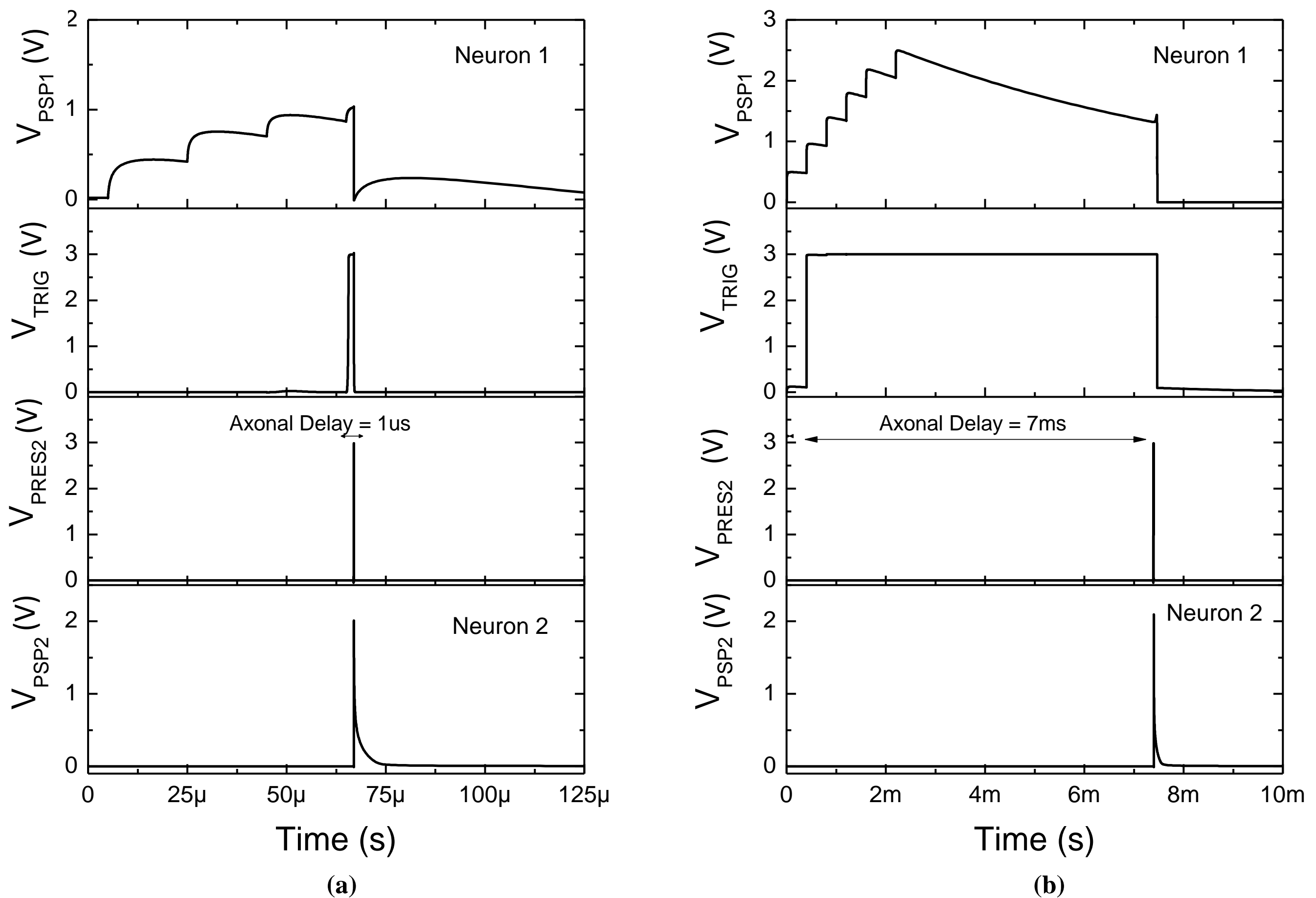

Figure 5.6 - Simulation of circuit in Figure 5.5. A delay of 1us is introduced in (a). The circuit is then reset. V_{W1} = 0.85V, V_{W2} = 0.7V, V_{LEAK1} = 0.22V, V_{LEAK2} = 2.2V, V_{LEAK3} = 0.25V. The delay in (b) is 7ms, all voltage are equal, other than V_{LEAK1} and V_{LEAK2}, which are set to 0.2V and 2.55V respectively.

References

[1] D. Purves, Neuroscience, 2 ed.: Sinauer, 2001.

[2] N. J. W. Russell, "Axonal conduction velocity changes following muscle tenotomy or deafferentation during development in the rat," Journal of Physiology, vol. VOL 298, pp. 347-360, 1980.

[3] L. A. Jeffress, "A place theory of sound localization," Journal of Comparative and Physiological Psychology, vol. 41, pp. 35-39, 1948.

Chapter 6: VLSI Issues

6.1 Introduction

In previous chapters, the synapse, neuron and axonal delay circuits have been studied in isolation, in order to evaluate and characterise their operation. However, the success of any large scale implementation depends not only on the ability of individual cells to replicate their biological counterparts, but also on how well the cells can be scaled up. Hundreds of thousands of interconnected cells may be required for a given application. From a hardware perspective, this raises a number of issues, which are discussed in this chapter.

Section 6.2 considers the effects of device variability, Section 6.3 covers issues associated with the scalability of the technology, where a standard neural cell is proposed. Section 6.4 discusses weight storage and training methods using a standard benchmark exclusive-OR (XOR) circuit built from the functional blocks described in earlier chapters. The latter section draws on the contributions from other researchers in an EPSRC funded project [1]. Conclusions are drawn in Section 6.5.

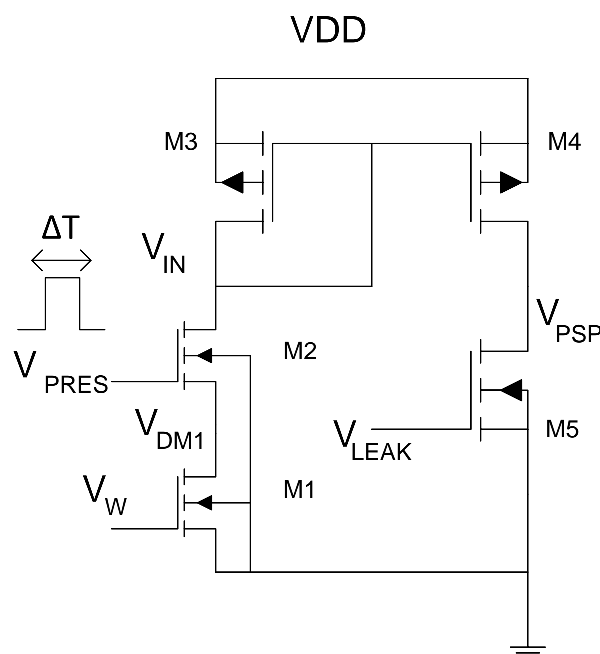

Figure 6.1 - Neuron circuit with two terminal (static) synapse.

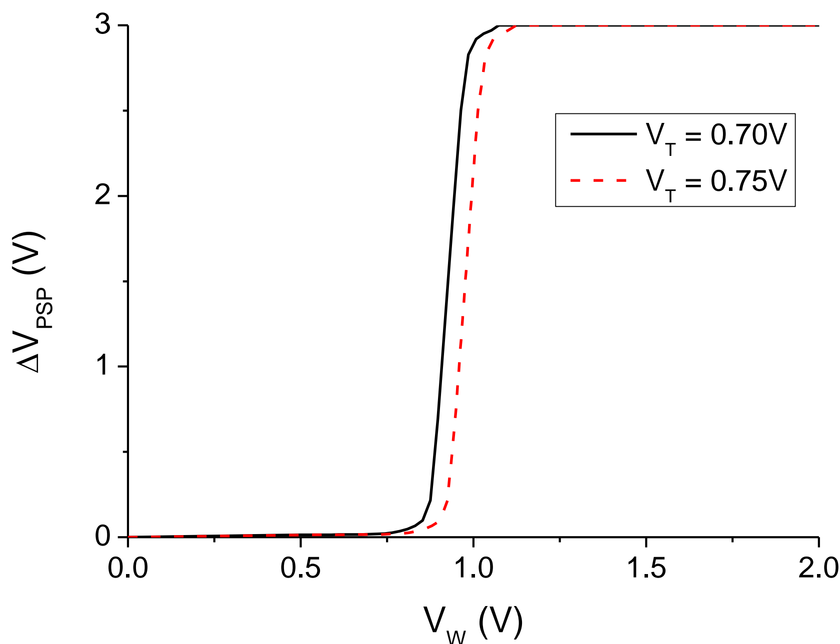

Figure 6.2 – Simulation results showing the effect of a threshold voltage shift of 50mV.

6.2 Device variability

Variation in the physical characteristics of fabricated silicon devices is inevitable. The effects of variability are not limited to the devices described in this thesis, any implementation of neural hardware, particularly those where analogue circuitry is used, will face similar problems. In this section, two kinds of variability will be discussed. The first is variability between separate chips, the second is variability among devices on a single chip.

The first issue is the simpler of the two with which to deal. For example, consider the neuron circuit, shown in Figure 4.1. A universal threshold voltage shift of 50mV between two sets of chips would manifest itself as shown in Figure 6.2. The ΔV_{PSP} vs. V_W characteristic would be shifted by 50mV along the x-axis from its 'typical' position. Clearly, if two such circuits were operated using the same voltage levels, the results would be considerably different.

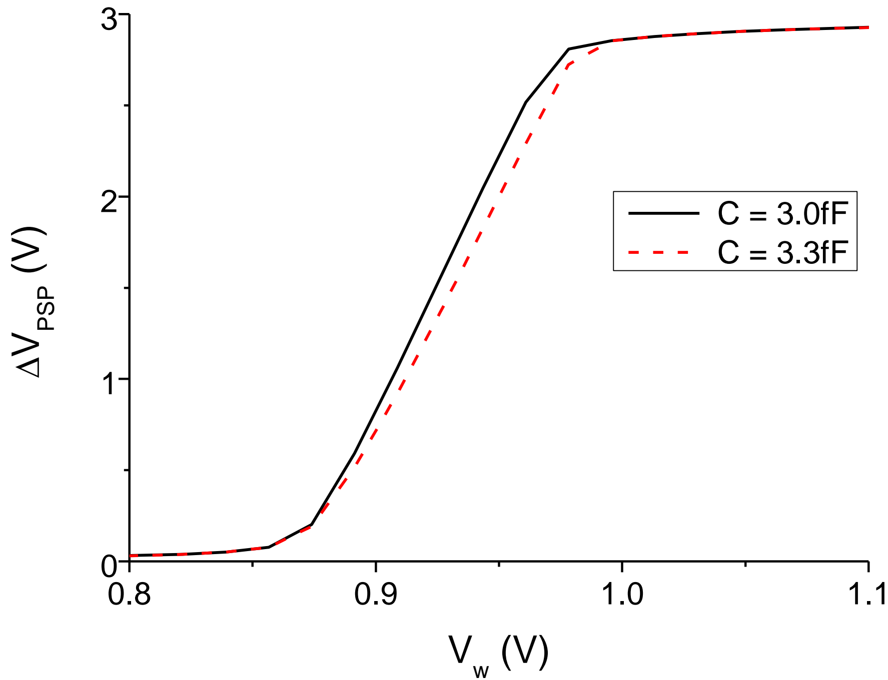

Figure 6.3 - Simulation results showing the effect of an increase in C_{PSP} of 10%.

Assuming that suitable test structures are in place on each chip to facilitate accurate threshold voltage measurement, this problem can be solved through the application of a DC bias voltage, equivalent to the shift in threshold voltage, to the V_W device. In addition, there are a number of techniques available where feedback circuitry is used to apply substrate bias to compensate for process variations [2, 3].

Another, slightly more complicated example, is that of a variation in capacitance. Consider a 10% increase in capacitance at the V_{PSP} node, implemented by attaching an additional capacitance of 0.3fF to the node. The value of ΔV_{PSP} depends on C_{PSP} according to:

$$\Delta V_{PSP} = \frac{Q_W}{C_{PSP}} \qquad (6.1)$$

The results would be a smearing of the ΔV_{PSP} vs. V_W characteristic, as demonstrated in Figure 6.3 where the dynamic range of the neuron is increased. Again, given that the characteristics of the neuron will be known in advance, it would be possible to modify any weight update system used to take into account this difference. A 100mV increase in the value of ΔV_{PSP} would require a V_W increase of 4mV for the original case and 5mV when the capacitance is increased. As seen in Chapter 4, Section 4.3.1, C_{PSP} has a voltage dependence:

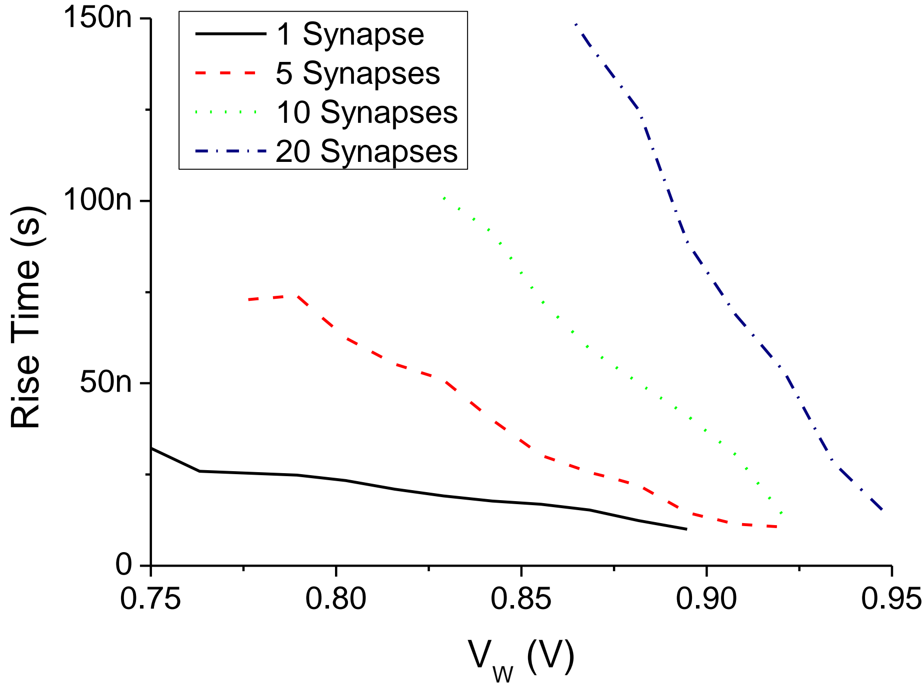

Figure 6.4 - Simulation results showing effect of increasing C_{VIN} through additional synapses on the rise time of V_{PSP}. Input pulse width is 10ns.

$$C_{PSP} = \sqrt{\frac{qN_A\epsilon_{si}\epsilon_0}{2(\varphi_b + V_{PSP})}}A + C_P \tag{6.2}$$

This will also cause some variation in the value of ΔV_{PSP}, but be counteracted in a similar fashion, as any changes in C_{PSP} can be fed back into the training algorithm.

The fall time of the PSP is dependent on C_{PSP} according to:

$$t_{fPSP} = \frac{\Delta V_{PSP} C_{PSP}}{I_0 e^{\left(\frac{qV_{LEAK}}{mkT}\right)}} \tag{6.3}$$

If the capacitance increases, the fall time will increase proportionally. A suitable increase in the value of V_{LEAK} would counteract the increase in C_{PSP}, leaving t_{fPSP} unchanged.

The second type of variability is more difficult to counteract. Unanimity between separate device on a single chip cannot be guaranteed. However, from a strictly biological point of view, variation between individual cells is not uncommon. Different neurons can have different decay periods, respond differently to identical synaptic inputs and have varying time constants for facilitation and depression [4]. Even a single neuron can react differently over time in response to similar inputs [5-7].

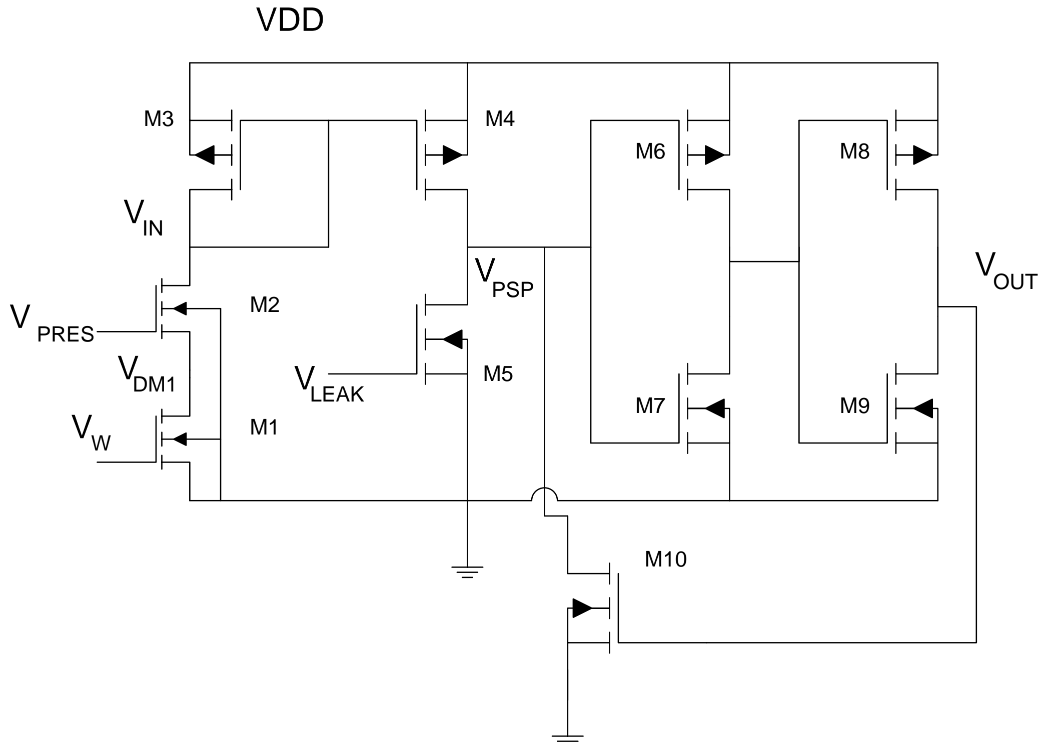

Figure 6.5 - Neuron with triggering and reset circuitry. Transistor dimensions: M6/M8 – 3.5µm x 0.35µm, M7/M9 – 0.4µm x 3.5µm, M10 – 0.4µm x 0.35µm.

This suggests that large scale neural systems will be able to function effectively despite variations between individual cells, assuming that suitably robust learning algorithms are used. Even still, it is advantageous to be able to deal with potential issues in advance.

Given that the characteristics of each individual device cannot be measured, or that it is impractical to do so, the techniques described for dealing with chip wide variations cannot be used for this scenario, although it may be possible to use some of the substrate biasing techniques mentioned earlier, the effects of possible variations are more effectively solved through considerations at the design stage. Several steps can be taken during layout creation to reduce the possibility of unwanted variations. Device matching, dummy structures, symmetry, large transistors and the use of identical wiring paths are all common techniques used to minimise variability [8].

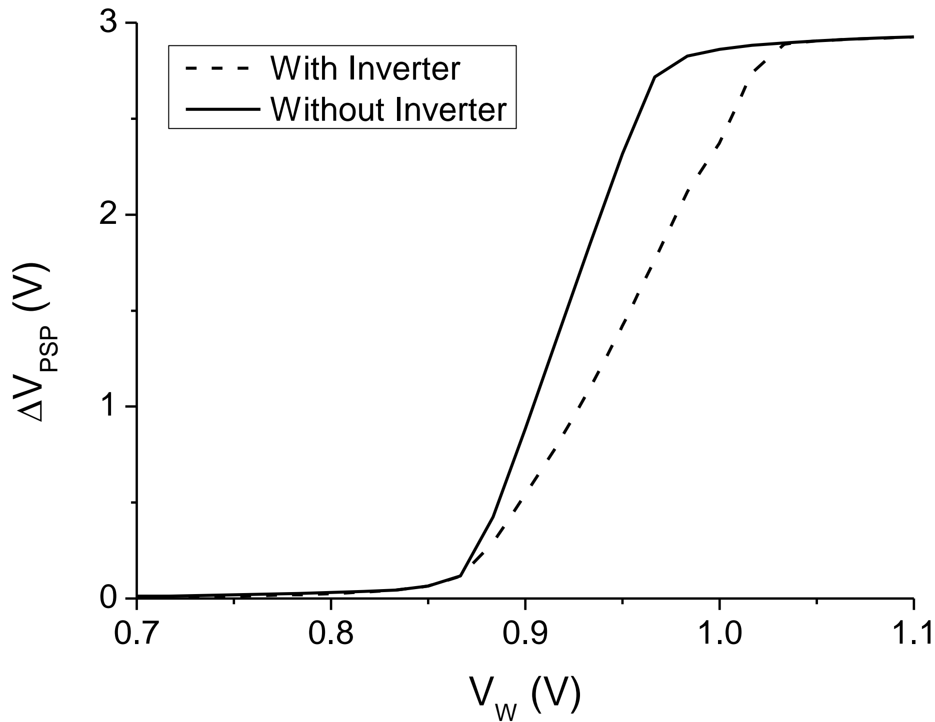

Figure 6.6 - Simulated results with and without triggering circuitry showing ΔV_{PSP}.

6.3 Scalability Issues and Standard Neuron Cell

Scaling up the neural circuitry requires some thought to be given to the connectivity of the devices. There are two components of connectivity to be considered. The number of synapses connected to a neuron is one and sets the value of C_{VIN}, the capacitance associated with the V_{IN} node; The other is the output capacitance, C_{PSP}, which is dependent on the choice of output circuitry. Higher levels of connectivity also increase the amount of silicon area required for layout, as, in a fully connected system, the number of metal interconnects required increases dramatically as the number of devices is increased [9].

Connecting additional synapses to the neuron increases the value of C_{VIN}, which affects the rise time of the PSP, as it has a dependence on C_{VIN}, given in Chapter 4, Section 4.2.2. ΔV_{PSP} and the fall time are not affected. The change in rise time is illustrated in Figure 6.4 which plots the rise time against V_W for 1, 5, 10 and 20 synapses.

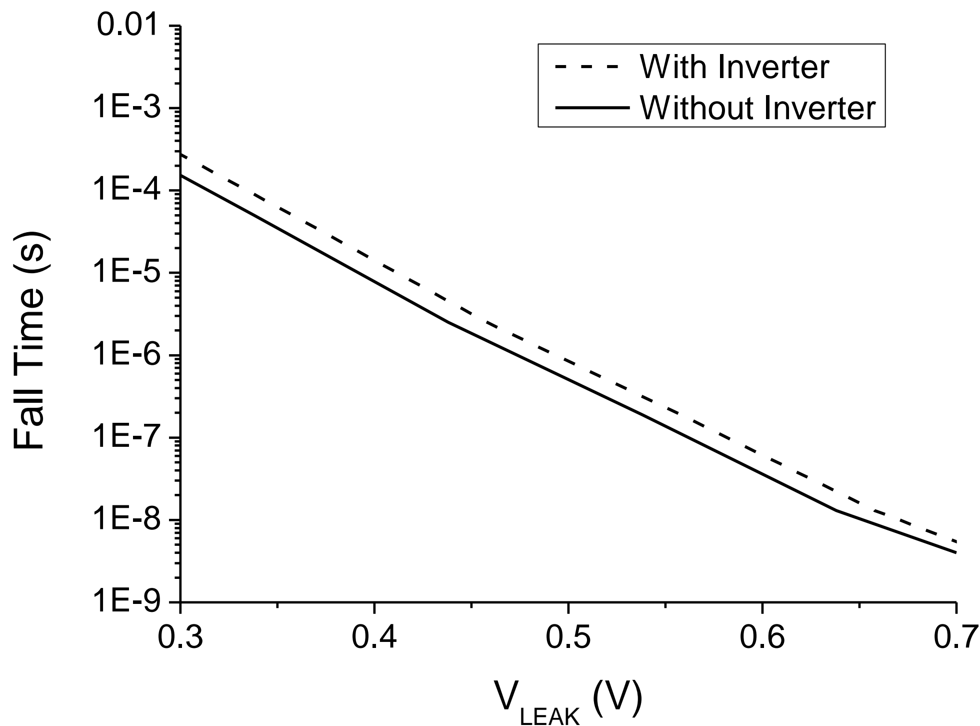

Figure 6.7 - Simulated results with and without triggering circuitry showing fall time, when ΔV_{PSP} = 1V.

The effects of increases in C_{PSP} have been described in the previous section – the dynamic range and fall time are increased. Two possible output configurations have been considered in this thesis, the standard neuron circuit shown in Figure 4.1, and the triggering circuitry, shown in Figure 6.5.

The addition of triggering circuitry introduces an additional capacitive loading to the node. Figure 6.6 and Figure 6.7 show the differences in ΔV_{PSP} and the fall time for the two sets of circuitry. The value of C_{PSP} is also increased if inhibitory synapses are connected to the neuron circuit, as described in Chapter 4, Section 4.5.

While there is no set limit on the level of input and output connectivity, it is useful from a design point of view to define a 'standard' neuron cell, with a fixed number of inputs and uniform output circuitry. Not only does this mean that each cell will operate in a similar fashion, it also simplifies the layout creation process, as a single, modular design can be used to create larger networks of neurons. To this end, the configuration shown in Figure 6.8 is proposed. It consists of the neuron circuit with 20, three-terminal, depressing synapses (10 excitatory and 10 inhibitory), connected to the triggering circuitry.

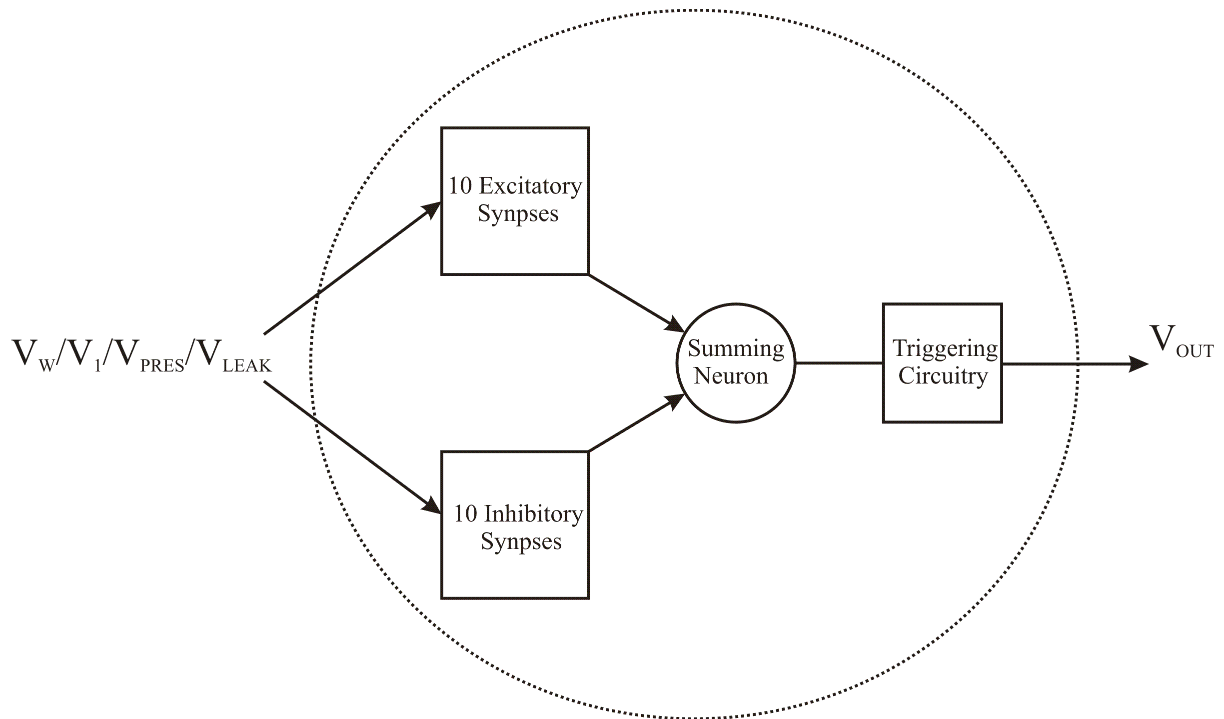

Figure 6.8 - Proposed standard neural cell.

Having twenty synapses provides a reasonable level of connectivity, while ensuring that the amount of metal interconnects required does not dominate the layout of the cell. If more than twenty synapses are required, it is possible to ‘chain’ multiple blocks of synapse together at the layout stage, where a single summing neuron receives inputs from several different sets of synapses.

Assuming each synapse shares V_{LEAK}, V_1 and V_{PRES} terminals, the cell requires up to 23 input voltages, depending on the number of active synapses. The layout for the standard cell is shown in Figure 6.9, which was created by a project collaborator [1]. The metal interconnect lines, at either side of the cell, consume approximately half of the total area, with dimension of 87.3µm x 32µm. A chip of dimension 1mm^2 would be able to fit up to 360 of these standard cells (360 neurons, 7200 synapses). In reality, some of the silicon area will be consumed by additional control circuitry and wiring. If it is assumed that such additional components were to consume 25% of the available area, it would be possible to fit 270 standard cells (270 neurons, 5400 synapses) in the remaining space.

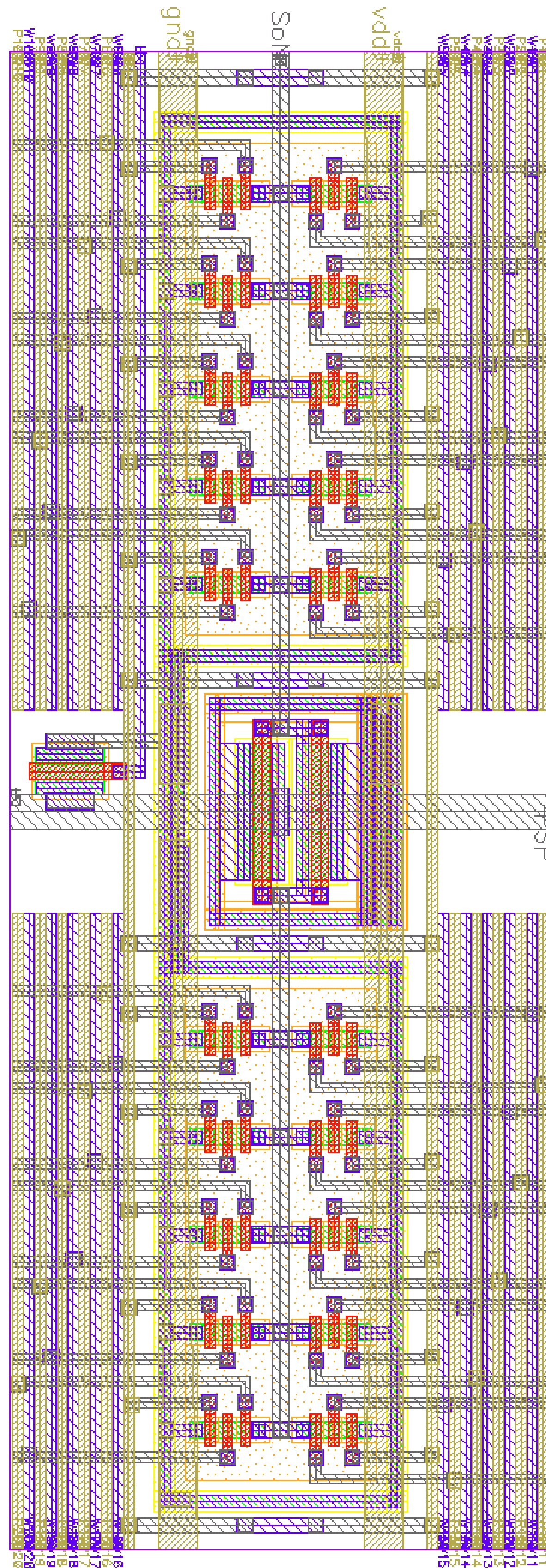

Figure 6.9 - Layout of standard neural cell, containing 20 synapse and a neuron. Dimensions are 87.3µm x 32µm.

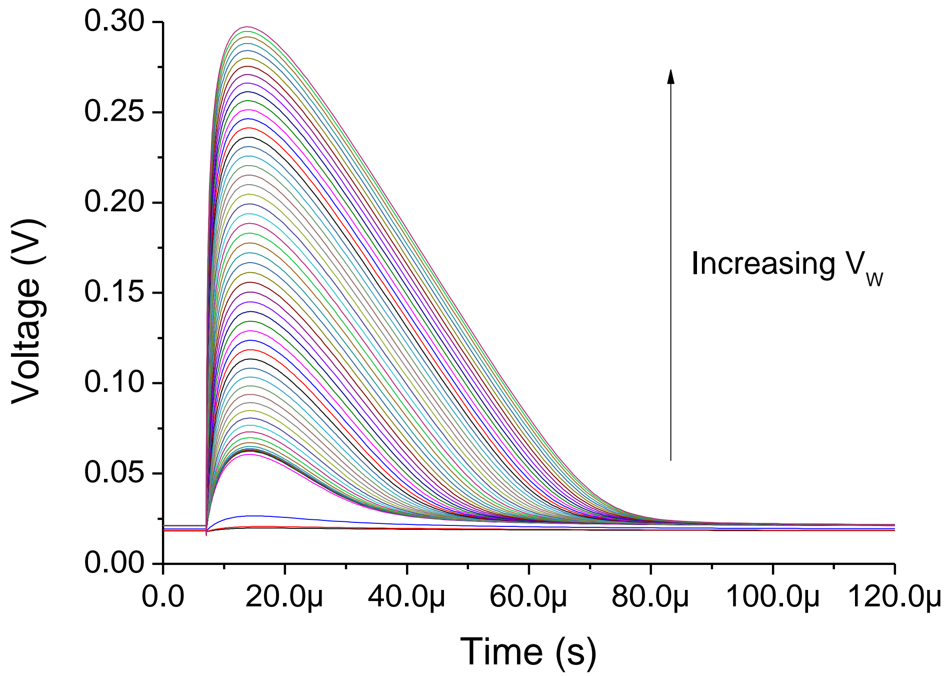

Figure 6.10 - Simulated PSPs from standard neural cell. V_{LEAK} = 0.3V, V_P = 0.3V. V_W step size is 50mV

Having established a standard neural cell, it is possible to create an empirical model which describes its operation. The approach was to undertake appropriately chosen simulations to allow the building of empirical models for ΔV_{PSP}, the rise time and the fall time as a function of $V_W/V_1/V_{LEAK}$. For fixed values of V_1 and V_{LEAK}, the operation of the entire cell can be expressed solely as a function of V_W. Simulated PSPs generated by the standard cell in response to a single synapse firing are shown in Figure 6.10, for $V_{LEAK} = V_P = 0.3$V.

From this, a model for the PSP can be generated:

$t < t_{rPSP}$

$$V_{PSP} = (AV_W + B)t^2 + (CV_W + D)t + E \tag{6.4}$$

$t > t_{rPSP}$

$$V_{PSP} = (FV_W + G)e^{-\frac{(t-t_{rPSP})^2}{2(HV_W+J)^2}} + K \tag{6.5}$$

Parameters A-K are fitting parameters extracted from the simulated results, depending on the value of V_1 and V_{LEAK}.

6.4 Proposed Circuit for XOR Benchmark Problem

Ideally, weight storage and learning circuitry would be integrated into the neural cells on a single chip. For example, in order to implement spike time dependent plasticity (STDP), it is necessary not only to have some mechanism for physically increasing or decreasing the weight voltage, circuitry is also required which can determine the order in which pre- and post-synaptic signals occur. While such circuits do exist [10-12], it is not possible to include an STDP circuit for each synapse, due to the relatively large area required. Given that suitable circuitry is not currently available, it is necessary to consider off-chip methods of storing and updating weights. Provided sufficient pins are available on a fabricated chip, an arbitrary number of adjustable control voltages can be supplied.

The issue of training is more complicated. Online training, where the network weights are adjusted in real time in response to the pattern of inputs and outputs, requires the implementation of some form of learning rule (Hebbian learning or back propagation for example). This would require the use of a PC or microprocessor to play the role of a 'teacher', which monitors and updates weights as necessary. Another option is to use off-line training, whereby the network weights required to solve a particular problem are decided upon in advance. This can be achieved through the use of simulations, where an optimisation technique is used to find the optimal weights.

Consider the sample network shown in Figure 6.11, consisting of 2 input neurons, 10 hidden layer neurons and 1 output neuron. Each neuron can be considered to be the standard neuron cell, the operation of which is described by equations (6.4) and (6.5). The operation of the entire network can be simulated using Matlab, or a similar programming environment. For a given problem, statistical optimisation techniques can be used to find the optimum set of weight voltages.

A genetic algorithm is a particular type of optimisation technique which mimics the process of natural evolution to arrive at a solution. For a given problem, an initial population of data sets, in this case the weight voltages, are randomly generated. Each individual in the population is evaluated according to some fitness function, which assesses how well they solve the problem.

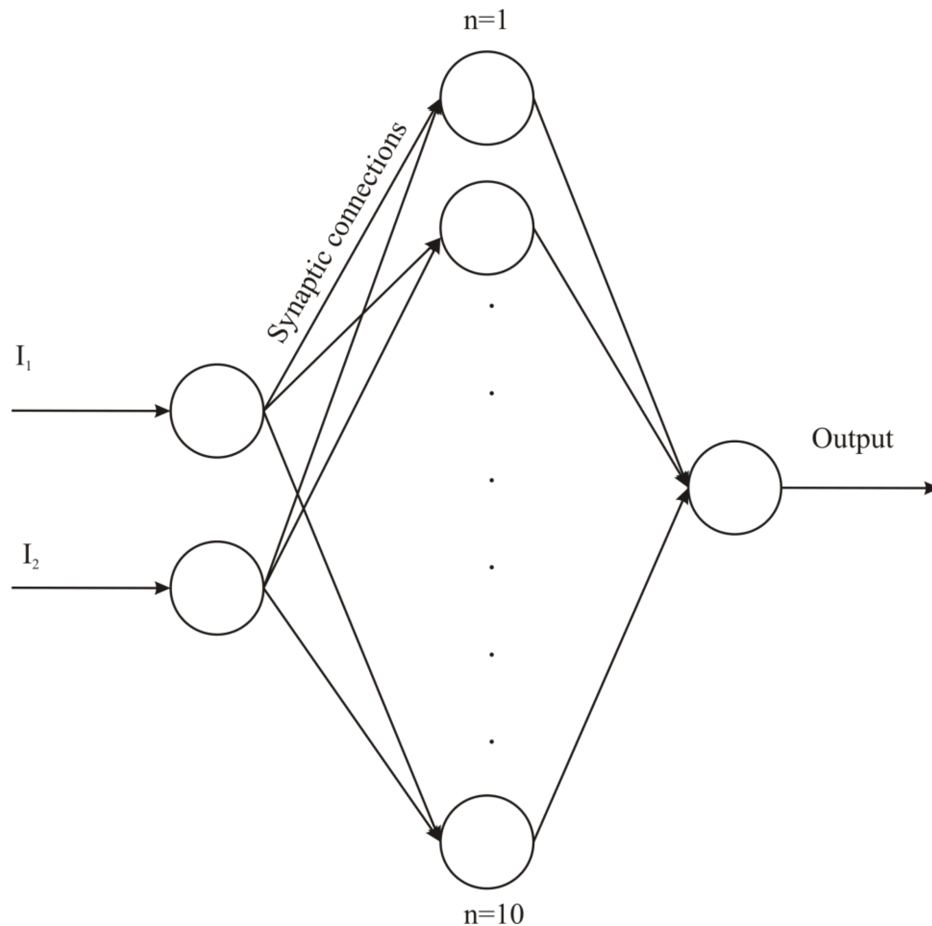

Figure 6.11 - 2-10-1 fully connected network of standard neuron cells.

Individuals are then ranked according to their fitness, with the highest ranking selected to be propagated to the next generation of solutions; where they are heuristically modified, generally through either a recombination or random mutation, in an attempt to more fully explore the solution space. This process can be repeated for a set number of generations, or until a suitably accurate solution is found. Providing a solution to the problem exists, the genetic algorithm will be able to find it, given enough time. Assuming that the algorithm is able to find a suitable set of weights, the chosen values can then be applied to a hardware version of the neural network, to evaluate its real world performance. A genetic algorithm is well suited to this type of problem as it can handle data sets consisting of multiple variables over a large search space, in complex problem domains, without any understanding of how the system itself operates.

A number of benchmark problems have been adopted in neuromorphic engineering; for example and in order of complexity, the exclusive-OR (XOR) problem, the Iris flower data set and the Wisconsin breast cancer data set.

I_1	I_2	Output
0 (10us)	0 (10us)	0 (30us)
0 (10us)	1 (20us)	1 (40us)
1 (20us)	0 (10us)	1 (40us)
1 (20us)	1 (20us)	0 (30us)

Table 6.1 - Truth table for exclusive-or problem. Equivalent spike times are given in brackets.

An example of how the GA approach could be used to solve the exclusive-or (XOR) problem, the truth table for which is given in Table 6.1, is described below. As the network is designed to work with spiking inputs, it is first necessary to map the 0s and 1s to spike firing times, which are chosen relative to an arbitrary start time, t = 0s. In this case, a '0' at one of the inputs is mapped to a spike at 10us and a '1' to a spike at 20us. For the output, 0 and 1, spiking times of 30us and 40us are chosen. This means that when a 0 is desired at the output, a spike should be generated at 30us; when a 1 is desired, a spike should be generated at 40us. An example of the desired situation is shown in Figure 6.12, which shows inputs firing at 10us and 20us, corresponding to (0,1), and an output firing at 40us, corresponding to a 1. As previously stated, the entire network can be simulated using Matlab. The predicted output spike firing times can be measured for each set of weights. The fitness can be evaluated by comparing the measured firing time with the desired firing time, for the four different combinations of inputs. While it would be possible to implement the theoretical model developed in previous chapters in Matlab, more accurate results will be generated if the empirical model developed in Section 6.4 is used.

Initial work by project collaborators has shown that, for the network shown in Figure 6.11, a set of weights can be found which will solve the XOR problem with a SSE (Sum of Squared Errors) of less than 3% [13]. Work is ongoing to produce a test chip containing the network, the operation of which can be compared against these initial findings.

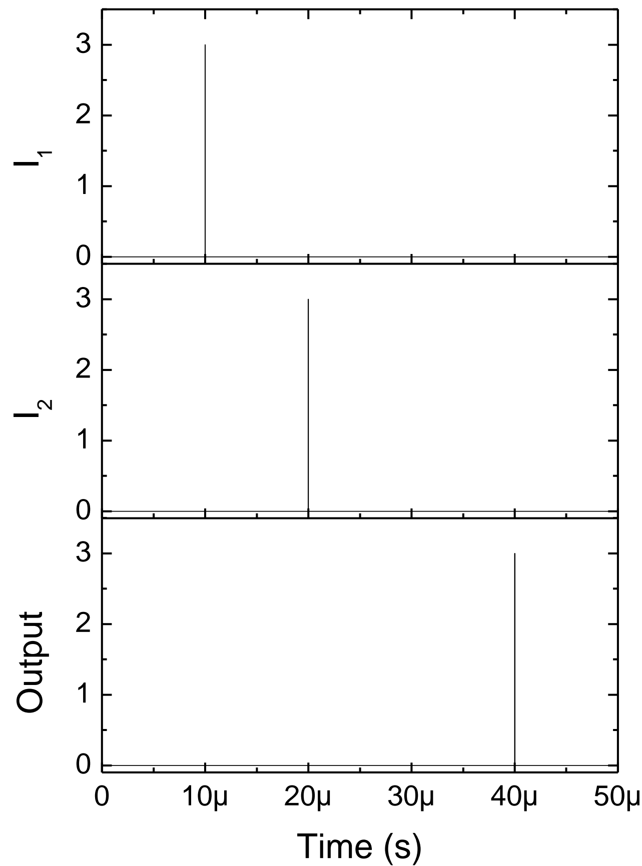

Figure 6.12 - Illustration of desired outcome for an input of (0,1). A output spike fires at 40us, corresponding to a 1.

6.5 Conclusions

A number of the issues which arise when scaling to large numbers of neural cells have been considered in this chapter. Where possible, potential solutions have been outlined with examples given. However, until fabricated chips containing such networks are produced, the exact effects cannot be fully investigated. Further work is required on the creation of large scale networks. To this end, a standard neuron cell has been proposed, several of which can be joined together in modular fashion to create networks of spiking neurons. A possible application of the neuron cell, to solve the XOR problem, has also been outlined.

References

[1] S. Huang and T. Dowrick, Internal report, EPSRC Project EP/F05551X/1 'A biologically plausible spiking neuron in hardware.', 2009-2001.

[2] L. A. P. Melek, M. C. Schneider, and C. Galup-Montoro, "Body-bias compensation technique for subthreshold CMOS static loqic gates," in Proceedings - 17th Symposium on Integrated Cicuits and Systems Design, SBCCI2004, Pernambuco, 2004, pp. 267-272.

[3] G. Paci, D. Bertozzi, and L. Benini, "Effectiveness of adaptive supply voltage and body bias as post-silicon variability compensation techniques for full-swing and low-swing on-chip communication channels," in Proceedings -Design, Automation and Test in Europe, DATE, Nice, 2009, pp. 1404-1409.

[4] R. H. S. Carpenter, Neurophysiology, 4th ed.: Arnold, 2003.

[5] W. E. G. Gerald M. Edelman, W. Maxwell Cowan, Synaptic Function: Wiley-Interscience, 1987.

[6] C. F. Stevens and A. M. Zador, "Input synchrony and the irregular firing of cortical neurons," Nature Neuroscience, vol. 1, pp. 210-217, 1998.

[7] P. Kara, P. Reinagel, and R. C. Reid, "Low response variability in simultaneously recorded retinal, thalamic, and cortical neurons," Neuron, vol. 27, pp. 635-646, 2000.

[8] J. S. Christopher Saint, IC Mask Design: Essential Layout Techniques, 1 ed.: McGraw-Hill Professional, 2002.

[9] F. Tuffy, L. J. McDaid, V. W. Kwan, J. Alderman, T. M. McGinnity, J. A. Santos, P. M. Kelly, and H. Sayers, "Inter-neuron communication strategies for spiking neural networks," Neurocomputing, vol. 71, pp. 30-44, 2007.

[10] A. Bofill-i-Petit and A. F. Murray, "Synchrony detection and amplification by silicon neurons with STDP synapses," IEEE Transactions on Neural Networks, vol. 15, pp. 1296-1304, 2004.

[11] G. Indiveri, E. Chicca, and R. Douglas, "A VLSI array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity," Neural Networks, IEEE Transactions on, vol. 17, pp. 211-221, 2006.

[12] H. Tanaka, T. Morie, and K. Aihara, "A CMOS circuit for STDP with a symmetric time window," International Congress Series, vol. 1301, pp. 152-155, 2007.

[13] A. Ghani, University of Ulster, Internal report, EPSRC Project EP/F05551X/1 ‘A biologically plausible spiking neuron in hardware.’, 2009-2011.

Chapter 7: Conclusions and Further Work

Neural hardware offers to be a promising alternative to conventional computational paradigms. Using conventional silicon processing techniques, it is possible to replicate the functionality of biological systems, with the aim of solving certain problems; for example pattern recognition, image processing and forecasting are particularly suited to neural circuit approaches. The inherently fault tolerant structure of the brain also provides inspiration for creating more robust, highly parallel systems, with built-in fault tolerance [1]. In this thesis, a number of devices have been presented with this aim. In this chapter, a summary of the work undertaken and ideas for further work are presented.

In Chapter 1, a review of the current state of neural networks in hardware was undertaken. A number of key requirements of neural circuitry were identified, namely the need for small area and highly scalable devices and circuit architectures, biological plausibility and low power consumption.

Chapter 2 gave a summary of the relevant device physics of MOS capacitors and transistors. Device parameters to be used throughout the thesis were extracted and compared to nominal values provided by AMS where appropriate.

A compact spiking synapse cell was presented in Chapter 3. A two-terminal device was shown to implement synaptic plasticity through an adjustable weight voltage, V_W. A more advanced three-terminal device introduces an additional control voltage, V_P, to allow for the implementation of additional neurological characteristics, namely depression and refraction. Both devices have a small silicon footprint, with dimensions of 2.1μm x 5.0μm and 2.1μm x 6.2μm when fabricated in a 0.35μm process, which is a smaller footprint than any similar device currently reported in the literature. A theoretical model for the operation of the synapse was developed and shown to be in good agreement with simulated and experimental results. Results obtained from simulations and measured from fabricated chips were analysed, to illustrate the ability of the device to implement the key synaptic features described, in a comparable manner to their biological counterparts.

In Chapter 4, it was shown how the synapse can be integrated with the neuron circuit. A current mirror is used to temporally sum multiple synaptic outputs. Biologically plausible post synaptic potentials are generated, with a controllable decay period implemented through

a MOSFET biased in subthreshold. A theoretical model was again developed, supported by physically measured and simulated results. Triggering circuitry compatible with the neuron circuit was also discussed, allowing the circuit to produce spiking outputs. It was shown how the synapse circuit can be operated in either an excitatory or inhibitory fashion, depending on how it is connected to the neuron circuit. At this stage, the triggering circuitry and inhibitory synapses have only been tested through simulations. Devices laid out on a test chip are required in future to confirm their validity.

A circuit block for the implementation of an axonal delay was discussed in Chapter 5, which requires fewer transistors than existing circuits with equivalent functionality. A subthreshold MOSFET feeding an inverter chain introduces a delay between a pre- and post-synaptic neuron. It was shown that varying the gate voltage of the MOSFET produced a variation of the delay time between several microseconds and tens of milliseconds. It was also shown, through simulations, how the synapse, axon and neuron could be combined together to create a functioning neural cell.

Chapter 6 addresses some of the issues which will be encountered when scaling the neural circuit blocks to become part of a VLSI implementation. The effects of device variability and scaling were considered and methods of preventing and dealing with the issues encountered were presented. A key issue is that of the area consumed by metal interconnects, as the area consumed can increase exponentially with the number of devices. A standard, modular, neural cell was also proposed, consisting of twenty synapses and a single neuron, striking a balance between a high degree of connectivity and the amount of area consumed by metal interconnects. It was estimated that it would be possible to have up to 270 separate neural cells on a single $1mm^2$ chip. Multiple cells can be multiplexed together to form larger networks, with arbitrary topologies. The challenges of weight updates and learning methodologies were also considered. While it would be preferable to have all of the training and weight update circuitry integrated with the neural cells on a single chip, there is currently an absence of such circuitry. In light of this, it is necessary to use offline training methods. An implementation of a network using 13 standard cells to solve the XOR problem was outlined, where a genetic algorithm is used to find the optimal weights, which can then be applied to the synapses using appropriate external circuitry.

The ability of the devices described in this thesis to produce biological plausible outputs has been confirmed by the theoretical analysis and results presented. Further work should

concentrate on the creation of networks of devices, such as the one proposed for the solution to the XOR problem, in order to investigate the potential of the circuitry to create functional, large scale neural networks. The design of a chip containing the standard neural cell and the XOR network is currently being undertaken. A number of other substantial challenges also exist. The most suitable methods of connecting large scale networks of spiking neurons together should be investigated. As previously stated, conventional metal interconnects can quickly dominate the chip area as networks are scaled up. Methods for the storing and updating of synaptic weights are also required. While it is possible to set the synaptic weights externally for small networks, this approach would not be practical when hundreds or thousands of synapses are involved. Currently, methods of implementing learning algorithms do exist, but the area consumed precludes their inclusion in large scale systems. Even techniques which are relatively simple to implement in software, such as STDP, introduce additional complexity when incorporated into a hardware system, due to the need for signal coincidence detection. In the future it may be necessary to develop alternative learning algorithms which can be tailored to the strengths and weakness of neural hardware..

References

[1] S. Furber and S. Temple, "Neural systems engineering," *Journal of the Royal Society Interface,* vol. 4, pp. 193-206, 2007.

Appendix – Test Chip Design and Fabrication

A1.1 Introduction

All prototype chips were fabricated in a three metal 0.35µm n-well process from Austria MicroSystems (AMS). Fabrication was coordinated through the Europractice service and performed at the IMEC facility in Belgium. Layouts were created and verified using the Cadence software package configured for the 0.35µm AMS process. Two separate prototype chips were produced. The first was received in May 2008 and the second in June 2009.

This appendix details the procedure of preparing the prototype chips for fabrication and subsequent testing. Creation of circuit layouts, routing, electrostatic discharge (ESD) prevention, output buffering, chip contents, packaging, PCB design and the experimental setup are considered in detail.

A1.2 Circuit Layouts

Initially, customisable layouts were created for the synapse, neuron and axon cells, shown in Figure A13 - Figure A16. Having done this, all further layouts could be completed by placing these cells and adjusting the transistor dimensions where necessary. Large area (100µm x 100µm) MOS capacitors and MOSFETs using standardised layouts provided by AMS were included for the purposes of extracting device parameters.

Bondpads were included at output nodes. A bondpad is a 70µm x 70µm arrangement of the three metal layers stacked upon each other with multiple vias connecting the layers. The top metal layer of the pad is left exposed and can either be probed directly or wire bonded. The predefined layout for a bondpad is provided by AMS and is shown in Figure A17. The capacitance of such a pad can be conservatively estimated from a back annotated simulation to be approximately 100fF.

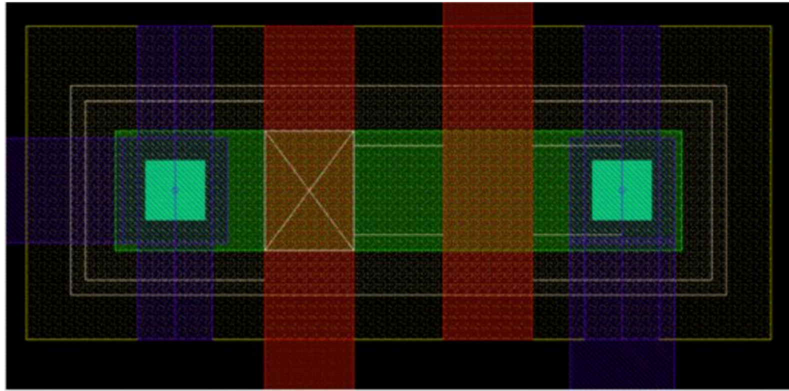

Figure A13 - Two gate synapse layout (2.1µm x 5.0µm).

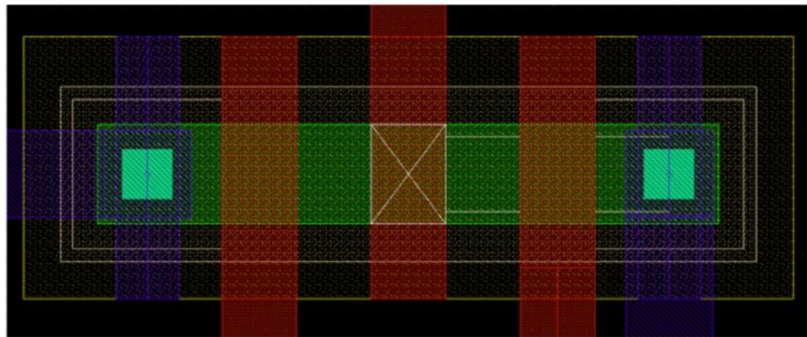

Figure A83 - Three gate synapse layout (2.1µm x 6.2µm).

Voltages are supplied to the chip through bondpads at the periphery of the chip. The supply rail and input voltage connections are taken from the bondpads and routed in a grid structure around the chip. Primary horizontal and vertical voltage rails are implemented on metal layers 35µm wide. At this width a DC current of 35mA and a peak AC current of 1050mA can be carried by the voltage rails[1]. The maximum expected DC/AC current levels for the whole chip are approximately 10uA and 25mA. The V_{DD} and ground connections are routed onto the chip from multiple bondpads. This creates parallel connections which reduces the overall resistance of these rails. Voltages are supplied to the individual neuron/axon circuits by routing narrower metal paths from the primary lines to the circuit nodes. The final layouts for the two test chips are shown in Figure A87 and Figure A88.

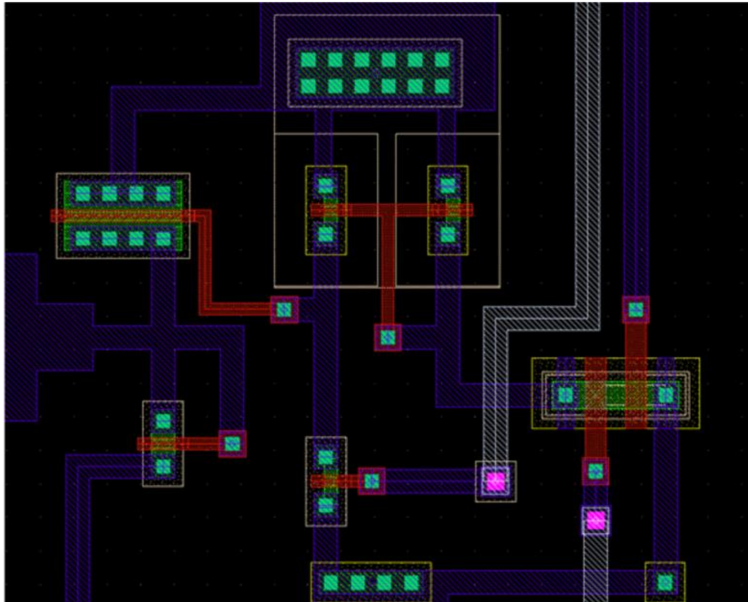

Figure A84 - Neuron layout (17.5µm x 19.4µm).

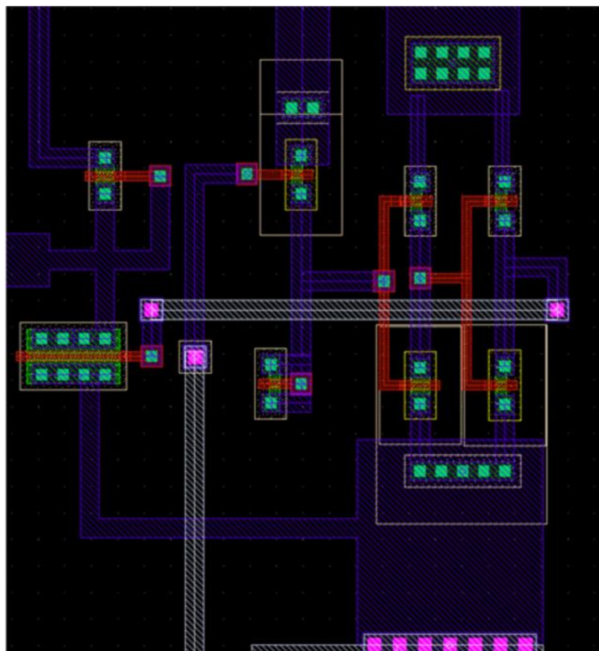

Figure A16 - Axon layout (17.2µm x 22.3µm).

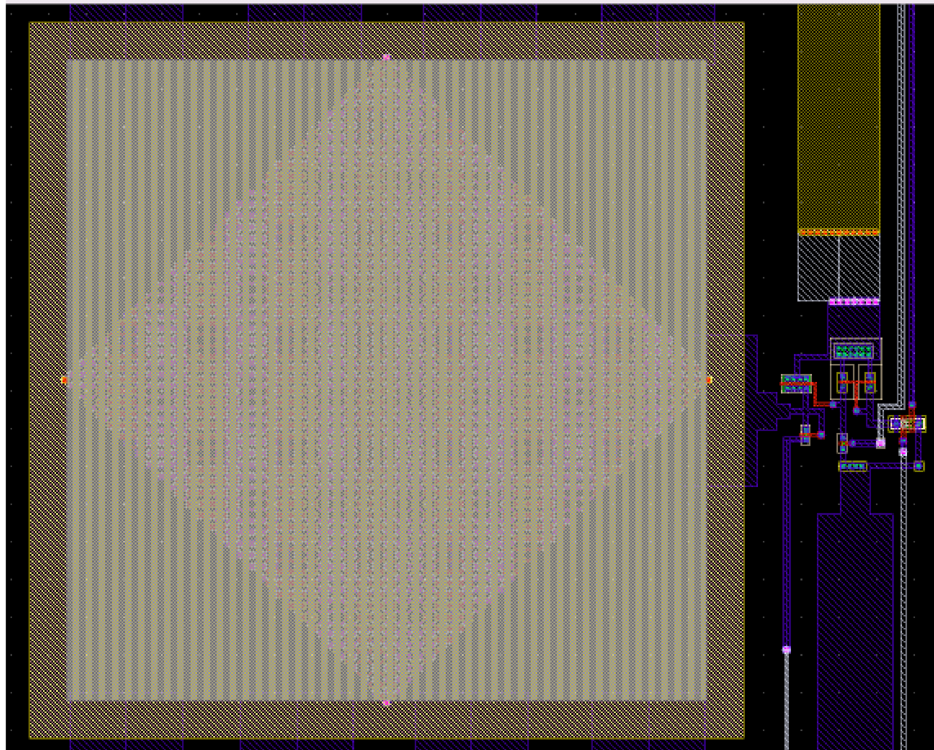

Figure A17 - Output bondpad connected to neuron circuit.

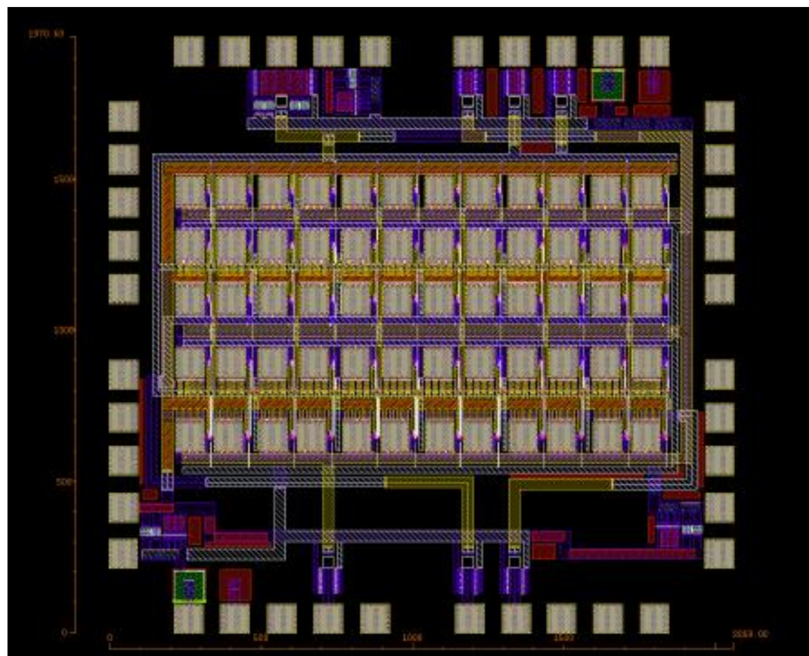

Figure A87 - Layout of 1st test chip. Dimensions are 2.059mm x 1.970mm. Area = 4.056mm^2.

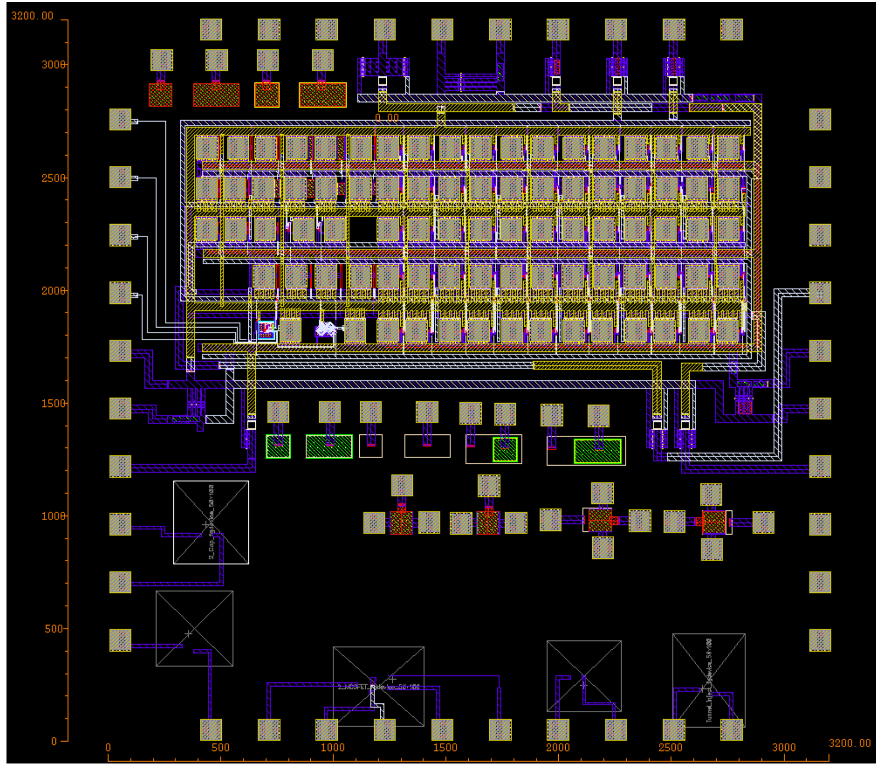

Figure A88 - Layout of 2nd test chip. Dimensions are 3.200mm x 3.200mm. Area = 10.240mm^2.

A1.3 Output Buffers

In order for the synapse/neuron/axon circuit to effectively drive the capacitance of an output pad, it is necessary to include output buffering circuitry. Figure A89 shows how a source follower on split supply rails is used as an output buffer. In the situation shown, the buffer is being used to measure V_{PSP}. There is a corresponding output buffer measuring the value of V_{IN} and one for the axon circuit. The buffering circuitry should not significantly alter the operation of the circuit being measured.

V_{SS} must be set such that M6 and M7 are always operating above threshold. The worst case scenario corresponds to $V_{PSP} = 0V$. At this point, M6 will be at threshold when:

$$-V_{OUT} = V_T \tag{A6}$$

M7 will be at threshold if:

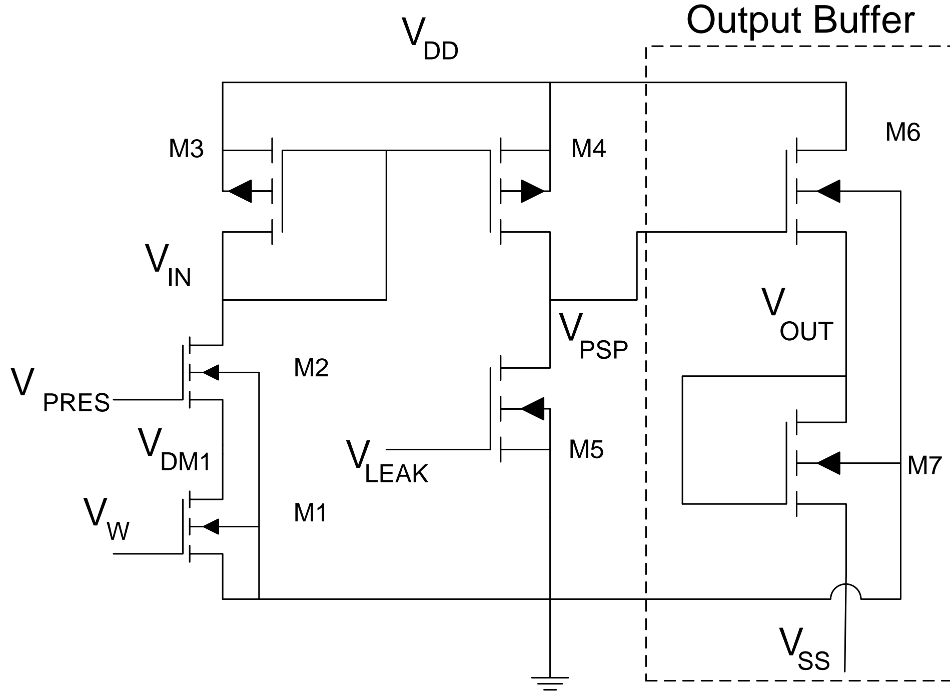

Figure A89 - Source follower connected to the V_{PSP} node acting as an output buffer. Split supply rails are used for the buffer.

$$V_{OUT} - V_{SS} = V_T \tag{A7}$$

Combining (2.24) and (A7) gives:

$$V_{SS} = 2V_{OUT} \tag{A8}$$

M6 will have a substrate bias equal to V_{OUT}, while a negative V_{SS} will introduce a substrate bias to M7. The substrate bias effect shifts the threshold voltage by an amount:

$$\Delta V_T = \gamma\left(\sqrt{V_{SUB} + 0.82} - \sqrt{0.82}\right) \tag{A9}$$

The dimensions of M6 and M7 must be chosen such that the output buffer is able to match the slew rate of the original signal, taking into account the capacitive loading of the output pad. Estimates for the required rising/falling slew rates are 1V/ns and 0.1V/ns. With C_L = 100fF, this corresponds to currents of 100uA and 10uA. Given that the rise time is set by M6 and the fall time by M7, this suggests M6 should have an aspect ratio ten times that of M7.

An estimate for the resting value of V_{OUT} can be found by equating the currents in M6 and M7 for V_{PSP} = 0V, where both are saturated. The effect of substrate bias on M6 and M7 is ignored:

$$\frac{\beta}{2}\frac{W_{M6}}{L_{M6}}(V_{PSP}-V_T-V_{OUT})^2=\frac{\beta}{2}\frac{W_{M7}}{L_{M7}}(V_{OUT}-V_T-V_{SS})^2 \tag{A10}$$

Simplifying, substituting V_{PSP} = 0V and rearranging for V_{OUT}:

$$V_{OUT}=\frac{\left(\sqrt{\frac{W_{M7}}{L_{M7}}}-\sqrt{\frac{W_{M6}}{L_{M6}}}\right)V_T+\sqrt{\frac{W_{M7}}{L_{M7}}}V_{SS}}{\left(\sqrt{\frac{W_{M7}}{L_{M7}}}-\sqrt{\frac{W_{M6}}{L_{M6}}}\right)} \tag{A11}$$

With V_T = 0.46V and taking $W/L_{M6} = 10W/L_{M7}$,

$$V_{OUT}=\frac{-1.45+V_{SS}}{4.16} \tag{A12}$$

Combining (A7) and (A12) gives V_{SS} =-1.36V for which M6 and M7 will always be above threshold. To ensure correct operation of the circuit, an operating value of V_{SS} = -1.5V was chosen.

Having chosen a value for V_{SS}, the effect of substrate bias in (A10) can be considered and a more accurate quadratic expression relating V_{OUT} to the value of V_{PSP} can be generated:

$$17.3V_{OUT}^2-(9.0-26.3V_{PSP})V_{OUT}+(1.4-3.1V_{PSP})^2-2.6=0 \tag{A13}$$

which can be solved to give V_{OUT} as a function of V_{PSP}.

The worst case current flow through M6/M7 will be for small values of V_{PSP}. Taking the case of V_{PSP} going from 0V to 0.1V. This corresponds to V_{OUT} going from -0.58V to -0.51V. The initial current through M6 is:

$$I_D=\frac{\beta}{2}\frac{W_{M6}}{L_{M6}}(V_{PSP}-V_{T0}+\Delta V_T-V_{OUT})^2 \tag{A14}$$

As the value of V_{OUT} at this point is negative, the threshold voltage of M6 is decreased, hence the + ΔV_T term which can be calculated using (A9). Substituting β = 170uA/V^2, V_{PSP} = 0.1V, V_T = 0.46V, V_{OUT} = -0.58V, γ = 0.56 (Value taken provided by AMS) and W_{M6}/L_{M6} = 10; (A14) gives I_D = 120uA, which exceeds the initial specification of 100uA.

When V_{PSP} is discharging, the current through M7 will be:

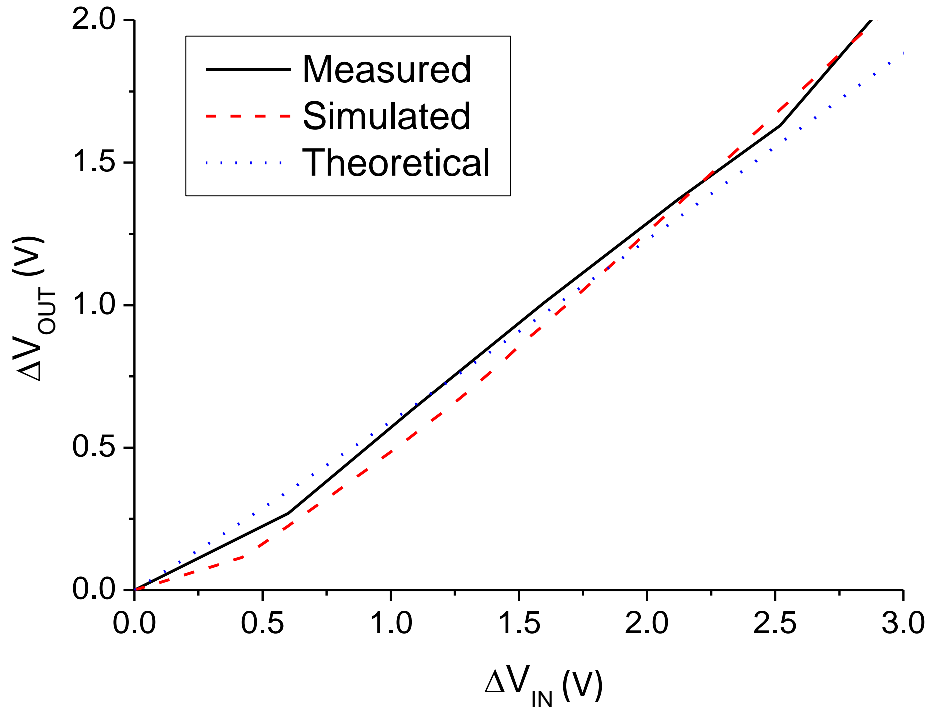

Figure A90 – Simulated, theoretical and measured buffer transfer characteristic.

$$I_D = \frac{\beta}{2}\frac{W_{M7}}{L_{M7}}(V_{OUT} - V_T + \Delta V_T - V_{SS})^2 \qquad \text{(A15)}$$

Again, the negative value of V_{SS} decreases the threshold voltage of M7. Taking V_{OUT} = -0.51V, V_{SS} = -1.5V and W_{M7}/L_{M7} = 1 gives I_D = 65uA, which again exceeds the initial specification of 10uA.

To extract the actual voltage at V_{PSP} from the output of the buffer, it is necessary to have an accurate expression for the gain of the buffer. While (A13) can be used to give approximate values, more accurate results will be obtained if the transfer characteristics of the buffer is physically measured. Figure A90 plots the transfer characteristic of the buffer (ΔV_{OUT} vs. ΔV_{IN}); values measured from the fabricated chips are shown alongside simulation results and theoretical values. There is reasonable agreement between the simulated and measured values across the entire voltage range. The theoretical curve has a similar gradient to the other results, except for low values of V_{IN}, where a mismatch can be seen. At low values of V_{IN}, M6 and M7 will be operating near the threshold voltage, with lower currents, slightly reducing the gain of the buffer. Curve fitting tools were used to produce an expression for the gain of the buffer, based upon the results measured from the fabricated chips:

$$\frac{V_{OUT}}{V_{IN}} = 0.21 + 0.45\left(\frac{(V_{OUT})^{1.71}}{0.27(V_{OUT})^{1.71}}\right) \quad (A16)$$

Where experimental results have been presented throughout this thesis, the values shown are those calculated using (A16), rather than the measured voltages from the output buffer.

A1.4 ESD Protection

Unforeseen electrostatic discharge (ESD) events can have disastrous consequences for ICs, ranging from increased leakage currents to complete breakdown of dielectric structures. Thin gate oxides are especially prone to damage. The Human Body Model (HBM) represents the ESD from human contact with an IC and is the most commonly used model for IC development. HBM events are typically 1kV or more in magnitude [2]. Steps to guard against damage from ESD events is taken at the circuit design/layout level; AMS provides a set of design rules to achieve a minimum ESD protection rating of 2kV-HBM[3].

ESD Protection is required between the V_{DD} and ground rails and for each of the input and output pads present on the chip. The ESD design rules differ for input/output pads depending on whether the signal they carry is analog or digital. All of the input/output signals, with the exception of V_{PRES}, can be considered analog signals and utilise a common ESD protection scheme. Figure A91 shows this ESD protection circuitry; the ESD protection between V_{DD} and ground. Figure A92 shows the ESD protection for the V_{PRES} pad.

Two types of ESD protection devices are used. The first is a large area diode (D_1 – D_5). Under normal operating conditions, all diodes are reverse biased. The occurrence of an ESD event on one of the inputs or on the ground rail will forward bias the affected diodes and excess current is conducted towards the V_{DD} rail, protecting the internal circuitry. Figure A93 shows the standard layout of an ESD protection diode, as specified by AMS [3].

The second ESD protection device is the snapback device (S_1-S_3). The snapback device present in the schematic of Figure A91 forms a parasitic bipolar transistor between V_{DD} and ground.

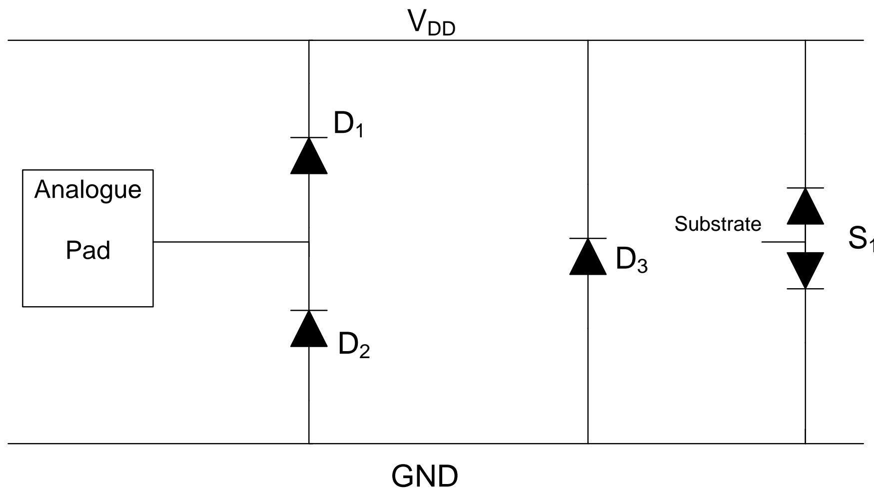

Figure A91 - ESD protection for input/output pad (D_1, D_2) and between V_{DD} and GND (D_3, S_1).

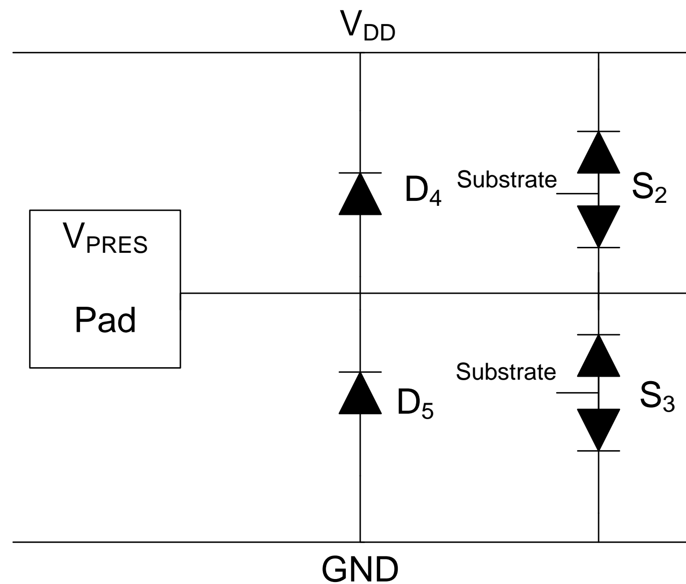

Figure A92 - ESD protection for V_{PRES} pad.

V_{DD} and ground function as the collector and emitter terminals respectively, the substrate functions as the base terminal. Under normal operating conditions, the collector-base junction is reverse biased below the breakdown voltage of the junction. An ESD event of sufficient magnitude will increase the reverse bias on the junction such that avalanche breakdown occurs. Holes move towards the substrate contact and the base-emitter junction becomes forward biased. As the base-emitter voltage reaches 0.7V, the parasitic transistor turns on and the ESD current flows to the ground rail. A layout of the snapback device is shown in Figure A94.

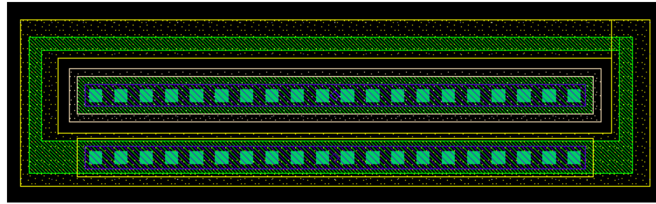

Figure A93 - Layout of an ESD protection diode.

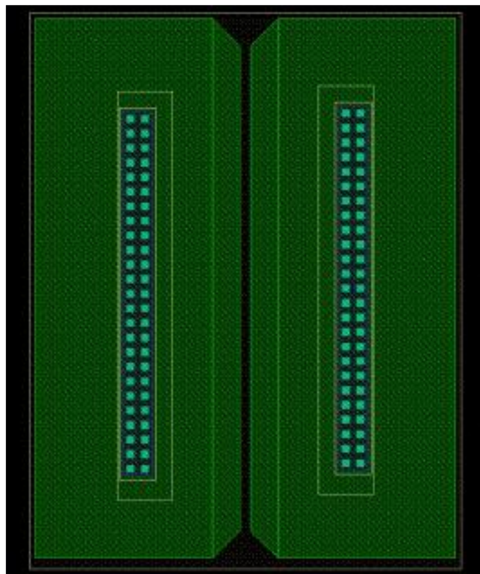

Figure A94 - Layout of a snapback ESD protection device.

A1.5 Chip Packaging

The final step of the IC fabrication process is the packaging of the chips. Each die is mounted onto one of the chosen package types (Dual In Line, Pin Grid Array, Ceramic Leaded Chip Carrier etc.). Where required, bondpads on the chip can be connected to the corresponding bondpins on the package.

The package which was chosen for the fabricated chips was the 40 pin DIL package. The 40 available pins can accommodate all of the input requirements for the chip and it is easily mounted onto a printed circuit board. Figure A95 shows a schematic view of the DIL Package. Figure A96 and Figure A97 show the correspondence between the bondpads on the chips and the pins on the package for the two fabricated chips. In addition to the connections between bondpads and pins, a connection is made from the package directly to the substrate for each chip.

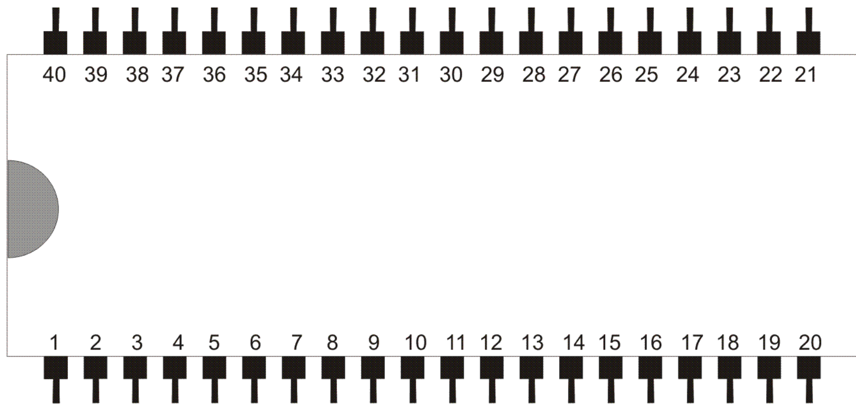

Figure A95 - Schematic view of 40-pin DIL package.

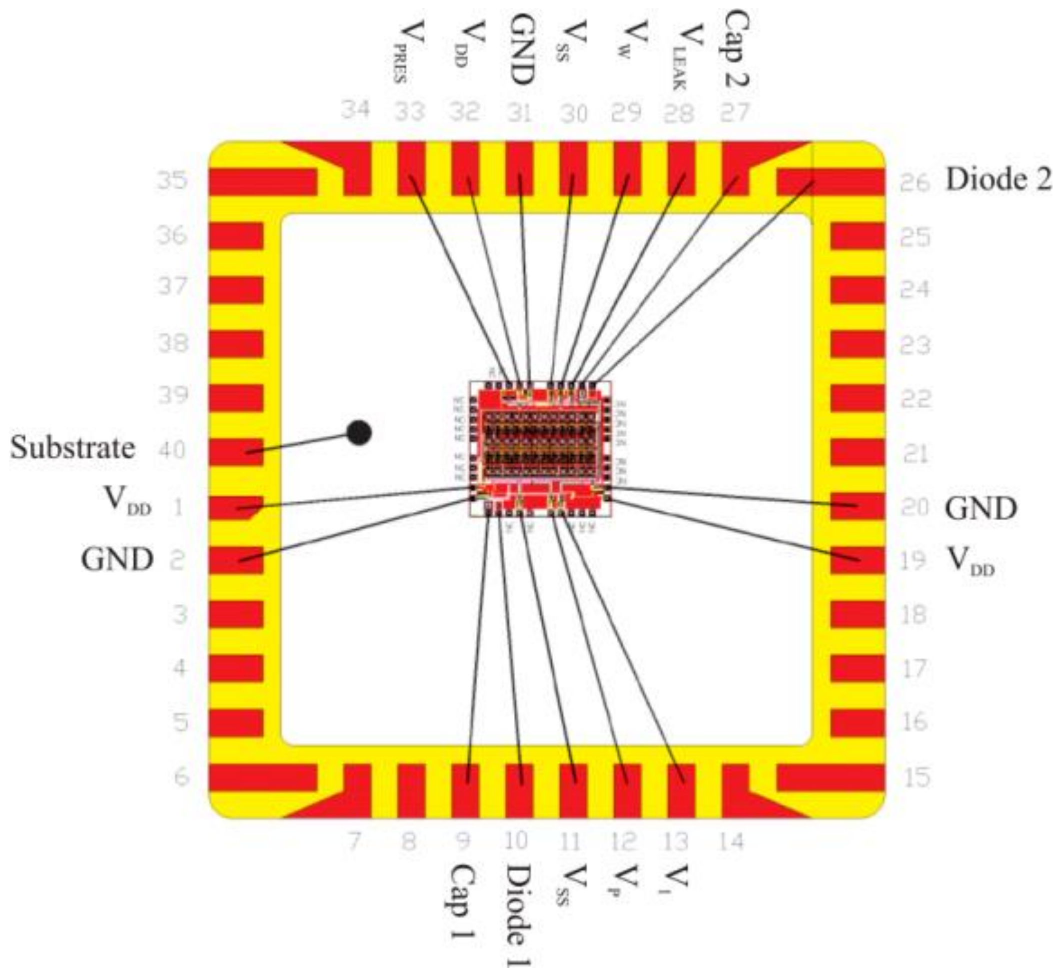

Figure A96 - Layout of 1st chip in 40-pin DIL package.

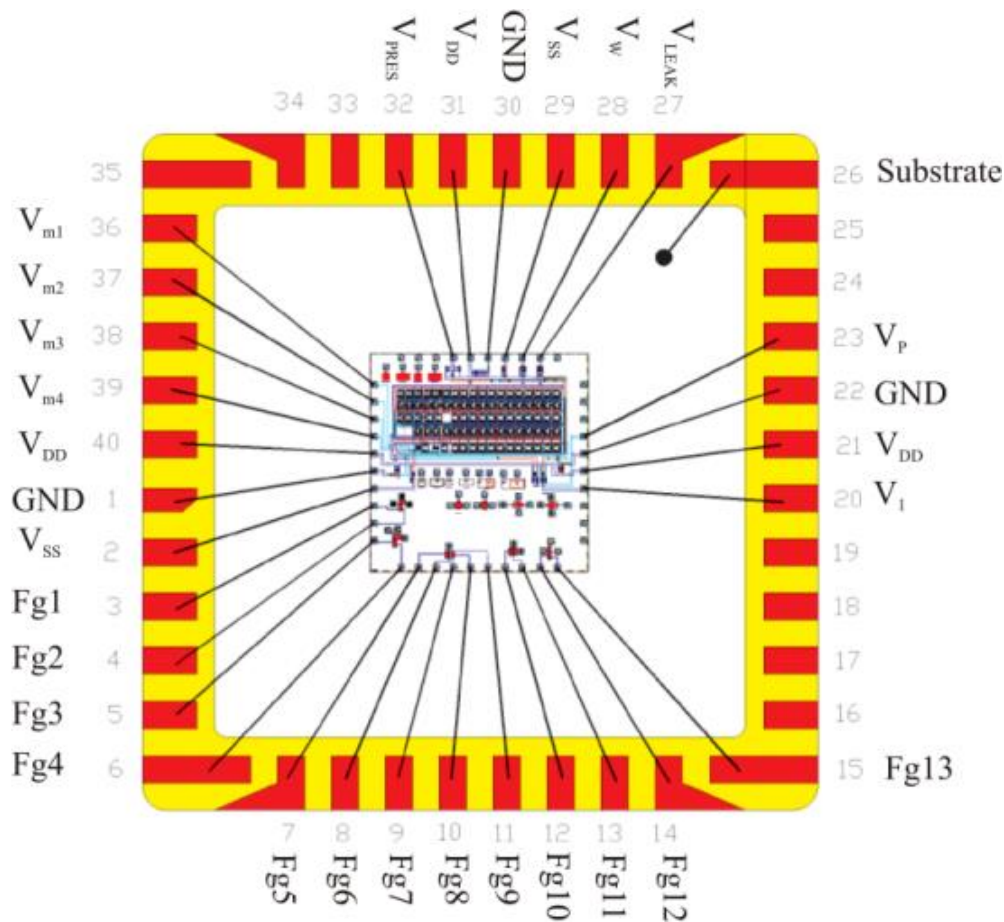

Figure A97 - Layout of 2nd chip in 40-pin DIL package.

A1.6 Measurement Setup

A packaged chip from the first round of chip fabrication is shown in Figure A98. A printed circuit board (PCB) was created to house the chip, Figure A99. A 40-pin zero insertion force (ZIF) socket was mounted to the board. To supply voltage inputs to the chip, several angled BNC sockets were also mounted; their input pins were routed to the appropriate pins on the ZIF socket. This process was repeated for the second set of fabricated chips.

DC voltages were supplied to the PCB using an HP416A Semiconductor Parameter Analyser, which can supply up to eight independent, controllable DC voltages. An Agilent 33220A Pulse Generator was used to supply the voltage pulse to the V_{PRES} terminal.

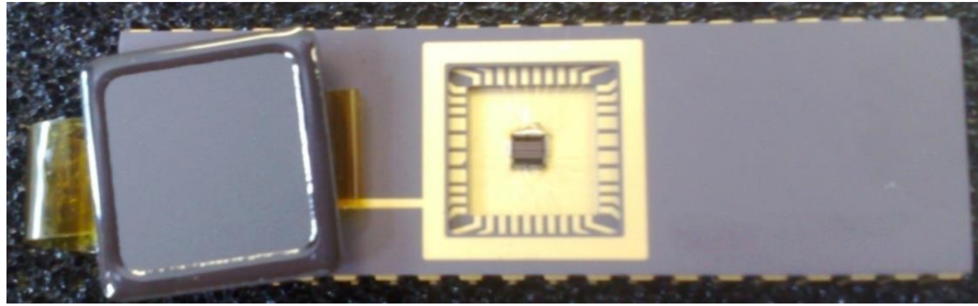

Figure A98 - Fabricated chip in 40-pin DIL package with taped lid.

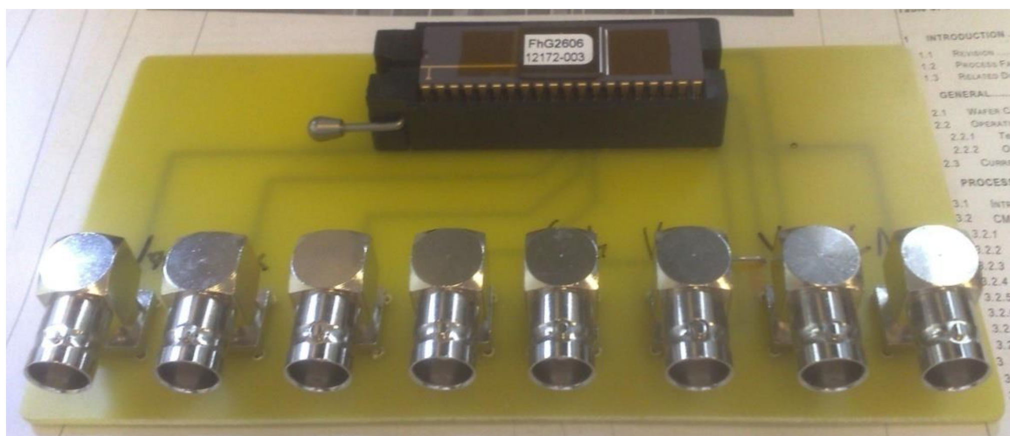

Figure A99 - PCB used to interface with chip.

Measurements were taken from the chip by direct probing of output pads using a probe station. Voltage waveforms were recorded using a Tektronix TDS1012 Oscilloscope which can be interfaced with a PC for the storage and analysis of waveforms.

References

[1] AMS, "0.35µm CMOS C35 Process Parameters," July 2007.

[2] O. Semenov, *ESD protection device and circuit design for advanced CMOS technologies*: Springer, 2008.

[3] AMS, "0.35µm ESD Design Rules," January 2007.