Biorealistic Learning on Memrisitive Network

Viet Cuong Vu

Principal Supervisor: Anthony J. Kenyon

Subsidiary Supervisor: Adnan Mehonic

Faculty of Engineering Sciences
University College London

This dissertation is submitted in fulfilment for the degree of
Doctor of Philosophy

July 2025

Declaration

I, Viet Cuong Vu, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Viet Cuong Vu
July 2025

Abstract

latex.

Impact Statement

latex.

Contributions

List of Publications

- V. C. Vu, D. J. Mannion, D. Joksas, W. H. Ng, A. Mehonic, and A. Kenyon, “Spiking Neural Networks with Silicon Oxide Memristive Devices in the Subthreshold Regime,” in 2025 IEEE 5th International Conference on Software Engineering and Artificial Intelligence, Jul. 2025, pp. 340–344.
- V. C. Vu, A. Kenyon, D. Joksas, A. Mehonic, D. J. Mannion, and W. H. Ng, “Spiking Neural Networks with Nonidealities from Memristive Silicon Oxide Devices,” in 2024 IEEE 24th International Conference on Nanotechnology, Jul. 2024, pp. 46–50.
- D. J. Mannion, V. C. Vu, W. H. Ng, A. Mehonic, and A. J. Kenyon, “Unipolar Potentiation and Depression in Memristive Devices Utilizing the Subthreshold Regime,” IEEE Transactions on Nanotechnology, vol. 22, pp. 313–320, 2023.

Conference Presentations

- "Unipolar Potentiation and Depression within Optically Active Memristive Devices Subthreshold Regime", EMRS Spring 2025.
- "Circuit-Based Modelling of Current Transients within the Memristive Devices Subthreshold Regime", MEMRISYS 2024.
- "A Compact SPICE Model for Current Transients within the Subthreshold Regime of Memristors", IEEE MetroXRAINE 2023.

Acknowledgements

latex.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Research Context

1.2 Motivations and Challenges

1.3 Thesis Outline

Chapter 2

Theoretical Foundations

2.1 Overview

Neuromorphic computing aims to bridge the gap between neuroscience and artificial intelligence by emulating the structural and functional principles of biological neural systems. At the heart of this vision is a pressing research question: how can we implement biorealistic learning on memristive networks to develop energy-efficient, scalable and adaptive computing architectures? Addressing this question requires a multidisciplinary approach, combining insights from neurobiology, electronic materials science and computational modelling.

Traditional von Neumann architectures [258], characterised by their separation of memory and processing units, struggle to efficiently handle tasks that the human brain performs effortlessly, such as pattern recognition, sensory integration, and decision making. In contrast, the brain achieves these feats with remarkable energy efficiency and adaptability, in part due to the tightly coupled nature of computation and memory in its neural circuits. Mimicking these biological characteristics in silicon and emerging nanotechnologies has become the guiding principle of neuromorphic engineering [215].

Recent advances in memristive technologies have reignited interest in neuromorphic computing. Memristors, or memory resistors, are two-terminal non-volatile devices that can emulate synaptic plasticity by adjusting their conductance based on the history of voltage and current applied. This property lend itself to naturally support learning rules such as Hebbian learning [89] and spike-timing-dependent plasticity (STDP) [176]. When arranged in crossbar arrays, memristors offer a promising platform for in-memory computation, which can significantly reduce the power and latency associated with traditional data transfer bottlenecks.

This chapter presents a comprehensive discussion of biorealistic learning mechanisms and their physical realisation on memristive networks. By grounding the discussion in the neuroscientific principles that underlie learning and cognition, the chapter aims to elucidate how these biological processes can be abstracted and implemented in hardware.

The chapter begins with an overview of the biological basis of computation, providing an essential neuroscience primer. It then moves to device-level considerations, discussing the properties of memristive devices and their integration into neuromorphic architectures. Throughout, the emphasis is on aligning computational models with biological fidelity, while navigating the constraints and opportunities offered by emerging nanotechnologies.

2.2 Neuroscience Primers

Computational neuroscience employs a computational methodology to elucidate the mechanisms underlying brain function. This entails not only identifying the computations performed by the brain but also understanding the interactions between brain elements, such as neurons and synapses, that facilitate these computations.

The brain is capable of performing a vast array of computations, with the fundamental units of the brain generally considered to be neurons and synapses. In the context of the nervous system, a synapse is defined as a structure that is capable of facilitating the transfer of an electrical or chemical signal from a presynaptic neuron to a postsynaptic neuron.

This section provides a concise overview of the relevant biological details, in addition to the concepts and models from computational neuroscience that are employed or expanded upon in this study. These details provide invaluable preliminary information for accurately modelling the implementation of silicon oxide device-based neuromorphic hardware and bio-inspired computing.

In addition to the fundamental biological details that are pertinent to the subject under discussion, the section provides information both at the neural level and at the network level. It is important to acknowledge that the models outlined in this study are comparatively rudimentary when placed in contrast to the substantial corpus of evidence that has been amassed on the neural system. This extensive body of evidence [119], constitutes the preponderance of neural data concerning these domains. Consequently, this section presents only the most fundamental biological facts relevant to the present work.

2.2.1 Neuron Anatomy and Electrophysiology

A neuron is a specialised biological cell that processes and transmits information through electrical and chemical signals [187]. They represent only one of the numerous cell types within the brain, yet they are the most frequently discussed due to their status as the primary computational entities. Their fundamental function is relatively straightforward: neurons receive input from other neurons, and if that input is sufficiently stimulating, they will fire an action potential (also known as a spike), which propagates to other neurons.

Figure 2.1 illustrates the basic structure of a neuron. Neurons can be subdivided into three principal parts: the dendrites, the cell body (soma), and the axon. Neurons receive input currents via their dendrites, which then transmit or channel this into the cell body, called the soma. When a neuron spikes, it sends current down its axon, which results in the release of neurotransmitter(s) at the synapses. These are connections from a neuron's axon to the dendrites of other neurons, and the neurotransmitter release causes dendritic input currents in these other connected neurons.

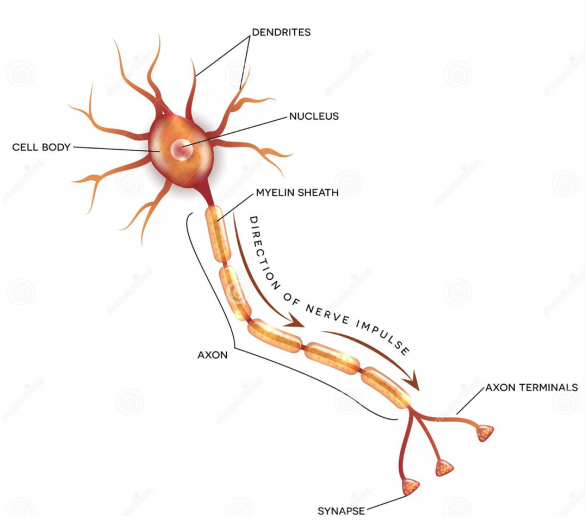

Fig. 2.1 Labeled diagram of the neuron, nerve cell that is the main part of the nervous system. A neuron's dendrites include synapses that allow it to accept input from other neurons. The dendrites carry current to the soma, which is where electrical charge is integrated. If the neuron membrane gets sufficiently polarised, an action potential (also known as a spike) travels down the axon. This causes neurotransmitters to be released at synapses, resulting in currents in the dendrites of postsynaptic neurons.

From a computational perspective, the soma represents the integration point for all incoming currents from dendrites [208], marking the initiation of the action potential generation

process (Figure 2.2). When a neuron is at rest, the soma exhibits a negative charge. This is referred to as the resting voltage and is maintained by ion pumps that regulate the concentration of ions (predominantly sodium, Na^{+}, potassium, K^{+}, and calcium, Ca^{+2}) within the cell.

As the currents arrive from the dendrites, they initiate a process of depolarisation of the cell [113]. Once the voltage within the soma reaches a sufficient level, it initiates the opening of voltage-activated sodium channels, which permit the influx of sodium ions into the cell, further depolarising it. This process persists until the electrical gradient resulting from the accumulation of sodium ions reaches a point where it is no longer in equilibrium with the chemical gradient caused by the imbalance of sodium within and outside the cell. This leads to a notable increase in the neuron's positive charge, exceeding the resting voltage.

Furthermore, this substantial depolarisation also activates voltage-gated potassium channels, which subsequently permit the release of potassium ions from the cell, thereby facilitating repolarisation. Concurrently, the sodium channels undergo inactivation. The opening of potassium channels ultimately results in the cell reaching a voltage below its resting level, a state known as hyperpolarisation. The sodium channels remain inactivated and the potassium channels remain open for a period of time following the spike.

The combination of these factors renders it almost impossible for the neuron to fire during this time; this is referred to as the absolute refractory period. The change in ionic concentrations within the cell is relatively minor during a single spike, but over the course of numerous spikes, the ion pumps are required to maintain the optimal concentrations of sodium and potassium. Other currents, most notably calcium currents, are present in some neurons.

The rapid depolarisation associated with an action potential not only causes an increase in the somatic voltage potential, but also results in partial depolarisation of the axon segments situated in closer proximity to the soma. This results in the opening of sodium channels in that part of the axon, which in turn causes further depolarisation and the opening of sodium channels in the subsequent section of the axon. In this way, the somatic spike triggers a voltage wave that travels down the axon, eventually leading to the release of neurotransmitter(s) from synaptic vesicles situated near the ends of the axon.

It has been established that all synapses located along a neuron's axon are responsible for the release of a singular neurotransmitter or a combination of neurotransmitters. This phenomenon is commonly referred to as Dale's principle. At the time of its development,

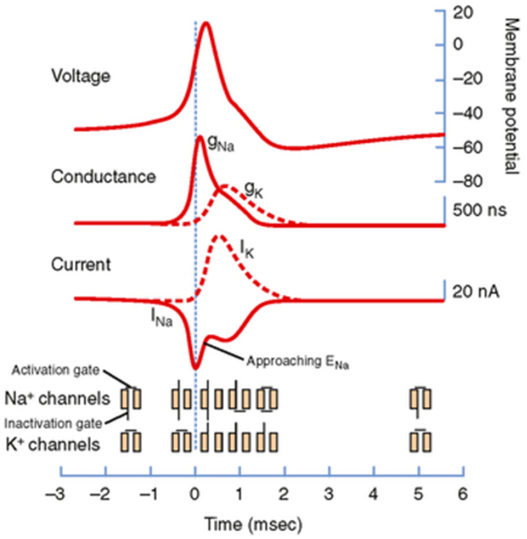

Fig. 2.2 The action potential and the underlying conductance and currents with respect to time [229]. It should be noted that the increased conductance for Na^+ (and its inward flow) is associated with the rising phase of the action potential, whereas the slower increase in conductance for K^+ (and its outward flow) is associated with repolarisation of the membrane and with afterhyperpolarisation. The reduction in I_{Na} before the peak of the action potential (even though G_{Na} is still high) is due to inactivation of the Na+ channels.

Dale's principle was based on the assumption that each neuron produced a single type of neurotransmitter. Nevertheless, evidence of cotransmission was only discovered subsequently [20]. It was understood that neurotransmitters can only be either excitatory or inhibitory,in relation to different postsynaptic cells.

2.2.2 Spiking Neuron Dynamics

In recent times, the number of available neuron models has proliferated. The models currently in use in the literature range from the simplest possible rate-neuron model, namely binary threshold units [232], to complex multi-compartmental models that account for detailed dendritic morphologies [177]. In the context of large-scale neural models aiming to reproduce high-level behaviours, single-compartment neuron models remain the prevailing approach.

These models treat the neuron as a single electrical compartment, combining the dendrites, soma, and axon. In contrast, multi-compartmental models represent the neuron as comprising multiple electrical compartments, with equations that describe the influence of activity in one compartment on that of another. By modelling the spike separately from the rest of the neural dynamics, it is possible to separate time scales, thereby avoiding the need for additional

computational resources to model the spike trajectory [1].

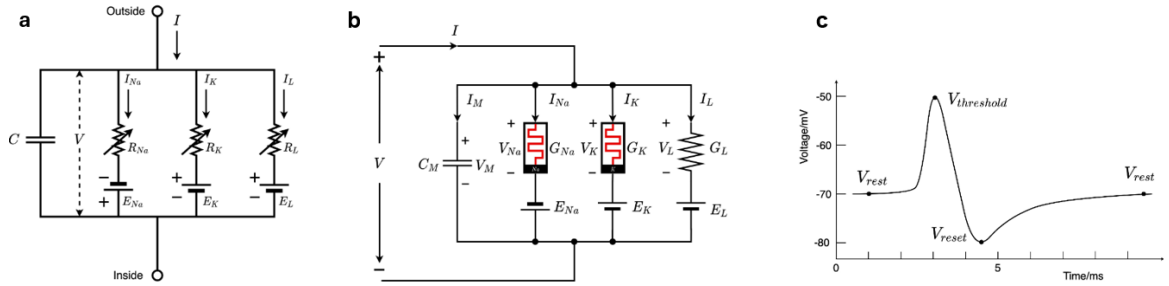

Fig. 2.3 Hodgkin-Huxley neuron model. (a) An equivalent circuit for the HH models [94]. (b) An equivalent circuit for memristive HH model [41]. (c) An action potential waveform, which demonstrates the resting, threshold, and reset potentials.

A significant proportion of the most influential findings in computational neuroscience are based on mathematically detailed models of neuronal functioning. One of the most renowned of these is the Hodgkin-Huxley model of the squid giant axon [94]. the Hodgkin–Huxley (HH) model is one of the most widely used, comprising a set of nonlinear differential equations that accurately approximate the electrical signals of neurons [41].

Figure 2.3(a) depicts the HH neural model, wherein the time-varying nonlinear conductor $R_{Na}(GNa)$ and $R_K(G_K)$ represent the sodium and potassium channels, respectively, while the linear conductor $R_L(G_L)$ simulates leak channels and C models the membrane of a neuron. The equations of the HH model are presented below:

$$C\frac{dV_m(t)}{dt} = I_C(t) + \sum_k I_k(t) \tag{2.1}$$

In this context, V_m represents the membrane potential. $\sum_k I_k(t)$ denotes the sum of the ionic currents flowing into the neuron. This can be formulated by three ion currents, as follows:

$$\sum_k I_k = C_m\frac{dV_m}{dt} + G_K n^4(V_m - V_K) + G_{Na}m^3(V_m - V_{Na}) + G_L(V_m - V_L) \tag{2.2}$$

$$\frac{dn}{dt} = \alpha_n(V_m)(1-n) - \beta_n(V_m)n \tag{2.3}$$

$$\frac{dm}{dt} = \alpha_m(V_m)(1-m) - \beta_m(V_m)m \tag{2.4}$$

$$\frac{dh}{dt} = \alpha_h(V_m)(1-h) - \beta_h(V_m)h \tag{2.5}$$

The reversal potentials V_K, V_{Na}, and V_L are the three parameters in question. The rate constants α_i and β_i, which depend on the membrane potential, describe the behaviour of the i^{th} ion channel. The maximal value of the conductance is represented by G_K, G_{Na}, and G_L.

Finally, the dimensionless quantities n, m, and h, which lie between 0 and 1, are associated with three ion channels. In order to achieve the optimal fit for human action potentials, the HH model is reduced by setting the leakage channel conductance to $G_L = 0$ [200].It has been demonstrated that G_{Na} and G_K are memristors [42], in the equivalent circuit in figure 2.3(b).

The integrate-and-fire (IF) neuron [141] constituted one of the earliest computational models of a neuron. This model was developed prior to the ability of researchers to measure the electrical and chemical changes occurring in a functioning neuron. It is based on the premise that the neuron membrane can be modelled as a capacitor that stores charge over time [1].

As the name suggests, the IF model exhibits two principal behaviours: The model integrates current over time, as would be expected of a capacitor, and fires when the voltage reaches a threshold. Furthermore, the model may or may not incorporate a leak term, which represents a resistor in parallel with the capacitor that permits the dissipation of charge over time. The model with a leak term is typically designated as the leaky integrate-and-fire (LIF) model. While the term "integrate-and-fire (IF) model" can be used interchangably.

To identify how the neuron's membrane voltage evolves over time and, based on this, to determine when the neuron spikes, the charge Q across a capacitor is represented by $Q = V \times C$, where V is the voltage across the capacitor and C is the capacitance. By differentiating this with respect to time, the membrane voltage *V(t)* of the neuron is:

$$C\frac{dV(t)}{dt} = J(t) \tag{2.6}$$

In this context, *J(t)* represents the input current to the neuron over time, whereas C denotes the membrane capacitance. The current here is the time derivative of charge. Equation 2.6 demonstrates that the IF neuron simply integrates the input current over time. It is still necessary to identify the point at which the neuron spikes.

This is achieved by defining a threshold voltage, V_{th}, which is exceeded when the voltage passes this threshold, resulting in the neuron firing. This is a fundamental principle in neurophysiology: once the neuron voltage passes a threshold, the neuron begins firing a

spike, and once this firing process begins, it is almost impossible to reverse.

Once a neuron has fired a spike, the membrane voltage is reset to the resting potential, V_{rest}. This phenomenon can be attributed to physiological resetting procedures. Following the occurrence of a spike in a neuron, other ionic currents, typically potassium, are initiated, leading to a restoration of the membrane voltage towards the resting potential.

The leaky integrate-and-fire (LIF) model [132] incorporates an additional physiological factor: Neuron membranes are not perfect capacitors; rather, they slowly leak current over time, pulling the membrane voltage back to its resting potential. Therefore, the membrane is modelled as a capacitor and resistor in parallel, which allows for the neuron to exhibit a degree of "forgetting": in the absence of any input, the membrane voltage will return to its resting potential [133]. The LIF dynamics are captured by the following equation:

$$C\frac{dV(t)}{dt} = J(t) - \frac{1}{R}(V - V_{rest}) \tag{2.7}$$

In this model, R represents the membrane resistance, and the remaining parameters are consistent with those of the IF model, with identical resetting procedure.

The LIF model comprises a number of parameters, including C, R, V_{rest} and V_{th}. It is possible to normalise the model in order to reduce the number of parameters while maintaining the full dynamics of the original model. In particular, the model can be manipulated so that the normalised voltage lies within the range [0, 1], with a normalised resting potential of zero and a normalised firing threshold of one. Initially, Equation 2.7 is multiplied by R to give:

$$\tau_{rc}\frac{dV}{dt} = RJ(t) - V + V_{rest} \tag{2.8}$$

$$\tau_{rc} = R \times C \tag{2.9}$$

$$\bar{V} = \frac{V - V_{rest}}{V_{th} - V_{rest}} \tag{2.10}$$

$$\bar{V_{rest}} = \frac{V_{rest}}{V_{th}} \tag{2.11}$$

By substituting $\bar{V}$ and $\bar{V_{rest}}$ into equation 2.3 to give:

$$\tau_{RC}(V_{th} - V_{rest})\frac{d\bar{V}}{dt} = RJ(t) - \bar{V}(V_{th} - V_{rest}) \tag{2.12}$$

$$\tau_{rc}\frac{d\bar{V}}{dt} = \frac{R}{V_{th} - V_{rest}}J(t) - \bar{V} \tag{2.13}$$

$$\tau_{rc}\frac{d\bar{V}}{dt} = \bar{J}(t) - \bar{V} \tag{2.14}$$

When the firing threshold for the new equation $\bar{V_{th}} = 1$, the voltage resets to $\bar{V_{rest}} = 0$, and $\bar{J}(t) = \frac{R}{V_{th} - V_{rest}}J(t)$. It can be observed that $\bar{J}(t)$ is merely a linear transformation of $J(t)$. Consequently, (2.14) retains the full dynamics of (2.7) for a scaled input, but with only one parameter, τ_{RC}.

It should be noted that both $\bar{V}$ and $\bar{J}$ are unitless quantities. Conventionally, the unitless space is employed exclusively, and the quantities are often referred to simply as V and J, despite the fact that they are not voltages or currents. This simplifies the mathematical representation, without limiting the generality of the models.

(2.14) provides an exact description of the circumstances under which the model neuron will spike in response to a given input current, *J(t)*. However, in some cases, it is sufficient to consider only the spike rate, that is, the number of spikes per second that the neuron will produce in response to a given input current.

In the case of the LIF model, it is possible to determine the analytical firing rate for a constant input current. This is achieved by calculating the inter-spike interval (ISI), which is the time between one spike and the next. The firing rate is then given by the inverse of the ISI. When a constant input current, $J(t) = j$, is provided, it is possible to solve (2.14) in order to find the neuron voltage over time.

$$V(t) = (V(0) - j)e^{\frac{-t}{\tau_{rc}}} + j \tag{2.15}$$

In the absence of spikes, the objective is to ascertain the time required for the voltage to increase from$V(0) = 0$ to $V(t) = 1$. This property will only occur if $j > 1$. Substitution into (2.15) and subsequent solution for t yields:

$$t = -\tau_{RC} log\left(-\frac{1}{j}\right) \tag{2.16}$$

Incorporating the refractory period and performing the inversion, the spike rate r for the LIF neuron is given by:

$$r = \begin{cases} \frac{1}{t_{ref} - \tau_{RC} log\left(1 - \frac{1}{j}\right)} & \text{if } j > 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.17}$$

The LIF model is one of the most widely utilised simplified neuron models [142]. The simple equivalent model is illustrated in Figure 2.4(a). In this model [230], a resistor R, connected in series with a DC source V_{rest}/V_{reset}, is connected in parallel with a capacitor C. A postsynaptic neuron receives a synaptic current $I(t)$, generated by presynaptic spikes.

A proportion of the current $I(t)$ flowing into C results in an increase in the membrane potential $V(t)$. The charge leakage occurs via resistor R. When $V(t)$ reaches a threshold value, the neuron generates a spike. Following the generation of a spike, the membrane potential is reset to the reset value.

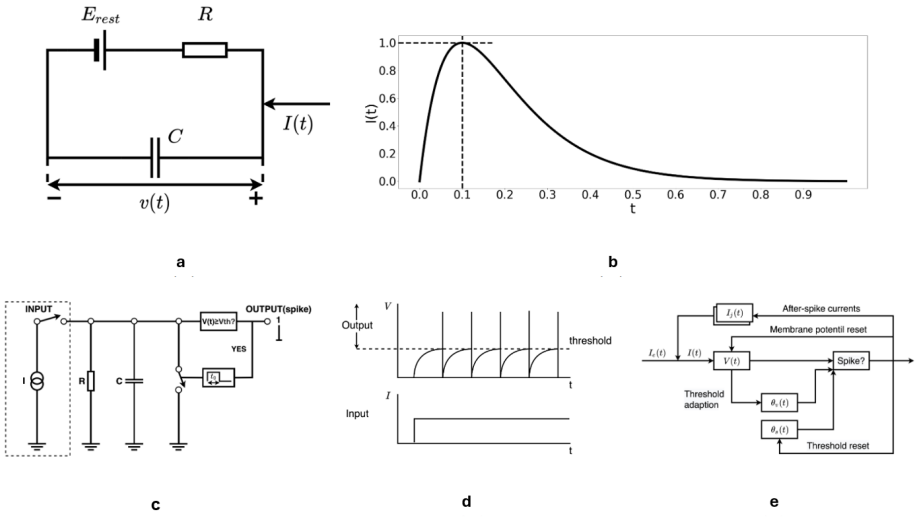

Fig. 2.4 The LIF neuron model. (a) Schematic diagram of the LIF electrical model. (b) Input current in the form of an alpha function, $\tau_\alpha = 0.1, I_0 = 1$ (c) The LIF model allows for the control of spiking behaviour through a comparison of membrane potential and threshold at each time step. Upon the triggering of a spike, a voltage-controlled switch discharges C for a duration corresponding to the refractory period t_0 [241]. (d) A simulation of constant firing frequency for DC current input in which t_0 is hidden. DC input current and output spikes are both shown. (e) A generalised LIF model with threshold control [244].

In the absence of *I(t)*, the voltage across C is eventually settled at V_{rest}, representing the cell's resting potential. During the refractory period t_0, a neuron is incapable of spiking. Figure 2.4(c,d) illustrates the LIF neuron dynamics for the case of a *DC* input current and a zero rest and reset potential, $E_{reset} = E_{rest} = 0$ [241].

The input synaptic current, *I(t)*, can be described by a time-varying alpha function. However, alternative functions may be employed, including "Instantaneous Rise and Single Exponential Decay," "Biexponential Functions," "Sawtooth," and "Pulse Function." The alpha synaptic current is modeled by Equation 2.15, and the resulting plot is shown in Figure 2.2(b).

Nevertheless, Figure 2.4(a) lacks a circuit for resetting the system when the threshold is reached. In order to evaluate the inequality $V > V_{threshold}$, it is necessary to use an active circuit, such as a comparator. Upon reaching the threshold, the membrane potential must be reset in accordance with the illustration in Figure 2.4(d). Therefore, the LIF model's generalized version necessitates additional overhead, as illustrated in Figure 2.4(e).

It is evident that neurons manifest considerable heterogeneity with regard to their dynamics, morphology and connectivity. They can summarily be categorised as follows: Excitatory neurons are responsible for promoting activity in connected neurons. In contrast, inhibitory neurons are responsible for suppressing activity, a process that is crucial for stability and rhythm. Finally, interneurons connect local circuits, thereby enabling complex computations.

2.2.3 Synaptic Transmission and Plasticity

Prior to model internal neural dynamics, it is essential to model the dynamics of the synapses that connect neurons to one another. Synapses exert a significant functional effect as a low-pass filter on the spikes that pass through them. A spike in the presynaptic neuron elicits an extended current pulse in the postsynaptic neuron. This pulse can be conceptualised as a low-pass filtered version of the presynaptic spike.

The simplest model of a synapse is that of a first-order low-pass filter. The impulse response of a filter describes the manner in which the filter responds to an infinitesimally short input of unit integral, which is called an impulse. This idealised impulse is also a reasonable model of a spike, and thus the impulse response also describes what the postsynaptic current will look like in response to a presynaptic spike. The impulse response of the first-order low-pass

filter is as follows:

$$h(t) = \frac{1}{\tau_s} e^{\frac{t}{\tau_s}} \tag{2.18}$$

The synaptic time constant, denoted by τ_s, is defined as the length of time over which the postsynaptic current is spread. Given that the impulse response is an exponential function, the exponential synapse model is therefore a suitable description.

$$h(t) = \frac{1}{\tau_s^2} e^{\frac{t}{\tau_s}} \tag{2.19}$$

It was determined that a second-order lowpass filter is a superior model for a synapse [165]. The impulse response of this filter is defined by (2.19). This function is referred to as the alpha function, and thus the model is designated as the alpha synapse model. Both of these models are based on the current generated by a spike in the postsynaptic neuron, which is a current-based synapse model.

One of the primary objectives of computational neuroscience is to ascertain the manner in which the brain represents—or encodes—information. To this end, researchers have put forth a multitude of potential coding schemes that neurons could utilise for information encoding. One key distinction between rate coding and temporal coding is the following dichotomy.

In a rate code, the sole pertinent measure is the firing rate (i.e. the number of spikes) of a neuron over a given period of time. An exemple of a rate code is motor neurons in the peripheral nervous system. The contraction of a muscle is contingent upon the number of spikes per unit time; thus, only the rate of motor neuron spikes is significant [75]. In a temporal code, the time of individual spikes is also a factor. For example, in the early auditory system, precise spike timing facilitates the localisation of sounds [31].

The precise definitions of rate and temporal codes remain contentious, with differing interpretations presented by various authors [46]. To illustrate, a neuron may discharge a number of spikes in rapid succession, followed by a period of quiescence. A second neuron may be observed to fire the same number of spikes, but in a more evenly distributed manner over a given period.

Both rate codes and temporal codes describe the encoding properties of individual neurons. Additionally, one may inquire about the coding properties of a group (also known as a population) of neurons. The concept of population coding pertains to instances where a

representation is distributed across numerous neurons within a population, such that the represented value cannot be decoded from the activities of a limited number of neurons.

The simplest method of extrapolating the concept of rate or temporal coding to multiple neurons would be to have numerous neurons all implementing the same code. In other words, all neurons will exhibit a similar firing pattern when representing a given value, due to their comparable tuning properties. This results in a significant degree of redundancy between neurons.

In contrast, population coding entails each neuron representing a distinct aspect of the represented value. To illustrate, if the objective is to represent head direction, there are neurons that represent a head that is fully turned to the left, others that represent a head that is fully turned to the right, and still others that represent a centred head. Additionally, there are neurons that represent values in between these three head directions. The direction in which a neuron is most active is referred to as its preferred direction.

Synapses are therefore known to play a dual role in the nervous system. They facilitate communication between neurons and serve as the primary locus of learning and memory. The magnitude of this influence, or 'synaptic strength' [14], is determined by the relative strength of the connection between the presynaptic and postsynaptic neurons. This phenomenon is known as synaptic plasticity.

It is important to note that plasticity can be categorised into several distinct types. Short-Term Plasticity (STP): Transient changes that last from milliseconds to seconds. Long-Term Potentiation (LTP) and Long-Term Depression (LTD): The phenomenon of sustained increases or decreases in synaptic strength over time.

The most significant model of synaptic plasticity is Hebbian learning, which can be succinctly summarised as follows: The hypothesis that neurons that fire together wire together has been proven to be accurate. A more precise, temporally-sensitive rule is Spike-Timing Dependent Plasticity (STDP) [302].

Mathematical Models can capture the effect of precise spike timing on synaptic weight updates. If a presynaptic neuron fires before a postsynaptic neuron within a short window, the synapse is strengthened; if the order is reversed, the synapse weakens. A common

representation for this is:

$$\Delta w = \begin{cases} A_+ \cdot e^{-\Delta t/\tau_+}, & \Delta t > 0 \\ -A_- \cdot e^{-\Delta t/\tau_-}, & \Delta t < 0 \end{cases} \tag{2.20}$$

Where $\Delta t = t_{post} - t_{pre}$ is the timing difference, A_+, A_- are learning rates, τ_+, τ_- are time constants for potentiation and depression. This asymmetric window is indicative of experimental observations and provides a biologically plausible basis for synaptic learning in hardware.

As a small primer, memristive devices offer an electronic analogue to synapses due to their tunable conductance and memory of past activity. When configured in crossbar arrays, these devices have the capacity to implement synaptic weight matrices directly in hardware, with updates governed by local voltage or current pulses.

Memristive STDP implementations frequently exploit device physics, where conductance change is contingent on pulse overlap:

$$\Delta G = f(V_{pre}, V_{post}, \Delta t) \tag{2.21}$$

where f is a device-specific function determined by material properties and pulse shapes.

It is important to note that memristors have the capacity to inherently facilitate the nonlinear, history-dependent behaviour that is characteristic of biological plasticity rules, such as STDP. To illustrate this point, the application of carefully timed voltage pulses to a memristor has been demonstrated to result in an increase or decrease in conductance, respectively, reminiscent of LTP and LTD.

2.3 Foundations of Neuromorphic Computing

Extensive research has been conducted in device physics and material science to explore innovative materials and techniques for memories and prolonged retention objectives [106]. The term "neuromorphic" was created by researchers to describe new technologies and systems that, in addition to being essential for the construction of massive AI computer networks, exhibit certain behaviours that can be compared to those of real synapses [50].

Soon after, the notion of using these novel nanoscale components as "memristors" gained popularity, with the underlying notion being that they could be utilised to produce synapses in deep neural networks and sustain their synaptic weights locally [111]. The hardware and technology described could enable neural networks to perform "in-memory computing" and exhibit advanced non-linear properties, mimicking the physics of biological synapses [215].

The research in this field aims to develop various types of volatile and non-volatile memristive electronics. Additionally, spike or pulse-based control systems are being created to elicit biologically realistic learning behaviours in memristive cross-bar arrays. The challenging task is to find the perfect artificial synapse, which requires investigation into different materials, tools, and techniques.

2.3.1 Memristor Fundamentals

Memristive devices, also known as memory resistors, are emerging as foundational elements in neuromorphic computing due to their ability to retain resistive states based on electrical history, thereby emulating biological synapses. The central purpose of these components is to facilitate both memory storage and computation within a unified, compact structure, thereby enabling the co-location of memory and processing elements that is vital for brain-inspired architectures.

The presence of symmetry in nature, which is believed to arise from a common origin, is remarkable. However, the traditional electromagnetic passive circuit components of resistor, capacitor, and inductor are inadequate for describing the characteristics connected by the symmetry of circuit theory. Leon Chua addressed this issue by introducing the concept of a memristor in 1971 [39], which couples flux linkage and charge as a circuit device:

$$M(q) = \frac{d\phi}{dq} \tag{2.22}$$

where M denotes the memristance, a quantity whose value is known to be dependent on the history of the current that has previously passed through the device. This phenomenon gives rise to a form of resistive memory, wherein the device retains a memory of its previous state. However, proof of resistive switching in the memristor model was not established until 2008 [236].

There are three fundamental circuit elements and four essential circuit variables in basic electrical circuit theory. It is evident that one component is absent to achieve symmetry. This device ought to function in such a way that charge and magnetic flux are interconnected,

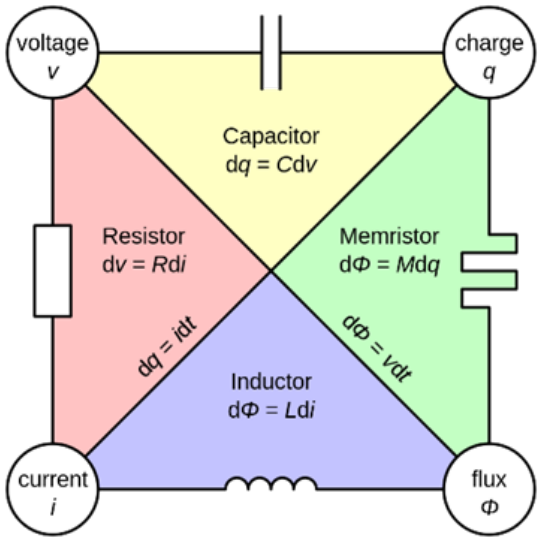

Fig. 2.5 Conceptual symmetries of resistor, capacitor, inductor, and memristor [53]. These four fundamental variables in circuit theory are depicted with their relationships. Each variable can be related to another via either a passive component or a well-known equation.

as illustrated in Figure 2.5. The link between the mathematical memristive model and a two-terminal resistive switching device is pivotal in this instance.

The concept of memristance is distinct from that of resistance in that it is dependent on charge, rather than being a constant value. The current is defined as the amount of charge flowing per unit time. Therefore, the expression for current can be written as:

$$q = \int_{-\infty}^{t_0} i(t)dt \tag{2.23}$$

where charge is the sum of current at a given time t_0.

This indicates that the memristance, being dependent on charge, is determined by the historical currents that have previously passed through the device. In the event of interruption to the current flowing through the device, the memory state persists until the current flow is restored. The device is evidently equipped with a type of memory known as a "memristor".

In physical terms, memristors are often modeled as two-terminal devices whose resistance varies due to the drift of ions or vacancies in a dielectric medium. The state-dependent

resistance can be written as:

$$V(t) = R(w(t)) \cdot I(t) \tag{2.24}$$

where $V(t)$ and $I(t)$ are the voltage and current at time t, $R(w(t))$ is the resistance depending on the internal state $w(t)$. This state $w(t)$ often represents physical quantities like oxygen vacancy concentration or filament length in resistive switching materials. $f(w,I) = \frac{dw}{dt}$ can be further defined as to how the internal state changes with input current.

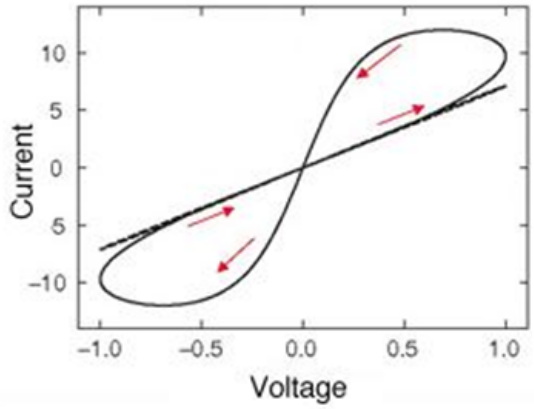

Fig. 2.6 The typical I-V characteristic of a memristor displays a pinched hysteresis loop resulting from the nonlinear relationship between current and voltage in memristance [274].

From a visual standpoint, the pinched hysteresis loop, which is characteristic of the devices and dependent on frequency, distinguishes these memristor devices from other components [40], as shown in Figure 2.6. This loop represents a prevalent and inherent phenomenon in the natural world. It is evident that as the voltage input frequency increases, the loop undergoes a reduction in size. When the frequency approaches infinity, the memristor can be approximated as a resistor.

Among the range of new non-volatile memory devices, the primary focus of this study is on memristor devices, including MRAM, PRAM, FeRAM, and RRAM [269]. Resistive switching, a reversible phenomenon of two-terminal elements, characterises the devices. Through electrical signalling, they change resistance in a non-volatile manner, with the process driving the resistive switching defined by the device's materials [186].

Resistive random-access memory (RRAM) is a device that uses resistance switching, where reversibility is attained through repeated application of appropriate stimuli, according to

[161]. Repeated application of suitable stimuli ensures reversibility. An RRAM cell comprises an insulating thin film (usually a metal oxide), sandwiched between two electrodes, within which resistance switching occurs.

The term "memristance" is favoured to express the general characteristics of these RRAM devices. The central hypothesis of this model is that memristance is a function of the total charge that has been passed through the device or that the integral of the applied voltage is consistent with certain experimental data. This can be used to toggle between different resistance levels. Although this ideal memristor model is often used in RRAM cells, it may not satisfy practical requirements.

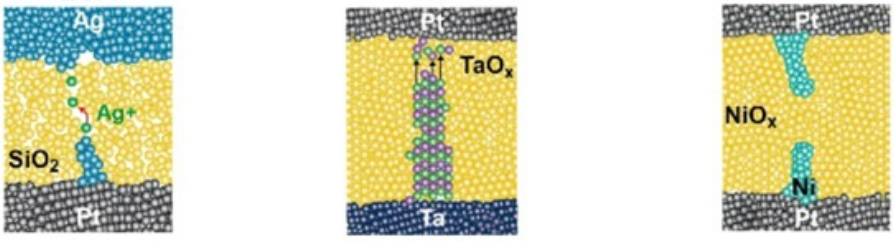

Fig. 2.7 The main RRAM types, from left to right: electrochemical metallization memory (ECM), vacancy change memory (VCM), and thermochemical memory (TCM). [79]

Although all RRAM devices operate on a metal-insulator-metal (MIM) architecture, their categorisation and analysis remain challenging. RRAM devices are loosely classified into two types based on their functional mechanisms: oxide-RAM (OxRAM) and conductive bridge RAM (CBRAM) [159].

However, the internal physical behaviour of RRAM devices greatly varies, making it difficult to obtain a unified picture of them. RRAM cells may be classified according to their switching mechanisms which are electrochemical metallisation (ECM), valence change mechanism (VCM), or thermochemical mechanism (TCM) [79]. The operation of the device is explained by one of the mechanisms depicted in Figure 2.7.

In normal operation, the state of the memristor can be objectively designated as having high resistance in the "OFF" state and low resistance in the "ON" state, with a substantial difference in resistance levels. The shift from the high resistance state (HRS) to the low resistance state (LRS) is referred to as "Set," while the reverse is called "Reset." An electroforming step is generally necessary to convert the device from pristine to switchable, whereby the former

tends to exhibit higher resistance.

A hallmark of biorealistic learning is the ability to adjust synaptic strength continuously over time in response to neural activity. Analog (or gradual) switching in memristors is thus essential. Instead of binary ON/OFF transitions, these devices exhibit incremental conductance changes under controlled voltage pulse conditions. Consider a simplified state evolution model:

$$\frac{dw}{dt} = \mu \cdot I(t) \tag{2.25}$$

Where μ is a mobility parameter dependent on the device materials and structure. For voltage pulses of controlled amplitude and duration, this allows precise modulation of conductance:

$$\Delta G \propto \int I(t)dt = Q \tag{2.26}$$

Here, Q is the total charge transferred, which accumulates over spike events. By shaping input pulses (in terms of rise time, width, or height), one can encode temporally-dependent plasticity rules such as STDP directly in hardware.

Despite their initial promise, memristive devices have been found to exhibit significant non-idealities [80]. Device-to-device variability, encompassing factors such as conductance range, switching thresholds, and cycle-to-cycle behaviour, has the potential to vary across devices that appear to be nominally identical.

Non-linear phenomena, such as conductance updates, frequently exhibit saturation or asymmetric responses to positive and negative pulses. It is important to note that drift and retention loss, such as those occurring over time, may be attributable to the effects of relaxation.

To address these issues, neuromorphic systems often employ redundancy, error-tolerant learning algorithms, or closed-loop calibration techniques. Furthermore, some variability may be biologically realistic: synapses in the brain are not perfectly precise either, and stochasticity can enhance learning generalization and robustness.

For design and testing purposes, compact models of memristive devices are essential. These range from physics-based models to empirical abstractions. A popular framework is the

linear ion drift model [23], applicable to early TiO_2-based devices:

$$w(t) = w_0 + \frac{\mu_v R_{ON}}{D} \cdot \int_0^t I(\tau)\, d\tau \tag{2.27}$$

Where μ_v is the ion mobility, R_{ON} is the low resistance state, D is the device thickness. For practical simulations, window functions are often added to prevent unrealistic values of $w(t)$ outside the physical boundaries. A widely used modified form is:

$$\frac{dw}{dt} = \mu \cdot I(t) \cdot f(w) \tag{2.28}$$

Where $f(w)$ is a window function such as:

$$f(w) = 1 - (2w - 1)^{2p} \tag{2.29}$$

with p controlling the non-linearity near the boundaries. These models allow researchers to prototype neuromorphic algorithms and circuits in software before hardware realization, enabling design-space exploration and validation under realistic conditions.

2.3.2 In-memory Computing Paradigms

Neuromorphic computing signifies a paradigm shift of the manner in which information is processed and stored; this is inspired directly by the architecture and function of biological neural systems. Conventional computing systems compartmentalise memory and processing units, a configuration that engenders energy and velocity inefficiencies due to incessant data movement. Neuromorphic systems are designed to co-locate memory and computation by leveraging distributed, parallel architectures that emulate the brain's functionality.

The origin of neuromorphic engineering can be traced to the pioneering work of Carver Mead in the 1980s [180], who proposed using analog electronics to mimic the function of neurons and synapses. Since then, the field has grown to encompass both analog and digital implementations of brain-inspired circuits.

The fundamental principle of neuromorphic computing is the translation of key neurobiological principles into hardware. Event-driven processing is a key feature of neuromorphic circuits, which, like biological neurons, only activate when necessary, thereby significantly reducing power consumption.

As an illustrative example, synapses (which are implemented by resistive memory elements) function as both computational units and memory stores. Neuromorphic systems have been demonstrated to exhibit plasticity and the capacity for real-time, local learning through the utilisation of biologically plausible rules, such as the STDP (synaptic tagging with depolarisation-dependent plasticity) rule.

Neuromorphic architectures can be broadly classified into two categories [26]. Digital neuromorphic systems are comprised of digital circuits which simulate the behaviour of neurons, i.e. their spiking. Notable examples of this include IBM's TrueNorth and Intel's Loihi. These chips implement large networks of spiking neurons with programmable connectivity and plasticity. Analog/Mixed-Signal Systems have been shown to exhibit a greater degree of similarity to the continuous dynamics of biological neurons and synapses. Memristive arrays frequently fall into this category, offering a physical substrate for analogue computation.

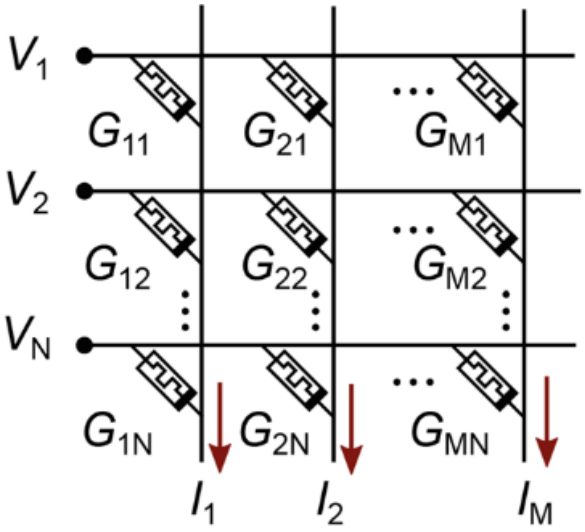

Fig. 2.8 The memristor-based crossbar architecture with a single memristor array and a constant-term circuit [248]. The resistive components are located at the connections of the word and bit lines. When voltages **V** are applied to the word line, the resistive element at the junction of the i^{th} word line and the j^{th} bit line generates $V_i \times G_{i,j}$ units of current, assuming zero wire resistance, as per Ohm's law. The currents created by each individual element are then aggregated along the bit lines using Kirchhoff's current law.

A key architectural unit is the crossbar array, in which vertical and horizontal metal lines intersect at memristive devices. This structure supports efficient matrix-vector multiplication—the fundamental operation in neural networks. Let V be a voltage vector applied to input rows, and G be the conductance matrix of memristive elements. The output current

vector I on the columns is given by:

$$\mathbf{I} = \mathbf{G} \cdot \mathbf{V} \tag{2.30}$$

This analog computation occurs in a single step without requiring data movement between separate processing and memory units, thus offering significant energy efficiency. Linear algebra and vector-matrix products rely heavily on multiplication and addition. These procedures can be carried out by using fundamental circuit laws.

Consider a resistive element with conductance G (the reciprocal of resistance). If a voltage V is supplied to it, the current I flowing through it will be equal to $V \times G$. This indicates that conductance G functions as a multiplicative factor, as per Ohm's law. For a circuit with several branches, each carrying a current I_i. At the intersection of these branches, the total current flowing through it will be $I = \sum_i I_i$. This indicates that currents are combined together, as per Kirchhoff's current law.

Once multiplication and addition are possible, higher-level operations may be performed using specialist circuits. For vector-matrix products, a resistive crossbar array can be used, which is a two-dimensional grid of conductive wires with resistive components at each intersection. A crossbar array's output currents are essentially the product of a voltage vector and a conductance matrix. Consider a vector-matrix product, $\mathbf{y} = \mathbf{x}^\intercal \mathbf{W}$. Where $\mathbf{x}$ can be translated to voltages $\mathbf{V} = k_V \mathbf{x}$, $\mathbf{W}$ to conductances $\mathbf{G} = k_G \mathbf{W}$, and generate outputs $\mathbf{y}$ from currents $\mathbf{I} = \mathbf{y} k_V k_G$, where k_V and k_G are positive constants.

Crossbars can compute products of voltage vectors and conductance matrices due to their structural design. This is because the structure controls which voltage-conductance pairs are multiplied and which consequent currents are combined together. These circuits have two sets of wires: word lines and bit lines. Voltages $\mathbf{V}$ are applied to the word lines, and currents $\mathbf{I}$ are measured along the bit lines. A resistive element located at the junction of the i^{th} word line and the j^{th} bit line has a conductance of $G_{i,j}$.

When V_i is applied to the i^{th} word line, the device generates a current of $V_i \times G_{i,j}$ (assuming no wire resistance). The currents generated in the jth bit line are added together to provide a current of I_j. This current is calculated by taking the dot product of voltage $\mathbf{V}$ and the j^{th} column of the conductance matrix $\mathbf{G}$. Given that the j^{th} element of a vector-matrix product is just the dot product of the vector and the j^{th} column of the matrix, the vector containing

all output currents may be concisely expressed as $\mathbf{I}^\intercal = \mathbf{V}^\intercal\mathbf{G}$.

The individual application determines which resistive devices are used in the crossbar array. Weights $\mathbf{W}$ are repeatedly updated during neural network training, necessitating the ability to alter the conductances in the crossbar array numerous times. In contrast, during inference, the weights are fixed, allowing the conductances to be set after initial programming. Regardless of the conditions, the conductances will be unique to the network, requiring the ability to change them at least once.

Memristive devices, or memrisitors, are differentiated by their ability to change their conductance in response to electrical inputs. As a result, they make an excellent choice for crossbar-based linear algebra accelerators. The choice of memristor depends on whether the crossbar array is used for training or inference. The former is far more difficult and would demand memristors that can be repeatedly programmed in a linear fashion. Given these complications, much research on memristive crossbars has been on inference.

Even in the absence of nonidealities, any memristor will have a restricted range of conductance values that it can be configured to. This is a hurdle when attempting to represent real numbers with solely positive conductances G. To demonstrate, if the range of attainable conductances is $G \in [G_{off}, G_{on}]$, the crossbar array can only represent matrix values up to $w \in \left[\frac{G_{off}}{k_G}, \frac{G_{on}}{k_G}\right]$. Since G_{off} is a positive number, hence only positive w may be expressed.

One potential option is to employ differential pairs, in which the matrix element w is represented as the difference between two conductances, $G+$ and $G-$ [116]. The two conductances can be chosen symmetrically around the 'average' value $G\pm = G_{avg} \pm \frac{k_G w}{2}$, where $G_{avg} = \frac{G_{off}+G_{on}}{2}$. The two sets of conductances can be represented by independent conductance matrices $\mathbf{G}+$ and $\mathbf{G}-$, which are assigned to different bit lines of the crossbar array [130].

The bit lines will then generate independent sets of currents, which may be represented as vectors $\mathbf{I}+$ and $\mathbf{I}-$. Vector-matrix products are linear, thus the result may be calculated by subtracting $\mathbf{I}-$ from $\mathbf{I}+$. In reality, the 'positive' and 'negative' bit lines are frequently arranged near to one another, which helps to mitigate the detrimental effects of line resistance, a significant non-ideality [115].

The learning process in neuromorphic systems can be categorised into three distinct approaches: supervised learning, unsupervised learning, and reinforcement-based learning [233]. Nevertheless, with respect to biorealistic implementation, unsupervised, local learning is most aligned with the biological model.

Biorealistic learning draws upon the empirical laws of synaptic plasticity observed in biological systems. Central among these are Hebbian Learning [123] which stated "Neurons that fire together wire together." This principle is often expressed in simplified form as:

$$\Delta w_{ij} = \eta \cdot x_i \cdot y_i \tag{2.31}$$

Where Δw_{ij} is the change in synaptic weight between pre-synaptic neuron i and post-synaptic neuron j, x_i and y_i are the activity levels of the respective neurons, η is the learning rate.

Hebbian learning is predicated on the premise that the strength of a synapse is enhanced by co-activity between pre- and post-synaptic neurons [66]. Alternatively, STDP is atemporally-sensitive variant of Hebbian learning, a concept that has already been covered in the preceding section. Homeostatic plasticity is a global mechanism that ensures that overall neural activity remains within functional bounds. This is analogous to metabolic regulation in biology.

These mechanisms are often implemented using local circuit rules. For instance, in memristive implementations of STDP, pulse timing determines the net change in conductance of a memristor. The result is a physical device whose behavior embodies the learning rule itself. In digital systems, learning involves weight updates of the form:

$$w_{ij} \leftarrow w_{ij} + \eta \cdot \delta_j \cdot x_i \tag{2.32}$$

Where w_{ij} is the synaptic weight from neuron i to neuron j, η is the learning rate, δ_j is the error signal at the output neuron j, and x_i is the activation of input neuron i. In contrast, memristive learning avoids explicit error backpropagation and instead uses local learning rules where the change in conductance ΔG depends on spike-timing and voltage:

$$\Delta G \propto f(\Delta t_{ij}) \cdot g(V_{pre}, V_{post}) \tag{2.33}$$

Here, $f(\Delta t_{ij})$ reflects the STDP window and $g(V_{pre}, V_{post})$ models the effect of voltage pulses on device conductance. Memristors thus act as "plastic synapses" whose weights evolve in real time, guided by the temporal correlation of pre- and post-synaptic activity.

2.3.3 Encoding Plasticity in Memristors

As neuromorphic systems aspire to emulate biological intelligence, the implementation of biorealistic learning—learning mechanisms that faithfully reproduce the behavior of biological synapses and neurons—has become central to the development of memristor-based architectures. This section explores how learning rules inspired by neuroscience, such as spike-timing-dependent plasticity (STDP), Hebbian learning, and homeostatic regulation, can be embedded into memristive networks.

To implement STDP and other plasticity rules in hardware, researchers have developed pulse-pairing schemes that encode spike timing as overlapping voltage pulses applied to memristive synapses. These rules can be implemented physically using the conductance modulation behavior of memristors, which act as artificial synapses [24].

For instance, a presynaptic spike instigates a positive voltage pulse, whereas a postsynaptic spike precipitates a negative voltage pulse. The net voltage across the memristor depends on the temporal alignment of these spikes. If the pulses overlap constructively (e.g., pre before post), the net voltage exceeds a potentiation threshold, increasing conductance. If the order is reversed (post before pre), the net voltage may trigger depression.

Let the pulse shape be $V_{pre}(t)$ and $V_{post}(t)$, The effective voltage across the memristor is:

$$V_{mem}(t) = V_{pre}(t) - V_{post}(t) \tag{2.34}$$

Depending on $V_{mem}(t)$, the conductance $G(t)$ changes according to a windowed integration rule, often expressed as:

$$\Delta G = \int_{-\infty}^{-\infty} \gamma(V_{mem}(t))\, dt \tag{2.35}$$

Where $\gamma(\cdot)$ is a non-linear function mapping voltage to conductance change.

At the network scale, memristive synapses form dense connectivity graphs akin to biological networks. When configured with spiking neurons, the resulting system exhibits emergent

learning behavior. A typical learning architecture may include a spiking neural network (SNN) layer of leaky integrate-and-fire (LIF) neurons, memristive crossbar arrays that implement synaptic weights, spike-based learning circuits that detect relative spike timing and apply appropriate pulses.

Mathematically, for a neuron receiving inputs $x_i(t)$ through synapses $G_i(t)$, the membrane potential V_m evolves as:

$$C_m \frac{dV_m}{dt} = -\frac{V_m}{R_m} + \sum_i G_i(t) \cdot x_i(t) \tag{2.36}$$

Where C_m and R_m are are the membrane capacitance and leakage resistance respectively. When V_m exceeds a threshold V_th, the neuron spikes and resets. Synaptic updates follow:

$$\Delta G_i(t) = f(\Delta t_i) \cdot Pulse_{pairing}(x_i(t), y(t)) \tag{2.37}$$

With $f(\Delta t_i)$ as an STDP kernel and $Pulse_{pairing}$ as the hardware-driven pulse overlap function. An illustrative application of biorealistic learning is unsupervised pattern recognition. For instance, when exposed to MNIST digit images encoded as spiking input, memristive networks have been shown to learn digit prototypes using STDP.

In such systems, Each input pixel is connected to neurons through memristive synapses. The input is converted to spike trains based on intensity. Competitive mechanisms such as winner-take-all inhibit multiple neurons from firing simultaneously. STDP strengthens synapses associated with active neurons and temporally correlated inputs. Over time, distinct neurons specialize in responding to specific digit patterns, emulating feature selectivity observed in biological cortical areas.

Unbounded synaptic growth has the potential to destabilise learning. It is evident that biological systems utilise homeostatic mechanisms in order to maintain equilibrium between synapses and network stability. It is therefore imperative to acknowledge that analogous mechanisms are indispensable in memristive networks.

Synaptic normalization is a common strategy that involves the enforcement of a constraint so that the sum of synaptic weights for a neuron remains constant:

$$\sum_i G_i = G_{max} \tag{2.38}$$

Alternatively, weight decay is used to gradually reducing all synaptic weights over time, modeling biological forgetting:

$$G_i(t+\Delta t) = (1-\alpha)\cdot G_i(t) + \Delta G_i \tag{2.39}$$

Where α is a small decay factor. These mechanisms ensure that learning remains stable over long durations, enabling continual learning without catastrophic forgetting.

2.4 Architectures and System-Level Integration

Memristive networks offer distinct advantages. The elimination of the von Neumann bottleneck by in-memory learning is a significant development in this field. The subthreshold operation enables ultra-low-power computation, while the physical time integration closely matches biological computation timescales.

Nevertheless, challenges persist. It is important to note that variability and inconsistent device behaviour have the capacity to disrupt precise learning rules. Furthermore, write endurance on devices may degrade under repeated programming, and non-linear dynamics with real devices often do not match idealised learning models. Notwithstanding these challenges, the co-design of algorithms and devices, whereby learning rules are adapted to the characteristics of the device, facilitates the practical implementation of biorealistic learning paradigms.

While individual memristive synapses provide the foundational building blocks for biorealistic learning, realizing practical neuromorphic systems requires architectural integration at scale. This section explores how memristive networks are organized into hierarchical architectures, interfaced with complementary computing modules, and optimized for system-level performance. The goal is to demonstrate how the principles of biology-inspired learning translate into cohesive hardware systems that support advanced computation.

2.4.1 Hierarchical Modular Architectures

At the heart of memristive architectures lies the crossbar array, a grid of horizontal and vertical metal lines with memristors at each intersection. This structure enables massive parallelism and efficient matrix-vector multiplication (MVM), a cornerstone operation in neural computation.

Nonetheless, practical implementations of crossbars encounter certain issues, including sneak paths and unintended current flows through unselected paths. Additionally, line resistance can diminish accuracy in large arrays, and variability and noise can compromise the reliability of analogue computations. To mitigate these, selector devices, resistive isolation, and adaptive calibration algorithms are used to maintain accuracy and scalability.

The modular organisation of the neocortex has provided a useful model for the design of system-level neuromorphic architectures, which often adopt a hierarchical structure. Each module, also known as a "core", comprises an array of spiking neurons (for example, LIF neurons), a local synaptic crossbar array with plastic memristive elements, peripheral circuitry for spike generation, timing and routing, and optional local learning engines implementing STDP or Hebbian updating.

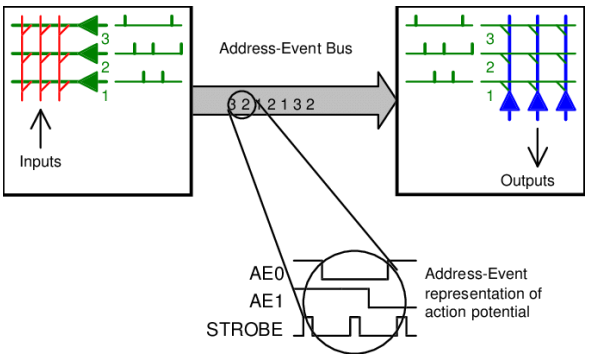

Fig. 2.9 The address-event representation [48]. Self-timed neurons on the sending chip generate trains of action potentials. The neurons request control of the bus when they generate action potentials. They are selected to have their addresses encoded and transmitted by the multiplexing circuitry. The transmission of addresses between the sender and receiver chips occurs in a sequential manner, with each address being transferred successively. The temporal stream is processed by the receiver to produce trains of action potentials, which are subsequently transmitted to their appropriate postsynaptic targets. The relative timing of events is preserved over the address-event bus to the destination, provided that the source neurons do not generate action potentials that are excessively close in temporal proximity.

Multiple such cores are connected via event-driven communication networks that mimic axonal signaling, typically using Address-Event Representation (AER) protocols. This allows the architecture to scale while preserving sparse, asynchronous activity similar to that found in biological systems (Figure 2.9). Each core processes and learns independently, while global integration arises through sparse spike-based communication.

One example is the Neurogrid architecture [128], where analog dendritic trees are combined with digital event processors. Memristive implementations extend this idea by replacing charge-based analog memory with conductance-based synapses. The equation governing this hybrid interface may resemble:

$$y_j = \sigma\left(\sum_i G_{ij} \cdot V_i + b_j\right) \tag{2.40}$$

Where G_{ij} is the conductance of the memristive synapse, V_i is the analog input voltage, b_j is the digital bias term, and σ is the thresholding or spiking function implemented digitally.

True biorealistic learning necessitates online, local learning, in contradistinction to traditional deep learning methods that rely on batch training and global error gradients. Memristive neuromorphic systems characteristically facilitate local learning through the implementation of pulse modulation circuits, which apply voltage updates in accordance with spike timing.

These systems also incorporate time-window detectors, which evaluate spike timing differences, and plasticity controllers, which adjust learning rates and homeostatic thresholds. The aforementioned engines have been demonstrated to exhibit both high levels of parallelism and low power consumption. This facilitates the capacity for real-time adaptation to dynamic inputs. For instance, a single synapse's conductance G may evolve via:

$$G(t+\Delta t) = G(t) + \Delta G = G(t) + f(\Delta t) \cdot V_{update} \tag{2.41}$$

Where $f(\Delta t)$ represents a learning window, and V_{update} is the programmed voltage pulse.

A number of experimental systems illustrate full-stack integration of biorealistic learning in memristive networks. The Intel Loihi 2 [202], equipped with emerging resistive memory, exhibits event-driven learning in a modular spiking architecture. The IMEC Dot Product Engine utilises analogue crossbars for the purpose of real-time pattern recognition [299]. NeuroSim+ is a simulation platform that validates large-scale memristive architectures with STDP and fault-tolerance [35]. These systems generally comprise thousands of spiking neurons and millions of synapses, thereby facilitating the execution of complex tasks such as speech recognition, visual scene analysis, and robotic motor control.

Memristive architectures are regarded as optimal candidates for edge AI applications due to their compactness and low power consumption. These systems are capable of performing

learning and inference directly on-device, thereby eliminating the need for cloud computation. This paradigm shift necessitates the incorporation of architectural features such as local learning with minimal supervision, non-volatile memory for state retention without power, and energy harvesting compatibility for autonomous operation.

From a biological standpoint, this phenomenon can be likened to the manner in which organisms acquire knowledge in an unsupervised, embodied context. This observation serves to reinforce the bio-alignment of memristive neuromorphic architectures.

2.4.2 Hardware-Software Co-Design

While memristive hardware presents novel capabilities for energy-efficient, biologically plausible computation, it is important to note that its full potential is only realised through careful co-design with software. In the context of biorealistic learning, and particularly in the domains of spiking neural networks (SNNs) and continuous-time recurrent systems, there is a necessity to reconsider conventional software stacks. This is so that the full potential of memristive systems, in terms of their dynamics, constraints and strengths, may be realised.

The majority of biorealistic models employ spiking neuron models, such as the leaky integrate-and-fire (LIF), Izhikevich, or Hodgkin-Huxley formulations. These models use time-dependent differential equations to simulate membrane dynamics:

$$\tau_m \frac{dV(t)}{dt} = -(V(t) - V_{rest}) + R_m I(t) \tag{2.42}$$

Where $V(t)$ is the membrane potential, V_{rest} is the resting potential, $I(t)$ is the input current from presynaptic neurons, τ_m is the membrane time constant, and R_m is the membrane resistance.

In order to implement such dynamics on memristive hardware, it is necessary for software frameworks to generate input spike trains, emulate conductance changes over time, encode time delays and refractory periods, and schedule event-driven computation in an efficient manner. Frameworks such as Brian2 [231], NEST [76], and BindsNET [86] offer front-end abstractions for defining neuron and synapse models. From a hardware perspective, translation layers facilitate the mapping of these models to control signals, which in turn manipulate memristive devices through the utilisation of pulse sequences and voltage control.

Local learning rules, such as STDP or Hebbian updates, must be compiled into device-level programming protocols that adjust synaptic conductance values in response to spike timing. A significant challenge pertains to the process of quantization, wherein memristive devices exhibit constrained resolution, typically ranging from 4 to 8 bits. This limitation dictates the mapping of continuous learning gradients onto discrete conductance states during the learning update process.

The physical constraints of memristive arrays—such as array size, non-idealities, and fixed connectivity—call for efficient placement and routing algorithms to distribute large neural models across hardware. The key design constraints here are the fan-in/fan-out limits with physical wiring impose a restriction on the number of connections that can be established between neurons. The synaptic locality dictates that connections are most efficacious when mapped to nearby memory cells. Event congestion, characterised by the influx of spikes, can lead to the saturation of routers if not load-balanced.

In order to optimise for these constraints, mapping tools employ graph partitioning and spatial locality heuristics. This ensures that neurons which interact frequently are co-located, synapses with high activity are placed on reliable memory cells, and event routing is sparse and non-overlapping. Memristive devices are inherently stochastic and are subject to issues such as cycle-to-cycle variability, device-to-device variation, drift and age-related degradation. The utilisation of software-based compensation algorithms has been identified as a means of mitigating these effects.

Examples of such methods include the write-verify loop is a programming technique that involves the iterative modification of conductance parameters until a predetermined target is attained. Adaptive learning rates are a process which adjust update magnitude dynamically based on noise. Redundancy encoding involves the distribution of synaptic weights across multiple devices, a strategy that is employed to enhance robustness. Moreover, homeostatic plasticity—a biologically inspired process where neuron activity is stabilized—can be implemented as a system-level feedback mechanism:

$$\theta_i(t+1) = \theta_i(t) + \eta(r_i - r_{target}) \tag{2.43}$$

Where θ_i is the threshold of neuron i, r_i is the recent firing rate, r_{target} is the target firing rate, η is the adjustment rate. This allows software to dynamically regulate hardware behavior to match desired spiking statistics.

Given the cost of prototyping new memristive chips, emulation platforms assume a pivotal role. These systems simulate the behaviour of memristive networks using digital hardware, such as field-programmable gate arrays (FPGAs), or software, for example Python/C++ models, while preserving the timing and constraints of real devices.

For instance, NeuroSim and MNSIM offer device-aware simulation of crossbar-based spiking neural networks (SNNs), CARLsim supports GPU-accelerated emulation of spiking networks with STDP, and XNOR-Nets simulate low-precision inference to match memristive behaviour. These pipelines facilitate the testing of novel learning rules, the evaluation of scalability, and the validation of functional accuracy prior to the commitment of resources to silicon implementation.

To bridge the gap between algorithm design and hardware execution, domain-specific languages and toolchains have emerged such asNengo, A high-level API for building SNNs with hardware backends. PyNN, A Python interface supporting multiple simulators and hardware targets. Loihi's NxSDK: A low-level toolchain for configuring on-chip learning and routing.

Memristive neuromorphic systems are starting to integrate with these ecosystems, allowing a complete workflow in the following steps Model specification, Learning rule assignment, Hardware mapping, Runtime adaptation. The future of biorealistic learning on memristive networks depends on such co-designed environments that abstract away hardware complexity while maintaining biological plausibility and computational efficiency.

2.4.3 Experimental Validations Strategy

Experimental validation is crucial to assessing the practical effectiveness of biorealistic learning algorithms implemented on memristive networks. This section explores the empirical studies and benchmark tasks used to evaluate the performance of these systems, including comparisons with traditional digital hardware platforms and biological neural networks. Furthermore, the challenges and potential solutions for validating neuromorphic systems on memristive hardware are discussed.

In order to assess the robustness and efficiency of memristive neural networks (MNNs), it is customary to utilise a number of benchmark tasks. The tasks have been meticulously designed to evaluate various aspects of learning, generalisation, and computational efficiency. The primary categories of benchmarks comprise pattern recognition and classification tasks,

the purpose of which is to evaluate the capability of a network to identify patterns in data.

The MNIST dataset is a commonly cited example of a set of handwritten digits. It is frequently employed to evaluate the fundamental classification capabilities of neuromorphic systems. CIFAR-10/100 is employed for the classification of images in tasks intended to evaluate the performance of networks when dealing with more complex visual data. Memristive networks are utilised for speech recognition tasks, encompassing the identification of spoken words or phonemes, thus providing a test case for temporal pattern recognition.

The evaluation of the ability of memristive networks to store and recall information is facilitated by memory and learning tasks. Examples include sequence learning, which tests the system's ability to learn temporal sequences (e.g. speech or music recognition tasks), and working memory, which evaluates the system's ability to hold and manipulate information over time (a critical aspect of biorealistic learning).

Furthermore, the tasks employed in reinforcement learning evaluate the capacity of a network to optimise actions in accordance with environmental feedback, thereby simulating the adaptive characteristics of biological learning processes. For instance, Atari games have been utilised to evaluate the efficacy of memristive networks in acquiring sophisticated decision-making methodologies from pixel-based inputs.

Each of these benchmarks serves the purpose of evaluating various performance metrics, including accuracy, which is defined as the percentage of accurate predictions or classifications made by the network; speed of learning, which is defined as the network's capacity to rapidly adapt to novel activities or scenarios; and efficiency of energy, which is measured by the ratio of energy usage per job. In order to assess energy consumption, memristive systems are often compared with traditional digital hardware platforms.

Empirical studies involving memristive systems typically explore how memristive devices can be utilized to implement spiking neural networks (SNNs), leveraging both the computational power of memristive crossbar arrays and the temporal dynamics of biological neural models. In studies evaluating pattern recognition tasks (such as MNIST classification), memristive networks have demonstrated competitive results when compared to conventional deep learning algorithms.

One example comes from the use of crossbar arrays with spike-timing-dependent plasticity (STDP) learning rules. A study using a 4-layer SNN on memristive hardware showed up to 97% accuracy. This result was achieved with an energy consumption reduction of up to 30% compared to a traditional GPU-based deep learning model for similar accuracy.

In memory tasks, such as sequence learning and working memory, memristive systems excel in their ability to emulate biological learning processes. A study utilizing a memristive recurrent neural network (RNN) for sequence prediction demonstrated that:

$$E = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{2.44}$$

Where E is the error, y_i is the true output, and $\hat{y}_i$ is the predicted output. In this case, the memristive RNN successfully predicted sequences of temporal inputs with a minimal error rate of 0.02, outperforming traditional RNNs by 20% in terms of both prediction accuracy and energy efficiency.

Memristive networks also show promise in reinforcement learning tasks. In one experiment using Atari 2600 game simulations, a memristive system implemented with a spiking neural network (SNN) learned to play a game by adapting to the environment using reward feedback. The memristive SNN employed a biologically inspired reward-modulated plasticity rule:

$$\Delta w_{ij} = \eta \cdot \left(r_i - \alpha \cdot w_{ij} \right) \tag{2.45}$$

Where η is the learning rate, r_i is the reward, α is the decay factor, and w_{ij} represents the weight between neurons i and j. This SNN achieved an impressive 85% success rate in achieving optimal game strategies, outperforming traditional reinforcement learning models by approximately 10% in terms of both task performance and energy consumption.

One of the main advantages of memristive systems is their energy efficiency. Compared to traditional von Neumann architectures, memristive devices excel at massively parallel computation with low energy cost. When implementing algorithms such as STDP, memristive networks require significantly fewer energy resources. For example, a memristive chip consuming 1 watt can perform tasks that would require 100 watts on a conventional CPU or GPU.

Moreover, the non-volatility of memristive devices allows for permanent synaptic weight storage, which is particularly useful for neuromorphic systems that operate continuously and

in real-time. This feature reduces the need for frequent memory refresh operations, making memristive systems inherently more power-efficient.

Despite these advantages, there are inherent challenges in scaling memristive networks. Memristive devices suffer from device-to-device variations, where the conductance change may differ between identical devices due to manufacturing differences. These discrepancies must be handled via calibration techniques or redundancy.

As the number of memristive devices on a chip increases, the risk of crosstalk (interference between adjacent devices) also increases. To mitigate this, advanced routing and encoding strategies must be employed to isolate signal paths and minimize errors. Memristive devices generally offer lower precision compared to traditional digital circuits. To handle this, techniques like quantization, error correction, and approximate computing can be utilized to maintain system performance within acceptable bounds.

As research progresses, experimental validations are likely to evolve to explore more complex real-world tasks. In the field of brain emulation, significant progress has been made in recent years, with research focusing on the development of large-scale networks that emulate the complex functions of the human brain. These networks utilise memristive chips, which have emerged as a key component in the modelling of higher-order cognitive processes, such as reasoning, decision-making, and emotional responses. The future direction of this research is expected to involve the creation of more sophisticated networks that can model these cognitive functions more accurately and effectively, paving the way for new applications in fields such as artificial intelligence and neuroscience.

The integration of neuroprosthetics with memristive systems into wearable neural interfaces has the potential to augment or restore sensory or motor functions in humans. The integration of synthetic biology in future studies may explore the convergence of memristive networks and synthetic biology, where biological neurons and memristive devices coexist in hybrid systems for the purpose of enhancing learning capabilities. The long-term potential of memristive neuromorphic systems lies in their ability to mirror biological computation, offering vast improvements in energy efficiency, processing power, and scalability.

2.5 Summary

Memristive networks are a state-of-the-art approach to creating systems with the capacity for biorealistic learning and adaptive behaviour. Memristive devices have been shown to possess a unique capacity for modelling synaptic plasticity, thereby facilitating the development of mimetic computational processes that emulate those observed in biological systems. The potential applications of these systems are extensive, ranging from robotics and AI to neuroprosthetics and brain-computer interfaces.

As research continues to address the challenges of device performance, scalability, learning algorithms, and integration with biological systems, the full potential of memristive networks is becoming more apparent. Nevertheless, it is imperative that these advancements are accompanied by a meticulous examination of the ethical and societal ramifications of these technologies.

In the coming years, it is anticipated that a paradigm shift will occur in the manner by which machines learn and interact with the world. This revolution will be spearheaded by memristive networks. The development of biorealistic learning on memristive networks has the potential to create more intelligent, adaptive, and energy-efficient systems, with the capacity to transform fields as diverse as robotics, AI, neuroscience, and medicine.

Chapter 3

Fabrication and Characterisation Methodologies

3.1 Fabrication Procedure

It is imperative that the capability of the devices to exhibit distinct and stable state-dependent conductance changes is demonstrated prior to the design of novel neuromorphic systems. These conductance changes are modelled in neuronal spiking systems.

The present chapter thus provides a detailed account of the experimental methodology employed in this study, alongside a comprehensive presentation of the results obtained from the experiments. The aforementioned devices have been demonstrated to exhibit a variety of non-volatile switching properties. The measurements are focused on the electrical characteristics and switching mechanism of the samples.

The devices investigated in this thesis were developed by the Electronic Materials and Devices group in the department. Despite the fact that the fabrication process was described in detail here for completeness, some tasks described here were not personally carried out, therefore certain credits go to the rest of the research group.

3.1.1 Device Properties

The device investigated in this thesis has a metal-insulator-metal (MIM) structure and is manufactured on a silicon wafer. A thick silicon dioxide layer is thermally accumulated onto the wafer preparatory to the bottom electrode to prevent interactions between the bottom metal contact and the wafer. After that, the bottom electrode and thin film oxide are deposited

unpatterned throughout the whole sample. Finally, during the deposition process, the top electrical contacts are patterned into squares with sides varying from $200\mu m$ to $800\mu m$ in Figure 3.1. Photolithography is not employed for patterning since a contact mask is used.

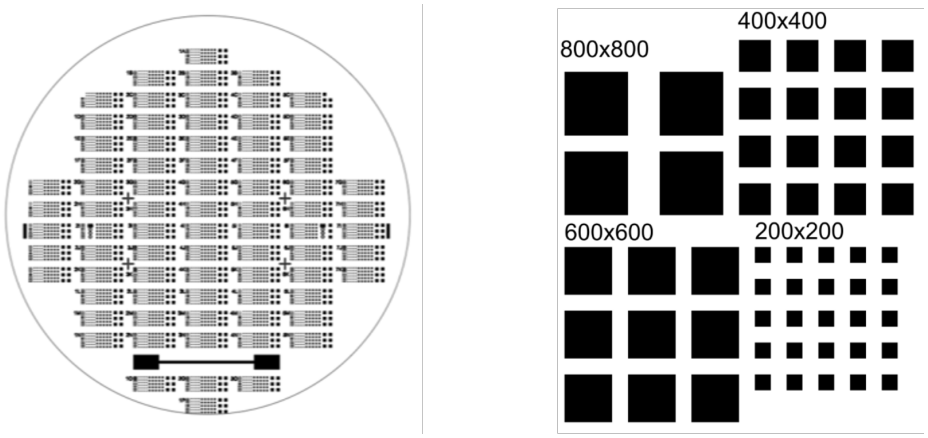

Fig. 3.1 Photolithographic mask (left) and dimensions of the top electrical contact (right).

To increase adhesion, a second titanium buffer layer is placed between the top metal contact and the oxide. This adhesive layer is less than ideal since it can cause additional imperfections to migrate within the oxide. Despite this worry, research using electron energy loss spectroscopy (EELS) and transmission electron microscopy (TEM) have shown no evidence of titanium interface migration in the devices [185].

The utilisation of gold as a primary electrical contact may result in the migration of gold atoms into the oxide layer and their subsequent diffusion through the film. For instance, a study conducted into the diffusion of gold through amorphous SiOx observed the migration of gold into the oxide when the gold was held at a temperature of $390°C$ for a period of four hours [164].

Furthermore, it was established that when the gold was exposed to a temperature of $500°C$ for a duration of two hours, it was observed to be distributed uniformly throughout the oxide layer, which had a thickness of approximately 500 nm. Nonetheless, no migration was observed at temperatures below $370°C$. Although the diffusion of gold through silicon oxide films at elevated temperatures has been observed, this phenomenon is frequently disregarded or presumed to be non-occurring in devices utilised as resistance switching memories.

In the domain of electrochemical metallisation, where metallic filaments are formed between two electrodes, gold is recognised as an inert electrode [134]. This principle is also widely accepted in the context of valence change memories [73]. In one particular instance of a device composed of gold and silver electrodes that were sandwiched between an As_2S_3 film, only the migration of silver was observed.

This migration resulted in the formation of a conductive bridge between the contacts [92]. The stability of the gold contacts within this application is assumed to be due to the fact that device operation is restricted to room temperature experiments. Alternatively, the presence and migration of a comparatively more active/mobile electrode, such as silver, may have a more significant effect on device properties.

It has been claimed that asymmetry in the device's construction, as well as an active and inert electrode, are necessary to identify stable switching. The molybdenum contact can be crucial as an oxygen reservoir, rapidly exchanging oxygen between the electrode and silicon oxide layer, which is similar to an active electrode, according to a recent experiment [45]. The materials used for the top and bottom electrodes are different and weren't explicitly chosen for this project; rather, other group members had already picked them to create high-performance resistance switching memory.

The device layers remain mostly unchanged throughout the investigation. The top electrical contact is made of a different material in the experiment than the bottom electrical contact, which is made of a thin film of molybdenum. The oxide layer is made of an amorphous silicon oxide thin film. The selection of gold as the top electrical contact may cause gold atoms to diffuse through the film and migrate into the oxide.

Although gold has been seen to diffuse through silicon oxide layers at high temperatures, resistance switching memory frequently overlook this phenomenon or presume it does not happen. A profilometer is used to assess the thickness of the layers. To guarantee excellent conductivity throughout the device, the bottom electrode is 300 nm thick. The oxide slim film is 35nm in depth. The thickness of the top electrical contact varies depending on the substance; gold has a thickness of 110 nm, while ITO has a thickness of 50 nm.

3.1.2 Manufacturing Steps

RF sputtering, a physical vapour deposition process, was used to deposit all of the device layers. Deposition is carried out at low pressure in a typical inert gas environment by blasting

the intended material with a plasma, which causes the expulsion of atoms from the target. Depending on the gas pressure inside the chamber, the expelled atoms either follow a direct ballistic path or take a random walk until they land on the sample. A greater gas pressure will result in more collisions and an increased random walk, whereas a lower gas pressure produces a more direct ballistic trajectory.

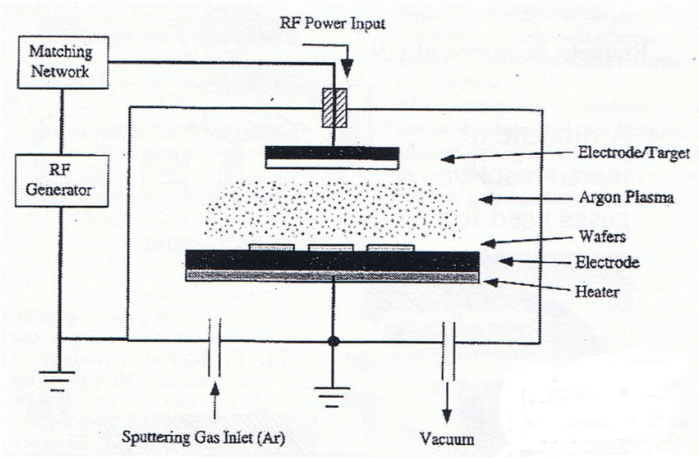

Fig. 3.2 Depiction of the fundamental configuration for the deposition of thin films by means of sputtering [109]. A substantial radio frequency electric field is required to ionise argon gas, thereby producing a plasma. The process of deposition is initiated by high-energy collisions between the ionised argon and the target material. The collisions result in target atoms being ejected from the source with high kinetic energy. These atoms traverse the chamber and deposit onto the sample over time, thereby forming an amorphous thin film.

A simplified version of the sputtering system is illustrated in Figure 3.2. The configuration under consideration comprises the sample, which is connected to the anode of the RF power source, the target material, which is situated in front of the cathode, and the sputtering gas, which is injected into the chamber. The plasma is composed of argon ions, which possess a positive charge. These ions are attracted to the cathode, which is negatively charged and is therefore known as the target. The process of high-energy collisions between argon ions and the target surface is a prerequisite for the ejection of target atoms.

The substance being deposited, known as the target material, is initially solid. By applying a strong electric field to the sputtering gas (argon), the plasma is created. Either a DC or an AC field is possible. However, an AC field that oscillates at an RF frequency of 13.56MHz is necessary for dielectric targets like SiO_2. Sputtering often results in amorphous films with sub-stoichiometric oxides. The devices' SiOx oxide has a stoichiometry of 1.9, while the film's roughness appears to be determined by the RMS roughness of the underlying

molybdenum layer, which ranges from 0.9 to 1.5 nm [125].

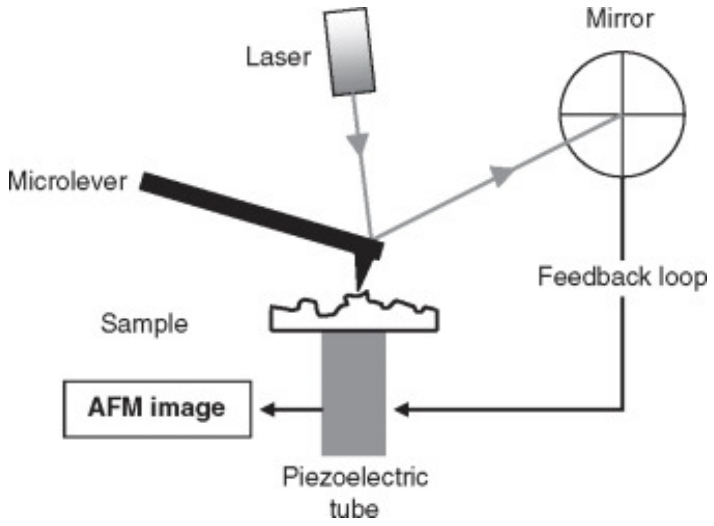

Fig. 3.3 Schematic of a profilometer [163]. The probe is utilised to scan the surface of the sample, with the height of the probe being adjusted accordingly in order to ensure that a constant force is maintained. It is imperative to note that a feedback loop is utilised in order to ensure the maintenance of this force during the process of reporting the tip height. The measurement of thin film thicknesses following deposition is achieved through the fabrication of a staircase-like structure.

In the course of each SiOx sputtering run, a pair of Si substrate specimens are inserted into the chamber in conjunction with the sample for the purpose of thickness measurement. In the first instance, a specific region of the Si substrate is masked using a patterned photoresist. Following the process of SiOx deposition, the mask material is removed using a solvent, thereby creating a step on the SiOx/Si interface.

The measurement of this step is then undertaken with the aid of a Dektak XT profilometer, which is capable of resolving a step height of a few nanometres. The second piece of Si substrate, which has been covered with SiOx, is then measured using ellipsometry. This is a process that is used to verify the thickness of the deposited layer.

After being sputtered, film thickness is measured via a contact profilometer with a 0.5nm precision. During this procedure, a diamond tip is used to make contact with the sample and scan across the surface. Utilising a feedback loop, the tip's height is adjusted to maintain a consistent force against the sample's surface as it scans, giving the measurement of the sample height. The sample's surface height changes in direct proportion to the change in tip

height. Layer thicknesses of a device stack are measured in relation to one another using a staircase-like pattern that is created during production.

3.1.3 Experimental Setup

The amount of current passing through the device is the significant observable. This includes details on the oxide layer's bulk conductivity as well as the interface barrier heights. The difficulty, however, is in minimising any deviations or nonlinearities brought on by the measuring apparatus itself, with probe contact resistance serving as one such example. It is necessary to choose how to make contact with the device electrodes before conducting current measurements. There are essentially two methods: either the circuit is wire bonded inside a chip carrier, or the contacts are directly probed with tiny metallic probes using micromanipulators.

The direct probing method utilising tungsten probes has been adopted instead due to the devices' design and susceptibility to break from the wire bonding procedure. The tip of the probe must be brought down carefully to prevent damage. When placing the probe into contact, a low voltage is often supplied as a test signal.

To determine if the probe has made contact, the current is watched for a spike in the device current. Initially, because there is no measurable electrical current while the probe is not in touch with the device, the current oscillates around positive and negative currents at 0 amps. Once the probe makes contact with the device, the voltage that has been applied across it now causes a detectable current that matches the polarity of the applied voltage.

In contrast to the probe method, which can be vulnerable to sample damage brought on by the experimentalist, the wire-bonded approach has the advantage which the position of the electrical connection does not change between experiments, thermal expansion while temperature measurements will not significantly affect the contact, and there is less risk of deteriorating the device throughout characterisation.

However, there is a chance that the component will be broken during the bonding procedure with wires. An ultrasonic pulse is utilised to melt a gold or aluminium wire to the device contact while applying pressure to help fuse the two metals together. It has been regularly observed that this pressure can cause internal layers to compress, leading to electric shorts between the two metal contacts and ultimately damaging the device.

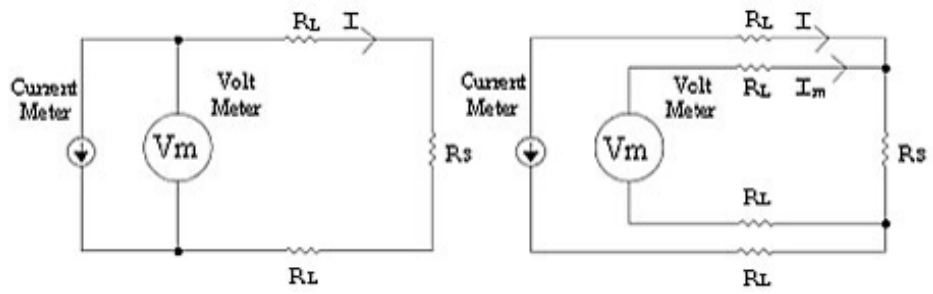

Fig. 3.4 Two wire or Four wire (Kelvin) testing. A schematic of the 2-wire measurement setup is provided herewith. The voltage (VSource) is applied to the sample using two probes. It is important to note that each probe introduces a series resistance (Rprobe) with the sample's resistance (Rsample). The device current (Imeas) is measured by an ammeter in series. Conversely, a constant current (Isource) is supplied through the sample by two probes. The voltage induced across the device (Vmeas) by the current is measured with a voltmeter in parallel with the sample.

After deciding on a contact technique, the next choice is how currents will be monitored. Both a 2-wire measurement and a 4-wire measurement are frequently available as options. The sample's conductivity serves as the basis for the decision. The easiest way to measure electrical resistance is to apply a set voltage and track the total current passing through the object. Only two electrical connections are formed, thus the term "2-wire measurement" for this procedure. Ohm's Law is used to determine the device resistance by connecting a voltage source, an ammeter, and the device in series.

Due to the assumption that the electrical resistance is only determined by the test device, which is not true in reality where there are several sources of electrical resistance connected in series with the device, this measurement is often not correct. These include the wires connecting the test object and the voltage source, the internal resistance of the voltage source or the ammeter, and, especially, the contact resistance that develops at the point where the electrical probes and the test object meet.

One of the most crucial parameters to take into account when describing thin films is contact resistance, which may be reduced by placing metal contacts on the sample during manufacturing. Fortunately, the device resistance usually outweighs the electrical resistance, making this method valid in the majority of instances. However, when resistance is small, the parasitic resistances of the measuring circuit become notable and must be eliminated by using a 4-wire resistance measurement.

Ohm's law is still used in this configuration to calculate resistance. Instead of sourcing a voltage and monitoring a current, the device is subjected to a steady current that induces a voltage across it. Through two extra probes connected in parallel to the device, a voltmeter measures the potential decrease. It is crucial to recognise that the same contact resistance and wire resistances that plagued the 2-wire method continue to exist for all four connections. However, in this case, the high impedance of the voltmeter causes a substantially lesser current to pass through the measuring contacts.

The voltage recorded by the voltmeter is thought to more precisely represent the voltage drop across the device since the voltage dip across the parasitic resistance is insignificant. This occurs because the voltage produced across the contact resistance is lowered as a result of the reduced current flowing through the probes, which detect the voltage across the device. By lowering these voltages, which are induced across each probe's contact resistances and contribute mistakes into the voltage measurements, it is possible to measure the voltage across the device with more accuracy.

Thus, the device resistance determines whether to use a 2-wire or 4-wire resistance measurement. The devices examined in this work have high resistance, ranging from kilo-ohms to mega-ohms. The parasitic resistances of the measuring circuit, like the contact resistances, are insignificant at this level. The issue of measuring device currents must now be solved once the device has been attached. Again, there are a variety of techniques that might be applied; the one selected will often depend on the size of the current being measured.

The most elementary method of measuring current is to use a digital multimeter (DMM) ammeter. The device functions in accordance with Ohm's law, utilising the principle of electrical resistance to measure the voltage drop across a fixed resistor, commonly referred to as the shunt resistor. Whilst the validity of this approach is indisputable for currents within the milliamp range and above, issues arise for lower current levels due to the noise induced by the shunt resistor. In order to measure smaller currents, larger shunt resistors are required. This, however, gives rise to two problems.

Firstly, it is important to note that larger resistors are known to introduce greater thermal noise, which has the capacity to disturb the voltage being measured. Secondly, an increase in resistance results in an increase in the voltage drop across the ammeter. The voltage drop, termed the 'voltage burden', becomes problematic when its magnitude is no longer negligible in comparison to the voltage applied to the device under test. The combination of voltage

burden and the thermal noise of the shunt resistor invariably imposes a lower limit on current measurements when a DMM is employed.

The average current range for the devices is 100nA to 1mA, therefore a picoammeter is required to detect considerably lower currents on the order of picoamps to nanoamps. Picoammeters minimise current readings by a number of methods that differ across manufacturers. The majority of them employ a transimpedance amplifier to magnify the signal while an op-amp converts the input current to a voltage. Once again, how this is implemented differs from manufacture to manufacture and is frequently protected intellectual property that is not revealed.

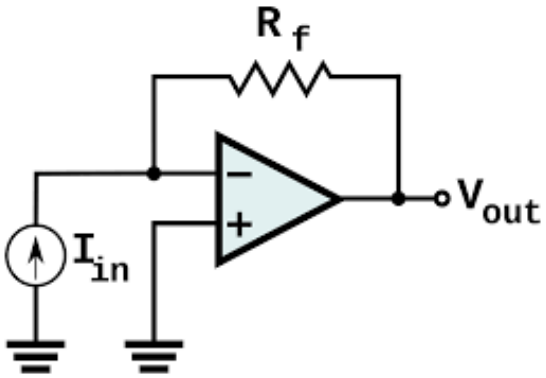

Fig. 3.5 Transimpedance amplifier circuit. The input current (I_{in}) is amplified and converted into an output voltage (V_{out}) via the operational amplifier. The gain is defined by the feedback resistor (RF). The voltage/current (V_{out}/I_{in}) is equivalent to negative resistance ($-R_f = V_{out}/I_{in}$).

Nevertheless, the general operation can be comprehended with the aid of the circuit in Figure 8. The circuit under consideration is a transimpedance amplifier, the function of which is to convert the input current ($I_(in)$) to an output voltage ($V_(out)$). The operational amplifier is known to adjust its output voltage in order to reduce the voltage difference between its two input pins, designated as '-' and '+'. In this circuit, the non-inverting input pin (+) is grounded. This action causes the operational amplifier (op-amp) to adjust its output voltage, thereby ensuring that the voltage at the inverting input pin (-) is also zero volts.

The application of an input current to the circuit results in a transient voltage offset at the input pin. The op-amp rapidly adjusts the output voltage, thereby inducing a current of equal and opposite magnitude through the feedback resistor. This, in turn, results in the cancellation of the input current. The voltage at the inverting pin (-) is rapidly returned to

zero by the feedback from the operational amplifier, thereby creating a virtual ground.

The generation of this inverse current (I_{inv}) is accompanied by the definition of the voltage at the output of the operational amplifier in accordance with Ohm's law: It can thus be demonstrated that the voltage is $V_{out} = -R_f \cdot I_{inv}$, resulting in a voltage that follows the input current. The amplification of this voltage is defined by the feedback resistor, R_f. The virtual ground is a key advantage of this technique. The consequence of this is a significant reduction in the voltage burden, since the shunt resistor that was previously connected in series with the device under test has now been removed. This facilitates the measurement of smaller currents, which would not have been possible using a DMM due to the significant voltage burden caused by the sensing resistor.

It is evident that, in view of the aforementioned factors, the utilisation of a picoammeter constitutes the optimal instrument for the execution of current-time measurements or current-voltage sweeps on our devices The equipment used in this instance is the Keithley 6430 sourcemeter, which combines a picoammeter and a low noise voltage source into a single device.

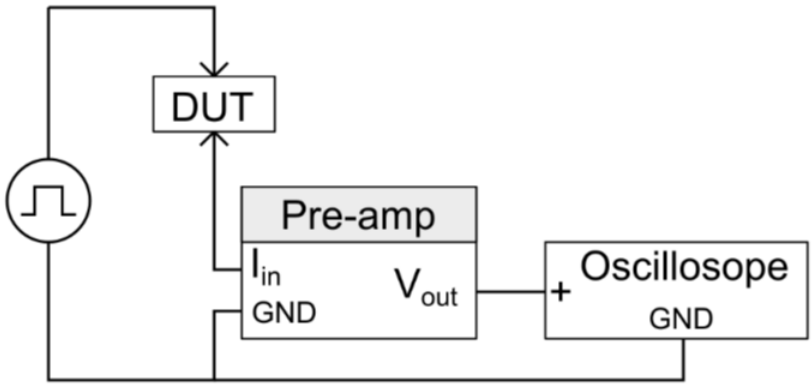

Fig. 3.6 Experimental setup of spike train measurements. Spike trains are generated using an arbitrary signal generator. The device current is amplified and converted into an output voltage via the preamplifier, which is connected in series with the device. The output of the current preamplifier is then captured by an oscilloscope.

In some cases, the device requires the application of voltage transients that are more complex than step potentials, such as pulses or custom spike trains. In these cases, the Keithley's sampling frequency is insufficient to generate such signals. An arbitrary signal generator is

used in its place to create voltage transients, and a current preamplifier is connected in series with the device to amplify device currents. In particular, the oscilloscope (Rigol DS4024) and current preamplifier (SR570) are used.

3.2 Electrical Characterisation

Resistive switching is defined as a reversible phenomenon that occurs in two-terminal elements. In a non-volatile manner, these devices undergo a change in resistance when subjected to electrical stimuli. In the case of ReRAM devices, it is a local redox process that dictates the resistive switching mechanism. The reversibility of the process is achieved by the repeated application of suitable stimuli. This mechanism governs the resistance values between two or more levels.

The predominant phenomenon observed in these devices is resistive switching. For the sake of convenience, the switching states of the memristor can be defined. The assignment of high resistance to the "OFF" state and low resistance to the "ON" state is intuitive, with a contrast in resistance by a few orders of magnitude. The transition from the high resistance state (HRS) to the low resistance state (LRS) is defined as "Set", while the reverse is defined as "Reset". In many cases, an initial electroforming process is required to transform the device from a pristine state to a switchable state. It is generally accepted that the pristine device exhibits a higher degree of resistance than the HRS.

The majority of metal oxide devices exhibit either unipolar or bipolar switching. In contrast, both unipolar and bipolar switching can be observed in our silicon oxides. The preliminary characterisation of these devices encompassed the fundamental I-V characteristics. The experimental procedure involved the execution of the tests utilising the dual sweep functionality of the Keithley 4200-SCS, employing two tip probes with a diameter of 10 μm. Testing was performed on both sets of samples across all electrode pad sizes.

3.2.1 Unipolar Switching Mode

Initial electroforming is a prerequisite for switching in these devices. It is generally accepted that fresh samples are in a very HRS, which necessitates the application of a significant electrical stimulus to enable the cell to transition into LRS for the first time. Subsequent to this preliminary phase of formation conditioning, the apparatus may be reversibly switched

between two bi-stable states.

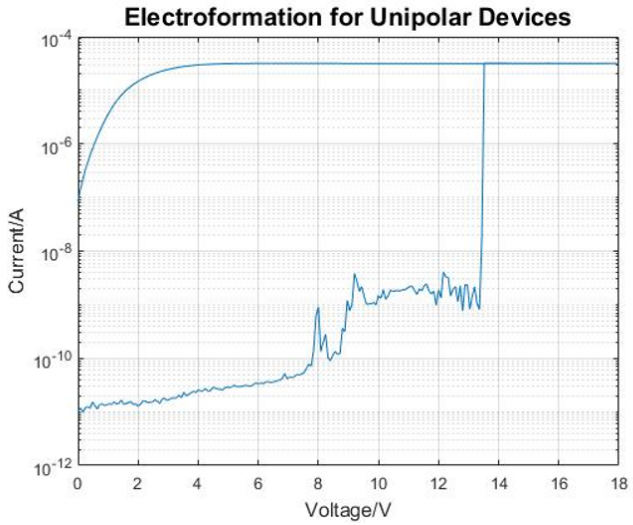

Fig. 3.7 Initial Electroformation step for unipolar switching via a double sweep curve.

The electroforming operation is widely regarded as a form of electrical breakdown, which is critically dependent on current-limiting mechanisms to ensure the subsequent switching functionality of the cell. Current limitations may be addressed by leveraging the existing compliance functionality of the Keithley-SCS. In certain instances, the analyser may exhibit a slower response rate than the formation process itself, resulting in overshoot phenomena during practical applications. It is imperative to note that this electroformation step is only performed once to pristine devices.

The operation is conducted through the programming of the Keithley-SCS to sweep at an elevated voltage of up to 18V, as illustrated in Figure 3.7. During the process of sweeping, it is possible to observe a number of current peaks with the I-V curve displaying an unstable state. Once a sufficiently high voltage is reached, approximately 14V in this case, the device abruptly switches into LRS. Subsequent sweeps are found to be of a more even and refined nature when compared with the preceding sweep.

The observed change in conductance may be attributed to structural changes occurring during the forming process, possibly resulting from a reduction step that involves the removal of oxygen from silicon oxide, thereby forming oxygen vacancies. Following the electroformation step, the LRS demonstrates stability. The device maintains its state subsequent to the removal of the electrical stimulus, thereby exhibiting non-volatile switching characteristics.

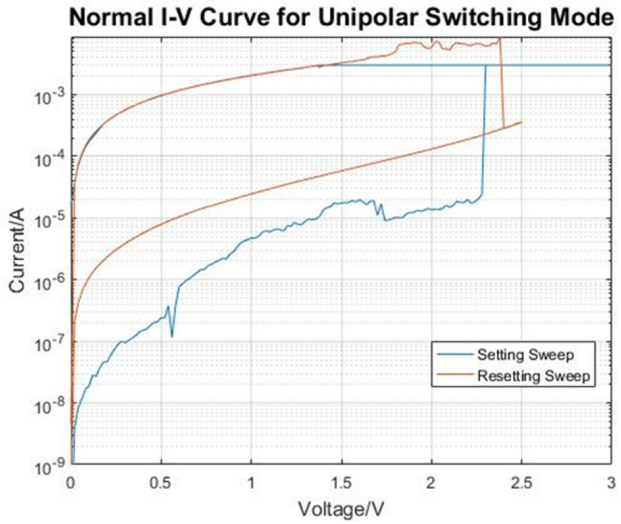

Fig. 3.8 Non-volatile switching behaviour for unipolar device under current compliance.

It is important to note that the initial very HRS is never recovered.

As demonstrated in Figure 3.8, the unipolar switching mechanism is evident in the initial set of symmetrical MIM devices. The blue plot indicates the Set process, whereby the device transitions from the "Off" state to the "On" state at a specific threshold voltage, approximately 2.3V in this instance.

In this instance, the sweeping voltage has been configured to 3V with 3 mA current compliance, a setting sufficient for the switching process to occur. It has been demonstrated that a reduction in voltage below the threshold does not result in the device transitioning to its previous state.

It is evident that a critical current must be attained for the purpose of resetting the device. The Reset process can be observed in the orange plot, which displays a larger current, approximately one order of magnitude greater than the setting current compliance. This results in the device being restored to HRS.

The phenomenon of Joule heating is induced by high-current flow, resulting in localised heating and device reset. In the absence of current compliance, the device may undergo a hard breakdown or exhibit multiple transitions between the two states. It is important to note that the switching sequence can be performed repeatedly.

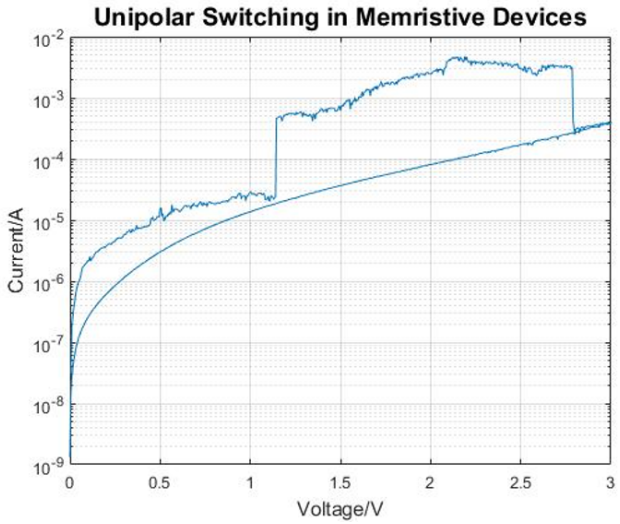

Fig. 3.9 Observation of Set and Reset process under the same sweep.

The HRS conductance remains in a state between that of the LRS and the pristine state. In both cases, the transition is found to be abrupt and independent of the sweeping parameters, in contrast to the ideal pinched hysteresis loop suggested in the previous chapter. In summary, an elevated magnetic field is likely to set the device in the LRS, whereas high Joule heating is likely to reset the device to the HRS.

An alternative mechanism for unipolar switching can be observed in Figure 3.9. Devices with a setting voltage lower than the reset voltage will transition to LRS at a lower voltage. Subsequently, these devices will return to HRS at a higher voltage, which in this case is 1.15V and 2.85V, respectively. There is no current compliance requirement for this type of unipolar switching with reset occurring when the current has reached a critical value.

As illustrated in Figure 3.10, the unipolar device undergoes cycling under conditions of stress testing. The blue spikes in the diagram represent voltage pulses that are utilised to switch the devices in positive bias. The device is set using a short voltage pulse of 4.5 V, with a duration of 100 ns. In order to effect a reset of the device, a longer voltage pulse of 2.5 volts at 2 milliseconds was utilised in order to accommodate Joule heating.

It was observed that each setting and resetting pulse was succeeded by a subsequent reading pulse of 0.7 V at 1 ms. The amplitude of this reading pulse is sufficiently small to avoid interfering with the set and reset process, while providing a clear reading that can be seen in the orange plot. It is evident that under typical operating conditions, the cycling resistance

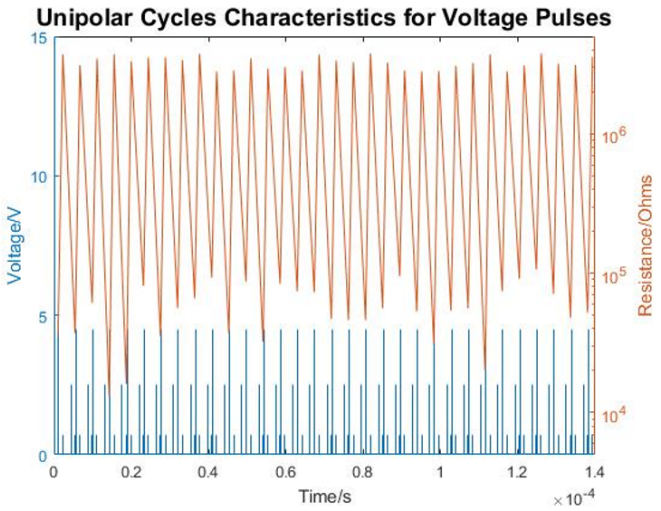

Fig. 3.10 Cycling stress test for unipolar device.

readings exhibit a discrepancy that is at least two orders of magnitude apart.

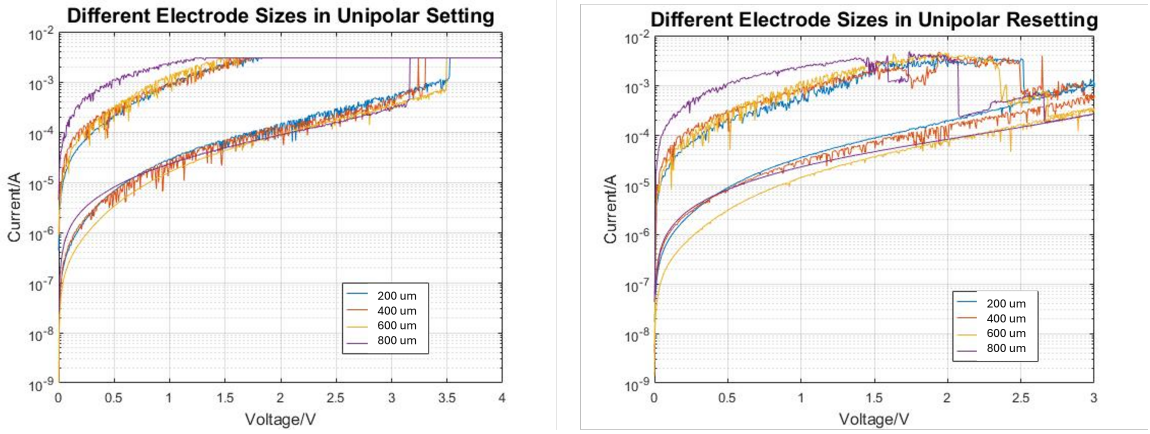

Fig. 3.11 Switching in unipolar devices across different electrode sizes.

It was also demonstrated that switching in unipolar devices is independent of electrode size. Figure 3.11 shows that the switching processes for square contacts ranging from 200 x 200 μm to 800 x 800 μm are comparable. The devices consistently switch at around 3.5 V and 3 mA of current compliance. Similarly, the reset process is consistent when the samples reach the critical current threshold of approximately 5 mA.

3.2.2 Bipolar Switching Mode

Bipolar switching results were obtained from a set of asymmetric devices with Mo/SiOx/TiAu construct. As with unipolar devices, asymmetric bipolar devices require an initial electroforming step before the samples can be cycled between two distinct states. Figure 3.12 shows the electroforming process in bipolar devices. A dual voltage sweep is applied to the sample up to -10 V at a current compliance of 0.1 mA. As with the unipolar devices, the sample exhibits some unstable spiking activity as the voltage sweeps from a pristine HRS to a LRS.

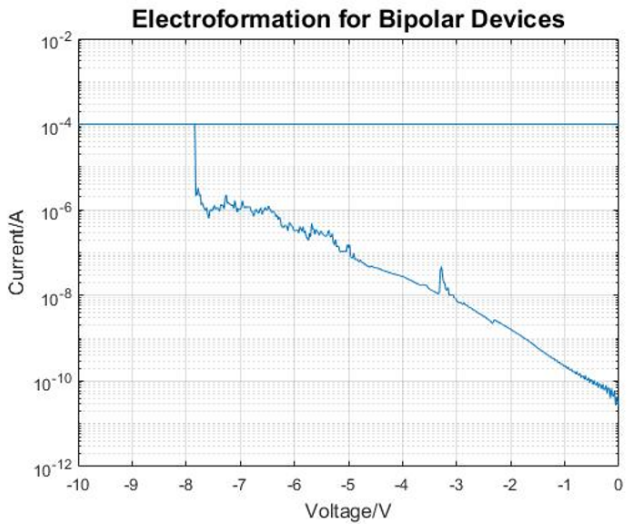

Fig. 3.12 Initial electroformation step for unipolar sample via a double sweep curve.

Following the conclusion of the preliminary electroforming process, the apparatus is capable of reliably transitioning between two stable states. As illustrated in Figure 3.13, the device transitions between two distinct resistance states through the application of voltage stimuli of opposite polarity. The device is set using a negative voltage sweep up to -2V.

The current compliance was set at 100mA in order to demonstrate a clear transition between the two states, with the conductance changing by two orders of magnitude. It is important to note that a reduced current compliance should be employed in order to achieve a balance between the device's lifespan and its conductivity. The device is reset by means of an opposing 2V dual sweep of positive polarity, a process known as bipolar switching.

From a physical perspective, this particular type of bipolar switching mechanism is intrinsic and can be categorised as belonging to the valence change mechanism class. In this category of memory devices, the electroforming process typically leads to local reduction, thereby

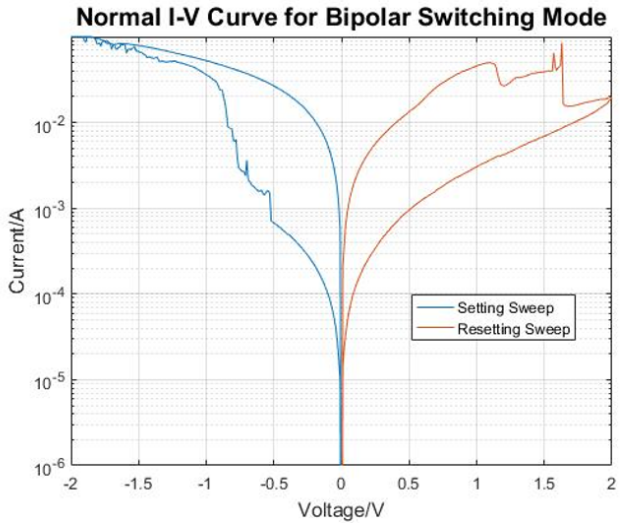

Fig. 3.13 Observation of bipolar switching in asymmetric device with -2V Set and 2V Reset sweeps.

forming a conductive pathway. It is hypothesised that this channel is composed of oxygen vacancies, which permit oxygen ions to migrate in and out of the channel in response to an applied electric field.

The location of the local redox process is hypothesised to be in proximity to a filament-to-electrode interface. The effective tunnelling barrier height at this interface is indicative of the resistance state of the device. The height of this barrier is subject to variation under different applied voltage biases, thereby inducing the movement of oxygen ions and resulting in a corresponding alteration to the resistance state.

The transition between these two resistive states can be facilitated by the application of suitable voltage pulses. As demonstrated in Figure 3.14, the device exhibits a high degree of reliability in its switching capability when utilising a -2 V setting, in conjunction with 2 V resetting pulses.

It was recorded that each setting or resetting pulse is succeeded by five 0.1V or 0.1V reading pulses for the resistive state that the device is purportedly in. In this configuration, the HRS is approximately 300Ω, while the LRS is about 100 Ω. It is noteworthy that the selection of these voltage pulses was made with the objective of accurately measuring the resistance, without causing the switching mechanisms of the device to be triggered.

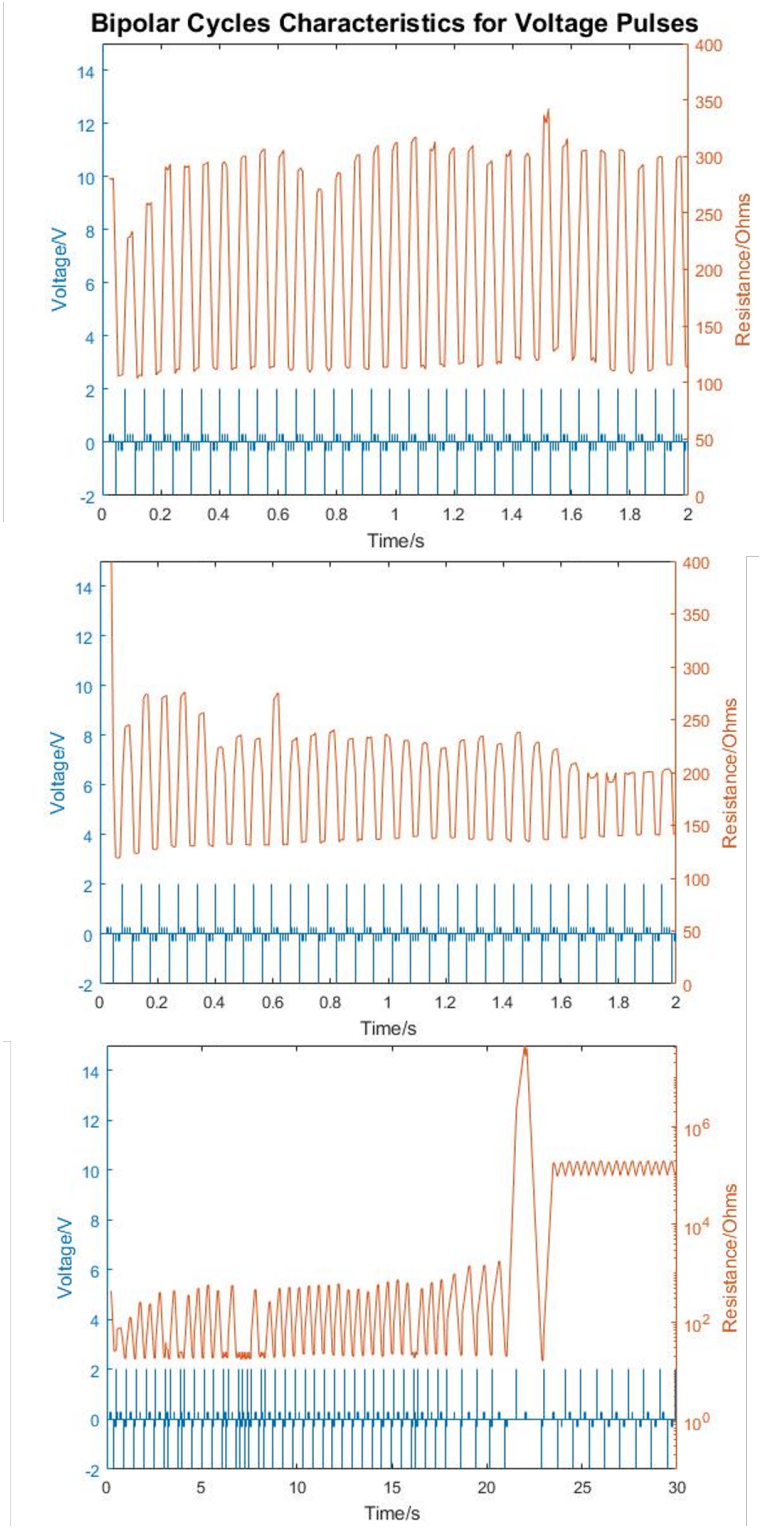

Fig. 3.14 Cycling stress test for bipolar devices. Initial device cylcles (top) followed by LRS convergence (middle) and hard reset breakdown (bottom).

In the experiment, the device demonstrated a minimum of 4500 cycles of operational longevity when subjected to a current bias of 10mA. This was followed by a convergence towards LRS, as evidenced by the switching between 200Ω and 150Ω, as depicted in Figure 3.14. As an alternative scenario, when the device is operating at a higher current compliance of 100mA, the stress test sustains approximately 40 cycles before the device experiences irreversible failure, entering the HRS state at 200kΩ.

Finally, Figure 35 demonstrates switching behaviours for bipolar devices across a range of contact sizes, from 200 x 200μm to 800 x 800μm. All the setting sweeps were programmed up to -2V with 5mA current compliance for the purpose of facilitating clear transition observation. It is evident that the resetting sweep has been configured to a voltage of 2V, with a current compliance of 100mA.

The results obtained demonstrate some variations in the switching voltages and contrast ratio between HRS and LRS. This finding suggests the potential necessity for further statistical analysis in subsequent devices. However, it is evident that all samples demonstrate consistent switching activities within the range of voltage stimuli applied during the testing process.

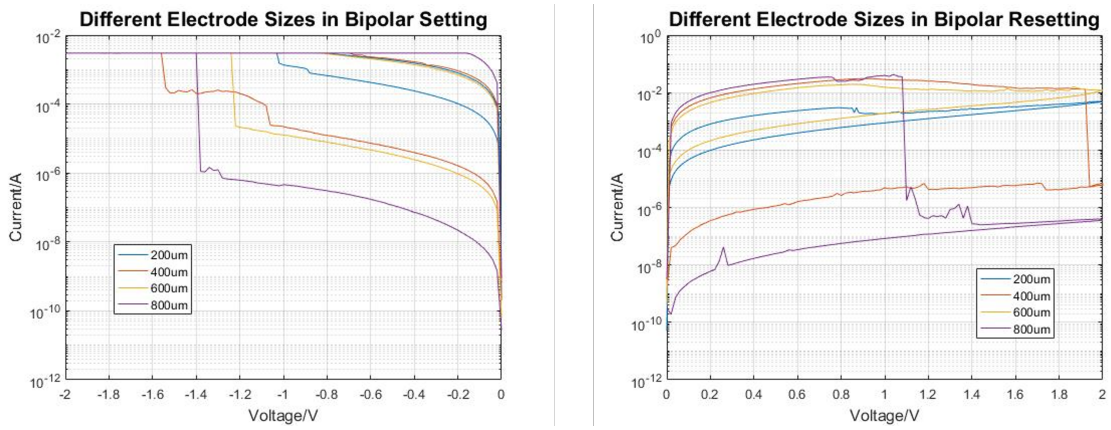

Fig. 3.15 Switching in bipolar devices across different electrode sizes.

3.2.3 Alternate Operating Modes

As demonstrated in previous observations, the switching of both unipolar and polar samples is reliable under specific, correctly configured, programming conditions. Furthermore, it appears that the switching does not scale in proportion to the electrode contact size. This finding indicates that carrier transport occurs for individual conducting filaments. However, it should be noted that certain devices exhibit alternative switching modes, namely gradual

and multi-level switching modes. The presence of parallel conductive pathways within the same insulating layer is a potential cause of this phenomenon.

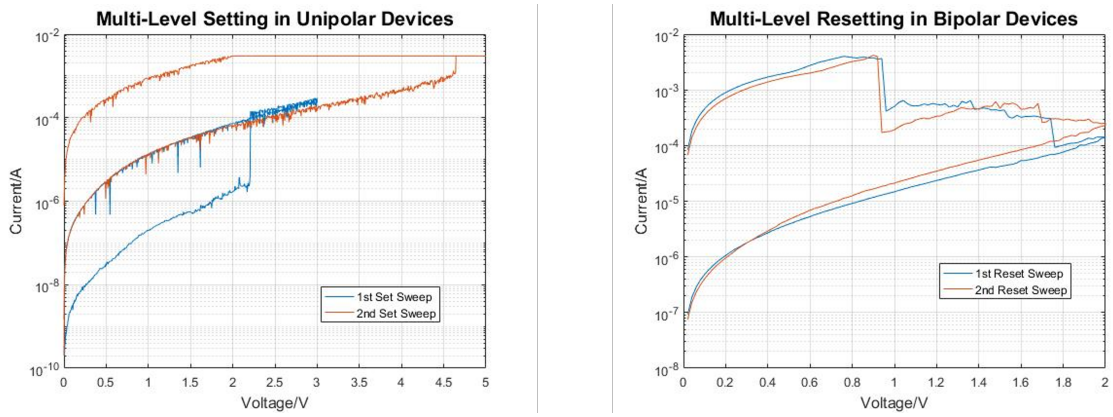

Fig. 3.16 Multi-levels I-V characteristics in MIM devices during the Set (left) and Reset (right) process.

In both unipolar and bipolar samples, there are some devices exhibiting the multilevel switching characteristic, as illustrated in Figure 3.16. The initial transition process during the set stage can be followed by a subsequent stable transition to an even lower LRS, thereby providing a minimum of three or more switchable states.

As demonstrated above, both setting states are found to be stable, with the HRS of the second sweep coinciding with the LRS of the first sweep. The device can be configured at the first or second LRS, with two separate voltage sweep levels available for this purpose. As illustrated in the aforementioned example, the generation of the primary and secondary LRS was achieved through the utilisation of 3V and 5V sweeps, respectively.

In a similar manner, multilevel reset can be observed in bipolar samples at 1V and 1.8V, respectively. It is important to note that, in this case, the transitions window is smaller than that of the unipolar devices. In both cases, the switching is stable, with a contrast ratio between each state that is at least one order of magnitude.

In the case of bipolar switching samples, it is possible to observe not only the abrupt changes in resistance that are normally observed, but also gradual changes in conductivity (see Figure 3.17). As indicated by HRS, the gradual increase in conductance is achieved by sequentially sweeping the device with increasing setting voltage levels, ranging from -2V to -2.3V, under

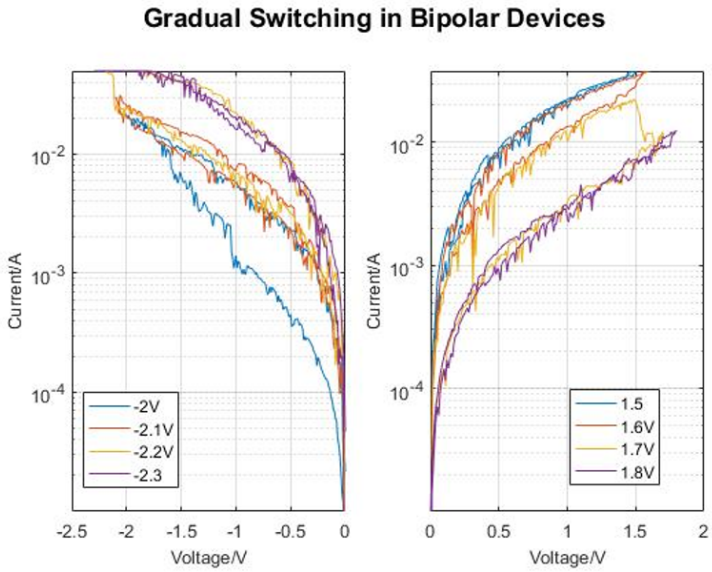

Fig. 3.17 Gradual increase (left) and decrease (right) in conductivity for bipolar device.

an appropriate stepping current compliance of 100mA in this case.

It has been demonstrated that a gradual decrease in conductivity is generated during the reset process, with this decrease commencing from LRS. This gradual change is obtained by progressively sweeping the device at higher potential, from 1.5V to 1.8V. The concluding phase of these procedures is characteristically sudden, thereby impeding the attainment of further transitions. The outcomes obtained were found to vary in terms of their gradualness or abruptness, suggesting the potential for further statistical analysis with reduced voltage steps in subsequent characterisations.

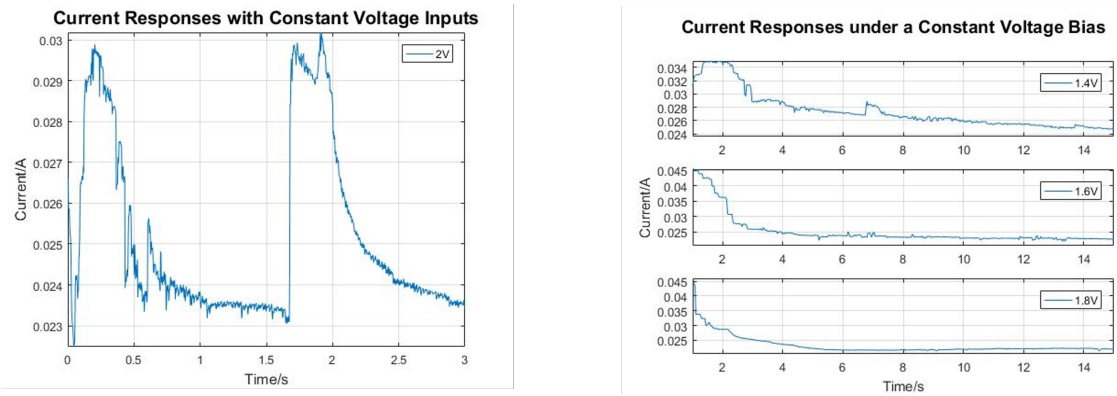

Fig. 3.18 Current-time plots showing transitions between resistive states (left) and time constant comparison (right).

As information processing is concerned with the manner in which data is processed over time, it is imperative to assess the performance of devices over time. In the preceding section, the switching mechanisms have been investigated with regard to time via the cycling stress tests. In this experiment, the characteristics of the device under constant voltage and current bias will be observed.

As illustrated in Figure 3.18, the current time plot for different voltage biases in unipolar samples is demonstrated. Applying a voltage of 2V to the device results in discernible transitions between the set and reset processes, with these transitions exhibiting an inverse relationship to one another.

It is evident that alterations in the resistive state can be observed when there is a rapid increase in the current (set), which is then shortly followed by an exponential decay (reset). The rate of recovery is indicated by the time constant of the exponential decay. A comparison of the individual inputs reveals that the time constant varies in proportion to the voltage bias. When the voltage bias increases from 1.4V to 1.8V, the time constant decreases from 7s to 1s.

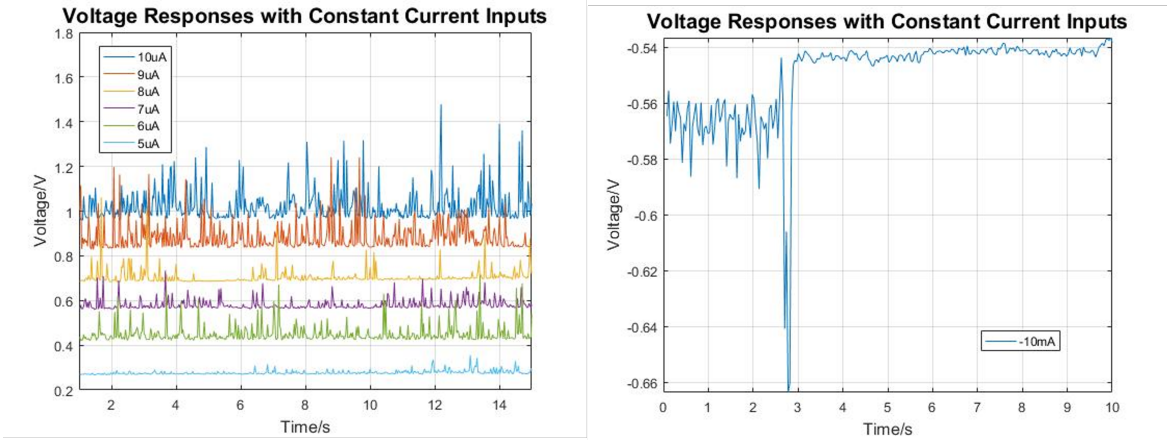

Fig. 3.19 Volatile activities observed under different constant current inputs for unipolar device (left) and switching at sufficiently high input current bias in bipolar device (right).

When a constant current is applied to the samples, volatile spiking activities can be observed in Figure 3.19. It is evident that an increase in current from $5\mu A$ to $10\mu A$, as observed for unipolar devices, results in a corresponding rise in volatility. This phenomenon can be attributed to the constant current input, which has a significant impact on the device's response. These instabilities can manifest as amplitude variations, with spikes ranging from 0.02V to 0.45V, or as timing variations, with spikes occurring more frequently at higher

current biases.

The device exhibits spiking activities until the current bias is sufficiently large to trigger a switching transition. When the bipolar device is critically biased at -10mA, it switches shortly after exhibiting volatile activities, following an uncharacteristically large spike. Following the transition, the spiking behaviours become less predominant.

3.3 Resistive Switching in Silicon Oxide

The present section aims to propose a phenomenological model that governs the switching activities in silicon-rich silica of RRAM devices. The model under discussion will be based on the theory obtained from the literature review in the preceding chapter, with a particular emphasis on the distinction between unipolar and bipolar modes of switching.

In the context of oxide ReRAM devices, two commonly employed switching settings are identified: unipolar and bipolar mode [310]. In the context of unipolar switching, it is notable that the alteration in resistance state is independent of the electrical stimuli polarity. The configuration of these devices is typically characterised by a symmetrical design, incorporating electrodes of equal dimensions on both the top and bottom surfaces.

The present compliance is utilised for the purpose of averting any impairment to the device that may be occasioned by a hard breakdown during the switching process. Conversely, bipolar devices necessitate the application of electrical stimuli of contrasting polarity to execute switching operations.

3.3.1 Conduction Mechanisms

For the devices and samples referenced in this study, the primary material utilised for the insulating layer is silicon dioxide, SiO_x. It has been reported that silicon dioxide has been doped with conducting ions, such as silver or copper, during the fabrication process in order to behave like ECM cells.

However, diffusion of metallic ions is generally undesirable in CMOS processing, as it can compromise the operations of neighbouring electronics. The present study is concerned with the intrinsic resistive switching property, irrespective of the electrode materials. Given that silicon-rich silica is predominantly employed in the insulating layer, its capacity for complete

CMOS-compatible processing is deemed to be highly favourable.

Fig. 3.20 : Conductive regions for filamentary switching (left) and interface switching (right).

In the context of bulk silicon oxide, the formation of a conductive filament within the insulating layer typically occurs during the electroforming process. This conductive filament is generally independent of electrode size, with the switching mechanism being dominated by a single filament. The switching process instigates a minor alteration to the filament. It has been established that this is independent of the insulating layer thickness. This is due to the fact that changes in resistance usually take place in a localised region.

The surface switching mechanism is comparatively under-researched in comparison to filamentary switching. The conductivity of this mechanism is found to be predominantly contingent upon the dimensions of the electrode. The primary driving force behind this mechanism is the formation of a Schottky tunnel barrier across the entire electrode interface and the insulating layer. Consequently, a switching layer is formed at the interface.

Electrical conductivity is defined as an intrinsic property that determines the extent to which a given material can oppose a flow of charge. The ideal insulator is characterised by a complete absence of conductance and infinite resistance. It is evident that the conductivity of the silicon oxide thin film, which is measured in several hundred nanometres, is suboptimal due to the presence of a finite amount of conductance. The conductivity of the semiconductor material is also contingent on a variety of external conditions. The aforementioned parameters encompass specific frequencies of light, temperature dependency and applied electric field.

$$E = \frac{V}{d} \tag{3.1}$$

In (3.1), the electric field strength, E, can be expressed as a function of the applied voltage, V, and the distance that the voltage is being applied across, d. Nevertheless, it is important to note that this fundamental estimate may not be applicable to the actual devices. The validity of assumptions regarding negligible oxide charges, voltage flat-band and small band bending

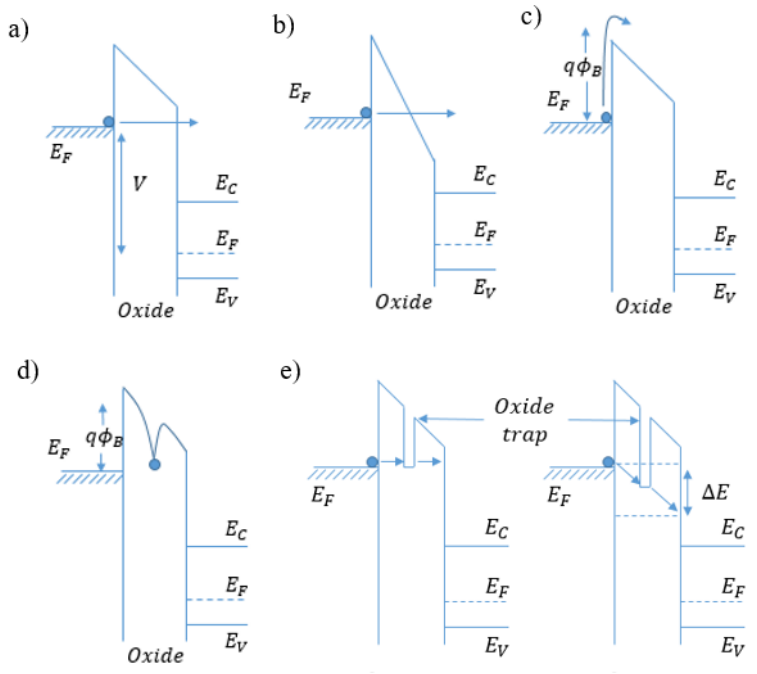

Fig. 3.21 Energy-band diagrams showing different conduction mechanisms: (a) direct tunnelling, (b) Fowler-Nordheim tunnelling, (c) thermionic emission, (d) Poole-Frenkel emission, (e) trap-assisted tunnelling [240]

may be called into question. In order to facilitate a more thorough analysis and identification of the switching procedure in the samples, additional conduction mechanisms in the insulator are considered.

$$J \propto E^2 exp\left[-\frac{4\sqrt{2m^*}(q\phi_B)^{3/2}}{3q\hbar E_i}\right] \tag{3.2}$$

$$J \propto V^2 exp\left(-\frac{b}{V}\right) \tag{3.3}$$

(3.2) displays the tunnelling current density's dependence on the electric field and voltage, applied as appropriate, while being independent of temperature. In the context of the aforementioned equation, ϕ_B denotes the tunnelling barrier height, E represents the insulator electric field, m^* is defined as 0.42m, which is the carrier effective mass for silicon oxide, $\hbar$ is the reduced Planck constant, q is the electric charge, and b is a constant of proportionality.

In the presence of a strong electric field, conventional tunnelling is the predominant conduction mode for insulating materials. The tunnelling process is a consequence of quantum mechanical effects, with the electron wave function having a finite probability of penetrat-

ing through a potential barrier of finite height. Conventional quantum tunnelling refers to the direct passage of an electron through the entire width of a barrier. Alternatively, Fowler-Nordheim tunnelling refers to the electron tunnelling through only part of this height.

$$J \propto E \cdot exp\left(-\frac{\Delta E_a}{kT}\right) \tag{3.4}$$

$$J \propto \frac{V}{T} \cdot exp\left(-\frac{c}{T}\right) \tag{3.5}$$

(3.4) illustrates the ohmic current density as a function of electric field, as well as the voltage applied and the temperature. In this equation, ΔE_a denotes the activation energy, k is the Boltzmann constant, T is the temperature in Kelvin and c is a constant of proportionality. In the context of low fields and elevated temperatures, ohmic conduction exerts a predominant influence. This phenomenon entails the thermally induced excitation of carriers, thereby facilitating their transition between conductive states.

$$J \propto \frac{E}{T} \cdot exp\left(-\frac{\Delta E_a}{kT}\right) \tag{3.6}$$

$$J \propto \frac{V}{T} \cdot exp\left(-\frac{d}{T}\right) \tag{3.7}$$

Ionic conduction exhibits a comparable expression to ohmic conduction, yet it possesses a distinct activation energy and constant of proportionality. This process is typically characterised by the movement of ions across a material via defects in the crystal lattice of a solid. For an ideal insulator, ions cannot readily travel into and out of the material.

However, an applied electric field will result in the build-up of ionic carriers at the metal-to-insulator interfaces, thereby modifying the voltage distribution across the region. The elimination of the applied electric field will result in the retention of a significant internal field, thereby enabling the flow of an ionic current until equilibrium is achieved.

$$J = \frac{9\varepsilon_i \mu V^2}{8d^3} \tag{3.8}$$

$$J \propto V^2 \tag{3.9}$$

The phenomenon of space charge can be attributed to the injection of charge from the electrodes into the insulator, in the absence of compensating charges. The process involves the injection of charges into the dielectric from one electrode and their subsequent capture by the other. The Mott–Gurney law is delineated in (3.8) for space charge limited current in solid

and in the velocity-saturation regime accordingly.

In this equation, ε denotes the dielectric permittivity, μ is the carrier mobility, L is the material thickness, and $v = \mu E$ is the electron drift velocity. This conduction mechanism is predicated on the presence of a single type of charge carrier, the absence of intrinsic conductivity, and an electric field of zero magnitude at the cathode responsible for the injection of charge.

Further exploration will be directed towards other conduction mechanisms, including Schottky emission and Poole-Frenkel conduction, which will be examined in greater detail. Schottky emission, otherwise known as thermionic emission, occurs when the carriers receive thermal energy in excess of the potential barrier height. The phenomenon of Poole-Frenkel conduction occurs when trapped electrons are thermally excited into the conduction band.

3.3.1.1 Fowler-Nordheim Tunnelling

In the presence of elevated electric fields, quantum mechanical tunnelling emerges as the predominant conduction mechanism in insulating materials. This is a consequence of the process inherent in quantum mechanics, whereby the electron wave function is capable of penetrating a potential barrier.

This process is typically contingent on the electric field, irrespective of temperature. The phenomenon of direct tunnelling occurs when carriers traverse the entire width of the barrier. It has been established that, in the context of Fowler-Nordheim tunnelling, carriers only traverse a proportion of this width.

$$J = \frac{q^2 E^2}{8\pi\hbar\phi_B} exp\left[-\frac{4\sqrt{2m^*}(q\phi_B)^{3/2}}{3q\hbar E_i}\right] \tag{3.10}$$

$$J \propto \frac{4\pi q m^* kT}{\hbar^3} \tag{3.11}$$

The phenomenon of Fowler-Nordheim tunnelling is contingent upon the trapezoidal configuration of the potential barrier. In the presence of a substantial application of an electric field, an increased incidence of band-bending is observed. This results in a significant reduction in the effective width required for carriers to tunnel through. In the context of a thick oxide layer, this is the prevailing conduction mechanism for a metal oxide structure. Subsequent to the tunnelling process, the carriers are able to move freely between the conduction and

valence bands.

The identification of the mechanism for the device is possible through the rearrangement of (3.10) and the graphical representation of the Fowler-Nordheim plot of $ln(J/E^2)$ against $\frac{1}{E}$ for experimental I-V characterisations. The gradient of this straight-line plot is equivalent to $-\frac{4\sqrt{2m^*}(q\phi_B)^{3/2}}{3q\hbar}$, which can be rearranged to obtain the barrier height ϕ_B. The y-intercept, on the other hand, describes the geometrical efficiency of electron-field emission. The occurrence of this mechanism is contingent upon the product of the electric field and layer thickness exceeding the barrier height.

3.3.1.2 Poole-Frenkel Conduction Hopping

Conduction may also occur in the absence of quantum tunnelling through the insulator. In the context of materials characterised by a high density of structural defects, the movement of carriers is constrained in a manner that is distinct from the behaviour exhibited by tunnelling mechanisms. The presence of these structural defects also gives rise to the appearance of additional energy states, also known as traps, in the vicinity of the energy band edges. The function of these traps is to restrict the flow of current, and they achieve this by means of a capture and release process.

The Poole-Frenkel conduction mechanism is concerned with electrons trapped in these states. These trapped electrons can eventually amass sufficient energy via thermal fluctuations in the material to escape from the localized trap states. It is imperative to note that, in the absence of being captured in an alternative trap state, the electrons can ultimately reach the conduction band. It is evident that this mechanism is contingent on two factors: the application of an electric field and the presence of thermal energy. The electron derives its total energy from two sources: the electric field and thermal fluctuations.

$$J \propto E_i \cdot exp\left[-\frac{q(\phi_B - \sqrt{qE_i/\pi\varepsilon_i})}{kT}\right] \tag{3.12}$$

$$J \propto V \cdot exp\left[\frac{q}{kT}\left(2a\sqrt{V} - \phi_B\right)\right] \tag{3.13}$$

This conduction mechanism is typically driven by electron drift current, $J = qn\mu E$, where q is the electric charge, n is the carrier density, μ is the carrier mobility and E is the electric field. This current may be expanded into (3.12) with dependence on the trap depth ϕ_B, the permittivity of the insulator ε and the temperature T. A non-ideality factor m, varying

between 1 and 2, may be introduced to the equation to account for the fabrication process and the semiconductor materials used.

3.3.1.3 Thermionic Emission

The concept of thermionic emission can be explained through the utilisation of the Schottky diode as a theoretical model. In the majority of cases, Schottky diodes are constructed using a metal-to-insulator junction as opposed to a P-N semiconductor junction. This configuration frequently enables a low forward voltage drop and a rapid switching action. It is imperative to ensure a pristine surface for the purpose of facilitating intimate contact between the metal and the semiconductor surface during the fabrication process.

$$J = A^{**}T^2 exp\left[-\frac{q(\phi_B - \sqrt{qE_i/4\pi\varepsilon_i})}{kT}\right] \tag{3.14}$$

$$J \propto exp\left[\frac{q}{Kt}\left(a\sqrt{V} - \phi_B\right)\right] \tag{3.15}$$

(3.14) denotes the fundamental relationships underlying the thermally induced current. In the context of electrical engineering, the effective Richardson constant, denoted by A^{**}, is a critical metric that quantifies the electrical properties of a material. The insulator permittivity, represented by ε, and a constant of proportionality, denoted by a, are key inputs in the calculation of A^{**}.

This current is attributable to the thermally excited flow of charge carriers from a surface, which can be electrons or ions, over a potential barrier. It is imperative that the thermal energy of the carriers exceeds the material work function. The magnitude of the current density is found to depend quadratically on the temperature.

In the event of contact between a semiconductor and a metal surface, a Schottky barrier is produced. The metal functions as the anode, with the n-type semiconductor acting as the cathode. In the context of thermal equilibrium, the net flow of electrons is sustained until the two Fermi levels are equal. The phenomenon of electron flow gives rise to a depleted region on the interface. This depletion region is primarily composed of positive ions, which is a characteristic of n-type semiconductor material.

A layer of negative space charge is built up on the metal interface in order to maintain charge neutrality. This results in the formation of a potential barrier, with electrons migrating from the semiconductor interface to the metal interface. The Schottky diode typically functions

with a small forward bias of approximately 0.2V, while its reverse breakdown voltage is approximately 2V.

In the forward bias configuration, the semiconductor experiences a decline in its electrical potential relative to the metal, thereby reducing the interface barrier and facilitating enhanced electron flow through thermionic emission. Conversely, when the diode is under reverse bias, the potential barrier will increase, causing a very small number of thermal electrons to tunnel through the barrier. This effect persists until the reverse breakdown voltage is attained.

In many cases, it is necessary to establish a non-rectifying ohmic contact to facilitate the flow of current into the semiconductor interface, thereby enabling the carriers to move unimpeded across the junction in any direction. This objective can be realised through the process of quantum mechanical tunnelling across a potential barrier. The magnitude of this effect is modulated by the width of the depletion layer, which can be reduced by increasing the dopant concentration on the semiconductor.

It has been shown that, at elevated dopant concentrations, a substantial number of carriers can be permitted to flow, thereby establishing an ohmic interface. It is important to note that this doping concentration cannot be achieved by conventional means. Alternatively, a thin layer of metal can be evaporated on the semiconductor interface, allowing diffusion to take place and heavily dope the semiconductor material.

3.3.1.4 Trap Assisted Tunnelling

As outlined in preceding sections, the tunnelling process under discussion is predicated on a one-step tunnelling process. Furthermore, the presence of defects within the insulating layer can facilitate two or more tunnelling steps. The presence of structural defects or traps has been observed to occur during the fabrication process or as a result of exposure to high levels of electrical stress. The presence of traps has been demonstrated to result in the division of the energy barrier into multiple paths. This phenomenon is known to increase the probability that carriers will sequentially tunnel through thinner barriers.

$$J \propto exp\left[-\frac{8\pi\sqrt{2qm^*}}{3\hbar E}\phi_t^{\frac{3}{2}}\right] \tag{3.16}$$

The process of trap-assisted tunnelling can be categorised into two distinct classifications: the elastic process and the inelastic process. These processes may occur with or without loss of energy in the carriers. In the context of materials characterised by a high density of

structural defects, the probability of conduction is observed to be amplified in the presence of multiple oxide traps during the tunnelling process. A plethora of theoretical models have been advanced to explain the process of trap-assisted tunnelling. A simplified expression relating current density with trap barrier height ϕ_t is given in (3.16).

3.3.2 Switching Model Analysis

An analysis of the conduction mechanisms was performed on the results obtained from both unipolar and bipolar switching modes. The application of the equations obtained in the preceding sections facilitated the completion of appropriate curve fittings, which were utilised to assess the presence of the various conducting mechanisms within the samples. Employing the relationships previously delineated, it is feasible to extract the relative dielectric constant, ε_r, for Poole-Frenkel and thermionic emission, as well as the trap barrier height, ϕ_t, for trap-assisted tunnelling.

Furthermore, the graphical representation of Fowler-Nordheim plots is a possibility, albeit with significantly inferior fittings. The following section presents the fitting results for both unipolar and bipolar switching across all conducting mechanisms. The forward fit is representative of HRS in the Set case and LRS in the Reset case. Conversely, the reverse fit represents LRS in the Set case and HRS in the Reset case.

Table 3.1 Curve fitting results for the conduction mechanism analysis.

Conduction Mechanism	Process	Unipolar Device		Bipolar Device	
		HRS	LRS	HRS	LRS
Poole-Frenkel Hopping	Setting	$\varepsilon_r = 15$	$\varepsilon_r = 11$	$\varepsilon_r = 12$	$\varepsilon_r = 9.1$
	Resetting	$\varepsilon_r = 13$	$\varepsilon_r = 10$	$\varepsilon_r = 7.8$	$\varepsilon_r = 5.8$
Thermionic Emission	Setting	$\varepsilon_r = 1.4$	$\varepsilon_r = 0.9$	$\varepsilon_r = 1.5$	$\varepsilon_r = 1.0$
	Resetting	$\varepsilon_r = 1.3$	$\varepsilon_r = 0.2$	$\varepsilon_r = 1.9$	$\varepsilon_r = 1.5$
Trap-Assisted Tunnelling	Setting	$\theta_t = 0.26eV$	$\theta_t = 0.15eV$	$\theta_t = 0.21eV$	$\theta_t = 0.09eV$
	Resetting	$\theta_t = 0.41eV$	$\theta_t = 0.21eV$	$\theta_t = 0.14eV$	$\theta_t = 0.10eV$

The relative dielectric constants obtained for Poole-Frenkel conduction are higher than the theoretical values for silicon dioxide ($\varepsilon_r = 4$) and comparable to those for pure silicon ($\varepsilon_r = 12$). The findings indicate that the Poole–Frenkel mechanism may be feasible with the samples, despite encountering certain challenges. Thermionic emission has been demonstrated to align the relative dielectric constants with theoretical values.

The findings of this study demonstrate that the performance of trap-assisted tunnelling is optimal for both unipolar and bipolar devices, particularly at higher applied electric fields. It can be posited that the application of an electric field facilitates the tunnelling process of the carriers through the silicon oxide layer, which may be considered to be of greater thickness, or alternatively, that it raises the height of the trap barrier. The obtained fitting values are displayed in the following table.

The findings indicate the potential for all three conduction mechanisms to occur within the samples, in addition to ohmic conduction. These conduction methods may manifest in isolation or in conjunction with one another. It appears that the transportation of current is predominantly facilitated by means of trap-assisted tunnelling in the devices that have been tested thus far. The subsequent section will provide a synopsis of the conduction mechanism and switching model.

The material of particular interest in the samples is silicon-rich silica, which contains a high concentration of oxygen vacancies. This material is known to readily separate into silicon dioxide and silicon. The application of an electric field to the material has also been demonstrated to facilitate the segregation of silicon and oxygen ions. This process is further exacerbated by structural defects in the material, particularly in structures grown via sputtering.

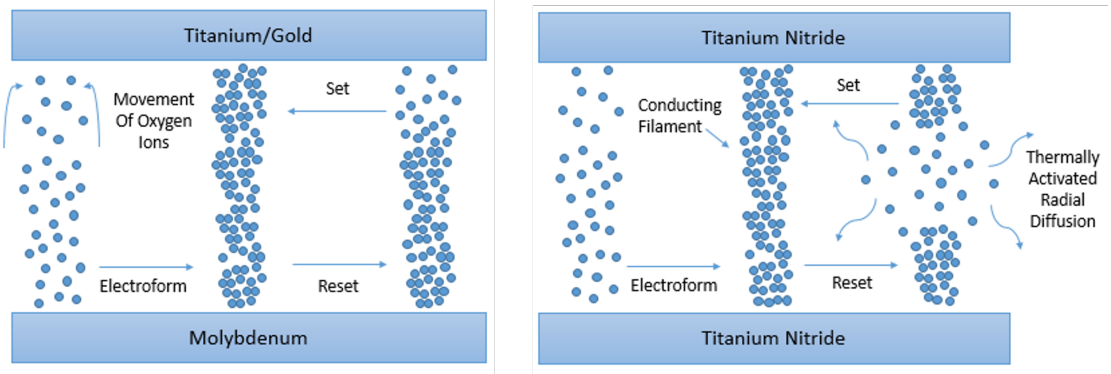

Fig. 3.22 Schematic of bipolar switching in asymmetric devices (left) and of unipolar switching in symmetric devices (right).

The phenomenon of bipolar switching in asymmetric devices can be described in qualitative terms as the movement of oxygen through a thin silicon oxide film (Figure 3.22). This phenomenon can be attributed to the deformation occurring in proximity to the electrode surface, which facilitates the injection of electrons and subsequent release of oxygen molecules. In the

absence of any external damage, the silicon oxide of the device contains oxygen vacancies, i.e. structural defects.

The defects under scrutiny were formed during the fabrication process and were distributed evenly throughout the amorphous structure. It has been established that, under ambient temperature conditions, the diffusion of these defects occurs at a relatively slow rate. This is attributed to the presence of a substantial diffusion barrier.

When a voltage bias is applied to the material, these structural defects can rearrange into different configuration, while additional defects are also generated. The diffusion barrier in silicon oxide is significantly reduced with the new configuration. This enables negative oxygen ions to travel faster towards the positive electrode.

These oxygen anions are eventually trapped at the metal-oxide interface. The anions then discharge and form oxygen molecules O_2. Under extreme conditions, enough oxygen molecules built-up at the interface can cause bubbles of oxygen to form and eventually burst at the electrode surface, in the form of super oxide.

In operating conditions, the movement of oxygen ions within the oxide facilitates the electrical properties of the device. When an electric field below electric breakdown is applied to the device, the configuration changes can be conceptualised as the electroforming process for RRAM. The alterations in the distribution of oxygen vacancies are terminated once an abrupt conductance change has occurred, thereby forming a conductive filament in the silicon oxide film.

It is hypothesised that trap-assisted tunnelling is the predominant mode of carrier transport for bipolar switching. This finding indicates that the conductive filament within the material is not continuous, but rather consists of a series of neighbouring oxygen vacancies. The lower barrier heights observed in the preceding section indicate the occurrence of electron conduction through oxygen vacancy defects. Once the conductive filament has been formed under controlled breakdown conditions, the subsequent diffusion of oxygen vacancies is known to control the switching mechanism in bipolar mode, possibly via a local redox process occurring at the metal-oxide interface.

The process of bipolar switching is primarily governed by the movement of oxygen ions in response to an externally applied electric field. Conduction in unipolar switching mode is

likely to share some similarity with bipolar switching. A significant distinction pertains to the reset process in these devices.

The conductive filament in unipolar switching samples can be conceptualised as exhibiting slightly greater continuity, thereby facilitating enhanced ohmic conduction. Consequently, a thermal effect is associated with the current passing through the filament during the switching process. The occurrence of a critical current threshold may result in the abrupt rupture of the conducting filament via thermally activated diffusion or Joule heating, thereby effecting a reset of the device.

In summary, it has been observed that resistive switching in metal oxides may be attributed to the presence of oxygen vacancies in the conductive filaments of silicon oxide. The role of the applied electric field is also of significance in the switching process, particularly in the context of bipolar switching, and in conjunction with Joule heating in the case of unipolar switching.

In the set process, oxygen anions are known to drift towards the electrode. This process results in the formation of positive oxygen vacancies, which facilitate the conduction of electrons. In the reset process, the oxygen ions residing near the metal-insulator interface are displaced into the vacancies. Alternatively, a sufficiently high local Joule heating can also overcome the vacancies binding energies. In either scenario, the conductive filament is ruptured and the device is reset.

3.4 Summary

The primary insulating layer material for the devices and samples examined in this study is silicon dioxide (SiO_x) [182], in conjunction with electrode materials such as silver (Ag) or copper (Cu). Silicon-rich silica, which is predominantly utilised in the insulating layer, exhibits considerable promise for fully CMOS-compatible processing. A silicon oxide electron injection model has been developed to facilitate a more profound comprehension of the characteristics of silicon oxide [71].

The amorphous silicon oxide structure exhibits $O-Si-O$ bonds, with a subset of these bonds featuring broad angle bonds, which possess the capacity to function as deep electron traps, with the capability to capture two electrons. The $Si-O$ bond is subsequently weakened once the broad bonds have collected both electrons. This reduction in energy requirement is

attributed to the minimisation of the force required to break the connection and create the Frenkel defect. The formation of these defects gives rise to a population of vacancies, which can then contribute to the formation of the conductive filament.

Following an investigation into the underlying physics of the materials and mechanisms that comprise new RRAM devices, it is advantageous to analyse general device behaviours in terms of overall $I-V$ characteristics, volatility, polarity dependence, and power consumption, with a view to enhancing application design. This analysis would facilitate improved application design. RRAM devices have been shown to respond in three distinct ways [153].

The classification of these as unipolar is predicated on the premise that they are set and reset with the same polarity. The term "bipolar" is employed to describe a device that has been successfully set and reset with opposing polarities. Finally, threshold switching occurs when a device transitions to its low resistance state within a certain voltage range.

Additionally, the term 'nonpolar' is employed to denote devices that are characterised by polarity independence, thereby enabling them to execute both unipolar and bipolar operations. As illustrated in Figure 3.23, these three groups can be summarised as follows.

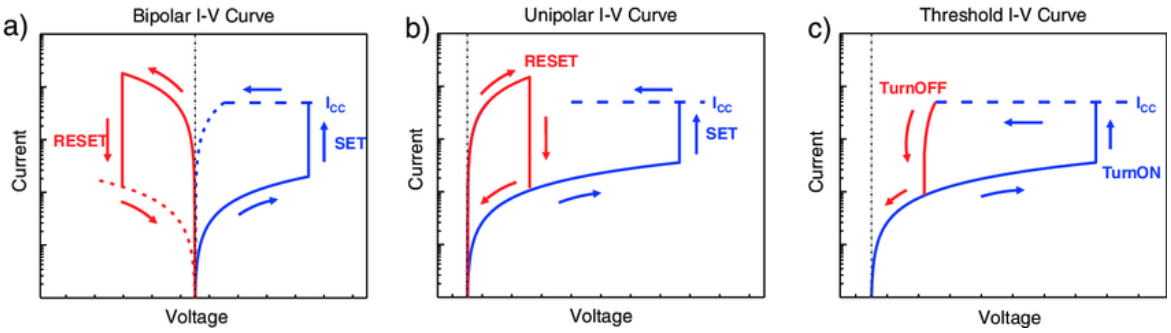

Fig. 3.23 Schematic I-V curves in: a) non-volatile bipolar memory switching mode, b) non-volatile unipolar memory switching mode; and c) volatile threshold switching mode. In the SET process, a compliance current is required to avoid hard breakdown [153].

Bipolar switching involves setting and resetting devices in opposing polarities. This behaviour has been observed in a variety of oxide materials [273]. The switching method requires current compliance during setup, but it can be reset with the same applied current compliance.

In comparison to the unipolar switching strategy, this results in a significantly simpler compliance system. The magnitude of the switching voltage for bipolar devices is generally

within the limits for digital chip integration. For example, the operational voltages can be sub $\pm 1V$ in both polarities [188].

Unipolar switching, like bipolar switching, has been demonstrated in a variety of oxide materials [108]. Under unipolar operation, electroforming and setting are field-driven processes, whereas resetting is typically a current-driven one. To eliminate rivalry between the two working systems, a current compliance must be imposed during the set process, as resetting is triggered by excessive currents. Once the device has been set, the current compliance can be increased or removed completely to allow a sufficient amount of current to flow and the device to reset.

The capacity of the unipolar device to function with a single supply rail allows it to be combined with digital integrated circuits. A rather sophisticated current compliance system, on the other hand, remains one of the main downsides of such devices, with the high reset current causing heating and power consumption difficulties, particularly when scaling big memory arrays [295].

Threshold switching occurs when a device flips to a low resistance state when the applied voltage reaches a threshold and then instantly resets when the voltage falls below a lower threshold, demonstrating hysteresis. [2]. This phenomenon was proposed as a result of heat dissipation after evaluating the varied thicknesses of the bottom electrode [29]. This implies that changing the size of the electrode during the fabrication process can result in different intended threshold switching behaviors.

Aside from the switching processes, several aspects must be addressed when developing RRAM. To start, OxRAM samples usually have a lower on/off conductance ratio in the range of 10s–100s and offer good retention of up to 10^{12} cycles[104], while the CBRAM on/off conductance ratio can be fairly high in the range of 10^3–10^6, but has limited endurance to less than 10^4 cycles [4].

The unpredictability of switching parameters resulting from the stochastic behaviour of oxygen or metal ions during ionic migration, as well as the variability in filament form from device to device and cycle to cycle within a single device, provide a significant difficulty for the design of RRAM cells [5]. As a result of these variations, the density of RRAM prototypes ready for commercialization is quite low at $\sim 4Mb$.

Recent research has made tremendous progress in demonstrating the potential of high-density devices with remarkable attributes and low power consumption of $\sim$0.1 pJ [289], better reliability with endurance of more than $> 10^{12}$ cycles and retention of > 10 years at $150^{\circ}C$ [97], higher density of 32 Gb via a simpler fabrication steps and stronger thermal characteristics [27]. This, together with the significant resistance variation not just between different devices but also inside and between programming cycles on the same device [193], has resulted in concerns with the repeatability of their electrical properties. These issues, taken together, have prevented RRAM from being commercialised, despite its many appealing characteristics.

Fortunately, the main deep learning-based neural processing techniques—such as regression, pattern recognition, and speech recognition—are random in nature and need less precise computation than deterministic traditional computing. As a result, compared to memory applications, variation in memory devices during manufacture has less of an influence on computation outcomes [166]. The more tolerated necessity for variation in neural applications, combined with recent technical advancement, has contributed to generate additional possibilities for RRAM devices to become suitable candidates for neural technologies.

Chapter 4

Current Transients in Memristive Devices

4.1 The Subthreshold Regime

The initial primary focus of this thesis is on the characterisation of silicon-based memristors, which are the fundamental components of memristive systems. Since the discovery of the memristor and its importance to replicating synaptic activity had such a profound impact on the field of neuromorphic engineering, investigating additional nanoelectronic components and behaviours in this context will lead to new neuromorphic computing applications.

This chapter investigates a phenomenon known as the "current transient" that has yet to be deliberately applied to the demand of neuromorphic computing. The current transient phenomenon can be similarly represented by the current flowing through a defective capacitor in response to a step potential to produce rich dynamics, both growing and decreasing in conductance, and can be beneficial in a computational device.

This chapter begins by documenting and characterising the current transients based on available literature. The experimental procedures used throughout the chapter were then described, and strategies were developed to aid in the characterisation of current transients. This provides a deeper understanding of the physical models underpinning the transients, allowing for the further development of an integrative memristive system based on silicon oxide samples that are already available.

4.1.1 Fundamental Properties

Fundamentally, the processes of capacitive decay and dielectric relaxation are ideal to define a capacitor's response to a step voltage. Applying a constant voltage across its terminals causes

the device current to decline until it ultimately comes to rest at a constant leakage current. This, however, is not always the case. When the voltage is applied for an extended period of time or at a high enough temperature, the current flowing through the device begins to grow as the oxide gets faulty and its resistance falls. This is known as oxide deterioration, an umbrella word for an oxide coating that becomes faulty over time as a result of environmental stress factors [78].

One specific form of this resistance degradation addressed here is the "current transient". This is distinguished by their characteristic form when plotted on a current-time graph as illustrated in Figure 4.1. When a voltage is applied to the device, the current grows swiftly to a maximum and then gradually decays, resulting in a distinctive peak in device current. Surprisingly, whereas oxide degradation is essentially a permanent impact, the change in oxide resistance that happens during a current transient is not; rather, it is volatile.

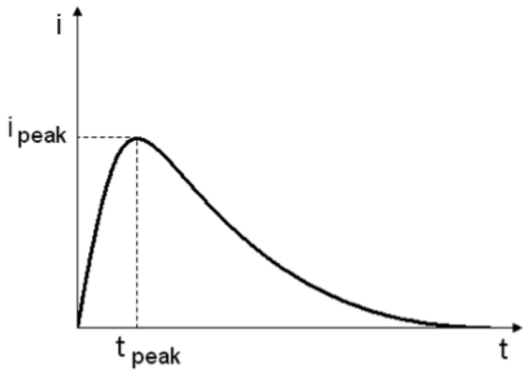

Fig. 4.1 Current transient caused by a change in temperature for a closed circuit [296], the plot is for illustration purpose only.

The most common differentiating feature is the combination of their distinct peak and their extended time periods during which they occur. Previously published papers documented peak periods ranging from 5 to 20 seconds, and the full transient is frequently seen for 10s to 100s of seconds [214]. They are both slow pace and long-lasting occurrences. Timescales of this magnitude are not usual for integrated resistor-capacitor (RC) circuits because they demand enormous capacitance, which necessitates physically much bigger capacitors [55].

One appealing aspect of the phenomena is the much longer time periods of the current transient's responses in contrasted with a traditional RC circuit. However, the dynamics of the transient are not fixed. They are known to be accelerated by increases in applied temperature [168], voltage [297], and when irradiated with a laser [152]. All of these variables can cause

the transient to accelerate, with the peak coming sooner in time. When pushed far enough, the peak disappears completely and only a decreasing current was observed, which is most likely owing to the instruments' temporal resolution.

Given that the features of transients have been similar across published research, it is worth noting that these studies have used a variety of oxide materials. However, despite being present in a variety of device setup, the current transient is not equally observed in every instance, and this is especially noticeable with regard to device polarity. Early investigations on current transients only detected transients in a single voltage polarity [167], however other devices in the literature now produce transients in both polarities [33].

Transients have been reported in a wide range of materials, implying that they are a generic phenomena that might occur in a wide range of MIM devices under the correct conditions. The current transient's characteristics deviate from the typical resistance switching behaviours reported in silicon oxide devices. They are analogous and volatile as compared to devices that display non-volatile binary switching when run in a normal resistance switching mode. Understanding the source and features of current transients is significant because they may have ramifications for a variety of research areas studying metal-oxide structures, such as transistor gates or the comparatively emerging subject of resistive switching memory.

4.1.2 Current Models

The most well-known model for explaining this behaviour is the Space Charge Limited Current (SCLC) hypothesis, which was created for describing a general space charge within a perfect insulator [172]. A variety of theories have been offered up to describe this phenomenon [140]. Since charged oxygen vacancies are thought to constitute the space charge in question, this initial model was first associated with them.

Along with the electronic current, an ionic current also flows as a result of the drift of charged defects under an applied bias. The electrical contact at the other end blocks the charged defects, causing them to build up. The Coulombic repulsion that results from this buildup prevents further migration, which in turn reduces the ionic current as it starts to oppose itself. Because of this, earlier research assumed that the transient behaviours was caused by a general space charge inside of an insulator without any trapping in order to develop the SCLC model.

The original study only considered in one dimension and made the assumption that the conduction is planar. This might represent an ionic current or an electronic current, depending on whether the charges and mobilities correspond to electrons or some other charged defect inside the oxide. Recent study has assumed that the latter is the case. The mobility is frequently thought to refer to moving oxygen vacancies inside the oxide [297], or in other instances, to moving ions [33]; in all cases, it is supposed that the current is ionic.

Originally, the current of the model is defined as:

$$j(t) = j_{cond} + j_{disp} \tag{4.1}$$

$$j_{cond}(x,t) = qn(x,t)\mu E(x,t) \tag{4.2}$$

$$j_{disp}(x,t) = \varepsilon \frac{\partial E(x,t)}{\partial t} \tag{4.3}$$

$$qn(x,t) = \varepsilon \frac{\partial E(x,t)}{\partial x} \tag{4.4}$$

where $j(t)$ is the total current, j_{cond} is the electronic conduction current, j_{disp} is the displacement current, q is the the charge of a single electron or the respective ion measured in Coulombs, n is the space charge concentration, μ is the oxide space charge mobility, E is the oxide local electric field, L is the oxide thickness, ε is the oxide dielectric permittivity, x is the position within the oxide ranging from 0 to L, j is either the electronic or ionic current depending on the space charge, and lastly the Poisson's equation is denoted with (4.4).

$$j(t) = \varepsilon \left(\frac{\mu}{2} \frac{\partial E^2(x,t)}{\partial x} + \frac{\partial E(x,t)}{\partial x} \right) \tag{4.5}$$

$$j(t) = \frac{\varepsilon\mu}{2L} \left(E_a^2(t) - E_c^2(t) \right) \tag{4.6}$$

$$j(t) = \frac{\mu Q(t)}{2L} \left(2E_a(t) - \frac{Q(t)}{\varepsilon} \right) \tag{4.7}$$

By substituting into the first equation, the device current density description can be derived to be (4.5). Assuming the voltage applied is fixed, the boundary conditions for the electric fields can be defined as $E_a(t) = E(L,t)$ and $E_c(t) = E(0,t)$ for the anode and the cathode respectively. Integrating (4.5) with respect to x from the cathode to the anode yield equation 4.6 which describes the current as a function of the electric fields across the electrodes.

Note that the applied voltage was assumed to be non-varying with time, $\int_0^L \frac{\partial E(x,t)}{\partial t} dx = \frac{\partial}{\partial t} \int_0^L E dx = \frac{\partial V}{\partial t} = 0$, in a single dimension while the problem is intractable for spherical and cylindrical geometries [140]. This equation can further be rewritten with respect to the

insulator total charge (4.7), with the relationship between the cathode and anode electric fields given as $E_a(t) = E_c(t) + \frac{q(t)}{\varepsilon}$, where $Q(t)$ is the insulator total charge.

$$\frac{E}{E_a^2} = \frac{\mu}{2L}\partial t \tag{4.8}$$

$$E_a(t) = \frac{V}{L}\left(1 - \frac{\mu V}{2L^2}\right)^{-1} \tag{4.9}$$

After deriving an equation describing the device current, the subsequent step is to obtain the equation of the SCLC model that connects the mobility of the space charge with the time at which the peak τ appears. It is assumed that this is the instant when the charge front reaches the anode of the device. The derivation starts with the assumption that the field at the cathode is zero, $E_c(t) = 0$.

The next assumption is that the conduction component of the current is zero when evaluated at the anode of the device, $j_{cond} = 0$, as no charge has reached the anode yet, leaving $j(t) = \varepsilon\frac{\partial E(t)}{\partial t}$. This, combined with (4.6) and the prior assumption that the field is zero at the cathode, results in a differential equation having a solution as provided by equation 4.9.

$$L = \mu\int_0^\tau E_a(t)dt \tag{4.10}$$

$$L = -2L\left[ln\left(1 - \frac{\mu V t}{2L^2}\right)\right]_0^\tau \tag{4.11}$$

$$\tau = \frac{2L^2}{\mu V}\left[1 - exp\left(-\frac{1}{2}\right)\right] \cong 0.787\frac{L^2}{\mu V} \tag{4.12}$$

After the field at the anode has been determined, it is now possible to calculate the time it takes for the front of the space charge to reach the anode. Assuming that the space charge's velocity is given by $\mu E(x,t)$, and the field is constant between the charge front and the anode, it can be assumed that the field experienced by the front of the space charge is equal to the field at the anode, $E(x,t) = E_a(t)$.

The time it takes for the charge front to cross the device can be determined by solving the integral (4.10), to get an analytical and approximated solution (4.11). This analysis has arrived at a significant equation of the SCLC model which has been widely used in the literature on current transients, frequently to determine the mobility of the space charge. To determine the equation's validity, it is necessary that the original model of the device's space

charge traversing holds true.

The prevalence of the SCLC model can be attributed to 4.12 and the appealing material properties it implies. For instance, it indicates the possibility of acquiring the mobility of the migrating defects that, when paired with the Einstein-Nernst equation, can disclose the activation energy of such defects. This equation has been used by researchers to determine the activation energies and mobility of different materials observed in current transients.

According to literature, the changes in the transient that occur due to laser illumination are evidence of accelerated mobility [152] or charge transitions within the defect [143]. However, if researchers intend to use this model to derive mobilities and secondary observations based on changes in mobility, the model's validity must be ensured. This is where problems may arise.

4.1.3 Alternate Models

Naturally, SCLC is not the only model used to explain these transients. While SCLC argues that the transient is the outcome of a substantial ionic current, it could also be inferred that the ionic currents are insignificant, and the transient results from a variation in the electronic conductivity of the bulk and interfaces [190]. This conclusion was reached by simulating the redistribution of vacancies, electrons, and holes within the BST oxide layer bounded by two Schottky contacts at the $BST - Pt$ interfaces. The simulation was conducted using a finite difference method, accounting for the redistribution of each particle species that is influenced by drift and diffusion.

$$j_k = -D_k \frac{\mathrm{d}C_k}{\mathrm{d}x} + \frac{Z_k}{|Z_k|}\mu_k C_k E, \text{ where } k \in \{n, p, V_{\ddot{o}}\} \tag{4.13}$$

The simulation iteratively establishes the particle redistribution and subsequently computes the total device current, which is based on the velocity and distribution of the particles. The total device current is composed of the electron, hole, and vacancy currents. For each particle species, the current comprises a diffusion and drift term, as outlined by the Nernst-Einstein (4.13), where k indicates particle species which can be electrons, holes or charged oxygen vacancies, j_k is the current density, D_k is the diffusion coefficient, C_k is the particle concentration, Z_k is the charge in Coulombs, μ_k is the mobility, and E is the local electric field. This method is beneficial since it allows for separation of the electronic and ionic currents, which is not feasible in an experiment.

Within the simulation, a rapid and asymmetric redistribution of electrons and holes is observed, resulting in a highly n-type interface at the negative electrode and the other slightly depleted from its resting electron concentration. Based on published data, this rate slows at the onset of the transient's peak. Following this redistribution of electrons is the migration of oxygen vacancies, which is a much slower process. Vacancies also migrate towards the negative electrode alongside the electrons.

This strongly impacts the internal electric field, causing some bending of the conduction band. Due to the lower mobility of the vacancies and their slower migration, it is found that the corresponding ionic displacement current is too small to be the cause of the increase in device current. Instead, the increase is caused by the modulation in electronic conductivity due to the change in vacancy distribution.

By varying the temperature of the simulated device, it was determined that the activation energy for increased conductance is 0.8eV, which is deemed reasonable for vacancies within a BST device [303]. However, as the peak of the transient has already taken place by the time the vacancies have redistributed to a considerable extent, it is advisable to refrain from asserting that the activation energy inferred from the peak characterises the migration of vacancies.

Instead, the published results suggest that this could be attributed to the migration of electrons and holes, which predominantly occurs prior to and during the peak. Other researchers have also reached a comparable conclusion [303]. For instance, electronic traps were observed with apparent activation energies ranging from 0.18-0.3eV. It was concluded that oxygen vacancy migration results in alterations of electron/hole concentrations, ultimately regulating bulk conductivity.

With two conflicting theories in existence, certain studies aim to determine whether the effect is electronic or ionic (SCLC). For instance, transients in some devices were analysed by considering both the SCLC model and the modulation of electronic conductivity [259]. It was found that the model of electronic conductivity produces a more credible value for the dielectric constant, suggesting it is the superior model.

However, a different study [55] aimed to determine whether the effect is due to ionic or electronic currents by assuming that electronic traps would be influenced by the oxide's crystallinity. It was discovered that almost identical transients were observed in the amorphous

and polycrystalline devices, suggesting that the effect is of an ionic nature, which is also further supported by a previous claim [297].

However, in a subsequent publication [56], it was determined that an alternative theory, based on the existence of shallow electronic traps, also favoured an electronic explanation. It seems that the issue of whether this effect is ionic or electronic has not been definitively resolved. Nonetheless, the prevailing trend in the literature has been towards mostly electronic conduction, which is altered by the movement of vacancies.

Retrospectively, the SCLC model has been found to be helpful, but it has not yet been fully validated. It has been observed that higher temperatures and applied voltages cause an increase in the rate of the transient as predicted, and peak time plotted against reciprocal voltage has been found to be linear in certain regimes. However, the original space charge limited current (SCLC) model can enables us to provide more detailed scrutiny.

Take the thickness dependence of peak time, as stated in (2.12), as an example. This phenomenon results from vacancies traversing the oxide from one electrode to the other, and it should lead to the peak time being proportional to the square of the device thickness, L. Although no research has specifically investigated this issue, published data from a related study can be utilised to provide a preliminary assessment of this dependence.

From previously published data [167], it seems that the timing of the peak is not affected by the thickness of the device. The studies investigate three oxide thicknesses: 900Å, 450Å and 200Å. An increase in peak time is anticipated. In comparison to the 200Å device, the 450Å device should have a peak timing 6 times greater, while the 900Å device should have a peak timing 20 times greater. However, no such increase is observed, as the evidence demonstrates.

Instead, it appears that the peak's height is affected, not its timing. In thinner samples, the peak appears more spread out, whereas in thicker samples, the peak is more pronounced. This disparity between the critical SCLC theory equation and experimental evidence raises concerns when applying this model, though it is not the sole concern.

Even more concerning is the predicted straight line fit for peak time versus reciprocal voltage plots, which is often utilized as the primary indicator of SCLC behaviour (2.12). Of the papers that present a straight line, many of them only have a few data points claiming a straight line. Additionally, these plots are reliant on measuring peak times, which is not an

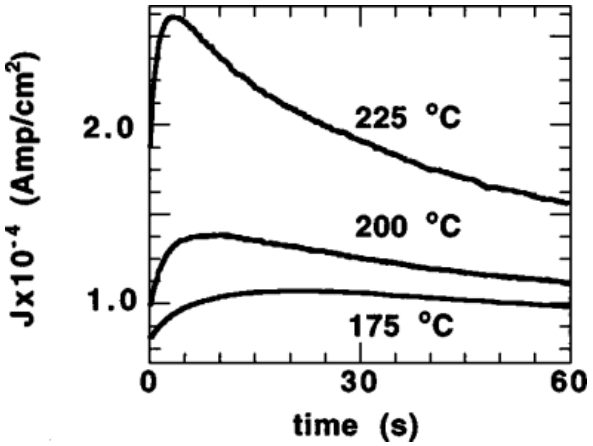

Fig. 4.2 Transients are plotted for BaSrTiO3 (BST) devices [18]. The peak at lower temperatures is significantly wider than at higher temperature.

easy process, as shown in Figure 2.2, where the shallowness of some curves obstructs the peaks.

This introduces a large degree of uncertainty for shallower transients, making the selection of peak times potentially subjective. What's more, some studies have shown that this linearity only occurs within a specific region while in others, it becomes exponential instead. Although these inconsistencies have not been studied thoroughly as yet, they do warrant a scepticism in applying the SCLC theory to the current transients.

The importance of verifying these models is exacerbated by their use in determining physical properties of devices. In some cases, these properties include mobility values and activation energies taken directly from the SCLC equations. However, further studies have delved deeper into this area. Naturally, these findings carry significant implications, particularly in the nascent field of defect engineering. The validity of the SCLC model applied in these contexts determines the certainty of these conclusions.

4.2 Current Transients Tuning

The device has a metal-insulator-metal structure as illustrated in Figure 4.3 and was originally developed for binary resistance switching applications [185]. The bottom metal contact is a molybdenum film of 280 nm thickness deposited via magnetron sputtering. The insulator layer is a slightly sub-stochiometric and amorphous layer of silicon oxide with a thickness of 35 nm, deposited via RF magnetron sputtering. The top metal contact is a 115 nm thick gold

film deposited via e-beam evaporation through a contact mask. The shape of the top contact is a square of $200 \times 200 \mu m$ defining the active area of the device. To improve adhesion of the gold contact, a 3 nm layer of titanium was deposited prior to the gold evaporation. This also serves as a gettering layer to seed the oxide layer with oxygen vacancies.

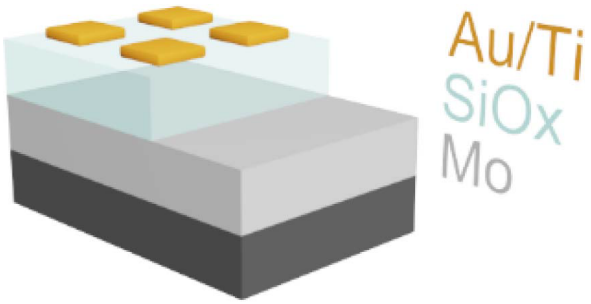

Fig. 4.3 Current Transient Device Structure. The device studied in this chapter has a metalinsulator-metal structure. The two electrical constacts are the molybdenum bottom contact and the gold top contact. A 3 nm titanium buffer layer is deposited prior to the gold, to improve the adhesion of the top contact and to serve as an oxygen getter.

4.2.1 Device Stressing

The method for reliably and repeatedly inducing the current transient phenomena is presented in this section. This not only enables the induction of current transients in a range of devices, but it also enables the progression of the transient from a barely detectable state to the dominant device activity. The approach described here has opened up the possibility of using the behaviour as a computational device while also enabling a more detailed characterisation of the behaviour.

Current transients were previously believed to be caused by oxide imperfections implying that if a device exhibits current transients, it only has to be designed as a faulty device. Fortunately, the fabrication and control of oxide defects is a current study area and is also known to as defect engineering. The act of applying sufficiently strong electric fields to a MIM device to cause a partial and reversible breakdown is known as electroforming. This is a common technique in memristor- and resistance-switching-memory (ReRAM)-based research, and it can also be simply seen in the silicon oxide devices utilised in this thesis.

This makes electroforming a good place to begin when trying to introduce defects. But it's not flawless. Device conductance often experiences huge discrete jumps while electroforming. When devices go through this process, they display a switching behaviour across two states

of resistance: one with high resistance and the other with low resistance. These discrete states are frequently highly stable. As a result, the qualities are significantly unlike from what are familiar known about the transient, which is in contrast analogous and volatile. This begs the issue of how electroforming can be considered an effective approach to create transients when the final device behaviour is so different.

Observing that the conductance of a device showing transients is more comparable to the HRS of an electroformed device than to the LRS can provide insight into the link between electroforming and a current transient device. This led to the idea that, if a slightly more subtle electroforming technique were utilised, that is, before the major switching event, the current transient may be created. Because the electroforming process itself seems to be the outcome of a positive feedback mechanism, it is difficult to achieve a delicate electroformation [134].

When a voltage flows to the device, charge is injected into the oxide, creating defects such vacancies that increase the device conductance. This procedure has inherent instability. Higher current densities brought on by the increase in conductance speed up the creation of defects, which in turn causes current densities to keep rising until a catastrophic breakdown happens. When high constant voltages (13V to 15V) are supplied to the device, it is easiest to see this exponential rise in conductance. A regulated, gradual, and delicate modulation of device conductance is difficult to perform when this positive feedback effect dictates the change in conductance.

It is obvious that electroforming's fundamental positive feedback must be avoided. This is accomplished by abandoning voltage-based electroforming in favour of constant current electroforming, often known as "stressing the device" to distinguish it from the more widely used electroforming. A comparable voltage is provided at the peak of continuous current straining as is done during electroforming.

The main distinction is that once the device begins breaking down, the voltage is reduced as a result of the current source's negative feedback. The time-dependent dielectric breakdown (TDDB) of oxides, which is caused by flaws in the insulator that increase conductance due to electron trapping, is frequently evaluated using the conventional approach of constant current stressing [137].

Instead of applying a set voltage, the device is provided with a constant current. This has the benefit of reducing the voltage across the device as it grows more conductive in order to maintain a constant current, which in turn delays the creation of oxide defects. This will go on once an equilibrium is established when the applied voltage is decreased to a level where formation no longer takes place, but is still great enough to keep the continuous current flowing. Now that the process presents negative feedback, it is considerably more conducive to causing subtle alterations in device conductance.

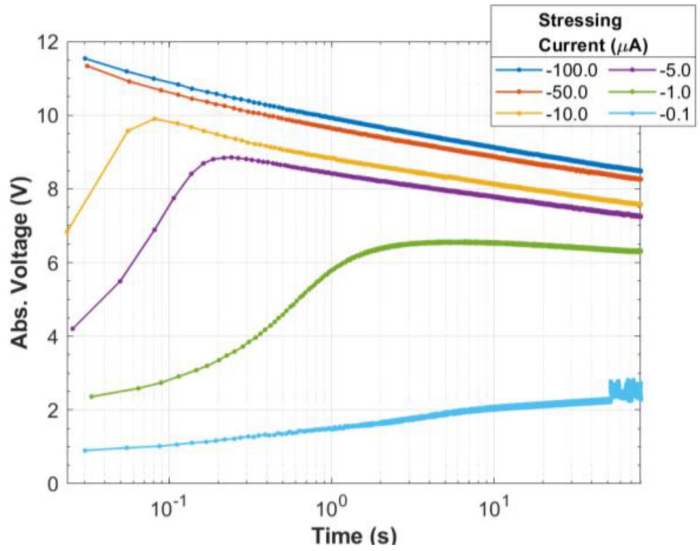

Fig. 4.4 The device's voltage is recorded while subject to constant current stress, with six devices tested at various current levels. The voltage across each device is plotted during the initial constant current stress for six devices, each with a different current magnitude.

Essentially, the devices are subjected to stress by grounding the molybdenum contact and applying negative currents to the gold contact, resulting in electron injection at the gold contact and hole injection at the molybdenum contact. The magnitude of the constant current is adjusted based on the desired behaviour, with values ranging from $-0.1\mu A$ to $-100\mu A$ for a duration of 100 seconds. During stressing, the voltage induced across the device gradually decreases, eventually reaching a steady state value as shown in Figure 4.4.

It's important to take into account that after being stressed, the device doesn't immediately display the current transient. The relaxation process can be accelerated by providing a positive potential to the above gold-titanium contact, but it must still be allowed to rest for 24 hours. When allowed to rest, the device will eventually show current transients. This delayed response is not unexpected as a result of a drifting defect with restricted mobility. While

the generated defects are still present, the relaxation time allows this to return to its resting distribution at thermal equilibrium. It is expected that the stressing results in significant drifting and defect generation.

4.2.2 Induced Transient

The established method enables the induction of current transients in devices, and it may be done so gradually. The prominence of the current transient can be altered when the devices are subjected to varying levels of stress by varying the current's magnitude. To be more precise, a device's current peak becomes more evident the more stressed it is. It was previously claimed that this was caused by the thinner devices' higher current densities masking the current transient [56].

Instead of looking at current densities, a possible reason of the prominence of current transients may be based on the total number of defects present in the oxide. It follows that thicker devices will have more defects overall if it is assumed that the concentration of defects in the oxide layer remains relatively constant throughout growth. In this instance, it is also assumed that more oxide defects will be produced the more aggressively the device is stressed. Therefore, it is possible that the total number of defects present in the oxide, rather than the conductance of the device, can be used to predict the transient's amplitude.

A device being stressed does not ensure that there will be current transients. As soon as electrons are introduced at the top gold-titanium contact, the device only displays current transients when it is stressed with negative currents. The response is different when electrons are injected at the molybdenum contact. The structural alterations that stressing causes in the device can be used as evidence for the reason only negative currents can cause current transients. As anticipated, the stressing process results in structural flaws at the electrical contact, as has been seen in the past when comparable devices were electroformed and switched repeatedly [271, 272].

Negative currents are used to stress devices, which involves grounding the molybdenum contact while applying negative voltages to the gold contact to inject electrons into the oxide. This causes a high number of tiny holes to form in the contact, which are often clustered together in one area. It is unclear where these traits came from. Since the specific area where the faults are present does not match the location of the probe on the object being examined, a mechanical explanation is ruled out.

When the device is strained with positive currents, or electrons are introduced into the molybdenum contact, noticeably distinct faults are generated. The overall contact looks to get rougher, not just the individual holes. The flaws in this instance are repeatable across devices and cover practically the whole contact. It should be emphasised, nevertheless, that these flaws are not necessary for observing current transients, and they were avoided in the devices examined here.

4.3 Transient Neuromorphic Behaviours

Current transients are difficult to define because they are elusive, sometimes mistaken for faults, and only seen in a small number of the devices in a batch. This is due in part to the difficulty of developing devices that reliably produce transients. It is ideal to be able to gradually create an effect and link its presence to any changes in device attributes in order to define a behaviour in detail. It is understandable that the analysis that could be done was restricted since published research on current transients relies on the irregular appearance of current transients. Therefore, it is obvious that the lack of a method to create transients is a crucial tool if this phenomena is fully comprehended.

4.3.1 Combined Potentiation and Depression

In essence, the behaviour of the current transient phenomenon can represent the ability to exhibit both analogue potentiation and depression under the same voltage polarity and amplitude. Potentiation and depression are two fundamental processes that occur within synapses and are the foundations of more complex learning rules. During potentiation, a synapse's conductivity increases, whilst during depression, it decreases.

Changes caused during potentiation or depression can also vary in volatility: they may persist and be long-term (Long-Term Potentiation (LTP) and Long-Term Depression (LTD)) or their effects can reverse over time and be considered short-term (Short-Term Potentiation). be regarded as short-term (Short Term Potentiation (STP) and Short Term Depression (STD)).

These fundamental synaptic behaviours have been replicated in a wide variety of electronic devices, predominantly memristors or ReRAM devices, which have been shown to exhibit both potentiation and depression over both long and short timescales [111]. However, in these examples, potentiation and depression often occur in opposite polarities. This necessitates that enhancing and reducing inputs should be capable of producing spikes of varying po-

larities, which could result in a rise in intricacy of the neuron circuits that govern the synapses.

It would be advantageous to construct networks that can operate on a single rail power supply similar to the majority of modern digital electronics. Achieving this would necessitate the use of devices in which potentiation and depression can occur solely as a result of a single polarity of voltage pulse. There are instances of devices displaying potentiation and depression within the same voltage polarity, including complementary ReRAM [126] or unipolar ReRAM devices [183]. Nevertheless, these devices demonstrate discrete binary switching behaviours, rather than the more biologically analogous analogue changes in resistance.

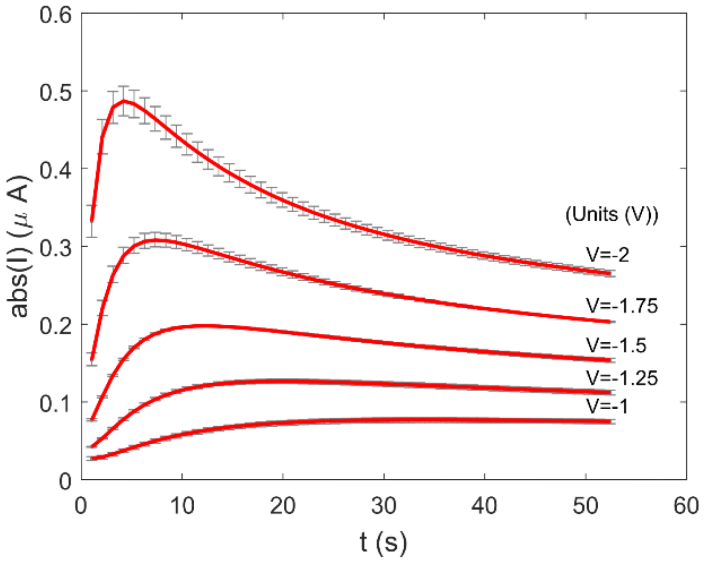

Fig. 4.5 The transient currents of amorphous silicon dioxide thin films were analysed for various voltages. The voltage is applied to the gold electrical contact while the molybdenum contact is grounded. For each voltage, the average of three trials is recorded and presented graphically, with error bars indicating the range of the three trials.

The device's potentiation and depression can be best demonstrated by applying a step potential. When a negative step potential is applied to the top gold electrode of the device, transient current is observed, as shown in Figure 4.5. Initially, a rapid potentiation occurs, taking just a few seconds, but it quickly reaches a peak beyond which the depression phase begins. The device conductance is gradually dominated by competition with depression, ultimately causing the conductance to fall below its original level. This occurs over a longer period of time, lasting tens of seconds and continuing for tens of minutes.

The transient current response of the device to a range of DC voltages is plotted in Figure 4.5. For each voltage the mean of 3 trials is plotted, with error bars indicating the maximum and minimum of all trials. At lower voltages, i.e. -1 V, negligible depression is observed, but it becomes progressively more prominent at larger voltages. Potentiation is observed for all voltages, including -1V, and appears to accelerate with increasing voltage.

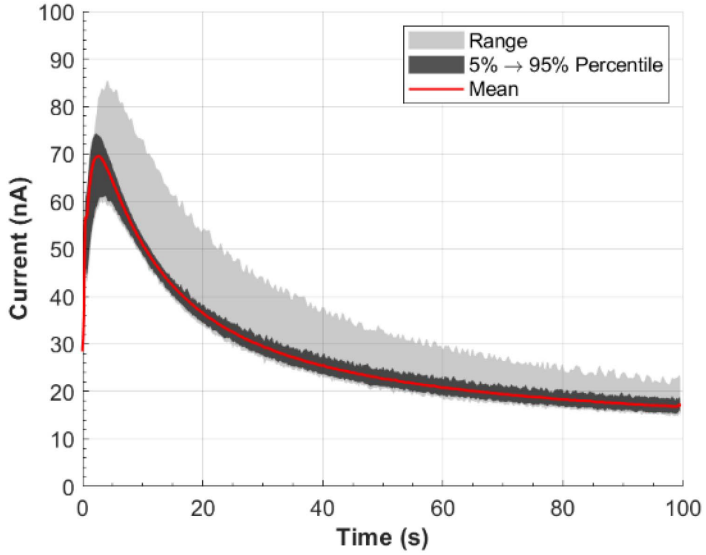

Fig. 4.6 The mean, range and 5th - 95th percentiles are presented graphically for 132 current transients that were induced by applying a -0.8 V step potential to the gold electrical contact while the molybdenum was electrically grounded. A single-order low pass filter with a cutoff frequency of 10 kHz was applied to the data to eliminate noise that may mask the variation between trials.

It is worth highlighting the repeatability of the device's response as presented in Figure 4.6, where a collection of 132 current transients induced by a step potential of -0.8 V applied to the gold electrical contact is displayed. Although resistance switching devices frequently display variable and stochastic responses, it is apparent that the device's performance is consistent and foreseeable, as demonstrated by the range (light grey) and the 5th and 95th percentiles (dark grey). The range has a wider spread than the 5th and 95th percentiles, attributed to bigger currents observed in the first three trials, resulting from a settling process across numerous trials.

This statement only holds true if the device is allowed to fully relax between trials. To achieve relaxation, both electrical contacts should be grounded and the device left to rest. The reset process is gradual, necessitating a resting period of one hour to guarantee complete

relaxation. Accelerated relaxation can be attained by applying a positive potential to the gold contact, as opposed to grounding it.

While this behaviour is noticeable during step potentials, it is crucial to examine whether the same behaviour can be reproduced when operating neuromorphic synapses using pulse trains. It was demonstrated that this is achievable by administering a sequence of Gaussian pulses to the appliance. The pulses have a full width half maximum (FWHM) of 20 ms, an amplitude of -3 V, a period of 300 ms, and are once more implemented on the gold contact. Note, while the pulses have a negative amplitude, the device current has been inverted in the following figures to enhance clarity.

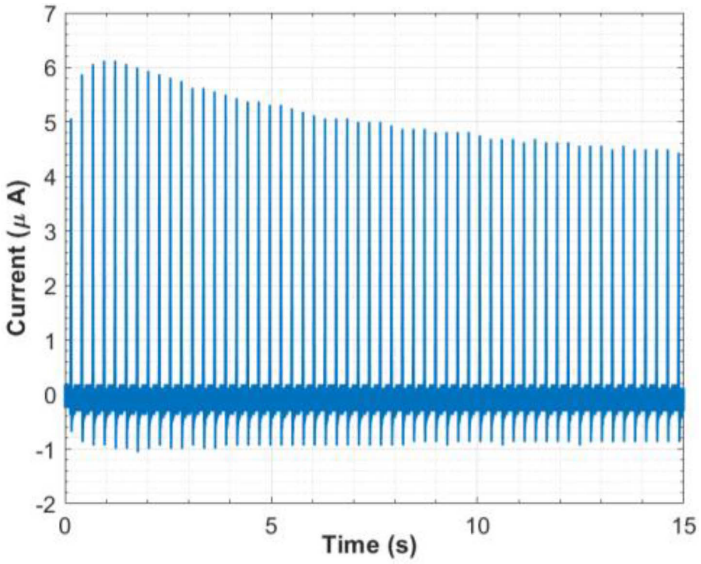

Fig. 4.7 The current that passes through the device while a spike train is applied is measured. The pulses are applied to the gold electrical contact while the molybdenum contact is grounded. Each pulse has a Gaussian form with a Full Width at Half Maximum (FWHM) of 20 ms, an amplitude of -3V, and a period of 300 ms. The magnitude of the current has been inverted for clarity. The device was electrically stressed with a constant current of $-100\mu A$ for 100 seconds.

The response to the series of Gaussian pulses is shown in Figure 4.7, indicating a behaviour similar to that in the DC measurements. An initial potentiation lasting about four pulses is followed by a longer depression period. This indicates that a combination of potentiation and depression can be achieved in both DC and spike train operation, promising the suitability of the device for use in spiking neural networks.

One possible advantage of the subthreshold regime lies in its substantial resistance value (approximately 10 MΩ). Typically, resistance switching devices applied for neuromorphic computing alternate between high resistance states (about 100kΩ) and low resistance states (about 1kΩ).

Nevertheless, in the subthreshold regime, our device remains within a range of resistances comparable to or even higher than the high resistance state of binary resistance switching devices. This indicates that the device may function with lower current consumption than a typical resistance switching device.

However, it must be acknowledged that inherent limitations exist. The broad range of resistance in switching devices reduces their sensitivity to noise or voltage fluctuations, while also enabling a wider range for programming. Thus, there exists a trade-off between current draw and resistance to noise. Additionally, the subthreshold regime poses the challenge of voltage dependency in depression mechanisms. As depicted in Figure 4.5, depression is seldom observed below -1V. This, in turn, constraints the operating voltage of subthreshold circuitry wishing to utilize depression dynamics.

4.3.2 Transient Tunability

The capability to choose between potentiation and depression is crucial for circuit designers who intend to exploit the device in neuromorphic circuits. The more adaptable approach is to regulate the magnitude of the applied pulses. Figure 4.5 demonstrates that lower voltages, such as -1V, display a certain level of potentiation but not depression. This is also true for low-voltage spike trains.

Figure 4.8 depicts the current response to spike trains of varying pulse amplitudes. Depression is not observed with lower voltages. It is possible that the absence of depression at lower voltages indicates a threshold electric field required to induce the change from potentiation. This implies overcoming an activation barrier to initiate defect drift.

On the contrary, depression may be the intended response. In this instance, the amplitude of impulses may be raised to the point where the initial potentiation is overcome. Figure 4.9 depicts the percentage upsurge in device conductance from its initial value. The experiment is replicated for spike trains of differing amplitudes. It can be seen that spike trains with an amplitude < -3.75V are largely potentiating, with the change in conductance remaining positive. However, when the amplitude is increased to -4.5 V, the spike train leads to negative

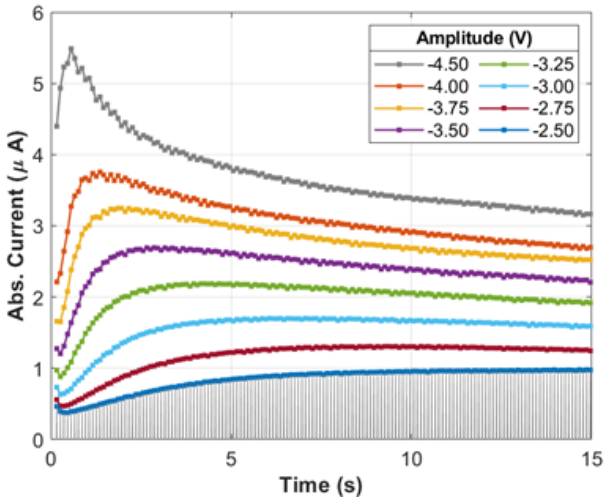

Fig. 4.8 The plot illustrates the device current for Gaussian spike trains with increasing amplitude. Grey-coloured Gaussian current pulses represent the smallest amplitude (-2.5 V). To improve clarity, only the peaks of each pulse are plotted for higher amplitudes. The pulses have a Full-Width at Half-Maximum (FWHM) of 20 ms and a period of 100 ms. Priorly, the device underwent electrical stress when exposed to a constant current of -50μA for 100 seconds.

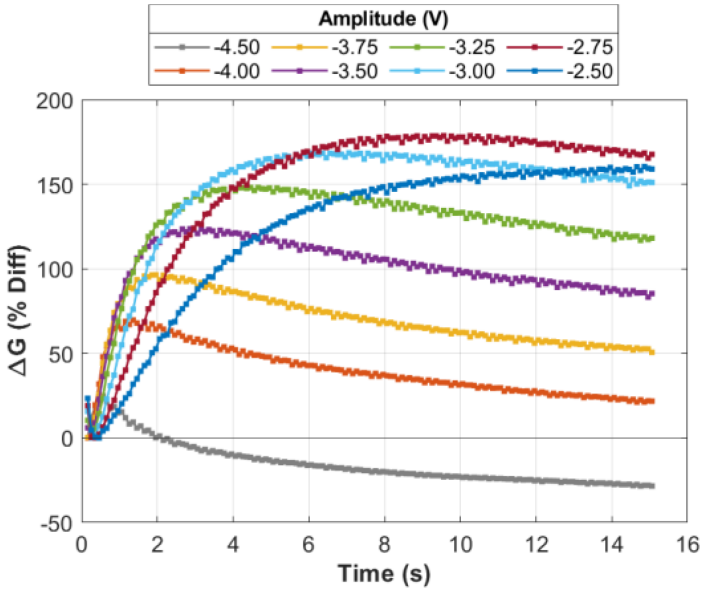

Fig. 4.9 The percentage difference in conductance caused by each spike train pulse is represented on a graph. Larger amplitudes mainly cause depression, whereas smaller amplitudes lead to potentiation. Before testing, the device underwent electrical stress through a constant current of -50μA for 100 seconds.

changes in conductance, resulting in depression.

An alternative and enduring method to select a particular behaviour is to modify the electrical stress that the device experiences after production. As outlined earlier, the devices undergo initial stress through the application of a constant current to the device. The current magnitude is modifiable, which, in turn, alters the degree of stress imposed on the device, thereby allowing adjustment of the device's current transient response. In figure 4.10, the response to stress is then plotted for six different devices, with each being subjected to varying currents ranging from -0.1 μA to -100 μA. Each device is subjected to the current for 100 seconds.

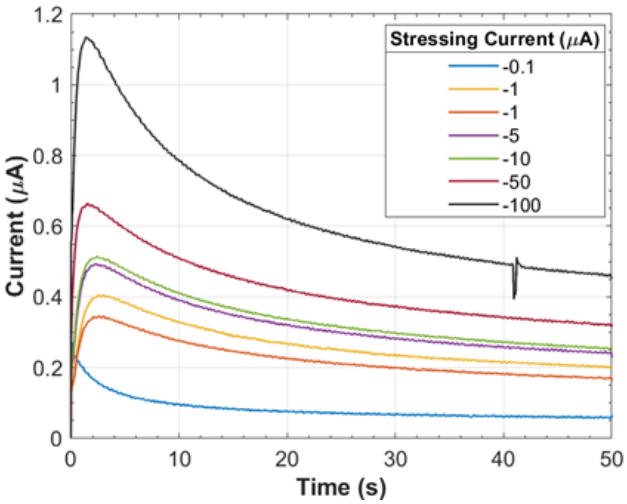

Fig. 4.10 The device current is plotted in response to a step potential of -1.25V for devices initially stressed to varying degrees. There is no potentiation in devices stressed to a lesser extent, while heavily stressed devices exhibit greater potentiation and depression. For clarity, the absolute value of the device current has been plotted.

The resulting current transients display a noticeable potentiation for higher stressing currents, accompanied by an increasing conductance. However, the device stressed with the smallest current, -0.1μA, shows no potentiation, only depression. Conversely, the device exposed to the maximum current, -100μA, potentiates up to almost 10 times its initial conductance.

These methods offer the chance to modify the level of potentiation taking place in the device, ranging from minimal to noticeable. It is crucial to observe that this adjustment is irreversible and would typically be determined during circuit manufacture. In contrast, the method of altering spike amplitude is adaptable and can be modified while in use.

Given the device's ability to display both an increase and decrease in conductance, each occurring at varying rates and with different relaxation rates, it is unavoidable that the steady state conductance will differ across various spike train periods. Before investigating the computational potential of this behaviour, it must first be confirmed.

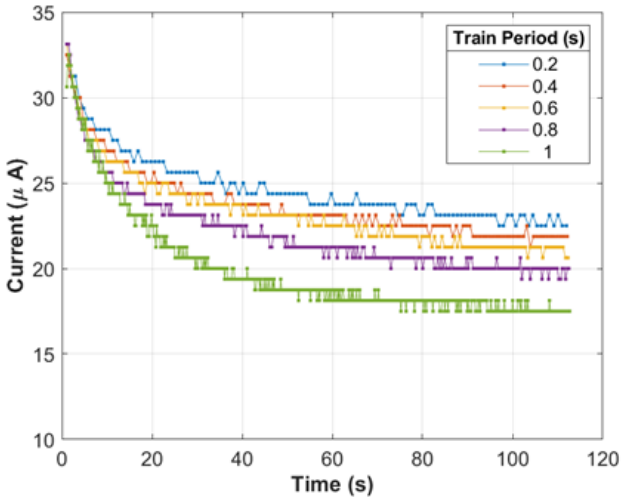

Fig. 4.11 The device's maximum current is graphed for each pulse applied to a 400 by 400 μm device. The pulses take the shape of a Gaussian curve with a FWHM of 50 ms and an amplitude of -2.25 V. For clarity, the absolute value of the device's current has been shown.

In Figure 4.11, a graph plots the device current produced by spike trains with varying inter-spike time periods. The device ultimately settles into a steady state conductance in response to each spike train, where conductance changes induced by the applied spikes are counterbalanced by the device's relaxation dynamics.

As anticipated, the steady state conductance is found to be dependent on the frequency of the pulses applied to the device - higher frequency pulses lead to decreased conductivity. These findings imply that this conduct could potentially be utilised to adjust the device resistance in response to spike trains of different frequencies.

4.3.3 Homeostasis Applications

The original aim of this study was to illustrate the potential application of distinctive memristive device behaviour for innovative computing architectures. One noteworthy application to emphasise is the neuron's capacity to adjust based on a firing frequency analogous to

homeostasis in biology [245].

In biological systems, homeostasis refers to the maintenance of preferred operating conditions or particular states [251]. An example of this is the regulation of body temperature. When applied to spiking neural networks, homeostasis plays a role in maintaining spiking activity to avoid excessive power consumption through unnecessary spike events. It can also protect against faulty neurons entering a chaotic, high-activity state. Without homeostasis, such events could cause the entire network to enter a chaotic state [174].

This process is also known as habituation, a type of homeostasis in which a change in input results in a temporary change in output that eventually diminishes to a steady-state value. There are already a few practical examples of this in recurrent neural networks, one of which resulted in a 20% improvement in classification accuracy on the MNIST dataset when compared to a neural network without homeostasis.

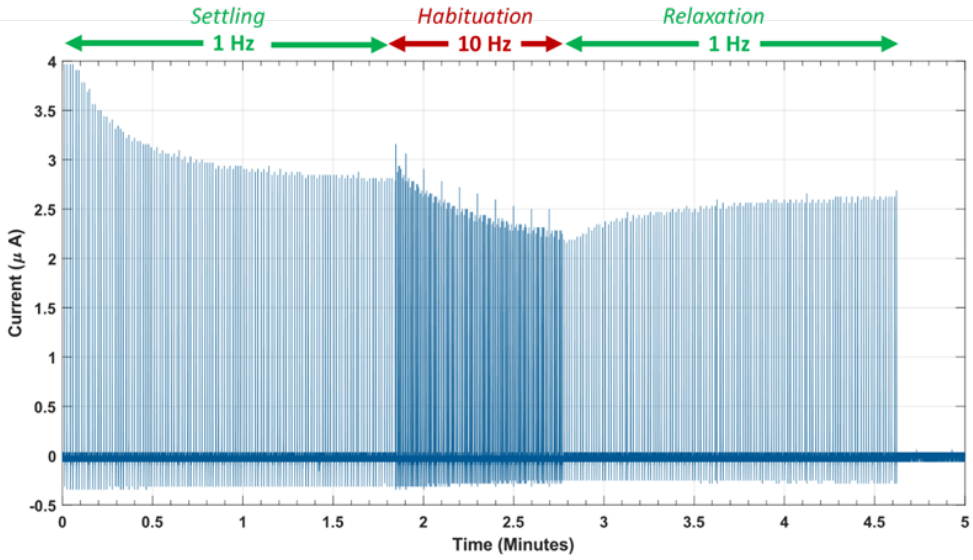

Fig. 4.12 Response of the device to varying frequencies of spike pulses. The device was initially operated with a sequence of 1Hz spike pulses, followed by a period of 10Hz spike pulses, before finally returning to 1Hz spike pulses.

The previous section showed that the steady state conductance of the device is reliant on the firing rate of the related neuron. The device conductance decreases at higher spiking frequencies, but has a greater steady state at lower frequencies. This behaviour is suitable for inhibiting the impact of suboptimal input neurons that have entered into a chaotic state of

high frequency.

Figure 4.12 demonstrates this concept in practice. The unit is connected to a signal generator that produces Gaussian pulses with a low frequency of 1Hz, which is the neuron's typical background activity. In these conditions, the device attains a steady-state conductivity. An input neuron is simulated as transitioning to a dysfunctional state of higher activity at t = 1.8 minutes, potentially caused by circuit damage. The frequency of the spike train is increased to 10Hz.

Without homeostasis, this mistaken input could cause the linked output neuron to reach a comparable state, which could then transmit through subsequent layers of the neural network. Luckily, when subjected to the high-frequency spike train, the conductivity of the device reduces due to the enduring behaviour of the current transient device. This decreases the overall current supplied to the linked output neuron, decreasing the chance of this high-frequency input triggering equivalent actions in the following neurons.

While the temporary suppression is advantageous for safeguarding the network against excessive firing, it should not be a permanent solution. If the input neuron recovers to normal functional conditions, the inhibition should be lifted, and the neuron should be allowed to resume its place in the network. At t = 2.8 minutes, evidence was presented to support this idea. With a 1Hz spike train, the sample recovers to the background level of activity. The device conductance responds to the fluctuation of current transient conductance changes by returning to its former steady state value prior to the neuron's faulty phase.

The ability of protective systems to return to their original state when normal operating conditions are restored is a fundamental characteristic of homeostatic processes. Although the concept of homeostatic habituation has been demonstrated with the current temporary device, there are still limitations. For instance, it is impossible to determine a specific and constant conductance value while the device is operating in its present transient state. If the device is electroformed similar to a regular memristor, it allows for the weight to be adjusted to a variety of analogue values.

To achieve homeostatic behaviour in a physical system, it may be necessary to connect the device in series with another device that is programmable. The good news is that the programmable device and the current transient device are architecturally identical and

manufactured using the same method. The devices can be combined on a single wafer since the only difference between them is the method in which they are electrically stressed.

4.3.4 Physical Implications

Initially it was thought that the transient was the consequence of an ionic current superimposed on an electronic current [172, 140]. The ionic current is the outcome of a field-driven drift of some space charge, which is often assumed to be charged oxygen vacancies in more recent research [214]. The variations in conductance seem to stem from two distinct modifications happening simultaneously in the sample. An ionic current could only account for this observation if the drifting space charge were capable of drifting at a constant velocity permanently.

A more appropriate explanation is that the transient is an electronic current that is altered by changes in the device's conductance, with two simultaneous changes causing the transient [190]. This is because the electronic conductance of an MIM device can be modified in several ways, some of which may occur simultaneously. The literature review identifies a general trend favouring an electronic explanation over an ionic explanation.

There are, however, numerous mechanisms through which the electrical conduction of an MIM device can be adjusted. An MIM device can be abstracted as a series combination of three distinct components: a metal-insulator interface, a bulk insulator, and a second metal-insulator interface. These components determine the device's conductance; however, in specific cases, one of these three components may be significantly more resistive than the others. In such circumstances, that particular component alone may provide a good estimate of the device's conductance.

So far, there has been no research to determine which layer, if any, dominates the conductance of a device that shows current transients. Most devices described in literature display rectifying behaviour, implying an interface-defined conductance. However, it is important to note that even if a device is rectifying, this does not necessarily indicate that the dynamics are always accurately described by an interface model.

In the reverse polarity, when the interface exhibits the least conductivity, the conductance of the device can be accurately captured by the interface model, since it is expected to have high resistivity. However, in the forward polarity, when the interface is most conductive, the bulk oxide may have a greater impact on describing the device conductance, necessitating a

bulk-based model. Without knowledge of which layers define the device conductance, it is necessary to investigate every layer within the device to determine the physical changes that could be causing the current transients.

The bulk oxide, where the crystalline oxide is highly insulating as a result of the oxide's large bandgap. At high voltage and low temperature, current through the oxide occurs predominately via tunnelling- either directly or via the Fowler-Nordheim mechanism. However, the presence of electron trap states within the oxide results in the possibility for higher current densities through the use of these traps as alternative conduction pathways.

This thesis examines a sample of amorphous silicon oxide. Such oxides contain a higher concentration of oxygen vacancies, which affect electronic conduction by acting as electron/hole traps. Additionally, these oxides have an efficient pathway to creating oxygen vacancies. In the amorphous phase, the silicon dioxide has wide $O-Si-O$ bond angles [57].

The sites can trap a maximum of two electrons in this broad bond, which decreases the energy barrier required to generate the vacancy [71]. The existence of oxygen vacancies and an efficient route to their generation results in the degradation of the oxide. Conduction takes place via vacancy trap sites despite the wide bandgap.

There is evidence of the formation of conductive bridges through the oxide via trap states according to TEM [195, 281] and CAFM [19] etching techniques. At higher temperatures, trapped electrons can be excited from one trap state within the bandgap to its neighbour, or across the oxide through the conduction band, via Poole-Frenkel conduction due to thermionic emission within these states [67].

$$\sigma_{PF} = q\mu_n n_{PF} = q\mu_n n_0 \left(\frac{-q(\Phi_B - \Delta\Phi_{PF})}{k_B T} \right) \tag{4.14}$$

$$j_{PF} = E\sigma_{PF} = Eq\mu_n n_0 \left(\frac{-q(\Phi_B - \Delta\Phi_{PF})}{k_B T} \right) \tag{4.15}$$

The Poole-Frenkel effect raises the conductance of an oxide, σ_{PF}, by increasing the amount of free electrons in the conduction band relative to its thermal equilibrium concentration, n_0 to n_{PF}. The applied field reduces the barrier height, Φ_B, necessary for an electron to be thermally excited from its trap state into the conduction band, causing this rise. For an applied field, E and equation (2.15) where $\Delta\Phi_{PF} = \sqrt{\frac{q^3 E}{\pi\varepsilon}}$, the drop in barrier height represents the increase in conductance produced by the reduction in barrier height [240]. Because this

impact is reliant on the thermal energy of the electron, the number of extra electrons is based on a Boltzmann distribution and is determined by the thermal energy, $k_B T$, of the electron.

It is evident that the conductance of a device, regulated by the Poole-Frenkel effect, may be influenced by a change in the carrier concentration of the trap sites, n_0, and a reduction in barrier height, which varies with the local electric field, E. Both of these factors will be considered when examining the model as a potential explanation for current transients.

Poole-Frenkel, however, is not the sole cause of conductivity through trap sites [52]. The transfer between proximate traps can occur via tunneling, termed trap-assisted tunnelling (TAT) [110]. This method has already been proven to sufficiently account for conduction in amorphous silicon dioxide sheets showing resistance switching behaviour [182]. One potential TAT model is founded on an inelastic tunneling process, ITAT [105].

Both the Poole-Frenkel and ITAT models are capable of explaining the alterations in conductance witnessed during the current transient. The rise in conductivity, which is probably due to charge trapping, can be effectively accounted for by these models. Take, for instance, the Poole-Frenkel model (4.15), in which the current of the device results from the trap population, n_0, and the likelihood of an electron making a thermal leap from the trap to the conduction band. The model details steady-state conditions, where the trap populations remain constant.

$$E(x) = \frac{V}{d} \tag{4.16}$$

$$E(x) = \frac{V}{d} - \frac{Qx}{\varepsilon d} \tag{4.17}$$

$$E(x) = \begin{cases} \frac{V}{d} - \frac{Qx}{\varepsilon\delta} & \text{if } x < \delta \\ \frac{V}{d} - \frac{Q}{\varepsilon} & \text{if } x \geq \delta \end{cases} \tag{4.18}$$

It is conceivable, though, that this term could be treated as time-dependent. At first, with the device grounded and at room temperature, the traps may be largely depleted, corresponding to the low conductance observed when the step voltage was first applied to the device. With the voltage now applied, however, traps that were previously depleted or had a low probability of being filled when grounded may now have a higher probability of being filled, corresponding to an increase in the probability of a trap being filled, n_0.

As the device becomes more conductive and higher current densities flow through it, the number of electrons trapped within the oxide increases. This eventually leads to a higher equilibrium or steady state value of n. A similar argument can be made for the ITAT model, which acknowledges that the probability of traps being populated varies with time and already accommodates transient effects. In order to observe an increase in conductance, the current flowing into the traps must exceed the current flowing out of the traps.

The decline in conductance may also be accounted for by positing an indirect interaction through the local electric field and a mobile defect drifting under the influence of the applied field, e.g. a charged oxygen vacancy. For instance, the Poole-Frenkel mechanism exhibits an exponential dependency on the local electric field to mitigate the trap depths faced by the trapped electrons. Decreasing the electric field would exponentially lessen the device's electrical current.

Mobile ions might have the ability to diminish the electric fields encountered by most of the traps. Consider an ideal insulator with a thickness of d, in which no space charge is trapped. In this case, the electric field would remain constant across the oxide layer (4.16). If a space charge, Q, is introduced into the oxide and distributed homogeneously, the electric field exhibits a gradient (4.17). The charge within the oxide decreases the field experienced by the traps.

However, when distributed across the oxide's thickness, the drop in potential is shared amongst many traps, resulting in less significant differences for each trap. If the charge is mobile and compelled to accumulate at one of the interfaces by an applied field, the decrease in the local electric field is focused near the interface. Consequently, numerous more traps now experience the lower electric field resulting in an overall decreased device current.

While there have been discussions about the potential explanations for the bulk effect, interface-based models may also explain the changes in conductance observed during the current transient. Memristor devices have previously demonstrated resistance switching, which interface models have explained. The models suggest the existence of a Schottky barrier at the metal-oxide interface, based on the rectifying nature of the device.

The Schottky-like barrier arises due to a defective oxide material behaving akin to either an n-type [69] or p-type [216] semiconductor. Upon contact with the metal, charge transfer takes place between the metal and defect sites within the oxide, caused by the offset of work

functions. Consequently, a space charge and electronic barrier are formed, hindering any additional charge transfer across the interface. The conductance of the device changes due to alterations in either the barrier height, which allows for larger thermionic currents, or its width, enabling greater tunnelling currents. It has been suggested that the barrier is controlled by one of two means.

One explanation is derived from interface states [221]. The interface states, also known as surface states, have an influence on the space charge within the interface [12]. These states may decrease or increase the barrier height or width depending on their type, whether they are n-type or p-type states [44].

If the concentration of these interface states is high enough, the barrier height may be solely defined by the interface states, leading to fermi pinning, rendering the metal/oxide work functions redundant. Charge trapping in these interface states could clarify the modulations within the barrier height or width. For instance, it was suggested that surface states of a metal-amorphous silicon Schottky barrier decreased the barrier height [279].

Alternatively, it has been suggested that the alteration in barrier height stems from a migration of oxygen vacancies towards the interface, determining the band alignments of the two materials [9]. Research has revealed that elevated levels of oxygen vacancies led to a larger bandgap in the bulk. This presents a model that revolves around the accumulation of vacancies, contributing to the expansion of the insulator's band-gap at the interface and thus an elevation in the barrier height.

To conclude, there are still several potential explanations that may account for the alterations in conduction that cause the current transient. However, none of them have been identified as the best explanation. If the device has bulk limitations, then the decline in conductance due to field driving could potentially be explained by a mobile space charge that reduces the local electric fields. This could apply to both a Poole-Frenkel and trap-assisted tunnelling model. Meanwhile, the increase in conductance driven by charge could be explicable through the population of traps required for electronic conduction to take place.

This applies to both Poole-Frenkel and trap-assisted tunneling models. Alternatively, if the conductance is limited by the interface, the increase in conductance due to the driven charge may be a result of trapping within surface states at the interface, which reduces the barrier height through Fermi-pinning. Meanwhile, the decay caused by the field can be explained by

charge defects drifting towards the top contact, modulating the band alignments of the metal contact and oxide. Equally, while these two groups are distinct - bulk or interface limited models - the answer may well involve a fusion of the two.

4.4 Summary

This chapter demonstrates a device that can exhibit potentiation and depression of conductance under the same voltage polarity in the sub-threshold regime. The methodology explains how to induce this behaviour within standard resistance switching devices and how to select specific dynamics. The presented methodologies will allow researchers to replicate the behavior in their respective devices and instruct circuit designers on how to fine-tune the behavior for their specific application.

The current transient property of a device that can be designed to generate conductance potentiation and depression under the same voltage polarity. The emphasis is on immediate applications and implications for spiking neural networks. The slow dynamics of conductance depression could be useful in establishing long-term homeostasis and habituation, yet the volatile potentiation may be more appropriate for Short-Term-Memory. Similarly, the convergence of these two actions on a solitary device may have unanticipated advantages in more integrated memristive systems.

In regards to spiking neural networks, implementing synapse weight update rules such as spike-rate-dependent plasticity (SRDP) [102] can be achieved through the subthreshold regime. In SRDP, synaptic weights are updated based on the frequency of the input signal. This is different from spike-timing-dependent plasticity rules, which adjust the weight according to time intervals between spikes from the presynaptic and postsynaptic neurons. The homeostatic behaviour, which is a direct function of the spike train frequency, has an instant impact on the weight update process between neural network layers.

Moreover, the volatile nature of conductance changes in the subthreshold regime resembles a forgetting process [149]. This phenomenon has been utilised to reduce synaptic weights without requiring inhibitory pulses, thereby simplifying circuit complexity. However, the effects of this forgetting process on network performance in the context of SNNs are yet to be comprehensively studied, and thus cannot be accurately gauged.

The applicability of this trait is restricted to circuits with slower dynamics due to its slow behavior, which may be a disadvantage in situations where speed is a concern. Fortunately, the majority of signals and monitoring techniques operate on a one-minute segment basis, making this behaviour beneficial for the future applications potential of the project [256].

If circuit designers intend to utilize this behaviour, it is imperative to address the physics that underpins it. Further verification is required for existing models like the SCLC model and other electronic explanations to develop accurate simulations/models and identify the fundamental limitations of behaviour. In order to advance research on this topic, future work should concentrate on discerning the location where these changes in conductance are transpiring, specifically at the interface or the bulk.

This could be accomplished by adjusting the electrode materials to potentially reveal interface effects, and modifying oxide thickness and area to potentially disclose bulk effects. This demonstrates potential evidence for conductance decay due to a process occurring at the interface, but it is not a definitive outcome.

Conducting further research in this field could assist in identifying which of the bulk or interface models are relevant. Once the location of the alteration has been identified, the findings regarding the driving force can suggest a physical model for the phenomenon. Charge trapping is expected to enhance the device's conductance, while a field-driven drift of defects should decrease it.

The capability of the subthreshold regime to generate both potentiation and depression with identical polarity voltage pulses is a distinctive behaviour that specifically deals with the current obstacles that circuits face when sourcing pulses of opposite polarities or when neurons have access to both ends of a synapse, resulting in intricate signal routing.

This operational regime has potential applications in the field of neuromorphic computing, where circuits can utilize the behaviour for single-rail power supplies, thereby simplifying circuit construction and layout. Furthermore, the intricate dynamics may offer opportunities for novel types of neuromorphic circuits, as noted in the discussion of homeostasis within individual synapses.

From a wider perspective, this operational approach additionally reinforces the concept of completely memristive circuits. Identifying and presenting an intermediate regime between

the pristine and electroformed states of MIM devices with qualitatively different characteristics has prompted consideration of how these behaviours may be combined. This raises questions regarding new computations that could be particularly suitable for enhanced signal processing applications in the future.

Chapter 5

Neuromorphic Modelling Framework

5.1 Optically Active Device

This chapter presents a model for resistance switching devices that demonstrates both analogue potentiation and conductance depression under the same voltage polarity, exploring the subthreshold regime with current transients, which has the potential to simplify neuromorphic circuits. The utilisation of an empirical SPICE model enables the simulation of these transients, with the process of validation achieved through the use of experimental data.

This provides parameters for simulations and facilitates applications such as fault mobility estimation and neuromorphic circuit functions. The enhancements made to the previous model include the integration of diodes to emulate Schottky-like contact and the incorporation of relaxation dynamics upon the removal of step potentials or the grounding of the device.

5.1.1 Modified Device Stack

A distinct set of devices with disparate top electrical contacts were characterised, one with conductive indium tin oxide (ITO) in lieu of gold. The bottom contact and oxide layer remained unaltered and consistent with those observed in the gold-contacted devices presented in the preceding section. When subjected to stress, the ITO-contacted device exhibited a distinct response compared to the gold-titanium contacted device. Instead of a gradual and smooth increase in conductance, the response was more erratic and chaotic.

The aluminium-contacted devices have yet to demonstrate the occurrence of current transients following the application of stress. In addition to failing to exhibit current transients, any increase in conductance induced by the constant current stress is also observed to be more

Table 5.1 Comparison of the modified device stacks.

Original Stack		Modified Stack	
Layer	Film Thickness (nm)	Layer	Film Thickness (nm)
Au	110	ITO	30
Ti	3	Ti	3
SiOx	35	SiOx	20
Mo	280	Mo	150

volatile than that observed in the other devices, with the devices returning to a high resistance state within a couple of hours.

The aluminium-contacted devices have yet to demonstrate the occurrence of current transients following the application of stress. In addition to failing to exhibit current transients, any increase in conductance induced by the constant current stress is also observed to be more volatile than that observed in the other devices, with the devices returning to a high resistance state within a couple of hours.

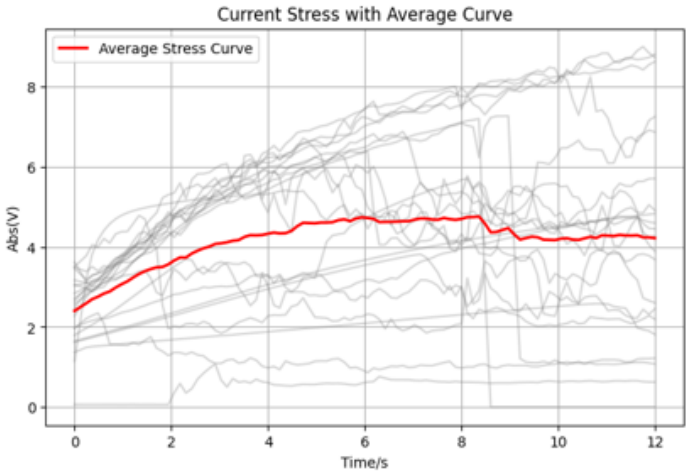

Fig. 5.1 Stressing responses of ITO top contacted device. With the objective to ascertain the stress responses of the ITO top-contacted device, the voltage across the device was monitored while it was subjected to a constant current of $-0.5\mu A$. It was not possible to apply larger currents. The response was observed to be less smooth when compared to that of the gold-contacted device.

In contrast, the ITO contacted devices did exhibit transients, but interestingly, only partially. A typical transient observed in ITO devices is plotted in Figure 5.1. It exhibits the initial increase in conductance as is typical with current transients, but does not then start reducing in

conductance. Instead, the device exhibits a chaotic spiking-like behaviour which, if observed for too long, will cause the device to switch to a low resistance state.

The observation of a partial current transient in ITO-contacted devices is a significant finding. As will be discussed in the following section, this evidence is indicative of the transient being the result of multiple simultaneous changes occurring in the device. Furthermore, it supports the hypothesis that the top metal-insulator interface plays a role in generating transients.

The hypothesis that alterations to specific interfaces of the device can influence the characteristics of the current transient is supported by findings in tantalum oxide-based devices [249]. In this study, a layer of Al_2O_3 was deposited between the tantalum oxide bulk and the titanium nitride electrodes, which reduced the prominence of the current transient in the absence of the buffer layer.

To illustrate, the gradual decline in conductance of the transients is exclusive to the gold-contacted device, indicating that it is either due to the characteristics of the metal-insulator interface or the disparate responses to stressing at this interface that determine whether the decaying behaviour is manifested.

In contrast, the initial increase in conduction is observed in both the ITO and gold-contacted devices. This suggests that the behaviour is less affected by the top metal-insulator interface and may be located in the bulk oxide layer or at the bottom metal-insulator interface. The disappearance of the slower decay in conductance with the change in top electrode may provide insight into the physical model describing the current change.

5.1.2 Conductance Variation Mechanisms

The initial step is to ascertain the location within the device stack where alterations are taking place that are responsible for the observed reduction in conductance. The absence of decay occurring concurrently with the alteration of the top electrode suggests that the causal factor responsible for the observed conductance decay is situated at the interface between the top electrode and the amorphous silicon dioxide.

Given the slow dynamics of the change in conductance, it is plausible that a drift of some mobile defect is responsible. It is well established that silicon dioxide films are susceptible to the influence of alkali mobile ions [226], including sodium and lithium ions, which are all characterised by a positive charge [292]. The drift of these mobile charges can significantly

affect the potential drops at metal-oxide interfaces, as well as modulate barrier heights when allowed to accumulate.

If some positive mobile ion, regardless of the species, existed in the oxide of the device, it would be attracted to the top electrode, which is at a negative potential. This would cause an accumulation of positive space charge at the interface, which would in turn reduce the potential across the oxide. Nevertheless, it can be argued that alkali metals, such as sodium and potassium, are unlikely to be the cause of this positive space charge, given that they do not migrate at room temperature.

Instead, they require temperatures in excess of 100 degrees Celsius (212 degrees Fahrenheit) [47]. The current transients presented in this thesis are all observed at room temperature, which suggests the need for an alternative candidate to explain the mobile space charge, in particular one that is mobile at room temperature.

It is noteworthy that modelling of the temperature within analogous TaO_x devices has indicated the potential for increases in oxide temperature of up to 100°C with applied voltages of -0.7 to -1.8V due to Joule heating [224]. This would imply that if comparable effects were present during the current transient, then elevated temperatures within the oxide could be occurring and potentially facilitating the migration of alkali metal defects.

It seems plausible to suggest that the proton [95] is a likely candidate for positive ions that are mobile at room temperature. The presence of ionised hydrogen in silicon dioxide films has been repeatedly observed to be both stable and consistent [254]. It has been demonstrated that protons can influence the electronic properties of capacitor devices in which protons are trapped within the oxide [253]. Their long-term stability has been demonstrated through multiple cycles of migration between device electrodes [270].

It is commonly assumed that these ions are introduced during the growth of the oxide [252]. Furthermore, their concentration has been demonstrated to increase through annealing in an atmosphere at temperatures above 200 degrees Celsius [155]. However, their presence has also been introduced electronically via the electrolysis of water within the device and via radiation [277]. It is also noteworthy that their migration has been shown to occur repeatedly at room temperature.

This raises the question of why the accumulation of protons occurs exclusively in the gold-contacted devices, rather than in the ITO. Given that gold is an inert metal and is unlikely to be reduced by protons, the accumulation at the gold interface is to be expected. In contrast, there is a substantial body of evidence indicating that the ITO would be reduced in the presence of protons.

Although ITO contacts are often considered to be inert in certain electrochemistry scenarios, this is not always the case. Their reduction is, in fact, heavily dependent on the pH of the electrolyte. For instance, the reduction of the electrode has been observed on numerous occasions in acidic electrolytes [43, 220].

The reduction of ITO in the presence of acids has been demonstrated in both electrochemical experiments conducted at room temperature [266] and in instances where ITO has been exposed to a hydrogen plasma [10]. In one study, the application of negative voltages to an ITO electrode immersed in hydrochloric acid resulted in the formation of spherical structures at the grain boundaries of the ITO film, which exhibited a metallic-like appearance [88].

Following characterisation with Energy dispersive x-ray Spectroscopy (EDS) [99] and X-ray Diffraction (XRD) in a separate study [160], the spherical regions were found to be depleted of oxygen or exhibited only peaks of indium and tin, providing compelling evidence that these spheres were metallic. The same spherical structures were observed in ITO films exposed to a hydrogen plasma, which, when analysed with Auger spectroscopy, again revealed a lower oxygen concentration in the spherical regions.

The reduction of ITO by protons may provide an explanation for the absence of space charge accumulation in ITO-contacted devices. Instead of accumulating, the protons reduce the ITO, producing water as a byproduct, which would not contribute to a positive space charge. Furthermore, the reduction of the electrode may also elucidate the more erratic current-time response observed in ITO devices, as the electrode structure undergoes substantial alterations.

The potential for structural changes to occur in both the oxide and the metal contacts makes it challenging to determine the specific role each plays in modulating the device's conductance. In order to investigate the effect of the contact, it would be beneficial to fabricate and characterise devices with a variety of contact materials.

Any discrepancies in the observed behaviour between the devices could be ascribed to the metal or the metal-insulator interface, whereas any enduring effects could be attributed to the oxide. To further examine any behaviour attributed to the oxide, devices of varying oxide thicknesses could be fabricated. This may reveal a dependence on the oxide thickness, which could be supporting evidence for the oxide having a role in the changing device conductance.

However, it is important to exercise caution when drawing conclusions from this approach, as the change in oxide thickness will also modify the magnitude of the current density flowing through the device, potentially affecting the interfaces. The fabrication of devices with different oxide thicknesses and a variety of metal contacts is a future research direction.

If the hypothesis that the transient is the result of two separate changes is true, then it would suggest that devices could potentially be made to exhibit only one of these changes in isolation. Fabrication of such a device would provide strong supporting evidence for the hypothesis and a clear demonstration that the changes are separable. Modification of the top contact material appears to facilitate this separation.

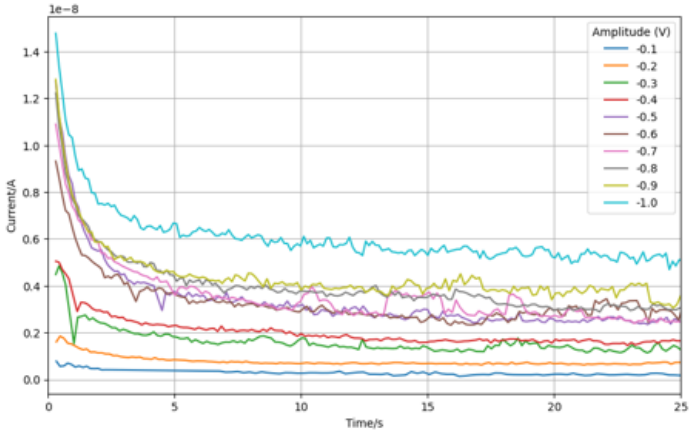

Fig. 5.2 The current-time response for a device with a conductive ITO top electrode. The current is generated in response to a step potential applied to the top contact with respect to the bottom. It exhibits only the increase in conductance and not the decrease. As the current increases, the noise level also rises until the device reaches a point of breakdown. The structure of the device is identical to that of the gold-contacted devices, with the exception of the change in the top electrode's material from gold-titanium to indium tin oxide (ITO).

Devices fabricated with a conductive ITO top electrode, in lieu of the gold-titanium contact, do not exhibit the anticipated decay in conductance; rather, they display only the initial increase. Figure 3.8 illustrates the current-time response of an ITO-contacted device. As

observed previously in the gold devices, the current begins to increase; however, it never reaches the inflection point. Instead, it continues to increase in conductance, becoming progressively noisier until the device undergoes breakdown.

The ITO device has undergone a comparable stressing process to that of the gold devices, which is outlined in the preceding chapter. In the initial stages, the devices exhibit only capacitive currents and possess a very high resistance. Subsequently, a constant current stressing procedure is employed to produce a more conductive device.

Following a period of relaxation, a repeatable transient is produced, provided that the applied voltage is maintained for a sufficiently brief duration to prevent breakdown of the device. A comparable phenomenon has been documented [192] in a variety of oxides and is frequently employed in the replication of short-term potentiation of synapses [300, 30].

Furthermore, this absence of slower dynamics may corroborate with previous findings [190], where the slower dynamics were postulated to be attributable to oxygen vacancies. It is conceivable that the ITO contact is more prone to exchange oxygen with the oxygen vacancy than the inert gold contact.

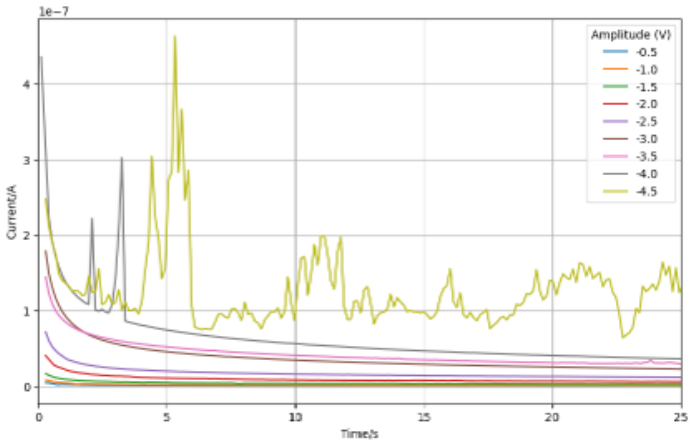

Fig. 5.3 instabilities of potentiation and depression on the amplitude of higher applied voltage pulses on ITO devices.

An alternative hypothesis is that the Au contact is diffusing through the oxide, whereas the ITO contact is not. Previous observations have shown that Au can form conductive bridges between two contacts through a thin film of *ZnO* [204]. Given that gold is known to diffuse in silicon dioxide films, this could be a possibility [164]. However, TEM analyses of the

devices studied in this thesis have not produced observable gold filaments, casting doubt on this hypothesis [185].

An additional potential explanation for the observed discrepancy in behaviour between the Au and ITO contacts is the possibility of differences in their respective work functions. While not directly measured on the samples in question, the work function of gold is reported to be between 4.9 and 5.2 electronvolts (eV) [247], while thin films of indium tin oxide (ITO) have been measured to have a work function between 4.25 and 4.28 eV [217].

This could result in a difference in work function of approximately 1 eV. Such differences would lead to offsets in the band alignments at the contact and oxide, which could affect which traps within the oxide the electrons are injected into. As discussed in the previous chapter, the disappearance of the slower decay in conductance with the change in top electrode is suggestive of proton migration playing a role in the slower decay in device current.

5.2 Empirical Model Fitting

This section presents the development of empirical models designed to track the rates of increase and decrease in conductance separately. The models are primarily intended to quantify the rate of change in conductance. The question of a physical model will be addressed in the following section.

5.2.1 Evaluation Approach

The models explored here are derived from the relaxation experiments that were previously outlined in the preceding chapter. It was evident that the gradual decline in device current delineated the upper limit of the maximum device current, while the initial surge in current approached this maximum but did not exceed it.

$$I(t) = f_{inc}(t) \times f_{dec}(t) \quad \forall \{f_{inc}(t) \in [0 \rightarrow 1] : f_{dec}(t) \in \mathbb{R}\} \tag{5.1}$$

This is represented by the product of two time-dependent functions, $f_{inc}(t)$ and $f_{dec}(t)$. The increase in current is analogous to a charging term, $f_{inc}(t)$, which rises from 0 to 1. Initially, this function defines the device current. However, as $f_{inc}(t)$ approaches 1, it then allows the function it is multiplied with to define the total current, in this case $f_{dec}(t)$.

Although this equation forms the basis of the empirical model, questions remain regarding the specific forms that $f_{inc}(t)$ and $f_{dec}(t)$ should take and the most appropriate means of comparing their effectiveness. The efficacy of each fitting equation is evaluated based on two criteria: the quality of the fit and the degree of realism of the fitted parameters in relation to the underlying physical system from which the model is derived.

The correspondence between each term of a given model and a physical property is contingent upon the physical system from which the model is derived. These properties are assigned a range of values that are deemed realistic. Values outside of this range may indicate that the assumed model is not applicable. The specific correspondence between physical properties and terms is model-specific and will be detailed later in conjunction with the model.

The discrepancy between the fitted equation and the original experimental data is referred to as the residual. This can often be a useful visual indicator of the quality of the fit. An optimal fit would manifest residuals that are centered around zero, exhibiting no systematic offsets or time-variant components.

The residuals in this form indicate that the fitted equation tracks the experimental data well, with the variances in the residuals around zero assumed to be a form of noise or variance in the original data. In contrast, a less optimal fit would exhibit systematic offsets that vary over time. This indicates that either an additional term is absent or the incorrect function has been selected.

However, while residuals are useful for visually assessing a fit, they are less so when larger datasets are being fitted. Therefore, only one residual for each model is assessed, which is representative of the model's performance. The same experimental data will be fitted for each model, thus ensuring an accurate comparison. In order to assess a model's goodness of fit across a whole dataset, a single numerical metric that can quantify the fit is preferred.

A more quantitative description of the goodness of fit is provided by the R^2 measure. The R^2 measure is a statistical tool that enables the comparison of the variance between the observed data points and the model's predicted values against the variance of the observed data and the mean of that data. In other words, it can be acknowledged that the most straightforward

model for predicting a dataset would be to assume the mean value of the dataset in all cases.

$$SS_{tot} = \sum_i (y_i - \bar{y})^2 \tag{5.2}$$

$$SS_{res} = \sum_i (y_i - f_i)^2 \tag{5.3}$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{5.4}$$

In this case, the variance is defined by (5.2), Where y_i is an individual datapoint and $\bar{y}$ is the mean average of the dataset, which is equal to the variance of the dataset. The variance of the dataset against the model's predictions can also be calculated with (5.3), Where f_i is model's predicted value at i^th index, for any other model developed.

If the variance is similar to that of the dataset, the model is no more than a simple mean value predictor, and thus the data fit is poor. An alternative indication of a good fit is provided by a model with a variance much less than that of the dataset. However, residuals should still be checked. The R^2 value shown in (5.4), defined by the ratio of the variance of the model to the variance of the dataset, is based on this premise and increases as the model's fit improves.

Notwithstanding the possibility of attaining an optimal fit, the values of the fitted parameters remain uncertain. To illustrate, a minimum in the fitting error could be achieved by a range of parameter values. This range is referred to as the confidence bounds, which can be interpreted as the range within which the fitting algorithm is certain the final value lies.

For example, 90% confidence bounds will define a range within which the algorithm is 90% sure the optimal value can be found. If a higher confidence is required, the range will generally increase. Consequently, there is a trade-off between certainty and specificity. In this work, the standard confidence threshold of 90% was employed.

The choice of model for a particular dataset is influenced by the magnitude of the confidence intervals. If a model results in fitted parameters with large confidence intervals, it can present a challenge when interpreting the results, particularly if the changes in these values are small. Consequently, when selecting a model for the analysis of a dataset, preference will be given to models with smaller intervals.

5.2.2 Model Definition

The current transients observed in amorphous silicon oxide devices appear to result from two distinct changes occurring within the device simultaneously, leading to both an increase and a decrease in conductance. The two changes in conductance exhibit a number of distinguishing characteristics.

Firstly, the increase in conductance occurs significantly faster than the subsequent decay. Secondly, in terms of volatility, the increase in conductance relaxes to its initial state within tens of milliseconds, whereas the decay in conductance can take hours to fully reset. Thirdly, in terms of their material dependence, the decay in conductance can be removed by changing the material of the electrodes [170].

This has led to the conclusion that the two changes are driven by different mechanisms and can exist in isolation, which will be detailed in the following sections. The SPICE models for each of these processes will first be presented separately, and then the two will be combined to obtain the final model.

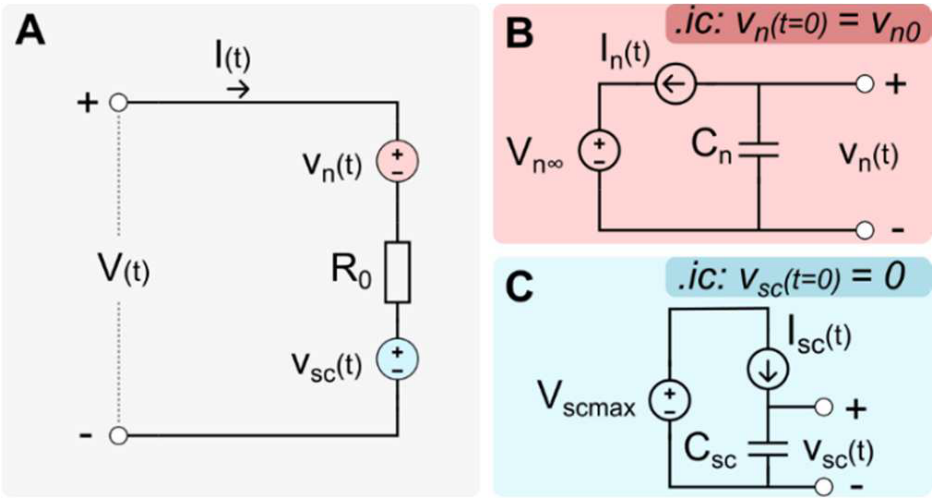

Fig. 5.4 Empirical SPICE Model diagram. (A) The SPICE model produces an output current, $I(t)$, given an input voltage, $V(t)$. The circuit consists of a single resistor, R_0, in addition to two voltage sources which act to reduce or increase the magnitude of the voltage applied to the resistor. These voltage sources are described by the sub-circuits highlighted in blue and red. (B) The voltage source, $V_n(t)$, has an initial value of V_{n0} and reduces over time leading to an increase in device current. (C) In contrast, the voltage source $V_{sc}(t)$ has an initial value of $0V$ and increases over time causing a reduction in the current flowing through the device. All capacitors are set to a value of $1F$, whereas the values of the resistors and voltage biases are obtained from fitting the experimental data.

The accelerated rise in conduction appears to be attributable to the phenomenon of charge trapping. This is corroborated by its shorter timescales, greater volatility, and experiments in which optically injected carriers were demonstrated to accelerate the process. One potential mechanism by which charge trapping affects device conductance is through modulation of the height of a Schottky-like barrier [44], which may be influenced by the population of interface states [240].

Although Schottky barriers are typically formed between metal-semiconductor interfaces, studies have identified analogous barriers within memristor devices [83]. The silicon oxide devices exhibit rectifying behaviour, indicating the presence of an interface barrier at one of the metal-insulator interfaces. Additionally, the filaments in the devices are composed of silicon-rich regions within the oxide, suggesting that the interface between these filaments and metal contacts may resemble a Schottky interface.

The Schottky barrier is modelled as a voltage drop, which acts to reduce the voltage across the active layer of the device, as illustrated in 5.4. As the height of the Schottky barrier diminishes, the potential across the active layer increases, resulting in a greater device current. The voltage drop resulting from the Schottky-like barrier is a function of time and is modelled using the sub-circuit illustrated in red in 5.4.

The magnitude of the voltage drop is represented within the sub-circuit by the voltage across the capacitor, C_n. The initial value of the aforementioned variable is discharged by the current source, V_{n0}. The rate of discharge is defined by the current source, $I_n(t)$, whose magnitude is proportional to the difference between the current voltage drop across the Schottky barrier and its final equilibrium value, $V_{n\infty}$.

$$I_n(t) = \alpha \times [V_n(t) - V_{n\infty}] \tag{5.5}$$

In this equation, α corresponds to the probability of trapping a carrier, while the voltage difference relates to the concentration of unpopulated states. The current source discharges the capacitor to a final equilibrium voltage, $V_{n\infty}$, where the trapping and de-trapping currents are equal. An additional leaking resistor, R_{leak}, is introduced to correctly adjust for the rectifying behaviour of the device.

There is compelling evidence that the observed decline in conductance can be attributed to the movement of charged ions within the oxide thin film. This hypothesis has been put forth in the majority of publications on such current transients [262]. This is largely

based on the observation that the decay in conductance occurs over a time scale of tens to hundreds of seconds, which is too long to be associated with the trapping of electrons or holes.

Furthermore, in accordance with the drifting defect hypothesis, it has been demonstrated that the process can be reversed by applying a voltage of opposite polarity, despite the device exhibiting significantly reduced currents in the opposite polarity. The model adhere to this approach and hypothesise the presence of migrating defects. Of particular significance is the assumption that the ionic current induced by the migrating space charge is negligible in comparison to the electronic currents, as predicted [190].

It is assumed that the electronic current flows through a conductive channel within the oxide. In our silicon oxide devices, this is a silicon-rich filament that forms during electrical stressing. Such filaments are common in memristors and have been observed in our devices using etching C-AFM techniques [19].

In order for an electronic current to be induced through these filaments, it is necessary that a potential be present across the channel. In our model, the migrating space charge modulates the current by reducing the potential experienced along the filament. It is probable that this space charge is a positively charged ion that has accumulated in proximity to the upper gold electrical contact. This assertion is corroborated by the observation that the impact of the accumulation of space charge can be negated by modifying the material of the top electrical contact from gold to indium tin oxide [170].

The precise nature of this ion remains uncertain. A substantial amount of oxygen migration and associated oxygen vacancies have been observed throughout the device when under bias [255], indicating that this could be a viable candidate. However, in other devices, hydrogen has been identified as a defining factor in device behaviour. Additionally, hydrogen has been detected in significant quantities within our devices [138]. Given the lack of certainty regarding the identity of the ion, we assume a generic space charge.

The effect of the space charge on the conductive channel can be described as follows. In the initial phase, the space charge exhibits a homogeneous distribution. Upon the application of a voltage to the device, a force is imparted on the space charge, resulting in its drift and accumulation at the device electrode. This accumulation effectively blocks the space charge from exiting the oxide.

This accumulation results in the formation of a region of higher space charge concentration, which consumes a portion of the potential applied across the device. This results in a reduction in the potential drop across the conductive channel. As the space charge accumulates, this voltage drop increases, meaning less potential is dropped across the channel and a reduction in device current is observed.

Eventually, the drifting force imparted on the space charge will reach equilibrium with the diffusion and Coulombic repulsion formed by the accumulated space charge, leading to a steady state condition. When the potential is removed, the space charge diffuses back to its original distribution.

The aforementioned process is modelled using the circuit illustrated in Figure 5.4. The conductive channel is represented by a fixed resistance, designated as R_0. The voltage across the conductive channel is defined by the applied potential at the terminals of the device, $V(t)$, and the voltage source, $V_{sc}(t)$, which represents the voltage drop caused by the accumulated space charge. This voltage source is time-dependent and is defined by the subcircuit shown in blue. As the voltage of this source increases over time, the voltage across the fixed resistor drops and the device current also reduces.

$$I_{sc}(t) = [V_{scmax} - V_{sc}(t)] \times [\mu\,(V(t) - V_{sc}(t))] \tag{5.6}$$

As illustrated in Figure 5.4C, the sub-circuit meticulously monitors the accumulation of the space charge and its concomitant voltage drop. The voltage across the capacitor is represented by the voltage drop. The charging of the capacitor is effected by the current source $I_{sc}(t)$, the result being the generation of a current in accordance with equation 5.6. The term is defined by the voltage applied across the device and μ, which symbolises the mobility of the space charge. The current source is able to draw charge from a voltage source, which represents the steady-state voltage drop. That is to say, the maximum voltage drop consumed by the accumulated space charge V_{scmax}.

5.3 The Combined Model

To complete the model, the subcircuits for both potential drops are combined as illustrated in Fig 5.4. The two voltage drops act upon a single resistor representing the conductive channel, R_0. In practice, this channel does not exhibit ohmic conduction and thus its resistance will have a voltage dependence. This is taken into account while collecting the meta-parameters.

5.3.1 Parameter Fitting

Table 5.2 Model parameter values.

Parameter	Function	Value
R_0	Exponential	$R_0(V) = 6.00 \times 10^6 \cdot exp(-1.07V)$
α	Linear	$\alpha(V) = 1.09V - 0.454$
$V_{n\infty}$	Constant	$V_{n\infty} = 0.35$
V_{n0}	Linear	$V_{n0}(V) = 1.03V - 0.391$
μ	Second Order Polynomial	$\mu(V) = -0.038V^2 + 0.142V - 0.057$
V_{scmax}	Exponential	$V_{scmax}(V) = 5.17 \times 10^{-2} \cdot exp(2.01V)$

The complete model comprises six parameters, which are listed in Table 5.2. These parameters will be fitted to the dataset plotted, which shows the current transients observed in a single device at various voltages ranging from 0.7 V to 1.1 V.

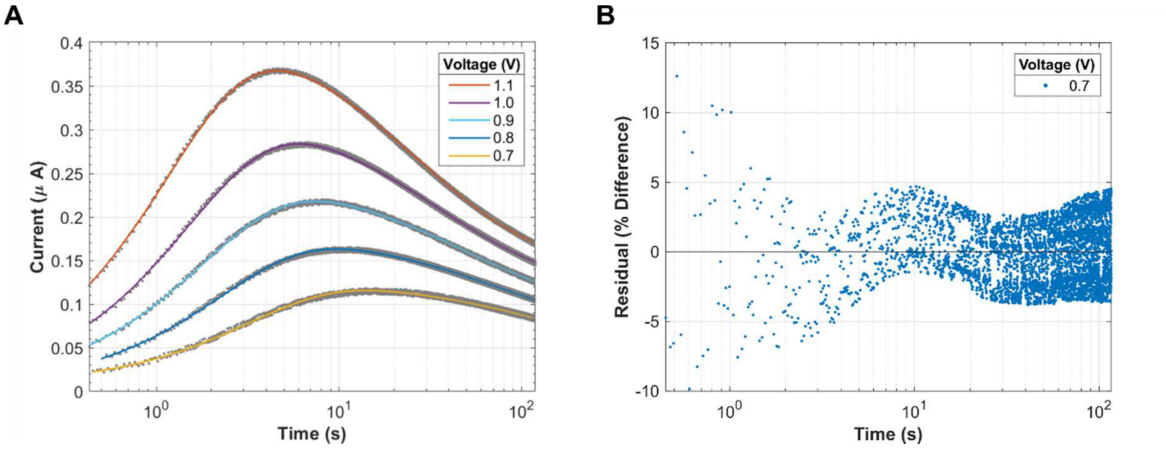

Fig. 5.5 Fitting performance of the SPICE model. (A) The model's response is compared with experimental data for a range of voltages. Experimental datapoints are indicated with gray points while model predictions are plotted with solid coloured lines. For each voltage case, the parameters in table 5.2 are refitted to minimise the mean-squared-error between the experimental and model's response. (B) The residuals between the model's prediction and experimental data is plotted for $V = 0.7V$ which exhibits the largest error of 13% at the onset of the transient which is predominantly due to noise within the experimental data. In contrast, the average percentage difference for the duration of the transient is 2% with absolute residuals never exceeding $5nA$.

Initially, the model is individually fit to each voltage trial and will then be generalised in a later stage. Fitting is carried out in MATLAB using the `fminsearch` function which employs the simplex search method [138] to minimise an error function which has been set to the

mean-squared-error.

The process of fitting individual voltage trials results in a set of parameters which exhibit voltage dependencies except for one exception, $V_{n\infty}$, which is spread around a central value for different voltages and so is assumed to be independent of the applied voltage. For this parameter we calculate the mean average of its value across the dataset and then set it as constant, 0.35V.

The prediction of the model when parameters are reoptimised for each specific voltage case is plotted in Fig. 3A. As illustrated in Figure 3B, an example residual for the case of 0.7 V is presented. This figure demonstrates the presence of suboptimal residuals, attributable to the diminished device currents. It is noteworthy that a maximum error of 13% is exhibited at the transient's onset. Nevertheless, these errors decrease quickly, resulting in a mean error of only 1.9% over the duration of the transient.

For all voltage cases, the absolute error remains less than 5nA throughout the transient. This appears to be a reasonable fit. However, a persistent oscillation is observed in the residuals of all voltage cases, suggesting that the model is missing some dynamics within the current transient. It is evident that this approach to fitting will result in an optimal fit, owing to the re-optimisation of parameters for each voltage case.

It should be noted that the model has not yet been generalised. In its present state, it is necessary to modify the SPICE model's parameters in accordance with the voltage applied to the device. The ideal scenario would involve the development of a generalized model, in which a single set of parameters is applicable to a range of applied voltages.

In order to generalise the model, it is necessary to modify it in order to account for the voltage dependency of each parameter. It is inevitable that this process will introduce error into the model's behaviour and increase fitting residuals. Nevertheless, the capacity to predict device behaviour over a range of voltages using a single set of parameters justifies this reduction in accuracy.

For each of the voltage-dependent parameters, the parameter values are fitted to a suitable function, as indicated in Table 5.2 and illustrated in Fig. 5.6A-E. The extraction of the coefficients that describe the voltage dependency, for example the gradient and offset of a single order polynomial, is achieved from these functions. The term 'meta-parameters' is

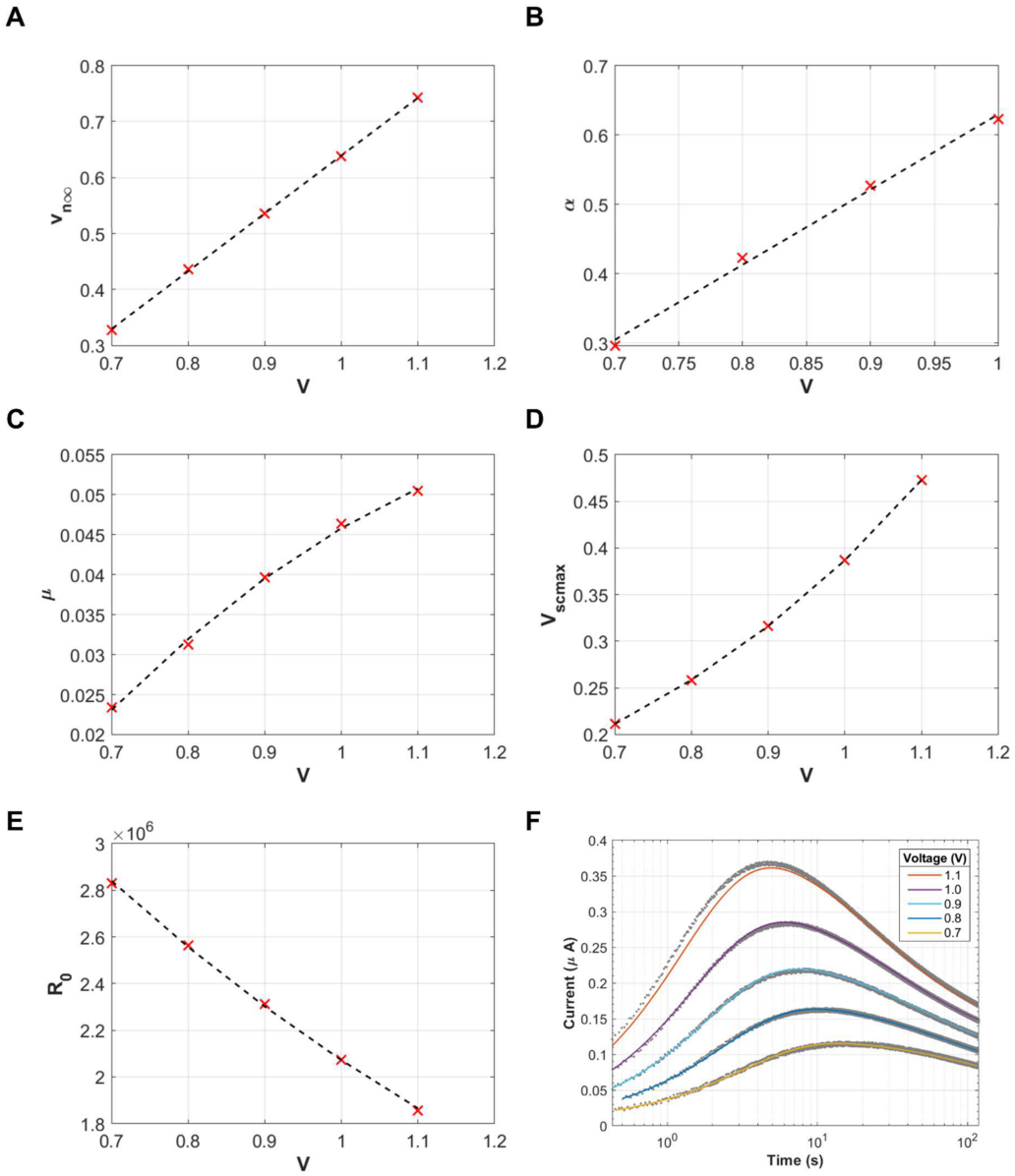

Fig. 5.6 Fitting of the model's meta-parameters. (A-E) The fitting for each of the model's parameters is plotted using the equations described in Table 5.2. These functions' coefficients are referred to as meta-parameters and are used in the generalised model to produce a black box model that reproduces the current-time response to any voltage between 0.7 and 1.0 V. (F) The generalised model's response is plotted alongside the experimental data. This model uses the previously described meta-parameters. However, the generalised model breaks down at larger voltages, i.e. at 1.1 V. We assume this is due to the additional stressing of the oxide that occurs at higher voltages and device currents.

employed to denote these coefficients, as they encompass the parameters obtained previously from fitting the experimental data.

The predictions of the generalised model are illustrated in Figure 5.6F. The discrepancies between the experimental and simulated device currents for each voltage curve have increased in comparison to the voltage-specific parameters. However, the advantage of the generalised model is that its parameters no longer require refitting for different applied voltages.

A substantial departure from the established model is observed for the case of 1.1V. As demonstrated in Figure 5.6F, the model reliably forecasts a reduced device current during the initial phase of the transient. This phenomenon can be attributed to the elevated applied voltage and the resultant currents, which induce electrical stress within the oxide.

It has been established that these devices exhibit an increase in conductivity in response to electrical stress [171]. This can be rectified by the introduction of a stressing term, which serves to augment the number of traps within our representation of the Schottky interface. The discharge current, as outlined in equation 5.5, is modified to encompass a term that facilitates the charging of the capacitor.

$$I_n(t) = \alpha \times [V_n(t) - V_{n\infty}] - \sigma I(t) \tag{5.7}$$

This modification is employed to simulate the creation of additional trap sites. The term is proportional to the current flowing through the device, σ, which represents the probability of defect creation for a given magnitude of current. This process ultimately yields the result depicted in equation 5.7.

The result of introducing this additional stressing term is illustrated in Figure 5.7. The two models are contrasted both with and without the stressing term, and it is evident that the introduction of the stressing term, where $\sigma = 5.61 \times 10^5$, improves the model's performance in the first half of the transient. It is important to note, however, that the model is no longer generalized because this stressing term does not apply to smaller voltages.

The SPICE model proposed in Figure 5.4 can therefore be generalised for silicon dioxide devices for voltage ranges within 0.7V and 1.0V. It is imperative that the model be modified to account for the additional stressing occurring within the oxide. This can be achieved by introducing a stressing term, as outlined in equation 5.7. Nevertheless, the validity of this

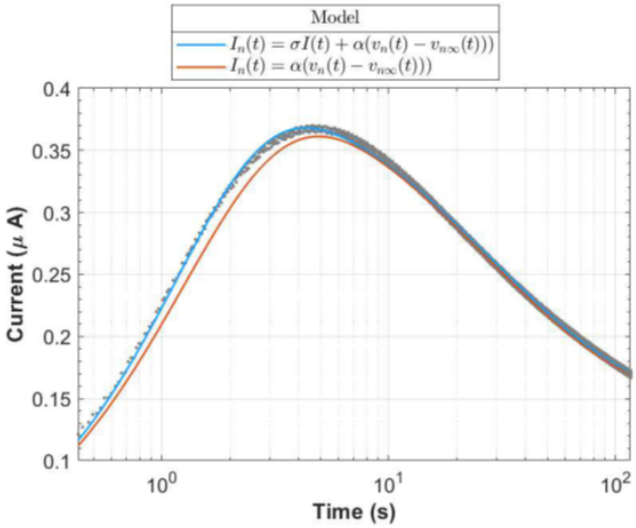

Fig. 5.7 The improvement of SPICE fit at higher voltages. Two versions of the The following generalised models are plotted for 1.1 V. As indicated in the legend, the model incorporating an additional stressing term, $\sigma = 5.61 \times 10^5$, demonstrates enhanced performance at elevated voltages in comparison to the model devoid of the stressing term. The necessity for this term can be attributed to the hypothesis that stress is beginning to occur within the oxide at this higher voltage.

stressing term remains unvalidated for voltages in excess of 1.1V. Consequently, it is highly probable that additional modifications will be required to ensure accurate stressing.

5.3.2 Model Performance

In its current configuration, the model operates within a constrained voltage range to prevent significant electrical stressing of the device. For example, device characterization has been limited to voltages not exceeding 1.1V. If the voltage is increased beyond this threshold, stress-induced effects alter the shape of the current transient, rendering the existing model invalid. The appropriate voltage range can be determined empirically by performing multiple trials at a fixed voltage. If a gradual increase in conductance is observed across trials, it indicates the onset of stressing, and the voltage should be reduced accordingly

To extend the model's applicability to more general inputs—beyond simple step potentials—it is essential to account for the relaxation behavior of the current transient. This can be achieved by adding leakage components R_{leak} to the charge-trapping and space-charge subcircuits, allowing their associated capacitors to charge and discharge over time.

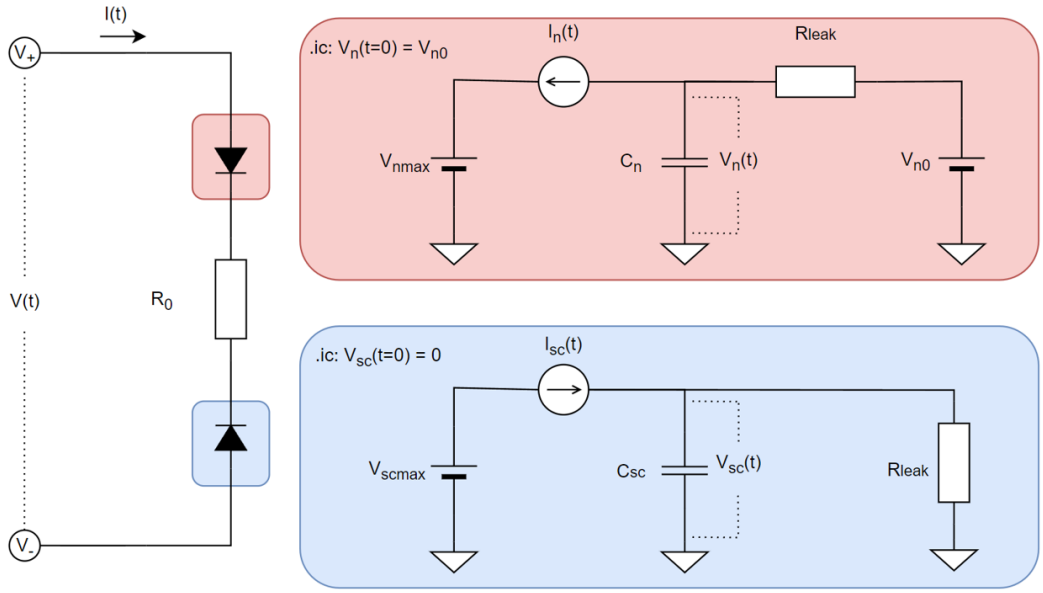

Fig. 5.8 The extended SPICE Model diagram. The improvement of SPICE fit at higher voltages. The SPICE model generates an output current, *I(t)*, based on an input voltage, *V(t)*. The circuit incorporates a single resistor, R_0, along with two diode models that serve to attenuate or amplify the voltage applied to the resistor. These diode models are represented by the sub-circuits highlighted in blue and red.

The enhanced model shown in Figure 5.8 builds upon earlier designs by replacing static voltage sources with diode-based elements that better replicate Schottky-like metal-insulator contact behavior. It also introduces dynamic relaxation features that emerge when input potentials are removed or the device is grounded.

$$I_d(t) = (I_0 + I_\delta V_n(t)) \times \left(e^{\frac{V(d_1,d_2)}{nV_t}} - 1 \right) \tag{5.8}$$

Where $I_d(t)$ is the time-dependent diode current, which captures how the current evolves over time due to transient phenomena or dynamic voltage input. I_0 is the reverse saturation current or the leakage current through the diode when reverse-biased. It is an intrinsic property of the junction and depends on factors like material properties and temperature.

$I_\delta V_n(t)$ is the modulated current term, a dynamic extension not present in the standard Shockley equation. It introduces a time-varying modulation of the saturation current, where I_δ is a scaling factor representing sensitivity of the diode to input modulation and $V_n(t)$ is a time-dependent voltage signal driven from an internal node. Together this term allows the diode behavior to adapt dynamically to changes in the input signal, which is crucial for

modeling transient memristive phenomena such as relaxation and adaptation.

$V(d_1,d_2)$ Voltage difference across the diode terminals in the SPICE model. This is equivalent to the V in the traditional Shockley equation. This voltage controls the exponential response of the diode. n Ideality factor that reflects how closely the diode follows the ideal Shockley behavior. An $n > 1$ accounts for recombination losses or non-idealities in real diodes and is often used to better fit experimental data. Finally, V_t is the thermal voltage.

To derive a form that better matches experimentally observed transient behavior from the original (5.5), it is assumed that $V_{n\infty} = V_{scmax}$, a simplification that assumes both voltages reflect the same saturation point for field-driven relaxation. Moreover, since the decay dynamics of $V_n(t)$ are influenced by the total device current, a modulation term by the main current $I_d(t)$ from (5.8) is introduced, leading to the modified expression:

$$I_n(t) = \alpha I_d(t)\,[V_{scmax} - V_n(t)] \tag{5.9}$$

This equation captures both the deviation from equilibrium and the feedback from the instantaneous device current, providing a more accurate representation of the charge relaxation behavior observed in subthreshold memristive devices.

To simplify the original space-charge current (5.6) expression for modeling in SPICE, the voltage driving the ionic drift as the terminal voltage is approximated as $V(d_1,d_2)$. Additionally, the modulation term $[V_{scmax} - V_{sc}(t)]$ can be treated as constant or absorbed into the mobility factor μ in certain operating regimes. This leads to the simplified form:

$$I_{sc}(t) = \mu V_{sc}(t) \cdot V(d_1,d_2) \tag{5.10}$$

which effectively captures the relationship between accumulated space charge potential and the applied terminal bias. These modified equations are better suited for implementation in compact SPICE models used to simulate transient behavior in neuromorphic circuits.

Table 5.3 Extemded model parameter values.

Parameter	Function	Value
I_0	Exponential	$I_0(V) = 2.39 \times 10^{-9} \cdot e^{-13.4V} + 1.03 \times 10^{-13} \cdot e^{-1.53V}$
I_δ	Power	$I_\delta(V) = 7.41 \times 10^{-15} \cdot V^{-2.91}$
n	Polynomial	$n(V) = -0.131V^2 + 1.01V + 0.615$

The initial equation fitting for each trial yielded small residuals, indicating a satisfactory fit, albeit only within the noise range and limited to individual voltages. The model was then generalised, with parameters plotted across various voltages. This resulted in a set of empirical functions (hyperparameters) that account for voltage dependency in table 5.3. This approach facilitates the simulation of blackbox circuits and consequently identifies novel diode parameters that augment the previous model's fit, as seen in Figure 5.9.

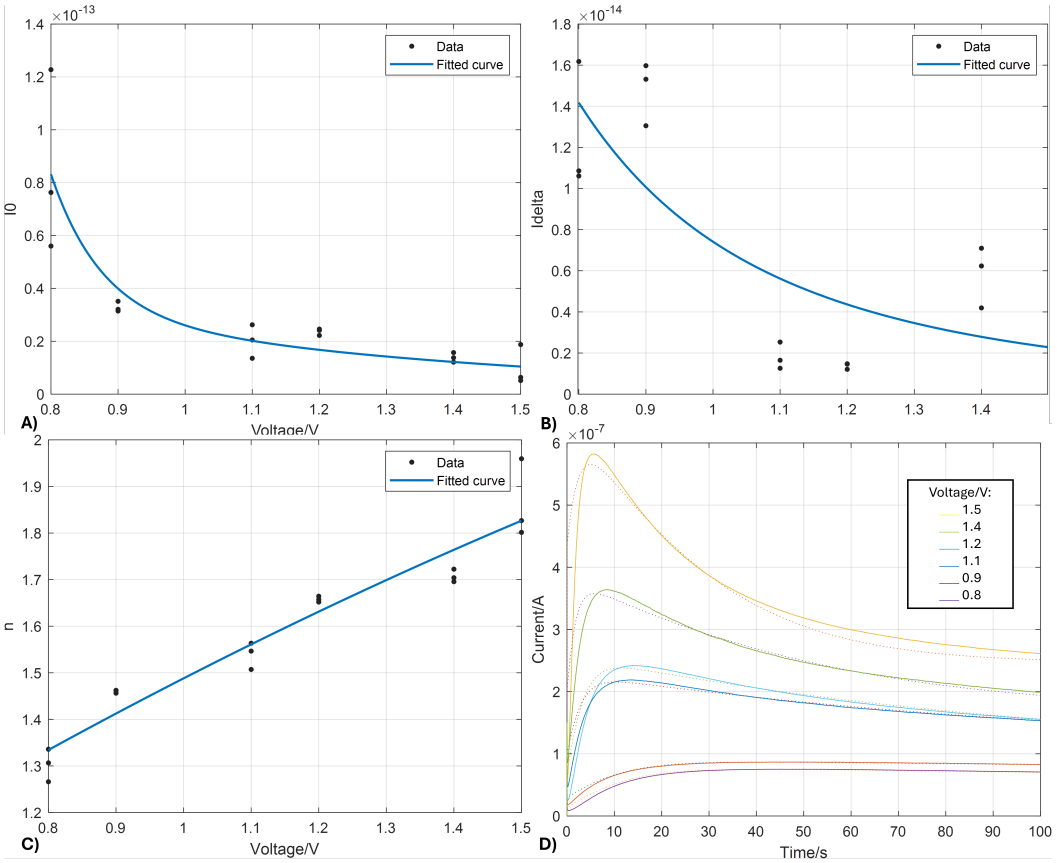

Fig. 5.9 Fitting of the extended model's meta-parameters. (A-C) The fitting for each of the model's parameters is plotted using the equations described in Table 5.3 that reproduces the current-time response at higher voltage. (D) The generalised model's response is plotted alongside the experimental data from 0.8V to 1.5V.

Finally, Figure 5.10 illustrates the behavior of the extended SPICE model incorporating relaxation dynamics, particularly under pulsed input conditions. Each pulse generates a sharp rise in current followed by a decay, reflecting the transient response of charge trapping and space-charge mechanisms.

Over successive pulses, the peak current gradually decreases, demonstrating the cumulative effect of relaxation when the step potential is removed between pulses. This decay mirrors experimentally observed behaviors in memristive devices operating in the subthreshold regime, where ionic and electronic processes relax during the off periods. The extended model captures this temporal evolution, offering improved accuracy for simulating pulsed neuromorphic stimuli.

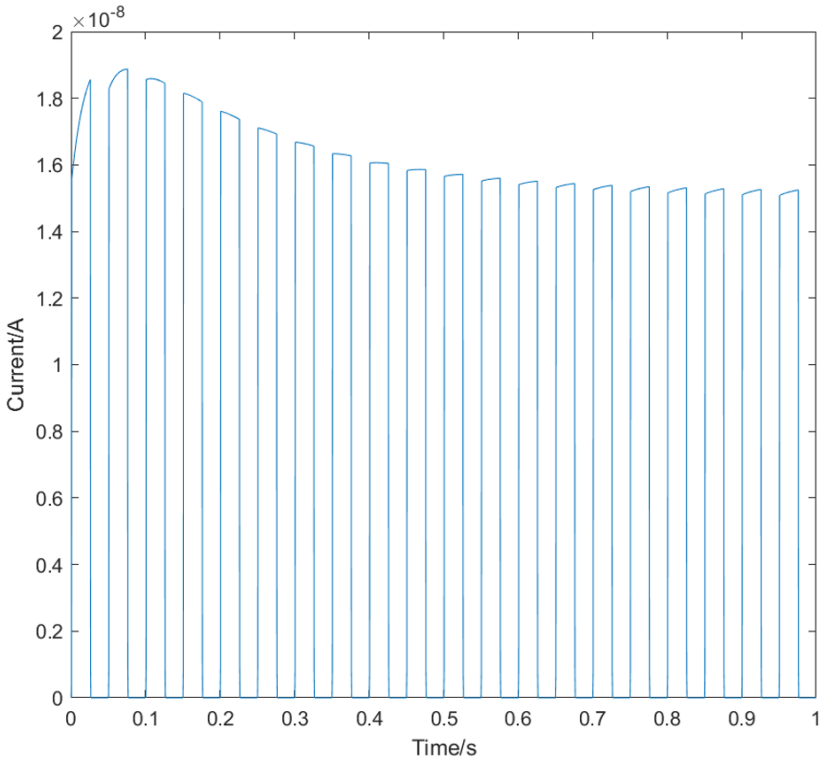

Fig. 5.10 Simulated current transient response under repeated voltage pulses using the extended SPICE model. The gradual decay in peak current illustrates the relaxation dynamics of charge trapping and space-charge effects when step potentials are intermittently removed, consistent with subthreshold memristive behavior.

5.4 Summary

This chapter has introduced a comprehensive framework for modeling current transients in silicon oxide-based memristive devices, with a particular focus on their relevance to neuromorphic applications. A key finding is the discovery that modifying the top electrode material—specifically, replacing gold with indium tin oxide (ITO)—leads to a significant change in transient behavior.

While gold-contacted devices exhibit both the characteristic increase and subsequent decay in conductance, ITO-contacted devices demonstrate only one phase strongly. This divergence provides strong evidence that the current transient is governed by two independent physical processes: a rapid increase in conduction attributed to charge trapping at Schottky-like barriers, and a slower decay driven by the accumulation of mobile ionic species, likely protons, near the top metal-insulator interface.

To accurately simulate these dual-phase transients, an extended SPICE model was developed that integrates two time-dependent voltage sources acting on a shared resistive path. The model incorporates leakage terms and relaxation dynamics, enabling it to replicate the experimentally observed response to voltage pulses with high fidelity. The inclusion of both fast and slow processes allows the model to track current evolution under biologically relevant pulse conditions, making it suitable for spiking neural network (SNN) simulations.

Through voltage-dependent parameter fitting and the addition of a stressing term at higher voltages, the model is generalised to predict device behavior across a practical operating range (0.7-1.5V). These results not only validate the physical hypotheses underpinning transient dynamics but also highlight the model's applicability in neuromorphic computing tasks, such as synaptic emulation and event-based learning.

Chapter 6

Biorealistic Computing

6.1 Spiking Deep Networks

Traditional neuromorphic circuits consist of two main components: neurons and synapses [180]. Neurons are active devices that act as decision-making units. When the input threshold is surpassed, an acceptable condition is met, and a voltage spike is generated. Synapses are usually passive elements that connect neurons with each other. This is also where primary calculations of the inputs are performed, as well as a potential storage mechanism.

The neurons and synapses are linked to create a neural network, which is used in all modern machine learning methodologies. Rather than responding to continuous signals, these networks communicate information through spikes. Spike trains convey data in both the form and timing of the spike, making them ideal for implementing synaptic learning principles. It is worth noting that this method adheres to the typical framework within the field, however, it presents a highly simplified conception of the biological system.

Modern artificial neural networks and neuro-computing architectures usually neglect the principles of neuroscience [206]. As a consequence, essential elements of the organic cerebral processing systems are either disregarded or overlooked. The aim of biorealistic approaches is to mimic the functions of computational cells using electronic devices. The creation of these circuits in CMOS was traditionally known as neuromorphic engineering, with backpropagating networks being a more direct influence from nature.

6.1.1 Neural Computing Nomenclatures

Deep artificial neural networks (DNNs), particularly convolutional neural networks, represent a significant triumph in the field of modern computer vision. They have demonstrated remarkable efficacy in recognizing a diverse array of objects within expansive, intricate images [135]. However, these networks have been engineered for and operate exclusively on rate-based neurons. The question of how they can be executed on spiking neurons represents an emerging frontier of investigation.

The question arises as to why such networks should be run in spiking neurons. There are two principal motivations behind the creation of deep spiking networks. The first is to enable the operation of some of the large CNN models that have recently demonstrated success in numerous object recognition and other tasks on spiking neuromorphic hardware. This will facilitate the development of energy-efficient systems capable of performing object recognition in real time on robotic platforms, where current technology is too energy-intensive to allow for deployment on mobile robots, for example.

The second motivation is to incorporate additional brain-like components into machine learning models. The field of neuroscience presents a multitude of distinctive challenges pertaining to the mechanisms of learning in the brain. These include the complexities associated with the nonlinear characteristics of neurons, particularly in relation to their firing thresholds, as well as the intricacies of spike-based communication and its inherent discreteness and variability.

Although the spiking deep networks presented in this chapter are not designed to be models of brain-like learning processes, the challenges addressed here are also faced by the brain. Some of the ideas presented in this section provide insights that motivate the development of more biologically plausible learning mechanisms.

Spiking deep networks facilitate the transmission of information between neurons in the form of discrete spikes. The initial distinction to be made regarding spiking networks is that they encompass an additional temporal dimension, which is not typically present in rate-based DNNs. In other words, a spiking neuron is a process that evolves over time, sometimes emitting spikes, sometimes not.

It is only possible to discuss this process over time; examining it in one instant provides little insight. In contrast, in rate-based networks, we typically present an input and can

instantaneously determine the activities of each subsequent neuron in the network, since they do not change over time.

A second notable distinction is that spikes are discrete and identical in nature. The sole information conveyed by a spike is the time at which it occurred. As previously discussed, this indicates that there are two principal categories of codes that a neuron can utilise: rate codes and timing codes.

In the case of a timing code, the focus is on the times of individual spikes. In contrast, if a rate code is employed, the focus shifts to the number of spikes occurring within a specific time window, or potentially the relative timing of spikes in relation to one another.

When examining the number of spikes within a specified time interval, it becomes evident that the resulting rate is inherently discrete. If there are n spikes within the t-second window, the firing rate can be expressed as n/t, where n is an integer. For a fixed window t, the firing rate can only assume a discrete set of values, specifically all the integer values of n.

A third distinction pertains to the inherent variability of the output of spiking neurons, which differs from that of their rate-based counterparts. In the event of a constant input, a rate neuron will provide a constant output, that is to say, the firing rate corresponding to that input.

It should be noted that rate neurons are capable of exhibiting internal dynamics, as exemplified by the adapting version of the rate-based LIF. Consequently, when presented with a constant input, the output of these neurons will not be constant. Nevertheless, their output remains considerably less variable than that of their spiking counterparts.

In contrast, spiking neurons will output a spike train, which, when filtered by a synapse, results in an oscillating signal whose variability depends on the firing rate. Consequently, even when presented with a constant input, a spiking network will exhibit variability in the inputs to each neuron and the outputs of the entire network.

DNNs are typically formulated as rate-based models, wherein the nonlinearity activity is understood to represent the firing rate of a neuron. To illustrate, a ReLU can be conceptualised as a neuron that is silent in the event that its input is less than zero, and whose firing rate increases in a linear fashion as the current rises above zero.

Moreover, cost functions associated with rate-based DNNs are based on the firing rates of the output neurons. In the context of classification, the chosen class is the output unit with the highest activity. One approach is therefore to treat spiking networks similarly and train the network so that the unit corresponding to the target class will have the highest activity.

It should be noted that this activity does not necessarily correspond to a neuron firing rate. Indeed, it is more likely to be the filtered, weighted sum over the spiking activities of the final layer of neurons in the network. Alternatively, the network can be trained so that the first neuron to spike will be the chosen class.

At the level of a single neuron, this distinction is no longer applicable. The activity of a neuron firing at a regular rate can be captured by its inter-spike interval, defined as the time between one spike and the next. Assuming the neuron is in a resting state, the inter-spike interval is equivalent to the time before the first spike of a neuron, plus the refractory period. The key design decision is how information is transmitted between neurons.

6.1.2 Memristive Frameworks

Modern deep learning algorithms are subjected to neuroscience-inspired restrictions by Spiking Neural Networks (SNNs) [243], which have shown notable increases in runtime efficiency [263]. Neuromorphic hardware has demonstrated considerable reductions in latency and energy consumption [284], by switching from full precision and fixed precision activations of artificial neuron models to temporally-encoded data representations collected by spiking neurons [305].

The considerable success of error backpropagation in training deep learning models has led to the development of numerous related training algorithms tailored for spiking neural networks (SNNs) [275]. These algorithms, which are guided by surrogate gradient descent, have been designed to address the non-differentiability of discrete spikes, which is a limitation of traditional gradient-based methods [198]. This proliferation of SNN usage is accompanied by the development of modular deep learning programming packages [86] that have optimised autodifferentiation for CUDA acceleration [203].

In parallel with these advances in training SNNs, the past decade has seen significant developments in brain-inspired devices, circuits, and architectures that integrate neuronal dynamics to enhance the hardware integration of SNNs and their constituent parts. Memristors and resistive RAM (RRAM) constitute a significant aspect of the exploratory research conducted

in the field of SNN implementation [39], as they serve as a natural conduit between SNN algorithms and accelerators [42]. They have been extensively utilized as both synapses and as spiking neurons.

At the ionic level, memristive synapses have been integrated into systems that naturally implement the spike-timing-dependent-plasticity (STDP) update rule using higher-order device dynamics [222], as evidenced by the literature [157]. An alternative application of ion-driven dynamics is the implementation of the memristor as a neuron, where nonlinear conductance evolution gives rise to abrupt switching that can be used to emit sudden voltage spikes [156].

This approach [11] is typically coupled with capacitive integration and has been referred to as a 'neuristor' [49], and a 'Memristive Integrate-and-Fire' (MIF) neuron [84, 121, 306]. Similarly, the leakage of ions through the membrane of biological neurons can be implemented using resistive dissipation [162], as in neuristors, as observed in nanowire networks [93, 307], or via the dynamic movement of ions in single devices [308].

At the architectural level, RRAM has been identified as a promising candidate for in-memory compute (IMC) architectures [151] due to its capacity to parallelise matrix-vector multiplication independently of time complexity when integrated as large-scale, modular arrays [60].

In contrast to the mapping of neurons, memristive synapses map neural network weights to device conductances. In general, RRAM IMC architectures are designed to be trained offline with weights mapped on-chip for inference and deployment [301]. Consequently, RRAM synapses should be stationary and only used for weight read-out. Higher-order dynamical behaviours of memristors are abstracted away and treated as non-idealities.

An additional challenge associated with RRAM-based IMC is the cost of communicating analog current signals along lengthy bit-lines and conversion into the digital domain. These issues have prompted the utilisation of binary activations in the form of spike-based IMC accelerators, which have been demonstrated to mitigate the challenges associated with mixed-signal computation by eliminating the necessity for extensive Analog-to-Digital Convertor (ADC) data conversion [61].

The majority of deep learning acceleration using memristors can be classified into one of the aforementioned categories: memristive neurons, memristive synapses that learn via

associative learning, and IMC accelerators. A limited number of designs have integrated memristive neurons and memristive synapses [268]. This is a praiseworthy achievement, as the intrinsic switching dynamics of memristive systems are leveraged to accomplish data-driven operations [242].

The consequence of allowing hardware to behave naturally is that a designer is no longer able to rely on synchronous, clock-driven processing and is susceptible to fault injections resulting from nonlinear ionic dynamics. Allowing the intrinsic dynamics of memristive hardware to 'teach itself' serves to exacerbate the challenges associated with training MSNNs.

This limitation has restricted the demonstration of MSNNs to unsupervised learning tasks that have been shown to solve simple, low-dimensional pattern recognition problems via local learning rules (typically STDP) and associative learning. Such tasks include the classification of a variety of characters and numbers, including a subset of the MNIST dataset.

A vast array of work has been conducted which integrates memristors with brain-inspired architectures [121]. This spans from low-level analogue action potential emulation to discrete spiking dynamics and non-spiking IMC processors [61]. The focus here is on prior work which uses nonlinear dynamics in memristive neurons together with memristive synapses, which also includes an associated demonstration of synaptic optimisation to achieve a data-driven outcome.

A fully memristive neural network (MSNN) is defined as an array that employs the nonlinear switching dynamics of memristors to trigger action potentials, with memristive weights utilized as neural network parameters. The 8×8 crossbar array presented [268] has been demonstrated to integrate a fully MSNN, including memristive synapses and neurons.

The synaptic array has been trained using unsupervised STDP to classify four letters in a 24-pixel grid. While the task achieved is considerably simple, the fully memristive experimental demonstration paves the way for the development of new training methods.

Another work employs the use of half-wave rectification [129], situated between crossbar arrays, to facilitate the processing of ReLU activation within the analog domain. Although not fully memristive nor a 'spiking' network, this approach offers a compelling illustration of the potential for successive analog activation transfer between RRAM crossbars, obviating

the need for intermediate data conversion.

This process bears resemblance to the transmission of analog action potentials between layers in biological systems. The training procedure employs gradient-based optimization, incorporating device non-idealities during the forward pass. This strategy has yielded a test set accuracy of 93.63% on the MNIST dataset.

In the referenced literatures, convolutional SNNs with memristors are employed [261], with both networks having undergone pretraining as non-spiking networks, which are then mapped or converted into the spiking domain [276]. Both networks exhibited satisfactory accuracy on the MNIST dataset; however, they did not demonstrate the capacity to process more complex, real-world data.

This discrepancy may be attributed to the significant disparities between the networks that underwent training and the MSNN that was implemented. A dense MSNN is adopted using a similar approach to that can be used here [54], and thus has minimal hardware requirements at run-time. The training process translates the switching dynamics of the memristive neuron into a firing rate, which may be the reason why a relatively low accuracy of 83.2% was achieved on the MNIST dataset.

The majority of these works present persuasive evidence utilising in-house fabricated arrays [191], either as standalone crossbars or as back-end-of-the-line (BEOL) integrated arrays with foundry-made chips. In contrast, the objective here is to utilise bespoke fabrication capabilities. Previously, memristors were employed solely in the forward pass, as their devices are not designed to be reprogrammed during inference. Consequently, their method does not necessitate switching to generate spiking dynamics.

Gradients can therefore be deterministically calculated partially off-chip. An alternative approach that harnesses memristive dynamics in the forward-pass computation in the network. Consequently, the MSNN approach can leverage the benefits of spike-based processing, such as sparse processing and lower data collision rates.

In order to facilitate and emulate the training process of memristive networks, a variety of valuable frameworks have been developed, each addressing specific niches within the field. These include MemTorch [139], NeuroSim [36], and the IBM Analog Hardware Acceleration Kit [213], which implement non-spiking networks that adopt mixed-signal bit-line

charge/current accumulation/summation processing.

In these simulators, memristive dynamics are accounted for during weight updates and otherwise fixed during inference. To complement these tools, NeuroPack [100] specifically targets the simulation of spiking networks, where memristive dynamics are also factored in during the weight update process and fixed during inference. Spiking dynamics are triggered by pulse-based input voltages.

In terms of hardware implementation, the conventional use of RRAM in circuits often necessitates a considerable amount of overhead to convert analogue currents into digital voltages, which in turn results in a significant power consumption [22]. In many instances, the power and area demands of the ADCs and digital-to-analogue converters (DACs) exceed the overhead brought on by RRAM, thereby negating the advantages of memristors. In contrast, spike-based approach eliminates the need for ADCs and DACs, thereby substantially reducing the cost of peripheral circuits.

6.1.3 Analogue Hardware Challenges

Novel computer hardware solutions that employ analogue devices still exhibit limited precision and unreliability. However, both physical and algorithmic techniques can be employed to mitigate these issues. In contrast to digital technology, the analogue approach inherently involves a degree of imprecision.

Analogue devices, such as RRAM, are susceptible to a number of issues, including stuck states, device-to-device variability and I-V nonlinearity. However, the development of advanced fabrication methods and circuit-level optimisations has enabled the mitigation of some of these non-idealities, while algorithmic techniques have also been shown to be effective in reducing their impact.

In contrast to the digital paradigm, where a multitude of physical imperfections are effectively concealed within a bit representation (either '1' or '0'), analogue electronics is confronted with significant challenges due to the intrinsic imprecision associated with non-discrete systems. Even with a minimal amount of non-idealities, it is challenging to encode information with perfect precision using an exact conductance value.

However, non-idealities do exist and can result in significant deviations from ideal behaviour. These include the device becoming stuck in certain conductance states, undergoing changes

in conductance over time, showing non-linear current-voltage characteristics, or displaying non-linear conductance modulation in response to voltage stimuli.

It could be argued that analogue computing's more fundamental challenge lies in its reduced precision compared to digital computing, especially when digital systems utilise 16 or more bits of representation. While these issues may be grounds for disqualification in many applications, this may not be the case for machine learning applications, which often employ reduced precision computing, even within digital systems.

In general, machine learning models demonstrate a degree of robustness to minor alterations, such as the presence of noise [38]. In the event of significant deviations from ideal conditions, hardware imperfections may result in a decline in accuracy. However, this does not necessarily render the system inoperable. It is, therefore, crucial to comprehend the impact of non-idealities and to ascertain how they can be effectively mitigated.

In the context of linear algebra applications, a proportional relationship between voltage and current (i.e. Ohmic behaviour) is the preferred option. This is due to the fact that Ohm's law is employed in the implementation of multiplication, as previously discussed. Nevertheless, exceptions to this linear relationship do arise, particularly in the case of high-resistance devices [185].

A number of approaches exist at the device and circuit level that facilitate the resolution or even circumvention of the issue of nonlinearity. During the fabrication of RRAM devices, the adoption of a hot-forming step can result in the generation of more linear characteristics [238].

In the case of individual device programming, the adoption of a transistor-to-resistor ratio (1T1R) architecture can facilitate the precise tuning of memristor conductance, despite the presence of any I-V nonlinearities [148]. Alternatively, a charge-based accumulation approach can be employed, wherein a constant voltage is applied, but the input is encoded into pulse width [7]. This eliminates the dependence on the shape of the I-V curve.

It is possible that some memristive devices may become fixed in a specific conductance state. This phenomenon has been observed following processes such as electroforming, as well as after several successful programming cycles [116]. In general, the greater the discrepancy between the intended and actual conductance, the greater the potential for adverse effects. Therefore, it is of paramount importance to identify methods for the prevention or mitigation

of faulty devices.

The overall effect of a device becoming stuck is contingent upon the behaviour of other devices, and thus this phenomenon can be employed to mitigate the negative effects. To illustrate, if a device becomes stuck, its negative effect may be counteracted by adjusting the conductance of another device in the differential pair [158]. On occasion, such an adjustment may occur accidentally, whereby both devices in a differential pair become stuck simultaneously.

As an alternative, if faulty devices can be identified prior to programming, more sophisticated mapping strategies can be employed. The most significant weights can be mapped onto crossbar rows and columns with the lowest incidence of stuck devices [72]. The most significant terms refer to the weights that could have the greatest impact on accuracy. One way to identify such weights is to calculate of sensitivity $\Delta w_{i,j} := -\eta \frac{\partial E}{\partial \Delta w_{i,j}}$, for each weight $w_{i,j}$, where E is the back-propagated loss at the current neuron and η is the learning rate.

The term 'limited dynamic range nonideality' is used to describe a situation whereby the $\frac{G_{on}}{G_{off}}$ ratio is relatively small, which can ultimately result in a reduction in effective precision. In the context of other non-idealities, such as device variability, limited dynamic range can result in a reduction in the number of distinguishable states that are available.

If each state is associated with a certain amount of absolute variability, it is evident that a larger dynamic range is preferable, as it allows for a more effective differentiation between those states that are less distinct. The impact of the dynamic range is highly dependent on the specific application. Should one desire to utilise analogue arrays for the storage of digital information, an enhanced dynamic range will facilitate a greater precision in the number of equivalent bits.

Nevertheless, when considering the acceleration of linear algebra operations (and, by extension, machine learning), such comparisons cannot be made with the same degree of ease. As these hardware accelerators are based on analogue computation, the concept of 'bits' – although potentially useful – does not apply directly. In analogue contexts, an error is defined as any deviation from the intended value. The magnitude of the error is the key factor in determining the severity of the mistake.

In the context of inference applications, a large dynamic range is not a crucial factor. If a naive mapping scheme is employed whereby the value of a weight is represented using a single conductance value, then the inaccuracies produced by this imperfect mapping can be addressed with a $\frac{G_{on}}{G_{off}}$ ratio of as low as 3 [184]. In other contexts, the impact of limited dynamic range cannot be evaluated without first understanding the nature of other non-idealities, namely the deviations they cause.

Line resistance represents a non-ideality that arises from the presence of non-zero interconnect resistances in crossbar arrays. In the event of its presence, this results in discrepancies from the ideal computation of vector-matrix products. While the impact on accuracy can be significant, there are both physical and algorithmic techniques that can be employed to mitigate it.

One of the most straightforward methods for mitigating the impact of line resistance is to enhance the ratio between the resistance of the devices and the resistance of the interconnecting wires. As resistance is inversely proportional to the cross-sectional area of the wire, one method of reducing interconnect resistance is to increase the width of the wires [147].

However, this can be challenging in dense arrays, and an alternative approach is to use more conductive materials, for example, $2nm$ platinum nanofins [207]. Another approach is to increase the resistance of the crossbar devices, although this can sometimes result in less stable device behaviour.

At the circuit level, a variety of techniques may be employed which utilise the systematic properties of line resistance in different ways. For instance, a technique designated as double biasing can facilitate a more symmetrical distribution of electric potentials within crossbar arrays, thereby attenuating the impact of line resistance effects [96]. As the size of the crossbar array increases, voltage drops tend to accumulate.

Therefore, splitting up the array into smaller units [285], or even organising them in three-dimensional structures [287] can help to mitigate this issue. In considering the specific applications for which crossbar arrays are to be employed, algorithms may be deployed to ascertain optimal mappings from software parameters to physical quantities, such as voltage and conductance. A nonlinear mapping from weights to conductances can be employed to counteract the detrimental effects of line resistance.

Alternatively, sensitivity analysis can identify the weights that are most sensitive, and thus map them closest to the applied voltages, where their contribution would be disturbed the least [3]. In the specific context of supervised learning, input intensities may be predicted, and the inputs with the highest expected intensity (as well as the corresponding weights) can be mapped closest to the outputs in order to minimise the negative effects of line resistance.

When training networks directly on crossbar arrays, i.e. in situ, linear adjustments of conductance are the preferred approach [21]. In order to ensure a linear response, the system must be modified physically. Some previous studies have proposed adjusting the device structure [278], typically by introducing additional layers [280]. An alternative approach is combining memristive devices with CMOS transistors, which help to improve the linearity [6].

Random telegraph noise (RTN) is defined as the occurrence of unpredictable switching between two or more discrete voltage levels in electronic devices [211]. This phenomenon is frequently observed in memristors. RTN is more commonly experienced in devices with higher resistance, which can impede the use of such devices for reducing power consumption or mitigating the effects of line resistance. To circumvent RTN or at least mitigate its effects, it is necessary to modify the fabrication process. For instance, some studies have demonstrated that non-filamentary devices can assist in reducing this type of noise [28].

Once the specific application where memristive crossbars will be employed is identified—for instance, classification using neural networks, a pertinent metric such as accuracy, may be optimised instead of attempting to address individual non-idealities. This methodology is more technology-agnostic, as the nature of non-idealities frequently differs between technologies. However, approaches that optimise the metrics pertinent to the application tend to be algorithmic and, thus, more readily transferable.

In the field of machine learning, averaging approaches have the potential to enhance the accuracy and robustness of models, particularly in situations where the memristive implementation is susceptible to non-idealities. One strategy is to utilise multiple networks in parallel and compute their average outputs. Additionally, stability over time can be a crucial consideration, as certain non-idealities, such as RTN, are stochastic in nature. By averaging over time, the effects of these non-idealities can be mitigated [260].

The statistical approaches employed in modern machine learning are based on the minimisation of deviations from ideal behaviour in the training data. This can be extended to

incorporate the non-ideal effects of the hardware on which the model will be implemented. In some instances, this can be achieved by introducing non-ideality-agnostic noise during training in order to enhance the robustness of the networks [290]. Alternatively, noise can be designed to reflect the nature of the non-idealities, thereby enabling the model to adapt more effectively to the various shortcomings of the hardware [101].

6.2 Nonidealities Simulation

Neuromorphic modelling is concerned with the creation of artificial systems that emulate the functionality of biological neural systems, particularly in terms of their physical implementation. The term was first used in the late 1980s to describe digital and analogue hardware that is organised in a more brain-like manner than traditional computer hardware [179].

One of the fundamental concepts underlying neuromorphic systems is parallel distributed processing. Neuromorphic systems arrange computations at the neural level, with a specific focus on facilitating rapid communication between neural processing units. This contrasts with other parallel distributed systems, such as graphics processing units (GPUs), which are typically optimized for independent parallel computations and exhibit limited communication between units.

6.2.1 Learning Rules

Synapses are capable of undergoing changes in their structure and function, a process known as synaptic plasticity. In neuronal systems, the strength of synapses undergoes changes in accordance with the occurrence of spikes in presynaptic or postsynaptic neurons, a process known as synaptic plasticity. Indeed, memory can be conceptualised as a vast neural network. In other words, the synaptic weight is a fundamental determinant of learning and memory processes.

Two principal forms of synaptic plasticity have been identified: long-term plasticity (LTP) [15] and short-term plasticity (STP) [311]. Synapses may undergo strengthening or weakening, and may also exhibit memory retention over a relatively long time, which is referred to as Long-Term Facilitation (LTF) or Long-Term Depression (LTD), respectively. If the change occurs within a relatively short time, it is referred to as Short-Term Facilitation (STF) or Short-Term Depression (STD).

The concept of synaptic long-term potentiation (LTP) has already been incorporated into the training process of deep neural networks (DNNs). This involves the concatenation of all synapse weights into a large multi-dimensional matrix, enabling the identification of the optimal weight matrix through error backpropagation. However, the mechanisms and learning rules in neuroscience are not identical.

One of the most celebrated learning rules in neuroscience is the Hebbian rule [88]. The most concise summary of this rule is: neurons that 'fire together, wire together' [223]. The Hebbian rule can be interpreted as a rate model defined by the neuron spiking rate. It is a local rule, and it requires neurons to be simultaneously active [74]. The general model for this local rule can be defined as follows:

$$\frac{dw_{ij}}{dt} = F\left(w_{ij}, M, v_j^{prev}, v_i^{post}\right) \tag{6.1}$$

Where w_{ij} is the synaptic weight, M is the effect of the neuromodulator, v_j^{prev} is the presynaptic neuron firing rate, and v_i^{post} is the postsynaptic neuron firing rate. A Taylor expansion of equation 6.1 with respect to the rate is:

$$\frac{dw_{ij}}{dt} = a_0(w_{ij}, M) + a_1(w_{ij}, M)^{prev} v_j^{prev} + a_1(w_{ij}, M)^{post} v_i^{post} + a_2(w_{ij}, M)^{corr} v_j^{prev} v_i^{post} + ... \tag{6.2}$$

In the absence of a spike at either the presynaptic or postsynaptic neuron, the effect is represented by a_0. When spikes occur solely at the presynaptic neuron, a_1^{prev} is the expansion coefficient. Similarly, when spikes occur exclusively at the postsynaptic neuron, a_1^{post} is the expansion coefficient. Finally, when spikes occur at both the presynaptic and postsynaptic neurons, a_2^{corr} is the expansion coefficient.

There are additional terms of higher orders, v_j^{prev} and v_i^{post} , which are represented by ellipsis. The coefficients are contingent upon the parameters w_{ij} and M. In accordance with varying parameters and conditions, the Hebbian rule can manifest as LTF or LTD. It is noteworthy that the Hebbian rule constitutes a set of learning rules, rather than a singular, fixed rule.

Another prevalent learning rule is spike-timing-dependent plasticity (STDP). The STDP process entails an increase or decrease in synaptic weight contingent on the time interval between pre- and postsynaptic spikes. The total weight change from neuron j to neuron i is

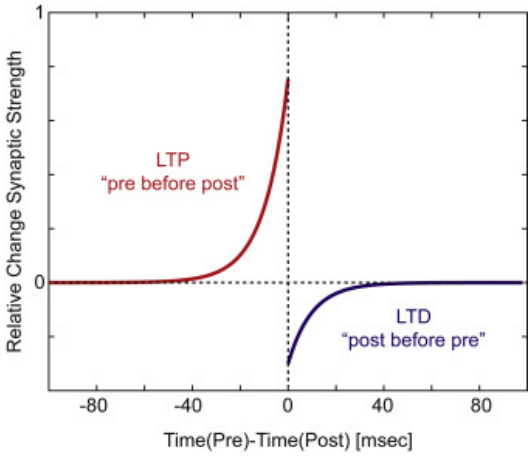

Fig. 6.1 Depiction of spike-timing-dependent plasticity (STDP). If the presynaptic spike occurs before the postsynaptic spike ("pre before post"), the synapse is strengthened (red, LTP, long-term potentiation). If the postsynaptic spike occurs before the presynaptic spike, the synapses are weakened (blue, LTD, long-term depression). Typically, two action potentials need to occur within at most a few tens of milliseconds for STDP to be recruited. [68].

defined as follows:

$$\Delta w_{ij} = \sum_{n}\sum_{f} W\left(t_i^n - t_j^f\right) \tag{6.3}$$

In this context, the term t_i^n denotes the spike times of postsynaptic neuron i, while t_j^f indicates the spike time of presynaptic neuron j. The variables n and f are used to count the pre- and postsynaptic spikes, respectively. The term $W(x)$ refers to the learning window of the STDP function. It should be noted that numerous variations of STDP exist, with one of the most common types being the pair-based variant. This is defined as follows:

$$W_+(x) = \left\{A_+(w)e^{-|\Delta t|/\tau_+}\right\} \quad \forall t_{post} \in t_{prev} < t_{post} \tag{6.4}$$

$$W_-(x) = \left\{A_-(w)e^{-|\Delta t|/\tau_-}\right\} \quad \forall t_{prev} \in t_{post} < t_{prev} \tag{6.5}$$

Where $|\Delta t| = \left|t_{post} - t_{prev}\right|$, the time of the postsynaptic spike is designated as t_{post}, while the time of the presynaptic spike is designated as t_{prev}. In most cases, $A_+(w)$ is positive, $A_-(w)$ is negative, and these values may be dependent on the current synaptic weight. $W_+(x)$ is associated with long-term facilitation (LTF), while $W_-(x)$ is associated with long-term depression (LTD).

By introducing $S_j = \sum_f \delta\left(t - t_j^f\right)$ and $S_i = \sum_n \delta\left(t - t_i^n\right)$, where S_j represents the spike train of the presynaptic neuron and S_i represents the spike train of the postsynaptic neuron. The pair-based STDP rule, as presented in previous equations, can be implemented by:

$$\frac{dx_j}{dt} = \sum_f \delta\left(t - t_j^f\right) - \frac{x_j}{\tau_+} \tag{6.6}$$

$$\frac{dy_i}{dt} = \sum_n \delta\left(t - t_i^n\right) - \frac{y_j}{\tau_-} \tag{6.7}$$

In this context, the notation t_j^f represents the spike time of the presynaptic neuron, while t_i^n denotes the spike times of the postsynaptic neuron. The variables x_j and y_i can be interpreted as a trace that each pre- and postsynaptic spike leaves, and respectively corresponds to the $e^{-|\Delta t|/\tau_+}$ and $e^{-|\Delta t|/\tau_-}$ terms to give:

$$\frac{dw_{ij}}{dt} = A_+ w_{ij} x_i(t) \sum_n \delta\left(t - t_i^n\right) + A_- w_{ij} y_i(t) \sum_f \delta\left(t - t_j^f\right) \tag{6.8}$$

Where the first term on the right-hand side denotes the pre-before-post effect, while the second term indicates the post-before-pre effect. It is noteworthy that for independent Poisson inputs, STDP models are related to rate models [74]. One may define it as a rate model as follows:

$$\frac{dw_{ij}}{dt} = \int_0^{+\infty} W(-s)\varepsilon(s)ds \cdot v_j^{prev} + \int_{-\infty}^{+\infty} W(s)ds \cdot v_j^{prev} v_i^{post} \tag{6.9}$$

The first term on the right-hand side is defined by the integral over the 'causal' part of the learning window, also known as the 'pre-before-post' relation. This integral is represented by the function $\int_0^{+\infty} W(-s)\varepsilon(s)ds$, where W is the weighting function and $\varepsilon(s)$ describes the time course of a Postsynaptic Potential (PSP) for $s > 0$. A comparison of Equation 6.2 with Equation 6.9 reveals that there are two terms defining coefficients $a_1(w_{ij}, M)^{prev}$ and $a_2(wij, M)^{corr}$, while the remaining terms are all zero.

Non-volatile memristors are capable of operating as long-term potentiation (LTP) synapses, as the memristance will not undergo alteration following each update. In a fully connected neural network, the number of synapses between two layers is given by the equation $m \cdot n$, where m is the number of neurons in the previous layer and n is the number of neurons in the next layer. Consequently, non-volatile memristors have been employed in crossbar arrays for vector-matrix multiplication, utilising Kirchhoff's current law.

Gradient backpropagation and STDP are two distinct mechanisms for learning. Gradient backpropagation is a widely adopted technique in the domain of deep neural networks (DNNs), whereas STDP is a prevalent approach in the field of spiking neural networks (SNNs). Both gradient backpropagation and STDP can be implemented using memristors. For gradient backpropagation [85], memristive crossbar arrays can be employed for both training and inference in neural networks [8].

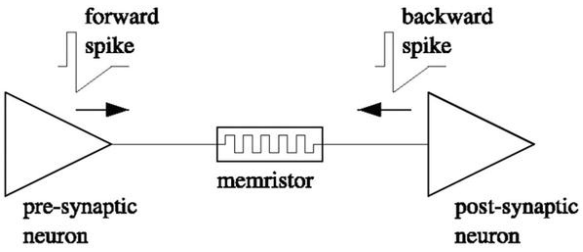

Fig. 6.2 Memristor between presynaptic and postsynaptic neurons [98].

The training process necessitates the utilisation of an external control unit, which introduces a greater degree of overhead than using the memristive crossbar alone. Furthermore, the STDP rule described above can be achieved by utilising memristors [173], and can be verified using SPICE models [288].

As illustrated in Figure 6.2, a memristor is situated between two neurons. The presynaptic and postsynaptic spikes will generate a voltage difference across the memristor, which will then cause a memristance update. Furthermore, the time interval between the presynaptic and postsynaptic spikes will result in varying changes in memristance.

The majority of research into the emulation of plastic synapses relies on non-volatile memristors for long-term potentiation (LTP), with few researchers focusing on the use of volatile memristors to mimic short-term potentiation (STP) synapses. The significance of volatile memristors is further underscored by their role in neural information processing, including functions such as motion detection, speech recognition, and working memory [77].

In contrast to long-term potentiation (LTP), short-term potentiation (STP) is dependent on the spiking activity of the presynaptic neuron. Let define the fraction $P_{release}$ as the amount of neurotransmitter released by the presynaptic neuron. It can then be shown that the synaptic weight is directly related to $P_{release}$, and that both STF and STD can be modelled with the

dynamics of $P_{release}$. STF can be defined as:

$$\frac{dP_{release}}{dt} = \frac{P_{release} - P_0}{\tau_F} + f_F(1 - P_{release}) \sum_f \delta\left(t - t^f\right) \tag{6.10}$$

Where t^f is the time for each presynaptic spike, P_0 is the resting value of $P_{release}$, τ_F is a time constant governing the recovery process, and f_F controls the facilitation degree. Similarly, STD can be defined as the following with D denotes for depression:

$$\frac{dP_{release}}{dt} = \frac{P_{release} - P_0}{\tau_D} + f_D(1 - P_{release}) \sum_f \delta\left(t - t^f\right) \tag{6.11}$$

The transient nature of volatile memristors allows them to emulate the STF or STD, as the memristance change is only retained for a brief period. The measurement of STF and STD is typically conducted through the use of Paired-Pulse Facilitation (PPF) and Paired-Pulse Depression (PPD). The PPF and PPD index is defined as $\frac{A_2}{A_1}$, where A_1 and A_2 are the absolute amplitudes of the EPSC or IPSC resulting from two successive presynaptic spikes.

This index is used to evaluate the strength of PPF or PPD. In 2018, a novel type of volatile memristor was proposed, and the PPF and PPD indexes were subsequently measured [237]. This is a two-terminal single-layered molybdenum disulfide (MoS_2) device, wherein the conductance change is achieved through Joule heating.

An alternative approach is to train the SNN using the supervised SRDP learning rule. This is more hardware-compatible than the backpropagation algorithm (BP) and overcomes the low accuracy of the unsupervised SRDP learning rule, which only considers local optimisation and ignores global errors [206]. The supervised SRDP rule can be described as follows:

$$\Delta_{ij} = \begin{cases} +1 & \text{if } f_{r_j} < f_{t_j} \;\&\; f_{s_i} > \text{ threshold} \\ -1 & \text{if } f_{r_j} < f_{t_j} \;\&\; f_{s_i} < \text{ threshold} \\ -2 & \text{if } f_{r_j} > f_{t_j} \;\&\; f_{s_i} > \text{ threshold} \end{cases} \tag{6.12}$$

In this context, f_s and f_r represent the output frequencies of the sensory and relay neurons, respectively. f_t denotes the teaching frequency, while threshold refers to the threshold of the output frequencies of the sensory neurons. The symbol Δ_{ij} indicates the pulses applied to the positive or negative synapses connecting the i_{th} sensory neuron and the j_{th} relay neuron.

When Δ_{ij} is equal to 1, a pulse is applied on the negative synapse to increase the synaptic weight, whereas when Δ_{ij} is equal to -1, a pulse is applied on the positive synapse to decrease the synaptic weight. In the event that the output frequency of a relay neuron is less than the teaching frequency and the output frequency of the sensory neuron is greater than the threshold, the neuron will learn the input pattern by increasing the synaptic weight $\Delta_{ij} = +1$. Conversely, if the output frequency of the sensory neuron is less than the threshold, the synaptic weight will be decreased $\Delta_{ij} = -1$.

In the event that the output frequency of the relay neuron is higher than the teaching frequency, this indicates that the neuron has either learned a feature belonging to an alternative class ($f_t = 0$) or has learned an excessive number of features ($f_t > 0$). In such instances, the neuron should erase these features by decreasing the synaptic weight $\Delta_{ij} = -2$.

In order to quantify the training effort, a loss function has been devised based on the discrepancy between the output frequencies of the relay neurons and the teaching frequencies. This can be expressed as follows:

$$l = \frac{\sum_j \left(f_{r_j} - f_{t_j}\right)^2}{\sum_j \left(f_{t_j}\right)^2} \tag{6.13}$$

$$Loss = \frac{\sum_i^n l_i}{n} \tag{6.14}$$

In this context, the variable l represents the error incurred when a single sample is input into the testing set. The term *Loss* denotes the total error after all samples in the testing set have been input. The variable f_{r_j} signifies the output frequency of the j_{th} relay neuron when the input signal originates from the synapses. The variable f_{t_j} represents the teaching frequency generated by the j_{th} relay neuron when the input signal is a teaching signal. The program terminates when the *Loss* value is sufficiently low.

6.2.2 Training Schemes

The discovery of spike-timing-dependent plasticity (STDP) mechanisms and the emergence of nanoscale non-volatile memory (NVM) devices have opened a new avenue towards the realisation of brain-inspired computing over the past decade [298]. Prior research suggests that STDP can be used to train spiking neural networks (SNNs) with resistive random-access memory (RRAM) synapses in-situ, without trading off their parallelism [212]. Furthermore, these devices have demonstrated low energy consumption for state transitions and a highly

compact layout footprint [294].

A neuromorphic system-on-a-chip (NeuSoC) architecture has been proposed, comprising of multiple-layer fully connected SNN [127]. In this model, the input layer encodes real-valued inputs into spatiotemporal spike patterns, which are then processed by the subsequent layers using STDP-based unsupervised or semi-supervised learning. Neurons in each layer are connected to the higher layers via synapses that hold the 'weights' of the SNN.

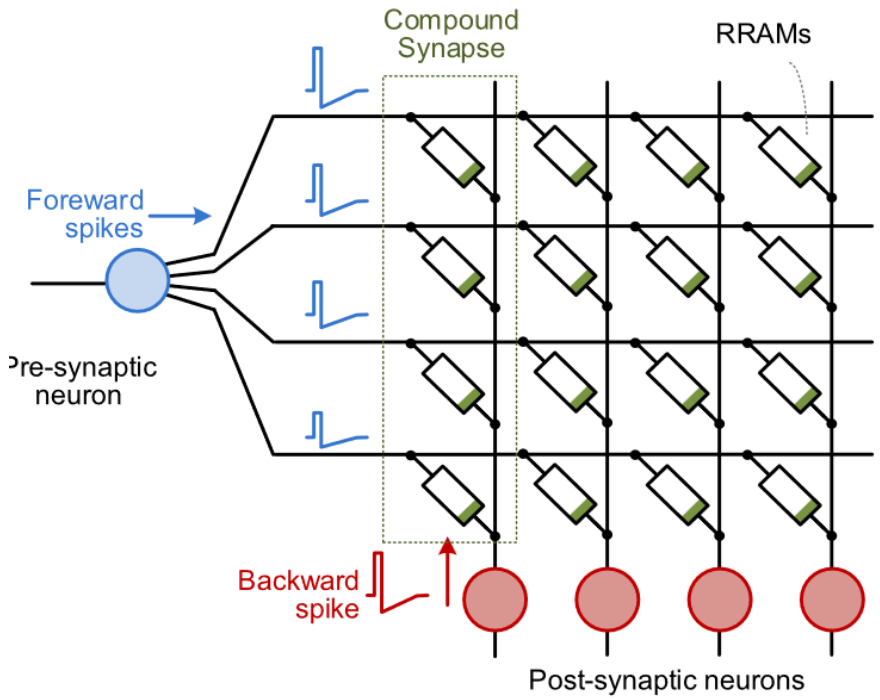

Fig. 6.3 A single layer of spiking neural network with RRAM synapses organized in crossbar architecture. Post-synaptic neuron connects with a pre-synaptic neuron with several RRAM synapses in parallel. Back-propagated spikes after dendritic processing modulate RRAM in-situ under STDP learning rule [283].

As illustrated in Figure 6.3, a RRAM crossbar array is employed to establish synaptic connections between the two neuronal layers. The spiking neurons in the second and subsequent layers implement competitive learning through a winner-take-all shared bus mechanism, whereby the neuron that spikes first for an input pattern inhibits the remaining neurons in the same layer [175].

A combination of STDP and WTA-based learning rules enables the synapses to be locally updated through the interaction of pre- and post-synaptic neuron spikes in Figure 6.2, thereby facilitating network learning in the form of fine-grained weight (or conductance) adaptation in the synapses.

The presented architectural configuration bears resemblance to existing memory architectures, wherein the devices are arranged in a dense two-dimensional array with the input and output neurons constituting the peripheral circuitry, laid out at a matching pitch with the memory array. This layout forms a repeatable motif that can be scaled to deeper SNN architectures.

The extension of larger 3D IC architectures using through-silicon-via (TSVs) represents a natural pathway for further scaling to very high integration density and network complexity. This can be achieved without resorting to the overhead incurred by asynchronous communication protocols such as the address-event-representation (AER) [112].

It is desirable to have non-volatile analog-like weights in order to achieve effective STDP learning [70]. However, the majority of practical [227], small-sized RRAM devices exhibit abrupt switching behaviour [293], which consequently limits the stable synaptic resolution to 1-bit (or binary, bistable) [239].

Furthermore, the switching probability and switching times of these devices typically depend upon the voltage applied across the device, as well as the duration of the voltage pulse [150]. In order to circumvent the binary resolution of these devices, a compound memristive synapse with multiple bistable devices in parallel was previously proposed as a means of emulating analog weights on average [16].

Theoretical studies have demonstrated that the STDP learning rule facilitates unsupervised local learning in spiking neural networks through the implementation of a Bayesian expectation-maximisation algorithm [199] and hidden Markov models [122]. Furthermore, a nonlinear STDP learning function, such as an exponentially shaped window observed in biological neural systems [246], is essential for ensuring the stability and efficiency of the computing process [228].

As previously stated, emerging memristive nanoscale devices are being considered as a means of enabling the realisation of large-scale neuromorphic hardware. Nanoscale implementations of memristors [136], including phase-change memory (PCM) [120], resistive random-access

memory (RRAM) [169], and spin-torque-transfer random-access memory (STT-RAM) [219], have been demonstrated to exhibit switching characteristics analogous to those observed in spike-timing-dependent plasticity (STDP) [154].

Furthermore, these devices enable the highly desired advantage of a small silicon area of $4F^2$ (F is the feature size of the semiconductor fabrication process) [32], ultra-energy-efficient operation of sub-pico-Joule per switching event, CMOS compatibility, and dense crossbar (or crosspoint) arrays and 3D integration.

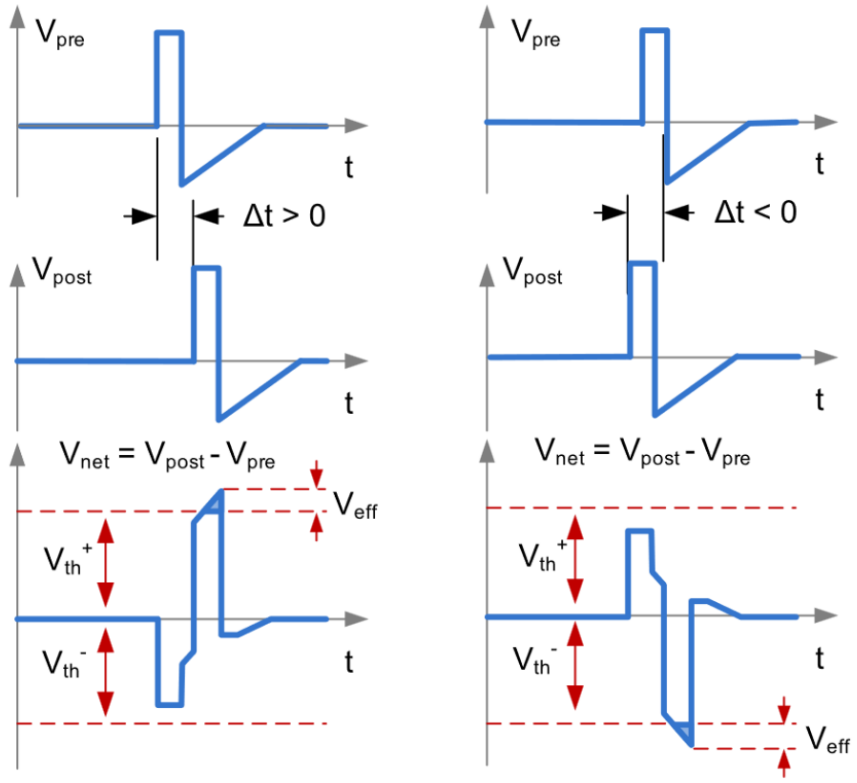

Fig. 6.4 A pair of spikes are applied across a synapse to create relative-timing dependent net potential V_{net}, of which the over-threshold portion V_{eff} may cause the RRAM resistance to switch [282].

In particular, RRAM devices exhibit characteristics that are analogous to those of biological synapses. Furthermore, RRAMs exhibit conductance levels comparable to synaptic strength (or weight), and their conductance can be modified by voltage pulses. Notably, RRAMs are capable of directly implementing spike-timing-dependent plasticity (STDP) through identical pair-wise spikes.

As illustrated in Figure 6.4, the net potential V_{net} created by a pre-synaptic spike and a late arriving post-synaptic spike, with $t > 0$, produces an over-threshold, V_{th}^{+}, portion V_{eff}, which then causes an increase in conductance in a typical RRAM. Conversely, a presynaptic spike and an earlier arriving post-synaptic spike with $t < 0$ produces V_{eff} that crosses the negative threshold, V_{th}^{-}, and thus causes a decrease in the conductance.

A memristor with analogue resistance is an optimal choice for STDP learning. However, experimental studies indicate that the majority of nanoscale implementations of memristive RRAMs exhibit a stochastic process in their filament formation, as well as an abrupt conductance change once the filament is formed [210]. The results demonstrate an abrupt transition from the high resistance state (HRS) to the low resistance state (LRS), as well as notable variations in the switching threshold voltages and HRS/LRS transitions, which tend to be stochastic in nature.

The intrinsic stochastic switching in RRAM therefore limits the stable synaptic resolution to 1 bit, or bistable behaviour. In order to unlock STDP-based learning in SNN hardware, a solution enabling non-linear, especially exponential, STDP learning functions with binary RRAMs is therefore highly desirable.

One proposed approach to enable the complexity of learnable tasks to be scaled up in fully MSNNs by directly applying gradient descent to the nonlinear state evolution of memristive neurons and synapses [304]. The modelling of both neurons and synapses in biological neural networks employs the use of memristors. The MIF neuron model has been devised with the objective of achieving distinct depolarisation, hyperpolarisation and repolarisation voltage phases with a minimal set of circuit elements.

Memristive synapses serve as interconnects between layers of neurons. This allows for the development of dynamical, time-varying memristive neurons, which are capable of learning and achieving significantly enhanced accuracy on data-driven tasks compared to previous reports on MSNNs [198]. By leveraging the analog spiking characteristics inherent to the MIF neuron model, the non-differentiability of spike-based activations is entirely circumvented.

In order to account for the hardware implementation of a fully MSNN, it is necessary to consider the voltage response of the MIF neuron, which must drive the memristive synapses. The synaptic conductances are correspondingly weighted by this response. This can be fully

integrated into a crossbar according to the following equation:

$$\begin{matrix} & \mathbf{G} & & & & \mathbf{V} & = & \mathbf{I} \end{matrix}$$
$$\begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,m} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,m} \\ \vdots & \vdots & \vdots & \vdots \\ w_{n,1} & w_{n,2} & \cdots & w_{n,m} \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_n \end{bmatrix} = \begin{bmatrix} I_1 \\ I_2 \\ \vdots \\ I_m \end{bmatrix} \tag{6.15}$$

where $\boldsymbol{I}$ is the output current vector, $\boldsymbol{V}$ is the input voltage vector, and $\boldsymbol{G}$ is the conductance matrix of the crossbar.

$$I_n = \sum_{i}^{m} v_i \times w_{n,i} \tag{6.16}$$

The output current vector is generated via bitline current summation, as illustrated in equation (6.16), and is directly driven into the input of the subsequent MIF neuron layer. Resistive loading may result in the attenuation of the output current in subsequent stages [265]; however, this can be accounted for through the utilisation of a scaling factor or the implementation of a buffer [264].

$$\tau_{syn} \frac{dI}{dt} = a - I \tag{6.17}$$

$$\tau_{syn} \frac{da}{dt} = W_i \cdot \sum_{n} \delta\left(t - t_i^n\right) \tag{6.18}$$

A considerable number of neural coding studies portray spike trains as a superposition of time-shifted Dirac delta pulses, represented by $\sum_n \delta\left(t - t_i^n\right)$. Consequently, this model of spikes is one idealisation. The spike train is employed to modulate a time-continuous alpha input current I, which has been modelled by the differential equation (6.17) presented in order to ensure compatibility with real, physical systems.

In this series of equations (6.18), a represents an internal state variable, τ_{syn} denotes a time constant that determines the shape of the alpha current, and W_i is the synaptic weight between a presynaptic neuron i and its associated postsynaptic neuron. It is widely accepted that such alpha waveforms correspond to the response from biological neurons in the sensory

periphery that respond with graded potentials [59].

$$I^{t+1} = \frac{a^t - I^t}{\tau_{syn}} + I^t \tag{6.19}$$

$$a^{t+1} = a^t - \frac{a^t + \sum_n \delta\left(t - t_i^n\right)}{\tau_{syn}} \tag{6.20}$$

In order to train a network of MIF neurons and synapses using gradient descent, the differential equations representing the MIF circuit dynamics are recast into discrete-time form (6.19, 6.20). This allows the memristive dynamics to be captured in a computational graph that evolves over time, in a manner analogous to that of a recurrent neural network (RNN).

In practice, SPICE-like simulators employ a variety of differential equation solvers, including the backward Euler method and the fourth-order Runge-Kutta method (RK4). In order to ensure compatibility with the BPTT algorithm, the differential equations are solved using the forward Euler method, which provides an explicit representation of the next time step based on present-time dynamics.

The intricate dynamics of the MIF neuronal network are now accounted for in the MSNN, which is unrolled in time in such a way that gradient descent can be used to optimise the memristive synapses as a function of the MIF evolution. The discrete-time solution can be illustrated as a directed, acyclic graph, where time flows in one direction.

In order to train a network, a loss function is calculated using the membrane potential v of the output layer at each step. It should be noted that the adjoint method [37] is normally not adopted, as an intermediate state is required in order to calculate the loss and guide the training process for each time step. The focus is on the spiking output at all time steps, rather than just on the final state of the system, which makes BPTT a more optimal choice.

The predicted MIF neuron is expected to spike most frequently by aiming to increase the membrane potential across time steps, while the incorrect target should be suppressed. Given that the membrane dynamics are continuous, it is possible to train a fully analogue MSNN, as has become the norm in the training of deep SNNs via error backpropagation [198].

The BPTT algorithm employs an iterative application of the chain rule from the output back to the leaf nodes in order to determine the optimal update direction for the network [62], with

previous demonstrations have treated w as the device conductance. This approach offers a more biorealistic representation and can be fully integrated using RRAM crossbars.

6.2.3 Conductance Mapping

The approaches described in the preceding sections pertained to memristive neural networks (MNNs) that had been trained through conventional methodologies without consideration of the non-idealities. The primary challenge lies in the discrepancy between the behaviour of conventional SNNs and that of MNNs.

The synapses of the former typically exhibit a consistent response and typically relate inputs to outputs in a linear fashion. In contrast, MNNs utilise synapses implemented using memristors, which can become fixed in a particular state, exhibit varying responses over time, due retention or random telegraph noise (RTN), and even produce nonlinear outputs, I-V nonlinearity.

A substantial proportion of the literature on addressing memristor nonidealities focuses on modifications at the hardware level. These may entail: Modifications to the device structure are a common approach, as evidenced by the literature [64]; additional circuitry is another avenue of exploration [6, 148]; programming and mapping schemes are also subject to modification [286]; in-situ retraining is a further strategy, as exemplified by the literature [34, 107, 267, 146].

However, these techniques may have some drawbacks. For example, pulse-width modulation [7], may minimise the effects of I-V nonlinearities; however, this comes at the cost of increased clock cycles [22]. Furthermore, hardware-level changes are often technology-specific, thus rendering them difficult to apply to a wide range of device types.

As an alternative, techniques that do not interfere with the hardware, such as committee machines (CMs), may be employed. These allow the performance of MNNs to be improved by combining them and require only the average of the outputs of crossbar arrays [115]. Additionally, modifications to ex situ training have been previously proposed.

Common approaches include adjusting the loss function [309] or introducing noise into the synaptic weights [87] or conductances that implement those synaptic weights [118], thereby enhancing the resilience of MNNs to the effects of non-idealities. It is evident that ex situ

training represents a promising methodology for enhancing the feasibility of MNNs.

However, previous approaches have only considered a limited number of non-idealities. For example, injection of noise into the conductances is insufficient in many situations, as the effect of non-idealities such as I-V nonlinearity cannot be represented by the disturbance of the weights alone. Furthermore, conventional weight implementations do not map naturally to resistive crossbar arrays, while training validation schemes are not well suited to the often stochastic nature of MNNs.

This research focuses on enhancing the efficacy of *ex-situ* training of memristive SNNs, while also evaluating the resilience of this approach. To this end, data from a SiO_x memristive device is employed, complemented by insights drawn from existing literature.

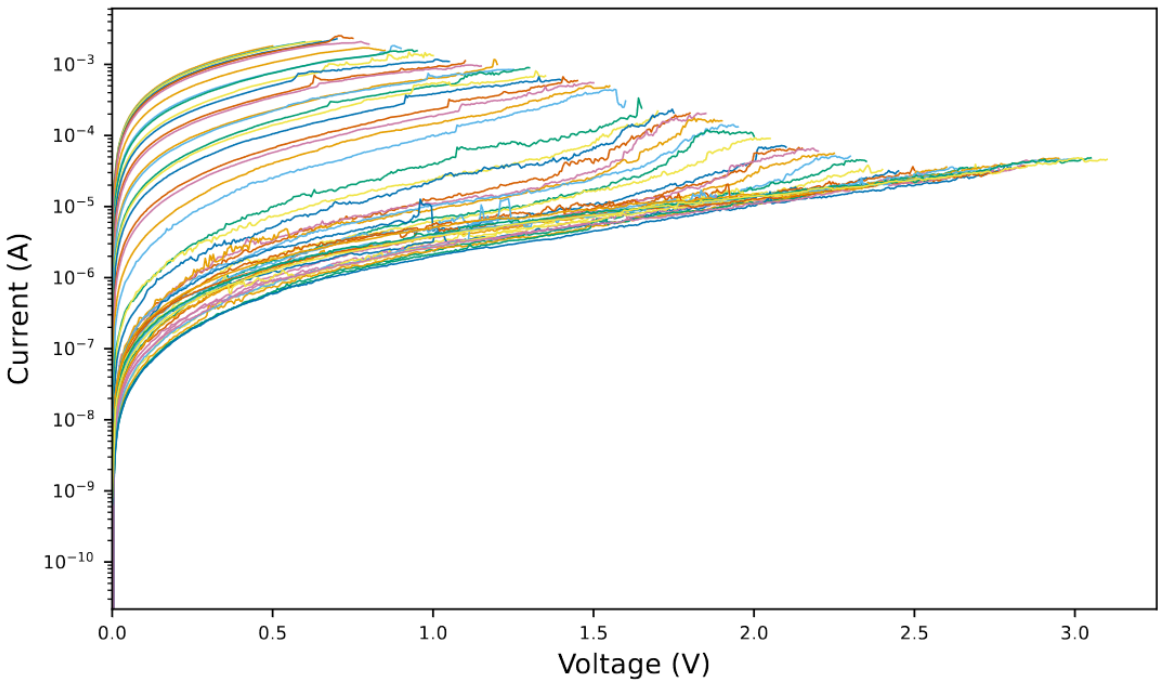

Fig. 6.5 Single sweeps of 53 resistance states of a SiO_x device. Different states were obtained by incrementing maximum voltage by 0.05V. Every seventh state shares the same colour; this does not indicate any other relationship between such states [13].

The SiO_x resistive random-access memory (RRAM) device comprised a $1\mu mSi/SiO_2$ switching layer positioned between a 100nm Mo bottom electrode and a 100nm Au top electrode. Furthermore, a 5nm Ti wetting layer was incorporated between the SiO_x and Au electrodes. Following electroforming and additional validation procedures, positive sweeps were con-

ducted on the SiO_x device, commencing from 0.5V and increasing by 0.05V in each run, thereby generating resistance states that spanned multiple orders of magnitude.

The majority of existing literature on ex-situ training of MNNs assumes a linear relationship between inputs and outputs at the level of individual devices. Non-linearities are only introduced at the level of the activation functions. In particular, outputs are assumed to be a function of the product of a vector of applied inputs and a matrix of weights:

$$y_j = f\left(\sum_{i=1}^{M} x_i, w_{ij}\right) \tag{6.21}$$

Where the outputs, $y_j \in \mathbf{y} \in \mathbb{R}^{1\times N}$ is calculated using the inputs $x_i \in \mathbf{x} \in \mathbb{R}^{1\times M}$, weights $w_{ij} \in \mathbf{w} \in \mathbb{R}^{M\times N}$, and a nonlinear activation function f. Non-Ohmic behaviour manifests itself in individual devices only, therefore these nonlinear effects can be separated from the other devices. To take I-V nonlinearities into account in ex-situ MNN training, the products in (6.21) can be replaced with a nonlinear function to give:

$$y_j = f\left(\sum_{i=1}^{M} g\left(x_i, w_{ij}\right)\right) \tag{6.22}$$

In order to account for the non-ohmic behaviour of memristors, Function g must be modified to reflect the specific characteristics of the mapping scheme between weights and conductances, as well as the non-ohmic device behaviour model, which is typically dependent on the type of devices employed [116].

The mapping scheme that relates weights and conductances is a crucial aspect to consider when designing memristive neural networks. In typical artificial neural networks, synaptic weights can assume any real value. Conversely, conductances are constrained to be non-negative.

To address this discrepancy, a neural network architecture was devised whereby twice the number of weights are trained, but each is constrained to be non-negative. This enables the association of weights with the conductances of individual devices, thereby creating a more natural mapping and facilitating adaptation to non-idealities. Additionally, it allows for more precise control over power consumption.

The aforementioned mappings are performed in an identical manner in the specific instance of fully connected synaptic layers in artificial neural networks, provided that conventional weight implementation methodologies are utilised. Both the inputs, designated as $x \in \mathbf{x}$, and the outputs, represented by $y \in \mathbf{y}$, are mapped onto voltages V and total output currents I, using the scaling factors k_V and k_I, where k_G is the conductance scaling factor:

$$V = k_V x \tag{6.23}$$

$$y = \frac{I}{k_I} = \frac{I}{k_V k_G} \tag{6.24}$$

$$k_G = \frac{G_{max} - G_{min}}{max|\mathbf{W}|} = \frac{G_{on} - G_{off}}{max|\mathbf{W}|} \tag{6.25}$$

Irrespective of the scheme employed, a minimum of two conductances are required to encode both positive and negative weights. Conductances G_+ and G_- are introduced into the "positive" and "negative" bit lines of the differential pair architecture, respectively. The proportionality of each weight is determined by the difference $G_+ - G_-$, with k_G serving as the constant of proportionality. This enables the encoding of any real number within a finite range. In a typical differential pair implementation, $G_{min} = G_{off}$ and $G_{max} = G_{on}$. In an alternate proportional mapping scheme, $G_{min} = 0$ and $G_{max} = G_{on}$ to give $k_G = \frac{g_{on}}{max|\mathbf{W}|}$.

The issues associated with proportional mapping schemes, such as the presence of unimplementable regions, are not the only obstacles to be overcome. Conventional differential pair realisation also presents design challenges. For example, infinite conductance combinations will produce the same difference [130], i.e. the same effective conductance.

This means that an arbitrary choice may have to be made of how to perform the mapping between weights and conductances. To illustrate this, consider the encoding of weights $W \in \mathbf{W}$. In this case, pairs of conductances G_+ and G_- may be picked symmetrically around the average value:

$$G\pm = G_{avg} \pm \frac{k_G w}{2} \tag{6.26}$$

$$G_{avg} = \frac{G_{off} + G_{on}}{2} \tag{6.27}$$

While multiple mapping schemes may yield the same conductance difference, some may prove more advantageous than others. The scheme in (6.26) can be beneficial in certain scenarios, as it typically reduces the number of conductances near G_{off} and G_{on}, which are often more challenging to achieve. However, the selection of the mapping scheme can be

explicitly linked to specific objectives.

The differential pair architecture can be employed to mitigate the effects of stuck devices. To illustrate, if G_+ is stuck in an undesirable state, G_- can be adjusted to minimise the negative effects. An alternative approach is to employ a simpler scheme that optimises a specific metric, such as power consumption.

This is illustrated in the simulations presented in this chapter, where conventional weights are used. In these simulations, a scheme that minimises power consumption is employed by ensuring that at least one of G_+, G_- is set to G_{off}:

$$G_+ = G_{off} + max\{0, k_G W\} \tag{6.28}$$

$$G_- = G_{off} - min\{0, k_G W\} \tag{6.29}$$

An alternative approach is to utilise two sets of non-negative weights, $W_{ij}^+ \in \mathbf{W}_+ \in \mathbb{R}_{\geq 0}^{M \times N}$ and $W_{ij}^- \in \mathbf{W}_+ \in \mathbb{R}_{\geq 0}^{M \times N}$, which are collectively referred to as double weights [124]. This method entails mapping each weight onto a single conductance in the aforementioned "positive" and "negative" bit lines, respectively.

Despite the nonnegativity of each weight, the differential pair architecture allows for encoding the negative contribution of the i_{th} input on the j_{th} output through the introduction of a subtraction operation in hardware. Only the nonlinearity-aware node function in (6.22) requires adjustment to give:

$$y_j = f\left(\sum_{i=1}^{M} g\left(x_i, W_{i,j}^+\right) - g\left(x_i, W_{i,j}^-\right)\right) \tag{6.30}$$

As all weights in $\mathbf{W} := [W_+, W_-]$ are non-negative, they can be related to the corresponding conductances in the same way: $W_\pm \in [0, max(\mathbf{W})]$ can be linearly mapped onto $G_\pm \in [G_{off}, G_{on}]$, thus avoiding the introduction of weight gaps:

$$G_\pm = k_G W_\pm + G_{off} \tag{6.31}$$

The initial benefit of utilising double weights is that the conductances are more directly exposed to the training algorithm. This enables the selection of combinations that achieve both optimal performance, for example in terms of loss and robustness. To illustrate, if specific non-idealities manifest more at lower conductance values, such as with programming

deviations [131], the training may be able to select pairs at higher values, while maintaining the same difference between G_+ and G_-. This allows the minimisation of the negative effects of non-idealities without the need to explicitly specify which conductance pairs should be selected.

When weights W are related to conductances G in a monotonically increasing fashion (in this case, linear), regularisation can be employed to influence the magnitude of both weights and conductances. This provides a means of utilising regularisation as a high-level tool for controlling power consumption. It has been proposed that the *L1* sparsification regulariser [82] should be employed, as this has the potential to not only enhance training, for instance, by preventing overfitting, but also to promote lower conductance values.

In lieu of manually adjusting the mapping scheme, the network designer may elect to prioritize low power consumption to a greater or lesser extent. This may be determined by, for example, adjusting the regularisation factor in *L1* regularisation and incorporated into the conventional hyperparameter tuning process [65], which is typically performed before deploying SNNs in the real world.

6.2.4 Nonidealities Calibrations

In the simulation presented in this chapter, a combination of experimental data from a SiO_x device, models based on simplified assumptions regarding breaking-down devices, and findings from the literature on programming variability were employed. I-V nonlinearities have been identified as a prevalent method for characterising deviations from ohmic behaviour in memristive devices.

$$\gamma \equiv \frac{G\left(2V_{ref}\right)}{G\left(V_{ref}\right)} \tag{6.32}$$

This approach involves the examination of two points on an I-V curve [145], as previously discussed (6.32). This equation defines conductance linearity, which serves as a means of quantifying nonlinear I-V behaviour. A conductance linearity value of 1 is indicative of ohmic behaviour, while any deviation from this value indicates I-V nonlinearity. While this metric can be useful for describing non-ohmic behaviour at different voltages, it is more challenging to utilise for modelling purposes [238].

$$I = V_{ref} G \left(\frac{V}{V_{ref}}\right)^{log_2 \gamma} \tag{6.33}$$

In order to construct an I-V curve from a given nonlinearity parameter γ, it is possible to assume that the equality in Equation (6.32) holds for all V, rather than just V_{ref}. This would result in a relationship between current and voltage, as shown in (6.32). In this equation, G represents the conductance parameter, which has a specific meaning in the context of V_{ref}, where the device produces the expected ohmic amount of current, which is equal to $V_{ref}G$.

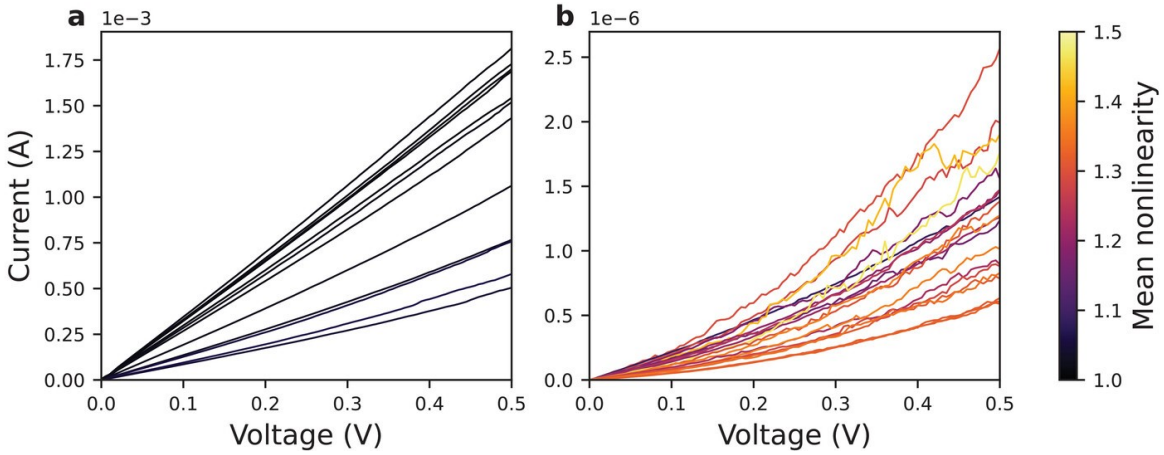

Fig. 6.6 I-V sweeps of a SiOx device are presented for two regions: a) low-resistance region with average resistance ranging from 284.6 Ω to 1003 Ω, and b) high-resistance region with average resistance ranging from 366.2 kΩ to 1.295 MΩ. Only the voltage range from 0.0 V to 0.5 V was considered for all curves. The nonlinearity parameter was calculated by dividing the current at 0.5 V by the current at 0.25 V [116].

The nonlinearity parameters γ were extracted from experimental data obtained from a SiO_x RRAM device. SiO_x devices are capable of undergoing resistance switching, which is characterised by a typical I-V switching curve. In order to achieve a wide range of resistance states and to analyse I-V nonlinearity, incremental positive sweeps were employed to gradually reset the device from the low-resistance state (LRS) to the high-resistance state (HRS). The low-resistance discrete states display greater linearity and exhibit minimal variability. Conversely, the high-resistance states are more nonlinear, and the nonlinearity is less predictable.

The value of V_{ref} from (6.32), was set to 0.25V, which is equivalent to half of the minimum switching voltage. In each case, the devices could be set to any conductance between $G_{off} = 1/R_{off}$ and $G_{on} = 1/R_{on}$. The states for the two groups were selected in such a way that the G_{on}/G_{off} ratio for high-resistance devices would be slightly larger. This approach ensures that any potential higher error rate in this group can be attributed to nonlinearity and its variability, rather than being attributed to limited dynamic range.

In order to guarantee more robust modelling, it was deemed necessary to take uncertainty regarding the degree of nonlinearity experienced by each device into account. Although there is a general tendency for high-resistance states to exhibit greater nonlinearity, the precise degree of nonlinearity at any given resistance state may be challenging to ascertain, as evidenced by the variability observed in the I-V curves depicted in the experimental data presented in Figure 6.6.

Consequently, for each device in either of the groups, the parameter γ was drawn from a truncated normal distribution. This distribution was truncated for γ values of less than 2, with the objective of ensuring realistic device behaviour and numerical stability of the simulations. The mean m_γ and standard deviation s_γ, both of which refer to the underlying normal distribution prior to truncation, were used to characterise the distribution.

It is not possible to simulate nonidealities such as I-V nonlinearity using conventional noise injection methods that merely disturb the conductance values. Consequently, a forward propagation function must be defined in order to reflect the nonlinear relationship between inputs and outputs.

$$g(x, w_\pm) = \frac{\left(k_G w_\pm + G_{off}\right) \times (2x)^{log_2 \gamma_\pm}}{2k_G} \tag{6.34}$$

In the proposed training function, as outlined in Equation (6.30), the aforementioned I-V nonlinearity is incorporated into the function g by combining Equations (6.25), (6.31), (6.33), using $k_V = 2V_{ref}$ in accordance with the definition in Equation (6.32), resulting in the form presented in Equation (6.34). This function is then implemented using Keras.

In order to conduct simulations involving training, it is essential to utilise an I-V model that incorporates stochasticity. As a specific instance of nonlinear behaviour may be acquired during training, it is unclear how beneficial such a model would be when applied to a different set of devices.

It can be seen, therefore, that the previous approach may not be sufficient, since the experimental I-V measurements were used as a lookup table for computing currents of devices in certain conductance states at certain voltages. There are, however, multiple physical models that could be employed for describing non-ohmic memristor behaviour.

It was decided that the Poole-Frenkel conduction model [116] would be the most appropriate to use, given that the underlying physical mechanism was deemed to be plausible for SiO_x devices. Furthermore, the model demonstrated excellent fit to the experimental I-V curves. Its simple analytical form also allowed for the incorporation of stochasticity by considering the uncertainty in certain parameters.

$$I = cVexp\left(\frac{2e}{k_B T}\sqrt{\frac{eV}{4\pi d\varepsilon}}\right) \tag{6.35}$$

The Poole-Frenkel model postulates that the current through a device can be described by (6.35), where I is the current, V is the voltage, c is a constant (with units of conductance), T is the temperature which is 20 degrees Celsius, d is the effective oxide thickness, and ε is the permittivity. Consequently, the parameter c and the product $d\varepsilon$ can be fitted.

In order to model the variability of c and $d\varepsilon$, an attempt was made to predict their values on the basis of observable variables. Deviations from any constructed trend would be indicative of the uncertainty in the values of these quantities. It has been demonstrated that memristive devices can behave differently in different resistance states [181], and this phenomenon has often been linked to the conductance quantum $G_0 = 2e^2/h$ [291].

The variables c and $d\varepsilon$ serve to reiterate some of the observations made regarding the I-V behaviour of the SiO_x device, thereby assisting in the assessment of the appropriateness of the Poole-Frenkel model. It was demonstrated that c behaves in a manner analogous to that of conductance, specifically as the reciprocal of resistance [114]. The primary distinction between the lower- and higher-resistance states is that the discrepancies from the trend are markedly more pronounced in the latter.

In (6.35), the product $d\varepsilon$ is indicative of the degree of nonlinearity. Given that this product appears in the denominator of an exponentiated square root, a smaller value of $d\varepsilon$ corresponds to a more nonlinear I-V curve. In LRS, this product is significantly larger. Although this parameter algebraically (6.35) can approximate linear behaviour (i.e. ohmic conduction), the values of $d\varepsilon$ are only plausible for the less conductive states proximate to and below G_0.

For higher-resistance states, this is less discernible. The $d\varepsilon$ values are considerably lower for all such states, yet the trend is not only plateaued but also more erratic. This unpredictability is indicative of the diverse range of colours observed in the curves presented in Figure 6.6. While the significant deviations from the trend line may limit the applicability of this

approach in conventional modelling scenarios, the inherent uncertainty is precisely what is required to assess the potential of ex-situ training in addressing unknown behaviours.

The individual data points were integrated into a statistical model that would facilitate the generation of as many data points as required. In order to evaluate the efficacy of the proposed training method, two distinct resistance regions were considered. The low-resistance range was constructed by means of interpolation of the model parameters between the lowest resistance state and five times that resistance. The high-resistance range was constructed by interpolating the model parameters between the highest resistance state and one-fifth of that resistance, to ensure that the dynamic range remained consistent.

The statistical model was employed to ascertain the output current of a given device in the following manner: the initial values of c and $d\varepsilon$ were interpolated from the pertinent fits using the resistance parameter R. Thereafter, c and $d\varepsilon$ were subjected to disturbance via a multivariate normal distribution, which took into account both sets of parameters. The covariance matrix was determined using the residuals of the fits, and the current I was determined using (6.35).

As previously stated, the selection of the output function g in ex-situ training with linearity-nonpreserving nonidealities is contingent upon both the mapping scheme and the intrinsic nature of the nonlinearity. The combination of a linear mapping between inputs and outputs (6.25), a double weights mapping (6.26), and the relationship between current and voltage characterised by c and $d\varepsilon$ (6.35); allows g to be expressed as follows:

$$g\left(x, W_{\pm}\right) = \frac{cx \cdot exp\left(\frac{2e}{kbT}\sqrt{\frac{ek_V x}{4\pi d\varepsilon}}\right)}{k_G} \tag{6.36}$$

$$\begin{bmatrix} c \\ d\varepsilon \end{bmatrix} = exp\left(-ln\left(k_G W_{\pm} + G_{off}\right)\mathbf{m} + \mathbf{b} + \mathbf{E}\right) \tag{6.37}$$

Where $\mathbf{m}$ is the slopes, $\mathbf{b}$ is the intercepts, and the error is drawn from a normal distribution $\mathbf{E} \sim \mathcal{N}_2\left(0, \Sigma\right)$, with 0 mean and standard deviation being the residual covariance matrix.

6.3 Inference and Classification

6.3.1 Simulation Configurations

6.4 Summary

Chapter 7

Homeostasis Optimisation

7.1 Optimisation Overview

Spike-based optimisation techniques seek to minimise the overall error of a network by optimising the times of individual neuron spikes. As the time at which a neuron spikes is a continuous value, continuous optimisation methods can be employed. However, the problem remains highly non-linear due to the potential for a small change in the input to a neuron (and thus a small change to the neuron's input weights) to push the neuron over its firing threshold, eliciting a spike and drastically changing the neuron's output [81].

The initial algorithm to perform supervised deep learning by optimising spike times was SpikeProp [17]. This algorithm makes the simplifying assumption that each neuron will fire at most one spike during the spiking interval. In the event that multiple spikes are fired, only the first is optimised. Additionally, each connection is composed of numerous synaptic terminals, each with a distinct synaptic delay and connection weight.

The authors demonstrate that their algorithm can solve the XOR problem and performs comparably to backpropagation, which has been optimized both with gradient descent (GD) and Levenberg-Marquardt (LM), on a number of small datasets (the largest of which has 36 input dimensions, six output classes, and 4,435 training examples).

The single-spike optimisation procedure and multiple connection weights per synapse have presented significant challenges in expanding this work to larger datasets. Another work [178] presented two methods to improve the rate of convergence of SpikeProp; however, the applications remain limited to small datasets.

Another publication put forth an alternative to the SpikeProp algorithm [194], which was designed with the specific intention of accommodating non-leaky integrate-and-fire neurons. The proposed method relaxes the restriction that a connection must be composed of many different discrete-delay elements, and instead employs a more standard network architecture, with one exponential-synapse connection between each pair of neurons. The networks were trained with both one and two hidden layers, achieving 2.45% and 2.86% accuracy on the MNIST dataset, respectively.

One challenge encountered by the author was that the dropout technique, the most commonly employed regularisation method, was ineffective in the network under consideration. This was because it frequently resulted in the complete cessation of neuronal firing. In the absence of an efficacious alternative regularisation method, the networks exhibited considerable generalisation errors, the training error for both networks was almost zero.

While earlier algorithms focused on optimising the initial spike of each neuron, another relaxed this constraint by employing a genetic algorithm to enhance multiple spikes from each neuron [234]. However, they limited their demonstration to relatively modest models comprising fewer than 10 hidden neurons. Genetic algorithms frequently encounter challenges in scaling to problems with numerous parameters, raising concerns about the scalability of this algorithm to datasets such as MNIST or CIFAR-10.

In a similar vein, another publication [144] also optimise over multiple spikes per neuron. The authors disregard the spiking discontinuity that occurs during backpropagation, instead treating the output of a neuron as a linear function of its inputs, which have been filtered by the membrane of the neuron in question. This enables the network to be run in spiking neurons during training, while still performing backpropagation without concern for discontinuities.

Furthermore, the refractory period of the neurons is disregarded, as it is deemed to be relatively brief in comparison to the time interval between spikes, and therefore has a minimal impact on firing rates. In order to enhance the performance of the network, lateral inhibition components are introduced; however, only the first-order derivatives caused by these connections are optimised.

Notwithstanding these simplifications, their method is still capable of learning appropriately on the MNIST task, achieving 1.30% error using standard SGD and 1.23% error using an

ADAM optimiser. While it remains to be seen whether this method can generalise to larger datasets in a tractable way, it introduces a number of new ideas for training spiking networks that will hopefully be improved upon by future work.

A different study introduce a novel method for making the spiking process continuous [103]. They introduce a gating function, *g(V)*, of the membrane voltage, *V*, that is greater than zero for voltages approaching the firing threshold and zero otherwise, with unit integral. The region where $g(V) > 0$ is referred to as the active zone.

In contrast to the conventional approach, whereby efferent synapses receive a current $\delta(t - t_k)$ upon the neuron crossing the firing threshold at time t_k, this method allows synapses to continuously receive current based on $g(V)\frac{dV}{dt}$. In the event that the voltage is situated outside the active zone, this term is rendered zero due to the fact that *g(V)* is equal to zero.

Conversely, should the voltage traverse the active zone (and thus the neuron spike), the integral of this term is equal to one. Ultimately, if the voltage enters the active zone but does not exceed the upper threshold (i.e., the firing threshold), the integral of the term will be a positive number between zero and one (this is analogous to a partial spike).

This induced synaptic current is nearly identical to the traditional spiking current $\delta(t - t_k)$ in the extreme cases (i.e., when the neuron is silent or firing at a relatively high rate), but continuous in the intermediate region. The authors may then achieve a gradient through the network at any given point in time by employing backpropagation, and optimise the entire network by utilising backpropagation through time.

The results presented focus on tasks that require a dynamic, temporal representation, such as predictive coding. This represents a distinct focus in comparison to the majority of other spike-based methods, which tend to prioritise static tasks (such as object classification). Consequently, a direct comparison is challenging.

As a consequence of the method rendering the neural nonlinearity differentiable, a multitude of distinct neuron models may be employed. The present paper utilises non-leaky integrate-and-fire and quadratic integrate-and-fire models, with a plethora of alternative models being equally compatible.

This, in conjunction with the capacity to optimise recurrent spiking networks effectively, renders this a potentially potent methodology for dynamic spiking networks. In the context of networks specialising in static tasks, it is postulated that the necessity for this method to optimise over potentially lengthy time series for each input stimulus would render it unsuitable for training large object classification networks.

To date, spike-based optimisation methods have yet to be applied to larger, deeper architectures such as convolutional neural networks (CNNs). This has precluded the implementation of spike-based training on any datasets that are either larger or more challenging than the MNIST dataset.

One of the challenges lies in the computational requirements of spike-based optimisation methods, which necessitate more computational resources than rate-based methods. This is due to the dynamic nature of the network and the iterative simulation required for each stimulus presentation.

The majority of existing software has been designed for static artificial neural networks (ANNs), and extending it to spiking networks is a complex undertaking. Consequently, researchers have employed rate-based optimisation methods to address larger datasets with spiking networks, which will be discussed next.

Rate-based optimisation methods operate under the simplifying assumption that all neurons are engaged in rate coding. Consequently, these methods are indifferent to the times of individual spikes, focusing instead on the number of spikes occurring within a given time period.

In the majority of cases, these types of methods utilise this simplifying assumption to replace the spiking neural process with a continuous-valued rate approximation. In the case of derivative-based methods, this rate approximation is then differentiated. Derivative-free methods, in contrast, circumvent the necessity of taking the derivative of this rate approximation, instead opting for an optimisation method such as Contrastive Divergence that does not require it.

Finally, function approximation methods approach the problem from a different angle. Rather than assuming a network of spiking neurons and attempting to identify rate approximations

to these spiking neurons, they select an arbitrary nonlinearity for training the network, and then employ spiking neurons to approximate this nonlinearity.

7.1.1 Derivative-based Methods

In 2010, a study pioneered the training of a spiking convolutional network [205]. A conventional convolutional neural network (CNN) was trained using the backpropagation algorithm with *tanh* units. In order to transform this into a spiking network, the *tanh* units were substituted with binary threshold units that incorporate a refractory period.

In particular, if the input to a neuron exceeds a specified threshold, it generates a spike, after which the neuron is unable to spike for a designated period. The refractory period causes the firing rates of the neurons to saturate. The authors posit that this emulates the corresponding *sigmoid* functions, namely the *tanh* function, used in the rate model.

However, they do not elucidate the precise mechanism through which the refractory period achieves this emulation. It may be presumed that the rationale is that the *tanh* function is a saturating nonlinearity, and thus having the neuron firing rate saturate makes the firing rate curve more similar to the *tanh* curve.

A more principled approach to the conversion of rate-based convolutional neural networks (CNNs) to spiking CNNs was adopted in 2015 [25]. They exploited the fact that rectified linear units (ReLUs) perfectly model the rate of a non-leaky integrate-and-fire neuron (IF). As ReLUs are currently the standard for deep networks, this allows networks to be trained with the same rate-based ReLUs and then replaced with spiking IF neurons at runtime.

Additionally, the max-pooling layers utilized by the network were replaced with average pooling layers. Average pooling layers only entail the summation and scaling of inputs, thereby enabling them to function effectively with spiking neurons, in a manner that is biologically plausible. In contrast, max pooling lacks a clear spiking analogue.

The process of taking the maximum of the binary spikes occurring at a given point in time (i.e., outputting one if any of the inputs are spiking, and zero otherwise) is not equivalent to taking the maximum across spike rates, as performed by traditional CNNs. Moreover, even taking the maximum across spikes is not biologically plausible, since there is no clear neuron-level mechanism for taking the maximum across incoming signals.

The same study also removed biases from the network, on the grounds that they would be challenging to implement in a spiking network [25]. However, it is not evident why biases would present a problem for spiking networks, particularly if targeting neuromorphic hardware (much of which supports biases). One advantage of removing biases is that it offers a somewhat more straightforward biological explanation, since the output of each neuron depends solely on the inputs from other neurons in the network.

In a similar approach to that employed previously, a novel method of normalisation for weight magnitudes was utilised [51]. One pivotal design decision in spiking networks is the trade-off between the firing rates of the neurons and the requisite integration time for the network to reach a decision.

Higher firing rates facilitate the expeditious transmission of information, thereby enabling more rapid responses and enhanced accuracy. However, this increased speed comes at the cost of greater energy consumption. On certain neuromorphic platforms, there are imposed limits on the maximum allowable firing rates. An increase in integration time permits the accumulation of greater quantities of information, thereby enhancing accuracy. However, this approach also results in slower responses and an elevated number of spikes per example.

It is therefore necessary to achieve a balance between accuracy, response time and energy efficiency by properly tuning the firing rates and integration time. Given that ReLUs—and, by extension, IF neurons—are scale-invariant, the firing rates of neurons in spiking IF networks can be set at any arbitrary value. In order to alter the firing rate of a neuron while preserving the integrity of the network, it is necessary to offset the increase in firing rate with a corresponding decrease in all connection weights originating from that neuron.

The study present two techniques for normalising network weights, namely model-based and data-based normalisation [51]. Both techniques are founded upon the principle that the input to any given neuron in a single timestep should not exceed the firing threshold of that neuron. Should this occur, it could result in the neuron being required to fire two spikes in the same timestep, which is not supported by the majority of neuromorphic hardware.

In other words, the objective of the normalization process is to guarantee that the instantaneous firing rates of all neurons never exceed the rate of the neuromorphic chip (simulation rate). In the model-based normalization technique, the maximum sum of positive input weights across all neurons in a layer is identified, and then all neuron inputs in the layer are

normalized using this sum.

This guarantees that no neuron can receive an input greater than one (the firing threshold) in a single timestep. The authors discovered that this technique can result in weights that are smaller than necessary. This is because the maximum possible input to a neuron is often never encountered in real data, which in turn leads to longer-than-necessary integration times.

In contrast, data-based normalization is contingent upon the examples within the training data. This ensures that for any given training example, no neuron will receive an input greater than one (the firing threshold). The authors discovered that data-based normalization reduced the integration time required by the network to achieve the same level of accuracy as an unnormalized network with a higher firing threshold. This paper presented the most advanced results to date for spiking networks on the MNIST dataset, achieving a test-set error of 0.88%.

In a study published in 2016 [63], the authors trained deep networks targeting the IBM TrueNorth neuromorphic chip [189]. In contrast to the other networks described herein, this network employs a basic binary threshold neuron, which generates a "spike" in response to a positive input at a given time step and remains quiescent otherwise. This can be considered to be equivalent to a rate-based binary threshold unit.

Furthermore, the network presents each image for a single timestep before resetting all neuron voltages and presenting the next image. As a result of these considerations, the network is classified as rate-based, rather than spiking. The authors trained very large networks, comprising 8 million neurons, on a number of tasks, including CIFAR-10 and SVHN. The results are at the cutting edge of the field when compared to spiking networks.

However, a more pertinent comparison is with rate-based networks, where the sole innovation is the use of binary threshold units, which are not commonly employed in deep networks. The network is capable of running on neuromorphic hardware (TrueNorth), but utilises a considerable number of neurons even for relatively modest datasets such as CIFAR-10. Consequently, it would be challenging to scale up to larger datasets such as ImageNet.

7.1.2 Derivative-free Methods

Derivative-free, rate-based optimisation methods employ rate approximations for the spiking neural process, yet do not necessitate the derivative of the rate approximation. The entirety of the aforementioned methods are founded upon the utilisation of the Contrastive Divergence

(CD) algorithm [90] for the training of deep networks comprising stacked RBMs.

In a study published in 2013 [201], a spiking network of LIF neurons using RBMs was developed. In order to train the network, the researchers employ the Siegert approximation to a noisy spiking LIF neuron as the rate neuron model. The Siegert approximation is an equation that describes the mean firing rate of an LIF neuron when presented with a number of input spike trains, where the spikes are distributed according to a Poisson process with a constant firing rate.

Subsequently, the output rates are normalised by the maximum firing rate (1/tref) in order to generate firing probabilities, and the hidden units are sampled from a binary distribution based on these probabilities. The authors employ this model to simulate the variability introduced by the use of spikes.

Consequently, spiking LIF neurons are introduced in place of the rate neurons, and the network is trained using Poisson input trains based on pixel intensities, resulting in an error rate of 5.91% on MNIST. A study in 2015 [235] further extend this work by demonstrating the method's resilience to reduced weight precision, achieving an error rate of 5.06% on MNIST with 11-bit fixed point weights running on the SpiNNaker platform.

In a similar vein, a 2013 model was also developed spiking RBMs using LIF neurons [197]. Their approach entails fitting the mean firing rate of a noisy LIF neuron to a sigmoid curve and subsequently training using this sigmoid curve. This restricts the range of possible parameters for the LIF neuron, as they must permit the firing rate to be accurately approximated by a sigmoid function. However, it permits the RBMs to be trained with conventional sigmoid activation functions, thereby facilitating a more straightforward and efficient training process.

They put forth an online variant of CD for spiking networks, based on an STDP learning rule. For each training image, two phases are conducted. The initial phase occurs immediately upon presentation of the image, wherein connections are updated based on the data (stimulus). Subsequently, following a designated period of time, the recurrent connections become active, thereby driving the network towards its reconstruction distribution.

This has an effect that is similar to that of Gibbs sampling in the CD algorithm. During this period, the gating signal on the derivative is reversed, resulting in a negative impact on the weights based on the reconstruction. This constitutes one complete cycle of the algorithm,

which is repeated for each training image over numerous iterations through the training data. This approach bears resemblance to CD, hence the authors have designated it as "event-based CD." Employing this technique, the researchers achieved an 8.1% test-set error on MNIST in spiking neurons.

7.1.3 Function-approximation and Noise

Approximation methods for functions diverge considerably from the previously discussed approaches. In contrast to the approach of utilising a single spiking neuron to represent each node in an ANN, function approximation methods employ multiple spiking neurons for each node. These spiking neurons are employed to approximate the nonlinear function (e.g., the sigmoid function) that is being computed by the node.

There is only one documented instance of this method being employed [58]. The approximation of the sigmoid function computed by each node is achieved through the use of three spiking LIF neurons. The parameters for each neuron are randomly assigned, with the most significant aspect being the uniform distribution of the random bias current. This enables each of the three neurons to target a distinct portion of the sigmoid curve, in addition to the bias current assigned to the node during the training process.

Once the random parameters have been set, a scalar weighting is calculated for each neuron, determining the extent to which it contributes to the sigmoid function. This is the standard method for function approximation, but applied to a limited number of neurons in order to approximate a sigmoid function. This method enables a spiking network to approximate a rate-based network in a highly general manner.

One disadvantage is that it requires a greater number of neurons than nodes in the network, which makes it more costly in terms of neural resources than the aforementioned methods. The efficacy of this approach is contingent upon the ability of the neurons to accurately represent the node function within the operational domain. Sigmoid functions can be readily represented with LIF neurons. However, ReLUs are considerably flatter and do not undergo squashing, necessitating a significantly larger range of output values for accurate representation. This renders them less compatible with this method.

The transmission of discrete, constant-amplitude spikes, as observed throughout much of the human central nervous system, represents a higher-variance method of information transmission than the simple transmission of a scalar value, such as a voltage. From a rate-based

perspective, the increased variance can be conceptualised as 'noise' surrounding the firing rate, which can be considered the 'signal'. The firing rate is defined as the mean value of a spike train signal.

If a spike train is passed through a low-pass filter, the result can be viewed as an estimate of the mean with additional time-varying noise superimposed. The quantity of noise is contingent upon the filter employed; the elimination of a greater proportion of high frequencies will yield a superior estimate with diminished noise, yet if the fundamental firing rate signal is also time-varying, then its higher frequencies will also be removed. In a spiking network that is receiving a sequence of images and outputting a response for each in real-time, this results in a longer response time for each image.

It is observed that a considerable number of rate-based spiking neural networks do not take into account the variance caused by spikes. Amongst those that do, two distinct approaches have been identified. The first involves modelling the impact of spike noise on the mean neural firing rates, with these revised mean rates then being employed during the training process. The second approach entails incorporating stochastic elements into the model during training, with the objective of ensuring that the probability distribution of the training output rates reflects, to some extent, the variance resulting from the spikes employed during testing.

An exemplar of this initial approach is the utilisation of the Siegert model to account for the effects of incoming spikes on the LIF neuron firing rate [201]. This approach is predicated on the assumption that neuronal inputs are uncorrelated and that spikes are governed by a Poisson process.

The Siegert model then provides the mean firing rate of the LIF neuron, given the mean firing rates of all inputs and their respective connection weights. One disadvantage of this approach is that the Siegert equation is complex, containing an integral with no closed-form solution. The integrand of this integral contains both the exponential and error (erf) functions.

Computing this function on a GPU (or even a CPU) would be challenging and time-consuming in comparison to the functions typically employed in neural networks. Additionally, this approach necessitates the calculation of two linear functions of the input neuron rates, one to determine the overall input mean and the other to ascertain the variance. Consequently, it requires twice the number of matrix-vector operations as a network without

the Siegert model.

The second approach, which involves incorporating stochastic elements into a model of spiking noise, has frequently been employed in a range of neural network contexts. However, this has not typically been done with the specific objective of running the final model in spiking neurons.

A foundational example is that of Boltzmann machines, which employ binary sampling on probabilistic units [91]. This enables units to communicate binary values with one another, which bears resemblance to spikes. However, if one considers each unit to be attempting to transmit its underlying firing probability (which is analogous to a firing rate), this represents a particularly severe form of noise.

Subsequent work significantly reduced the amount of noise by employing rectified linear activation functions with Gaussian noise on the neuron input [196]. An additional illustration of the incorporation of stochastic elements into a model is the denoising autoencoder [257]. This approach entails the introduction of noise into the model inputs, followed by an attempt to reproduce the original, noise-free version of the input.

The addition of noise to model inputs has frequently been employed as a means of enhancing the model's capacity to handle a diverse range of inputs, effectively constituting a form of data augmentation. Other forms of data augmentation, such as translation, rotation, or deformation of images, can also be regarded as a structured form of noise on model inputs, distinct from Gaussian noise.

Nevertheless, the incorporation of noise into hidden units, whether at the input or output level, remains a relatively uncommon practice in the context of nonlinear networks [209]. Despite extensive research, there seems to be a dearth of examples where this approach has been employed to model the inherent variability associated with spikes, with the aim of translating a rate model into a spiking neuron framework.

7.2 Homeostasis Regularization

7.2.1 Programming Variabilities

One of the most common non-idealities observed in memristive devices is the phenomenon of devices becoming stuck. This topic was previously explored in depth in the preceding chapter. In this chapter, a similar model is employed, in which devices may become stuck at G_{off} or G_{on}. For both types of simulations, a range of probability were used, indicating that any individual may become stuck in that state. Although this is a simple model, it is not data-specific, and thus could be combined with the nonidealities that were modelled using experimental data, specifically SiO_x nonlinearities.

A more realistic probabilistic model for describing stuck device behaviour in crossbar array can be introduced. This is achieved by randomly drawing from a sample of all memristors and setting the conductance to the closest achievable state. Although this method is relatively robust given the high number of devices, it is discrete in nature and therefore more difficult to apply during SNN training, where gradients need to be computed.

Furthermore, the method is limited in data in certain conductance regions, which may result in training that fits the weights to the behaviour of a single instance of a crossbar array. Consequently, a more continuous approach to picking the states at which the devices may get stuck in was adopted.

It is possible to display previously encountered data; in this case, all potentiation and depression cycles are shown for some of the devices. The majority of memristors are capable of achieving a high conductance range, with only a small proportion exhibiting limited conductance or remaining in a fixed state. In this chapter, the simulations were conducted with G_{off} and G_{on} defined as the median of minimum and maximum conductances, respectively.

Any device whose maximum range (i.e., the difference between the highest and lowest conductances achieved) was less than half the median range (Gon - Goff) was classified as stuck. A further simplifying assumption was made that any such device would be treated as fully stuck. This overestimates the effect of the variability because in reality some of the 'stuck' devices may still be tweaked, albeit within a narrower range.

The challenge in constructing a probabilistic model in this case is the generation of a probability density function (PDF) that accurately describes the conductance values at which

devices may become stuck. The selection of devices that may become stuck is a relatively straightforward process.

These devices can be chosen randomly, with the proportion of devices that fit the previously defined criteria for stuck behaviour being the determining factor. For instance, 10% of the devices may be deemed to exhibit the necessary characteristics. However, the conversion of the discrete mean conductance values of these devices into a probability distribution that can be applied to a wider range of situations represents a more challenging aspect of this process.

In order to produce a probabilistic model, it was decided that kernel density estimation (KDE) would be employed. KDE is a method of producing a probability density function (PDF) given a sequence of randomly distributed variables. Each point is usually approximated with a random distribution, which are then summed together.

As these distributions are typically identical, the only choices that have to be made are the type of distributions to be used, or the width of those distributions, which is more commonly referred to as "bandwidth" [250] in the context of KDE. In this chapter, it was decided to employ truncated normal distributions, truncated at zero to circumvent the issue of negative conductance.

The underlying normal distributions were set to have a mean equal to the mean conductance of faulty devices, with the objective of estimating the standard deviation of the underlying distributions. In order to achieve this, Scott's rule [218] was utilised. As a consequence of the clipping of conductance values below zero, a bias is introduced, whereby the probability density is underestimated in the vicinity of the boundary, even if the PDF is re-scaled after truncation [225].

To address this issue, a reflection method can be employed. Rather than performing re-normalisation, a second distribution was introduced, representing the reflection of the original normal distribution around zero. The negative part of this distribution was then clipped [117]. This approach ensures that the two truncated PDFs sum to one, while providing a more reliable estimate of the probability density near zero.

Furthermore, device-to-device (D2D) variability, which arises from inaccuracies during device programming, was also taken into account. As discussed in the preceding chapter, during the mapping of weights onto conductances, one may ultimately arrive at values that

differ from the desired outcome.

Lognormal distribution is a commonly utilized approach to model these discrepancies. For instance, it has been demonstrated that resistance deviations adhere to this trend, with the relative (and consequently, the absolute) magnitude of the deviations being more pronounced in the high-resistance state (HRS) compared to the low-resistance state (LRS).

7.2.2 Architecture Modifications

7.3 Biosignal Applications

7.4 Summary

Chapter 8

Conclusion

8.1 Contributions

8.2 Open Questions

8.3 Future Works

References

[1] Abbott, L. F. (1999). Lapicque's introduction of the integrate-and-fire model neuron (1907). *Brain research bulletin*, 50(5-6):303–304.

[2] Adler, D., Shur, M., Silver, M., and Ovshinsky, S. (1980). Threshold switching in chalcogenide-glass thin films. *Journal of Applied Physics*, 51(6):3289–3309.

[3] Agrawal, A., Lee, C., and Roy, K. (2019). X-changr: Changing memristive crossbar mapping for mitigating line-resistance induced accuracy degradation in deep neural networks. *arXiv preprint arXiv:1907.00285*.

[4] Ambrogio, S., Balatti, S., Cubeta, A., Calderoni, A., Ramaswamy, N., and Ielmini, D. (2014a). Statistical fluctuations in hfo x resistive-switching memory: part i-set/reset variability. *IEEE Transactions on electron devices*, 61(8):2912–2919.

[5] Ambrogio, S., Balatti, S., Cubeta, A., Calderoni, A., Ramaswamy, N., and Ielmini, D. (2014b). Statistical fluctuations in hfo x resistive-switching memory: Part ii—random telegraph noise. *IEEE Transactions on Electron Devices*, 61(8):2920–2927.

[6] Ambrogio, S., Narayanan, P., Tsai, H., Shelby, R. M., Boybat, I., Di Nolfo, C., Sidler, S., Giordano, M., Bodini, M., Farinha, N. C., et al. (2018). Equivalent-accuracy accelerated neural-network training using analogue memory. *Nature*, 558(7708):60–67.

[7] Amirsoleimani, A., Alibart, F., Yon, V., Xu, J., Pazhouhandeh, M. R., Ecoffey, S., Beilliard, Y., Genov, R., and Drouin, D. (2020). In-memory vector-matrix multiplication in monolithic complementary metal–oxide–semiconductor-memristor integrated circuits: Design choices, challenges, and perspectives. *Advanced Intelligent Systems*, 2(11):2000115.

[8] Ankit, A., Hajj, I. E., Chalamalasetti, S. R., Ndu, G., Foltin, M., Williams, R. S., Faraboschi, P., Hwu, W.-m. W., Strachan, J. P., Roy, K., et al. (2019). Puma: A programmable ultra-efficient memristor-based accelerator for machine learning inference. In *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 715–731.

[9] Asanuma, S., Akoh, H., Yamada, H., and Sawa, A. (2009). Relationship between resistive switching characteristics and band diagrams of ti/pr 1- x ca x mno 3 junctions. *Physical Review B*, 80(23):235113.

[10] Banerjee, R., Ray, S., Basu, N., Batabyal, A., and Barua, A. (1987). Degradation of tin-doped indium-oxide film in hydrogen and argon plasma. *Journal of applied physics*, 62(3):912–916.

[11] Bao, L., Zhu, J., Yu, Z., Jia, R., Cai, Q., Wang, Z., Xu, L., Wu, Y., Yang, Y., Cai, Y., et al. (2019). Dual-gated mos2 neuristor for neuromorphic computing. *ACS applied materials & interfaces*, 11(44):41482–41489.

[12] Bardeen, J. (1947). Surface states and rectification at a metal semi-conductor contact. *Physical review*, 71(10):717.

[13] Barmpatsalos, N. and Mehonic, A. (2021). SiOx memristor structure - Gradual RESET conductance modulation.

[14] Bastos, P. and Barbosa, R. (2022). Motor reserve: How to build neuronal resilience against ageing and neurodegeneration? *Revue Neurologique*, 178(8):845–854.

[15] Bear, M. F. and Malenka, R. C. (1994). Synaptic plasticity: Ltp and ltd. *Current opinion in neurobiology*, 4(3):389–399.

[16] Bill, J. and Legenstein, R. (2014). A compound memristive synapse model for statistical learning through stdp in spiking neural networks. *Frontiers in neuroscience*, 8:412.

[17] Bohte, S. M., Kok, J. N., and La Poutre, H. (2002). Error-backpropagation in temporally encoded networks of spiking neurons. *Neurocomputing*, 48(1-4):17–37.

[18] Boikov, Y. A., Goltsman, B., Yarmarkin, V., and Lemanov, V. (2002). Near-electrode model of transient currents in (ba, sr) tio 3 thin film capacitor structures. *Applied physics letters*, 80(21):4003–4005.

[19] Buckwell, M., Montesi, L., Hudziak, S., Mehonic, A., and Kenyon, A. J. (2015). Conductance tomography of conductive filaments in intrinsic silicon-rich silica rram. *Nanoscale*, 7(43):18030–18035.

[20] Burnstock, G. (2004). Cotransmission. *Current opinion in pharmacology*, 4(1):47–52.

[21] Burr, G. W., Shelby, R. M., Sidler, S., Di Nolfo, C., Jang, J., Boybat, I., Shenoy, R. S., Narayanan, P., Virwani, K., Giacometti, E. U., et al. (2015). Experimental demonstration and tolerancing of a large-scale neural network (165 000 synapses) using phase-change memory as the synaptic weight element. *IEEE Transactions on Electron Devices*, 62(11):3498–3507.

[22] Cai, F., Correll, J. M., Lee, S. H., Lim, Y., Bothra, V., Zhang, Z., Flynn, M. P., and Lu, W. D. (2019). A fully integrated reprogrammable memristor–cmos system for efficient multiply–accumulate operations. *Nature Electronics*, 2(7):290–299.

[23] Cai, W., Ellinger, F., Tetzlaff, R., and Schmidt, T. (2011). Abel dynamics of titanium dioxide memristor based on nonlinear ionic drift model. *arXiv preprint arXiv:1105.2668*.

[24] Campbell, K. A., Drake, K. T., and Barney Smith, E. H. (2016). Pulse shape and timing dependence on the spike-timing dependent plasticity response of ion-conducting memristors as synapses. *Frontiers in bioengineering and biotechnology*, 4:97.

[25] Cao, Y., Chen, Y., and Khosla, D. (2015). Spiking deep convolutional neural networks for energy-efficient object recognition. *International Journal of Computer Vision*, 113:54–66.

[26] Ceolini, E., Frenkel, C., Shrestha, S. B., Taverni, G., Khacef, L., Payvand, M., and Donati, E. (2020). Hand-gesture recognition based on emg and event-based camera sensor fusion: A benchmark in neuromorphic computing. *Frontiers in Neuroscience*, 14:637.

[27] Cha, E., Woo, J., Lee, D., Lee, S., Song, J., Koo, Y., Lee, J., Park, C. G., Yang, M. Y., Kamiya, K., et al. (2013). Nanoscale (10nm) 3d vertical reram and nbo 2 threshold selector with tin electrode. In *2013 IEEE International Electron Devices Meeting*, pages 10–5. IEEE.

[28] Chai, Z., Freitas, P., Zhang, W., Hatem, F., Zhang, J. F., Marsland, J., Govoreanu, B., Goux, L., and Kar, G. S. (2018). Impact of rtn on pattern recognition accuracy of rram-based synaptic neural network. *IEEE Electron Device Letters*, 39(11):1652–1655.

[29] Chang, S., Chae, S., Lee, S., Liu, C., Noh, T., Lee, J., Kahng, B., Jang, J., Kim, M., Kim, D.-W., et al. (2008). Effects of heat dissipation on unipolar resistance switching in pt/ ni o/ pt capacitors. *Applied Physics Letters*, 92(18):183507.

[30] Chang, T., Jo, S.-H., and Lu, W. (2011). Short-term memory to long-term memory transition in a nanoscale memristor. *ACS nano*, 5(9):7669–7676.

[31] Chase, S. M. and Young, E. D. (2006). Spike-timing codes enhance the representation of multiple simultaneous sound-localization cues in the inferior colliculus. *Journal of Neuroscience*, 26(15):3889–3898.

[32] Chen, A. (2016). A review of emerging non-volatile memory (nvm) technologies and applications. *Solid-State Electronics*, 125:25–38.

[33] Chen, L., Bang, W., Park, Y.-J., Ryan, E. T., King, S., and Kim, C.-U. (2010). Observation of space charge limited current by cu ion drift in porous low-k/cu interconnects. *Applied Physics Letters*, 96(9).

[34] Chen, L., Li, J., Chen, Y., Deng, Q., Shen, J., Liang, X., and Jiang, L. (2017a). Accelerator-friendly neural-network training: Learning variations and defects in rram crossbar. In *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2017*, pages 19–24. IEEE.

[35] Chen, P.-Y., Peng, X., and Yu, S. (2017b). Neurosim+: An integrated device-to-algorithm framework for benchmarking synaptic devices and array architectures. In *2017 IEEE International Electron Devices Meeting (IEDM)*, pages 6–1. IEEE.

[36] Chen, P.-Y., Peng, X., and Yu, S. (2018a). Neurosim: A circuit-level macro model for benchmarking neuro-inspired architectures in online learning. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 37(12):3067–3080.

[37] Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. (2018b). Neural ordinary differential equations. *Advances in neural information processing systems*, 31.

[38] Cheney, N., Schrimpf, M., and Kreiman, G. (2017). On the robustness of convolutional neural networks to internal architecture and weight perturbations. *arXiv preprint arXiv:1703.08245*.

[39] Chua, L. (1971). Memristor-the missing circuit element. *IEEE Transactions on circuit theory*, 18(5):507–519.

[40] Chua, L. (2019). Resistance switching memories are memristors. In *Handbook of memristor networks*, pages 197–230. Springer.

[41] Chua, L., Sbitnev, V., and Kim, H. (2012). Hodgkin–huxley axon is made of memristors. *International Journal of Bifurcation and Chaos*, 22(03):1230011.

[42] Chua, L. O. and Kang, S. M. (1976). Memristive devices and systems. *Proceedings of the IEEE*, 64(2):209–223.

[43] Ciocci, P., Lemineur, J.-F., Noël, J.-M., Combellas, C., and Kanoufi, F. (2021). Differentiating electrochemically active regions of indium tin oxide electrodes for hydrogen evolution and reductive decomposition reactions. an in situ optical microscopy approach. *Electrochimica Acta*, 386:138498.

[44] Cowley, A. and Sze, S. (1965). Surface states and barrier height of metal-semiconductor systems. *Journal of Applied Physics*, 36(10):3212–3220.

[45] Cox, H. R., Buckwell, M., Ng, W. H., Mannion, D. J., Mehonic, A., Shearing, P. R., Fearn, S., and Kenyon, A. J. (2021). A nanoscale analysis method to reveal oxygen exchange between environment, oxide, and electrodes in reram devices. *APL Materials*, 9(11).

[46] Dayan, P. and Abbott, L. F. (2005). *Theoretical neuroscience: computational and mathematical modeling of neural systems*. MIT press.

[47] Deal, B. E. (1974). The current understanding of charges in the thermally oxidized silicon structure. *Journal of the Electrochemical Society*, 121(6):198C.

[48] Deiss, S. R., Douglas, R. J., Whatley, A. M., and Maass, W. (1999). A pulse-coded communications infrastructure for neuromorphic systems. *Pulsed neural networks*, 6:157–78.

[49] Del Valle, J., Salev, P., Kalcheim, Y., and Schuller, I. K. (2020). A caloritronics-based mott neuristor. *Scientific reports*, 10(1):4292.

[50] Di Ventra, M., Pershin, Y. V., and Chua, L. O. (2009). Circuit elements with memory: memristors, memcapacitors, and meminductors. *Proceedings of the IEEE*, 97(10):1717–1724.

[51] Diehl, P. U., Neil, D., Binas, J., Cook, M., Liu, S.-C., and Pfeiffer, M. (2015). Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing. In *2015 International joint conference on neural networks (IJCNN)*, pages 1–8. ieee.

[52] DiMaria, D. and Cartier, E. (1995). Mechanism for stress-induced leakage currents in thin silicon dioxide films. *Journal of Applied physics*, 78(6):3883–3894.

[53] Du, C. (2017). *Metal Oxide Memristors with Internal Dynamics for Neuromorphic Applications*. PhD thesis, The University of Michigan, US.

[54] Duan, Q., Jing, Z., Zou, X., Wang, Y., Yang, K., Zhang, T., Wu, S., Huang, R., and Yang, Y. (2020). Spiking neurons with spatiotemporal dynamics and gain modulation for monolithically integrated memristive neural networks. *Nature communications*, 11(1):3399.

[55] El Kamel, F., Gonon, P., Ortéga, L., Jomni, F., and Yangui, B. (2006). Space charge limited transient currents and oxygen vacancy mobility in amorphous ba ti o 3 thin films. *Journal of applied physics*, 99(9):094107.

[56] El Kamel, F., Gonon, P., Yangui, B., and Jomni, F. (2007). Ionic and electronic defects in a-batio3 thin films studied by transient and steady state conductivity measurements. *physica status solidi c*, 4(3):1242–1245.

[57] El-Sayed, A.-M., Watkins, M. B., Shluger, A. L., and Afanas'ev, V. V. (2013). Identification of intrinsic electron trapping sites in bulk amorphous silica from ab initio calculations. *Microelectronic engineering*, 109:68–71.

[58] Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., and Rasmussen, D. (2012). A large-scale model of the functioning brain. *science*, 338(6111):1202–1205.

[59] Eshraghian, J. K., Baek, S., Kim, J.-H., Iannella, N., Cho, K., Goo, Y. S., Iu, H. H., Kang, S.-M., and Eshraghian, K. (2018). Formulation and implementation of nonlinear integral equations to model neural dynamics within the vertebrate retina. *International Journal of Neural Systems*, 28(07):1850004.

[60] Eshraghian, J. K., Cho, K., and Kang, S. M. (2021). A 3-d reconfigurable rram crossbar inference engine. In *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE.

[61] Eshraghian, J. K., Wang, X., and Lu, W. D. (2022). Memristor-based binarized spiking neural networks: Challenges and applications. *IEEE Nanotechnology Magazine*, 16(2):14–23.

[62] Eshraghian, J. K., Ward, M., Neftci, E. O., Wang, X., Lenz, G., Dwivedi, G., Bennamoun, M., Jeong, D. S., and Lu, W. D. (2023). Training spiking neural networks using lessons from deep learning. *Proceedings of the IEEE*.

[63] Esser, S. K., Appuswamy, R., Merolla, P., Arthur, J. V., and Modha, D. S. (2015). Backpropagation for energy-efficient neuromorphic computing. *Advances in neural information processing systems*, 28.

[64] Fang, Y., Yu, Z., Wang, Z., Zhang, T., Yang, Y., Cai, Y., and Huang, R. (2018). Improvement of hfo x-based rram device variation by inserting ald tin buffer layer. *IEEE Electron Device Letters*, 39(6):819–822.

[65] Feurer, M. and Hutter, F. (2019). Hyperparameter optimization. *Automated machine learning: Methods, systems, challenges*, pages 3–33.

[66] Frenkel, C., Legat, J.-D., and Bol, D. (2019). Morphic: A 65-nm 738k-synapse/mm^2 quad-core binary-weight digital neuromorphic processor with stochastic spike-driven online learning. *IEEE transactions on biomedical circuits and systems*, 13(5):999–1010.

[67] Frenkel, J. (1938). On pre-breakdown phenomena in insulators and electronic semiconductors. *Physical Review*, 54(8):647.

[68] Fröhlich, F. (2016). Chapter 4 - synaptic plasticity. In Fröhlich, F., editor, *Network Neuroscience*, pages 47–58. Academic Press, San Diego.

[69] Fujii, T., Kawasaki, M., Sawa, A., Akoh, H., Kawazoe, Y., and Tokura, Y. (2005). Hysteretic current–voltage characteristics and resistance switching at an epitaxial oxide schottky junction srruo3/ srti0. 99nb0. 01o3. *Applied Physics Letters*, 86(1).

[70] Gaba, S., Sheridan, P., Zhou, J., Choi, S., and Lu, W. (2013). Stochastic memristive devices for computing and neuromorphic applications. *Nanoscale*, 5(13):5872–5878.

[71] Gao, D. Z., El-Sayed, A.-M., and Shluger, A. L. (2016). A mechanism for frenkel defect creation in amorphous sio2 facilitated by electron injection. *Nanotechnology*, 27(50):505207.

[72] Gaol, D., Zhang, G. L., Yin, X., Li, B., Schlichtmann, U., and Zhuo, C. (2021). Reliable memristor-based neuromorphic design using variation-and defect-aware training. In *2021 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*, pages 1–9. IEEE.

[73] Ge, N., Zhang, M., Zhang, L., Yang, J. J., Li, Z., and Williams, R. S. (2014). Electrode-material dependent switching in taox memristors. *Semiconductor Science and Technology*, 29(10):104003.

[74] Gerstner, W., Kistler, W. M., Naud, R., and Paninski, L. (2014). *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press.

[75] Gerstner, W., Kreiter, A. K., Markram, H., and Herz, A. V. (1997). Neural codes: firing rates and beyond. *Proceedings of the National Academy of Sciences*, 94(24):12740–12741.

[76] Gewaltig, M.-O. and Diesmann, M. (2007). Nest (neural simulation tool). *Scholarpedia*, 2(4):1430.

[77] Ghanbari, A., Malyshev, A., Volgushev, M., and Stevenson, I. H. (2017). Estimating short-term synaptic plasticity from pre-and postsynaptic spiking. *PLoS computational biology*, 13(9):e1005738.

[78] Ghibaudo, G., Riess, P., Bruyere, S., DeSalvo, B., Jahan, C., Scarpa, A., Pananakakis, G., and Vincent, E. (1999). Emerging oxide degradation mechanisms: stress induced leakage current (silc) and quasi-breakdown (qb). *Microelectronic engineering*, 49(1-2):41–50.

[79] Goux, L. and Valov, I. (2016). Electrochemical processes and device improvement in conductive bridge ram cells. *physica status solidi (a)*, 213(2):274–288.

[80] Govoreanu, B., Redolfi, A., Zhang, L., Adelmann, C., Popovici, M., Clima, S., Hody, H., Paraschiv, V., Radu, I., Franquet, A., et al. (2013). Vacancy-modulated conductive oxide resistive ram (vmco-rram): An area-scalable switching current, self-compliant, highly nonlinear and wide on/off-window resistive switching cell. In *2013 IEEE International Electron Devices Meeting*, pages 10–2. IEEE.

[81] Gütig, R. (2014). To spike, or when to spike? *Current opinion in neurobiology*, 25:134–139.

[82] Han, S., Pool, J., Tran, J., and Dally, W. (2015). Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28.

[83] Hansen, M., Ziegler, M., Kolberg, L., Soni, R., Dirkmann, S., Mussenbrock, T., and Kohlstedt, H. (2015). A double barrier memristive device. *Scientific reports*, 5(1):13753.

[84] Hao, S., Ji, X., Zhong, S., Pang, K. Y., Lim, K. G., Chong, T. C., and Zhao, R. (2020). A monolayer leaky integrate-and-fire neuron for 2d memristive neuromorphic networks. *Advanced Electronic Materials*, 6(4):1901335.

[85] Hasan, R., Taha, T. M., and Yakopcic, C. (2017). On-chip training of memristor crossbar based multi-layer neural networks. *Microelectronics journal*, 66:31–40.

[86] Hazan, H., Saunders, D. J., Khan, H., Patel, D., Sanghavi, D. T., Siegelmann, H. T., and Kozma, R. (2018). Bindsnet: A machine learning-oriented spiking neural networks library in python. *Frontiers in neuroinformatics*, 12:89.

[87] He, Z., Lin, J., Ewetz, R., Yuan, J.-S., and Fan, D. (2019). Noise injection adaption: End-to-end reram crossbar non-ideal effect adaption for neural network mapping. In *Proceedings of the 56th Annual Design Automation Conference 2019*, pages 1–6.

[88] Hebb, D. (2002). The organization of behavior. 1949. *New York Wiely*, 2(7):8.

[89] Hebb, D. O. (2005). *The organization of behavior: A neuropsychological theory*. Psychology Press.

[90] Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554.

[91] Hinton, G. E. and Sejnowski, T. J. (1983). Optimal perceptual inference. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, volume 448, pages 448–453. Citeseer.

[92] Hirose, Y. and Hirose, H. (1976). Polarity-dependent memory switching and behavior of ag dendrite in ag-photodoped amorphous as2s3 films. *Journal of Applied Physics*, 47(6):2767–2772.

[93] Hochstetter, J., Zhu, R., Loeffler, A., Diaz-Alvarez, A., Nakayama, T., and Kuncic, Z. (2021). Avalanches and edge-of-chaos learning in neuromorphic nanowire networks. *Nature Communications*, 12(1):4008.

[94] Hodgkin, A. L. and Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, 117(4):500.

[95] Hofstein, S. R. (1967). Proton and sodium transport in sio 2 films. *IEEE Transactions on Electron Devices*, 14(11):749–759.

[96] Hu, M., Strachan, J. P., Li, Z., Stanley, R., et al. (2016). Dot-product engine as computing memory to accelerate machine learning algorithms. In *2016 17th International Symposium on Quality Electronic Design (ISQED)*, pages 374–379. IEEE.

[97] Hu, S., Wu, S., Jia, W., Yu, Q., Deng, L., Fu, Y. Q., Liu, Y., and Chen, T. P. (2014). Review of nanostructured resistive switching memristor and its applications. *Nanoscience and Nanotechnology Letters*, 6(9):729–757.

[98] Huang, A., Zhang, X., Li, R., and Chi, Y. (2018). Memristor neural network design. *Memristor and Memristive Neural Networks*, pages 1–35.

[99] Huang, C., Li, K., Tu, G., and Wang, W. (2003). The electrochemical behavior of tin-doped indium oxide during reduction in 0.3 m hydrochloric acid. *Electrochimica Acta*, 48(24):3599–3605.

[100] Huang, J., Stathopoulos, S., Serb, A., and Prodromakis, T. (2022). Neuropack: An algorithm-level python-based simulator for memristor-empowered neuro-inspired computing. *Frontiers in Nanotechnology*, 4:851856.

[101] Huang, L., Diao, J., Teng, S., Li, Z., Wang, W., Liu, S., Li, M., and Liu, H. (2021). A method for obtaining highly robust memristor based binarized convolutional neural network. In *INTERNATIONAL CONFERENCE ON WIRELESS COMMUNICATIONS, NETWORKING AND APPLICATIONS*, pages 813–822. Springer.

[102] Huang, P., Li, Z., Dong, Z., Han, R., Zhou, Z., Zhu, D., Liu, L., Liu, X., and Kang, J. (2019). Binary resistive-switching-device-based electronic synapse with spike-rate-dependent plasticity for online learning. *ACS Applied Electronic Materials*, 1(6):845–853.

[103] Huh, D. and Sejnowski, T. J. (2018). Gradient descent for spiking neural networks. *Advances in neural information processing systems*, 31.

[104] Ielmini, D., Nardi, F., and Cagli, C. (2010). Resistance-dependent amplitude of random telegraph-signal noise in resistive switching memories. *Applied Physics Letters*, 96(5):053503.

[105] Ielmini, D., Spinelli, A. S., Rigamonti, M. A., and Lacaita, A. L. (2000). Modeling of silc based on electron and hole tunneling. ii. steady-state. *IEEE Transactions on Electron Devices*, 47(6):1266–1272.

[106] Indiveri, G. (2021). Introducing 'neuromorphic computing and engineering'. *Neuromorphic Computing and Engineering*, 1(1):010401.

[107] Jain, S. and Raghunathan, A. (2019). Cxdnn: Hardware-software compensation methods for deep neural networks on resistive crossbar systems. *ACM Transactions on Embedded Computing Systems (TECS)*, 18(6):1–23.

[108] Jeong, D. S., Schroeder, H., and Waser, R. (2007). Coexistence of bipolar and unipolar resistive switching behaviors in a pt/ tio2/ pt stack. *Electrochemical and solid-state letters*, 10(8):G51.

[109] Jeong, J. H., Yang, H. W., Park, J.-S., Jeong, J. K., Mo, Y.-G., Kim, H. D., Song, J., and Hwang, C. S. (2008). Origin of subthreshold swing improvement in amorphous indium gallium zinc oxide transistors. *Electrochemical and Solid-State Letters*, 11(6):H157.

[110] Jiménez-Molinos, F., Palma, A., Gámiz, F., Banqueri, J., and López-Villanueva, J. (2001). Physical model for trap-assisted inelastic tunneling in metal-oxide-semiconductor structures. *Journal of Applied Physics*, 90(7):3396–3404.

[111] Jo, S. H., Chang, T., Ebong, I., Bhadviya, B. B., Mazumder, P., and Lu, W. (2010). Nanoscale memristor device as synapse in neuromorphic systems. *Nano letters*, 10(4):1297–1301.

[112] Jo, S. H., Kim, K.-H., and Lu, W. (2009). Programmable resistance switching in nanoscale two-terminal devices. *Nano letters*, 9(1):496–500.

[113] Johnston, D., Magee, J. C., Colbert, C. M., and Christie, B. R. (1996). Active properties of neuronal dendrites. *Annual review of neuroscience*, 19(1):165–186.

[114] Joksas, D. (2022). *Memristive crossbars as hardware accelerators: modelling, design and new uses*. PhD thesis, UCL (University College London).

[115] Joksas, D., Freitas, P., Chai, Z., Ng, W. H., Buckwell, M., Li, C., Zhang, W., Xia, Q., Kenyon, A., and Mehonic, A. (2020). Committee machines—a universal method to deal with non-idealities in memristor-based neural networks. *Nature communications*, 11(1):4273.

[116] Joksas, D., Wang, E., Barmpatsalos, N., Ng, W. H., Kenyon, A. J., Constantinides, G. A., and Mehonic, A. (2022). Nonideality-aware training for accurate and robust low-power memristive neural networks. *Advanced Science*, 9(17):2105784.

[117] Jones, M. C. (1993). Simple boundary correction for kernel density estimation. *Statistics and computing*, 3:135–146.

[118] Joshi, V., Le Gallo, M., Haefeli, S., Boybat, I., Nandakumar, S. R., Piveteau, C., Dazzi, M., Rajendran, B., Sebastian, A., and Eleftheriou, E. (2020). Accurate deep neural network inference using computational phase-change memory. *Nature communications*, 11(1):2473.

[119] Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S., Hudspeth, A. J., Mack, S., et al. (2000). *Principles of neural science*, volume 4. McGraw-hill New York.

[120] Kang, D.-H., Jun, H.-G., Ryoo, K.-C., Jeong, H., and Sohn, H. (2015). Emulation of spike-timing dependent plasticity in nano-scale phase change memory. *Neurocomputing*, 155:153–158.

[121] Kang, S. M., Choi, D., Eshraghian, J. K., Zhou, P., Kim, J., Kong, B.-S., Zhu, X., Demirkol, A. S., Ascoli, A., Tetzlaff, R., et al. (2021). How to build a memristive integrate-and-fire model for spiking neuronal signal generation. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 68(12):4837–4850.

[122] Kappel, D., Nessler, B., and Maass, W. (2014). Stdp installs in winner-take-all circuits an online approximation to hidden markov model learning. *PLoS computational biology*, 10(3):e1003511.

[123] Kempter, R., Gerstner, W., and Van Hemmen, J. L. (1999). Hebbian learning and spiking neurons. *Physical Review E*, 59(4):4498.

[124] Kendall, J., Pantone, R., Manickavasagam, K., Bengio, Y., and Scellier, B. (2020). Training end-to-end analog neural networks with equilibrium propagation. *arXiv preprint arXiv:2006.01981*.

[125] Kenyon, A. J., Munde, M. S., Ng, W. H., Buckwell, M., Joksas, D., and Mehonic, A. (2019). The interplay between structure and function in redox-based resistance switching. *Faraday discussions*, 213:151–163.

[126] Khan, S. A. and Kim, S. (2020). Comparison of diverse resistive switching characteristics and demonstration of transitions among them in al-incorporated hfo 2-based resistive switching memory for neuromorphic applications. *RSC advances*, 10(52):31342–31347.

[127] Kheradpisheh, S. R., Ganjtabesh, M., Thorpe, S. J., and Masquelier, T. (2018). Stdp-based spiking deep convolutional neural networks for object recognition. *Neural Networks*, 99:56–67.

[128] Khodagholy, D., Gelinas, J. N., Thesen, T., Doyle, W., Devinsky, O., Malliaras, G. G., and Buzsáki, G. (2015). Neurogrid: recording action potentials from the surface of the brain. *Nature neuroscience*, 18(2):310–315.

[129] Kiani, F., Yin, J., Wang, Z., Yang, J. J., and Xia, Q. (2021). A fully hardware-based memristive multilayer neural network. *Science advances*, 7(48):eabj4801.

[130] Kim, H., Mahmoodi, M., Nili, H., and Strukov, D. B. (2021). 4k-memristor analog-grade passive crossbar circuit. *Nature communications*, 12(1):5198.

[131] Kim, K. M., Yang, J. J., Strachan, J. P., Grafals, E. M., Ge, N., Melendez, N. D., Li, Z., and Williams, R. S. (2016). Voltage divider effect for the improvement of variability and endurance of taox memristor. *Scientific reports*, 6(1):20085.

[132] Knight, B. W. (1972). Dynamics of encoding in a population of neurons. *The Journal of general physiology*, 59(6):734–766.

[133] Koch, C. (2004). *Biophysics of computation: information processing in single neurons*. Oxford university press.

[134] Kozicki, M. N., Mitkova, M., and Valov, I. (2016). Electrochemical metallization memories. *Resistive switching: from fundamentals of nanoionic redox processes to memristive device applications*, pages 483–514.

[135] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

[136] Krzysteczko, P., Münchenberger, J., Schäfers, M., Reiss, G., and Thomas, A. (2012). The memristive magnetic tunnel junction as a nanoscopic synapse-neuron system. *Advanced Materials*, 24(6):762–766.

[137] Kusaka, T., Ohji, Y., and Mukai, K. (1987). Time-dependent dielectric breakdown of ultra-thin silicon oxide. *IEEE Electron Device Letters*, 8(2):61–63.

[138] Lagarias, J. C., Reeds, J. A., Wright, M. H., and Wright, P. E. (1998). Convergence properties of the nelder–mead simplex method in low dimensions. *SIAM Journal on optimization*, 9(1):112–147.

[139] Lammie, C., Xiang, W., Linares-Barranco, B., and Azghadi, M. R. (2022). Memtorch: An open-source simulation framework for memristive deep learning systems. *Neurocomputing*, 485:124–133.

[140] Lampert, M. A. and Schilling, R. B. (1970). Current injection in solids: The regional approximation method. In *Semiconductors and semimetals*, volume 6, pages 1–96. Elsevier.

[141] Lapicque, L. É. (1907). Louis lapicque. *J. physiol*, 9:620–635.

[142] Lapique, L. (1907). Researches quantatives sur l'excitation electrique des nerfs traitee comme une polarization. *Journal of Physiology, Pathology and Genetics*, 9:620–635.

[143] Lee, J., Schell, W., Zhu, X., Kioupakis, E., and Lu, W. D. (2019). Charge transition of oxygen vacancies during resistive switching in oxide-based rram. *ACS applied materials & interfaces*, 11(12):11579–11586.

[144] Lee, J. H., Delbruck, T., and Pfeiffer, M. (2016). Training deep spiking neural networks using backpropagation. *Frontiers in neuroscience*, 10:508.

[145] Lentz, F., Roesgen, B., Rana, V., Wouters, D. J., and Waser, R. (2013). Current compliance-dependent nonlinearity in tiox reram. *IEEE electron device letters*, 34(8):996–998.

[146] Li, B., Yan, B., Liu, C., and Li, H. (2019). Build reliable and efficient neuromorphic design with memristor technology. In *Proceedings of the 24th Asia and South Pacific Design Automation Conference*, pages 224–229.

[147] Li, C., Han, L., Jiang, H., Jang, M.-H., Lin, P., Wu, Q., Barnell, M., Yang, J. J., Xin, H. L., and Xia, Q. (2017a). Three-dimensional crossbar arrays of self-rectifying si/sio2/si memristors. *Nature communications*, 8(1):15666.

[148] Li, C., Hu, M., Li, Y., Jiang, H., Ge, N., Montgomery, E., Zhang, J., Song, W., Dávila, N., Graves, C. E., et al. (2018). Analogue signal and image processing with large memristor crossbars. *Nature electronics*, 1(1):52–59.

[149] Li, J., Xu, H., Sun, S.-Y., Liu, S., Li, N., Li, Q., Liu, H., and Li, Z. (2020). Enhanced spiking neural network with forgetting phenomenon based on electronic synaptic devices. *Neurocomputing*, 408:21–30.

[150] Li, Q., Khiat, A., Salaoru, I., Xu, H., and Prodromakis, T. (2014). Stochastic switching of tio 2-based memristive devices with identical initial memory states. *Nanoscale research letters*, 9:1–5.

[151] Li, Y. (2018). Analog computing using 1t1r crossbar arrays analog computing using 1t1r crossbar arrays. *Feb*, 28:1–81.

[152] Li, Y., Lei, Y., Shen, B., and Sun, J. (2015). Visible-light-accelerated oxygen vacancy migration in strontium titanate. *Scientific reports*, 5(1):14576.

[153] Li, Y., Long, S., Liu, Q., Lv, H., and Liu, M. (2017b). Resistive switching performance improvement via modulating nanoscale conductive filament, involving the application of two-dimensional layered materials. *Small*, 13(35):1604306.

[154] Li, Y., Zhong, Y., Xu, L., Zhang, J., Xu, X., Sun, H., and Miao, X. (2013). Ultrafast synaptic events in a chalcogenide memristor. *Scientific reports*, 3(1):1619.

[155] Lifshitz, N. and Smolinsky, G. (1989). Detection of water-related charge in electronic dielectrics. *Applied physics letters*, 55(4):408–410.

[156] Lim, H., Kornijcuk, V., Seok, J. Y., Kim, S. K., Kim, I., Hwang, C. S., and Jeong, D. S. (2015). Reliability of neuronal information conveyed by unreliable neuristor-based leaky integrate-and-fire neurons: a model study. *Scientific reports*, 5(1):9776.

[157] Lin, C.-Y., Chen, J., Chen, P.-H., Chang, T.-C., Wu, Y., Eshraghian, J. K., Moon, J., Yoo, S., Wang, Y.-H., Chen, W.-C., et al. (2020). Adaptive synaptic memory via lithium ion modulation in rram devices. *Small*, 16(42):2003964.

[158] Liu, C., Hu, M., Strachan, J. P., and Li, H. (2017). Rescuing memristor-based neuromorphic design with high defects. In *Proceedings of the 54th Annual Design Automation Conference 2017*, pages 1–6.

[159] Liu, J.-C., Hsu, C.-W., Wang, I.-T., and Hou, T.-H. (2015a). Categorization of multilevel-cell storage-class memory: an rram example. *IEEE Transactions on Electron Devices*, 62(8):2510–2516.

[160] Liu, L., Yellinek, S., Valdinger, I., Donval, A., and Mandler, D. (2015b). Important implications of the electrochemical reduction of ito. *Electrochimica Acta*, 176:1374–1381.

[161] Liu, Q., Long, S., Lv, H., Wang, W., Niu, J., Huo, Z., Chen, J., and Liu, M. (2010). Controllable growth of nanoscale conductive filaments in solid-electrolyte-based reram by using a metal nanocrystal covered bottom electrode. *ACS nano*, 4(10):6162–6168.

[162] Loeffler, A., Zhu, R., Hochstetter, J., Diaz-Alvarez, A., Nakayama, T., Shine, J. M., and Kuncic, Z. (2021). Modularity and multitasking in neuro-memristive reservoir networks. *Neuromorphic Computing and Engineering*, 1(1):014003.

[163] Mabilleau, G. and Sabokbar, A. (2008). In vitro biological test methods to evaluate bioresorbability. In *Degradation Rate of Bioresorbable Materials*, pages 145–160. Elsevier.

[164] Madams, C., Morgan, D., and Howes, M. (1974). Migration of gold atoms through thin silicon oxide films. *Journal of Applied Physics*, 45(11):5088–5090.

[165] Mainen, Z. F. and Sejnowski, T. J. (1995). Reliability of spike timing in neocortical neurons. *Science*, 268(5216):1503–1506.

[166] Malik, P. (2013). Governing big data: principles and practices. *IBM Journal of Research and Development*, 57(3/4):1–1.

[167] Manceau, J.-P., Bruyere, S., Jeannot, S., Sylvestre, A., and Gonon, P. (2007a). Current instability, permittivity variation with frequency, and their relationship in capacitor. *IEEE transactions on device and materials reliability*, 7(2):315–323.

[168] Manceau, J.-P., Bruyere, S., Jeannot, S., Sylvestre, A., and Gonon, P. (2007b). Metal-insulator-metal capacitors' current instability improvement using dielectric stacks to prevent oxygen vacancies formation. *Applied Physics Letters*, 91(13).

[169] Mandal, S., El-Amin, A., Alexander, K., Rajendran, B., and Jha, R. (2014). Novel synaptic memory device for neuromorphic computing. *Scientific reports*, 4(1):5333.

[170] Mannion, D. J. (2022). *Current transient phenomena in silicon oxide resistance switching oxides: characterisation and computational applications*. PhD thesis, UCL (University College London).

[171] Mannion, D. J., Vu, V. C., Ng, W. H., Mehonic, A., and Kenyon, A. J. (2023). Unipolar potentiation and depression in memristive devices utilizing the subthreshold regime. *IEEE Transactions on Nanotechnology*, 22:313–320.

[172] Many, A. and Rakavy, G. (1962). Theory of transient space-charge-limited currents in solids in the presence of trapping. *Physical Review*, 126(6):1980.

[173] Maranhão, G. and Guimarães, J. G. (2021). Low-power hybrid memristor-cmos spiking neuromorphic stdp learning system. *IET Circuits, Devices & Systems*, 15(3):237–250.

[174] Marder, E. and Goaillard, J.-M. (2006). Variability, compensation and homeostasis in neuron and network function. *Nature Reviews Neuroscience*, 7(7):563–574.

[175] Markram, H., Gerstner, W., and Sjöström, P. J. (2012). Spike-timing-dependent plasticity: a comprehensive overview. *Frontiers in synaptic neuroscience*, 4:2.

[176] Markram, H., Lübke, J., Frotscher, M., and Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic aps and epsps. *Science*, 275(5297):213–215.

[177] Markram, H., Muller, E., Ramaswamy, S., Reimann, M. W., Abdellah, M., Sanchez, C. A., Ailamaki, A., Alonso-Nanclares, L., Antille, N., Arsever, S., et al. (2015). Reconstruction and simulation of neocortical microcircuitry. *Cell*, 163(2):456–492.

[178] McKennoch, S., Liu, D., and Bushnell, L. G. (2006). Fast modifications of the spikeprop algorithm. In *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, pages 3970–3977. IEEE.

[179] Mead, C. (1990). Neuromorphic electronic systems. *Proceedings of the IEEE*, 78(10):1629–1636.

[180] Mead, C. and Ismail, M. (1989). *Analog VLSI implementation of neural systems*, volume 80. Springer Science & Business Media.

[181] Mehonic, A., Buckwell, M., Montesi, L., Garnett, L., Hudziak, S., Fearn, S., Chater, R., McPhail, D., and Kenyon, A. J. (2015). Structural changes and conductance thresholds in metal-free intrinsic siox resistive random access memory. *Journal of Applied Physics*, 117(12).

[182] Mehonic, A., Cueff, S., Wojdak, M., Hudziak, S., Jambois, O., Labbé, C., Garrido, B., Rizk, R., and Kenyon, A. J. (2012a). Resistive switching in silicon suboxide films. *Journal of Applied Physics*, 111(7):074507.

[183] Mehonic, A., Cueff, S., Wojdak, M., Hudziak, S., Labbé, C., Rizk, R., and Kenyon, A. J. (2012b). Electrically tailored resistance switching in silicon oxide. *Nanotechnology*, 23(45):455201.

[184] Mehonic, A., Joksas, D., Ng, W. H., Buckwell, M., and Kenyon, A. J. (2019). Simulation of inference accuracy using realistic rram devices. *Frontiers in neuroscience*, 13:593.

[185] Mehonic, A., Munde, M. S., Ng, W., Buckwell, M., Montesi, L., Bosman, M., Shluger, A., and Kenyon, A. (2017). Intrinsic resistance switching in amorphous silicon oxide for high performance siox reram devices. *Microelectronic Engineering*, 178:98–103.

[186] Mehonic, A., Shluger, A. L., Gao, D., Valov, I., Miranda, E., Ielmini, D., Bricalli, A., Ambrosi, E., Li, C., Yang, J. J., et al. (2018). Silicon oxide (siox): A promising material for resistance switching? *Advanced materials*, 30(43):1801187.

[187] Mel, B. W. (1994). Information processing in dendritic trees. *Neural computation*, 6(6):1031–1085.

[188] Menke, T., Meuffels, P., Dittmann, R., Szot, K., and Waser, R. (2009). Separation of bulk and interface contributions to electroforming and resistive switching behavior of epitaxial fe-doped srtio 3.

[189] Merolla, P. A., Arthur, J. V., Alvarez-Icaza, R., Cassidy, A. S., Sawada, J., Akopyan, F., Jackson, B. L., Imam, N., Guo, C., Nakamura, Y., et al. (2014). A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science*, 345(6197):668–673.

[190] Meyer, R., Liedtke, R., and Waser, R. (2005). Oxygen vacancy migration and time-dependent leakage current behavior of ba0. 3sr0. 7tio3 thin films. *Applied Physics Letters*, 86(11).

[191] Molter, T. W. and Nugent, M. A. (2016). The generalized metastable switch memristor model. In *CNNA 2016; 15th International workshop on cellular nanoscale networks and their applications*, pages 1–2. VDE.

[192] Moon, K., Lim, S., Park, J., Sung, C., Oh, S., Woo, J., Lee, J., and Hwang, H. (2019). Rram-based synapse devices for neuromorphic systems. *Faraday discussions*, 213:421–451.

[193] Moore, G. E. (2006). Lithography and the future of moore's law. *IEEE Solid-State Circuits Society Newsletter*, 11(3):37–42.

[194] Mostafa, H. (2017). Supervised learning based on temporal coding in spiking neural networks. *IEEE transactions on neural networks and learning systems*, 29(7):3227–3235.

[195] Munde, M., Mehonic, A., Ng, W., Buckwell, M., Montesi, L., Bosman, M., Shluger, A., and Kenyon, A. (2017). Intrinsic resistance switching in amorphous silicon suboxides: the role of columnar microstructure. *Scientific reports*, 7(1):9274.

[196] Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

[197] Neftci, E., Das, S., Pedroni, B., Kreutz-Delgado, K., and Cauwenberghs, G. (2014). Event-driven contrastive divergence for spiking neuromorphic systems. *Frontiers in neuroscience*, 7:272.

[198] Neftci, E. O., Mostafa, H., and Zenke, F. (2019). Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):51–63.

[199] Nessler, B., Pfeiffer, M., Buesing, L., and Maass, W. (2013). Bayesian computation emerges in generic cortical microcircuits through spike-timing-dependent plasticity. *PLoS computational biology*, 9(4):e1003037.

[200] Noble, D. (1962). A modification of the hodgkin—huxley equations applicable to purkinje fibre action and pacemaker potentials. *The Journal of physiology*, 160(2):317.

[201] O'Connor, P., Neil, D., Liu, S.-C., Delbruck, T., and Pfeiffer, M. (2013). Real-time classification and sensor fusion with a spiking deep belief network. *Frontiers in neuroscience*, 7:178.

[202] Orchard, G., Frady, E. P., Rubin, D. B. D., Sanborn, S., Shrestha, S. B., Sommer, F. T., and Davies, M. (2021). Efficient neuromorphic signal processing with loihi 2. In *2021 IEEE Workshop on Signal Processing Systems (SiPS)*, pages 254–259. IEEE.

[203] Pehle, C.-G. and Egholm Pedersen, J. (2021). Norse-a deep learning library for spiking neural networks. *Zenodo*.

[204] Peng, C.-N., Wang, C.-W., Chan, T.-C., Chang, W.-Y., Wang, Y.-C., Tsai, H.-W., Wu, W.-W., Chen, L.-J., and Chueh, Y.-L. (2012). Resistive switching of au/zno/au resistive memory: an in situ observation of conductive bridge formation. *Nanoscale research letters*, 7:1–6.

[205] Perez-Carrasco, J.-A., Serrano, C., Acha, B., Serrano-Gotarredona, T., and Linares-Barranco, B. (2010). Spike-based convolutional network for real-time processing. In *2010 20th International Conference on Pattern Recognition*, pages 3085–3088. IEEE.

[206] Pfeiffer, M. and Pfeil, T. (2018). Deep learning with spiking neurons: opportunities and challenges. *Frontiers in neuroscience*, page 774.

[207] Pi, S., Li, C., Jiang, H., Xia, W., Xin, H., Yang, J. J., and Xia, Q. (2019). Memristor crossbar arrays with 6-nm half-pitch and 2-nm critical dimension. *Nature nanotechnology*, 14(1):35–39.

[208] Polsky, A., Mel, B. W., and Schiller, J. (2004). Computational subunits in thin dendrites of pyramidal cells. *Nature neuroscience*, 7(6):621–627.

[209] Poole, B., Sohl-Dickstein, J., and Ganguli, S. (2014). Analyzing noise in autoencoders and deep networks. *arXiv preprint arXiv:1406.1831*.

[210] Prakash, A. and Hwang, H. (2016). Multilevel cell storage and resistance variability in resistive random access memory. *Physical Sciences Reviews*, 1(6):20160010.

[211] Puglisi, F. M. and Pavan, P. (2016). Guidelines for a reliable analysis of random telegraph noise in electronic devices. *IEEE Transactions on Instrumentation and Measurement*, 65(6):1435–1442.

[212] Querlioz, D., Zhao, W., Dollfus, P., Klein, J.-O., Bichler, O., and Gamrat, C. (2012). Bioinspired networks with nanoscale memristive devices that combine the unsupervised and supervised learning approaches. In *Proceedings of the 2012 IEEE/ACM International Symposium on Nanoscale Architectures*, pages 203–210.

[213] Rasch, M. J., Moreda, D., Gokmen, T., Le Gallo, M., Carta, F., Goldberg, C., El Maghraoui, K., Sebastian, A., and Narayanan, V. (2021). A flexible and fast pytorch toolkit for simulating training and inference on analog crossbar arrays. In *2021 IEEE 3rd international conference on artificial intelligence circuits and systems (AICAS)*, pages 1–4. IEEE.

[214] Saha, S. and Krupanidhi, S. (2001). Transient analysis in al-doped barium strontium titanate thin films grown by pulsed laser deposition. *Journal of Applied Physics*, 90(3):1250–1254.

[215] Saïghi, S., Mayr, C. G., Serrano-Gotarredona, T., Schmidt, H., Lecerf, G., Tomas, J., Grollier, J., Boyn, S., Vincent, A. F., Querlioz, D., et al. (2015). Plasticity in memristive devices for spiking neural networks. *Frontiers in neuroscience*, 9:51.

[216] Sawa, A., Fujii, T., Kawasaki, M., and Tokura, Y. (2004). Hysteretic current–voltage characteristics and resistance switching at a rectifying ti/ pr0. 7ca0. 3mno3 interface. *Applied Physics Letters*, 85(18):4073–4075.

[217] Schlaf, R., Murata, H., and Kafafi, Z. (2001). Work function measurements on indium tin oxide films. *Journal of Electron Spectroscopy and Related Phenomena*, 120(1-3):149–154.

[218] Scott, D. W. (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

[219] Sengupta, A., Al Azim, Z., Fong, X., and Roy, K. (2015). Spin-orbit torque induced spike-timing dependent plasticity. *Applied Physics Letters*, 106(9):093704.

[220] Senthilkumar, M., Mathiyarasu, J., Joseph, J., Phani, K., and Yegnaraman, V. (2008). Electrochemical instability of indium tin oxide (ito) glass in acidic ph range during cathodic polarization. *Materials Chemistry and Physics*, 108(2-3):403–407.

[221] Seong, D.-j., Jo, M., Lee, D., and Hwang, H. (2007). Hpha effect on reversible resistive switching of pt/ nb-doped srtio3 schottky junction for nonvolatile memory application. *Electrochemical and solid-state letters*, 10(6):H168.

[222] Serrano-Gotarredona, T., Masquelier, T., Prodromakis, T., Indiveri, G., and Linares-Barranco, B. (2013). Stdp and stdp variations with memristors for spiking neuromorphic learning systems. *Frontiers in neuroscience*, 7:2.

[223] Shatz, C. J. (1992). The developing brain. *Scientific American*, 267(3):60–67.

[224] Shen, W., Kumar, S., and Kumar, S. (2021). Experimentally calibrated electro-thermal modeling of temperature dynamics in memristors. *Applied Physics Letters*, 118(10).

[225] Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Routledge.

[226] Snow, E., Grove, A., Deal, B., and Sah, C. (1965). Ion transport phenomena in insulating films. *Journal of Applied Physics*, 36(5):1664–1673.

[227] Soni, R., Meuffels, P., Staikov, G., Weng, R., Kügeler, C., Petraru, A., Hambe, M., Waser, R., and Kohlstedt, H. (2011). On the stochastic nature of resistive switching in cu doped ge0. 3se0. 7 based memory devices. *Journal of applied physics*, 110(5).

[228] Sprekeler, H., Michaelis, C., and Wiskott, L. (2007). Slowness: An objective for spike-timing–dependent plasticity? *PLoS Computational Biology*, 3(6):e112.

[229] Squire, L., Berg, D., Bloom, F. E., Du Lac, S., Ghosh, A., and Spitzer, N. C. (2012). *Fundamental neuroscience*. Academic press.

[230] Stein, R. and Hodgkin, A. L. (1967). The frequency of nerve action potentials generated by applied currents. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 167(1006):64–86.

[231] Stimberg, M., Brette, R., and Goodman, D. F. (2019). Brian 2, an intuitive and efficient neural simulator. *elife*, 8:e47314.

[232] Stocks, N. (2001). Information transmission in parallel threshold arrays: Suprathreshold stochastic resonance. *Physical Review E*, 63(4):041114.

[233] Stone, J. V. (2019). *Artificial intelligence engines: a tutorial introduction to the mathematics of deep learning*. Sebtel Press Warszawa, Poland.

[234] Stromatias, E. and Marsland, J. S. (2015). Supervised learning in spiking neural networks with limited precision: Snn/lp. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

[235] Stromatias, E., Neil, D., Pfeiffer, M., Galluppi, F., Furber, S. B., and Liu, S.-C. (2015). Robustness of spiking deep belief networks to noise and reduced bit precision of neuro-inspired hardware platforms. *Frontiers in neuroscience*, 9:222.

[236] Strukov, D. B., Snider, G. S., Stewart, D. R., and Williams, R. S. (2008). The missing memristor found. *nature*, 453(7191):80–83.

[237] Sun, L., Zhang, Y., Hwang, G., Jiang, J., Kim, D., Eshete, Y. A., Zhao, R., and Yang, H. (2018). Synaptic computation enabled by joule heating of single-layered semiconductors for sound localization. *Nano letters*, 18(5):3229–3234.

[238] Sung, C., Lim, S., Kim, H., Kim, T., Moon, K., Song, J., Kim, J.-J., and Hwang, H. (2018). Effect of conductance linearity and multi-level cell characteristics of taox-based synapse device on pattern recognition accuracy of neuromorphic system. *Nanotechnology*, 29(11):115203.

[239] Suri, M., Querlioz, D., Bichler, O., Palma, G., Vianello, E., Vuillaume, D., Gamrat, C., and DeSalvo, B. (2013). Bio-inspired stochastic computing using binary cbram synapses. *IEEE Transactions on Electron Devices*, 60(7):2402–2409.

[240] Sze, S. M., Li, Y., and Ng, K. K. (2021). *Physics of semiconductor devices*. John wiley & sons.

[241] Tal, D. and Schwartz, E. L. (1997). Computing with the leaky integrate-and-fire neuron: logarithmic computation and multiplication. *Neural computation*, 9(2):305–318.

[242] Tang, Z., Chen, Y., Ye, S., Hu, R., Wang, H., He, J., Huang, Q., and Chang, S. (2020). Fully memristive spiking-neuron learning framework and its applications on pattern recognition and edge detection. *Neurocomputing*, 403:80–87.

[243] Tavanaei, A., Ghodrati, M., Kheradpisheh, S. R., Masquelier, T., and Maida, A. (2019). Deep learning in spiking neural networks. *Neural networks*, 111:47–63.

[244] Teeter, C., Iyer, R., Menon, V., Gouwens, N., Feng, D., Berg, J., Szafer, A., Cain, N., Zeng, H., Hawrylycz, M., et al. (2018). Generalized leaky integrate-and-fire models classify multiple neuron types. *Nature communications*, 9(1):709.

[245] Tien, N.-W. and Kerschensteiner, D. (2018). Homeostatic plasticity in neural development. *Neural development*, 13(1):1–7.

[246] Toyoizumi, T., Pfister, J.-P., Aihara, K., and Gerstner, W. (2004). Spike-timing dependent plasticity and mutual information maximization for a spiking neuron model. *Advances in neural information processing systems*, 17.

[247] Tran, R., Li, X.-G., Montoya, J. H., Winston, D., Persson, K. A., and Ong, S. P. (2019). Anisotropic work function of elemental crystals. *Surface Science*, 687:48–55.

[248] Truong, S. N. and Min, K.-S. (2014). New memristor-based crossbar array architecture with 50-% area reduction and 48-% power saving for matrix-vector multiplication of analog neuromorphic computing. *JSTS: Journal of Semiconductor Technology and Science*, 14(3):356–363.

[249] Tuller, H. L. and Bishop, S. R. (2011). Point defects in oxides: tailoring materials through defect engineering. *Annual Review of Materials Research*, 41(1):369–398.

[250] Turlach, B. A. (1993). Bandwidth selection in kernel density estimation: a rewiew. Technical report, Humboldt Universitaet Berlin.

[251] Turrigiano, G. G. (1999). Homeostatic plasticity in neuronal networks: the more things change, the more they stay the same. *Trends in neurosciences*, 22(5):221–227.

[252] Vanheusden, K., Karna, S., Pugh, R., Warren, W., Fleetwood, D., Devine, R., and Edwards, A. (1998a). Thermally activated electron capture by mobile protons in sio 2 thin films. *Applied physics letters*, 72(1):28–30.

[253] Vanheusden, K., Warren, W., Devine, R., Fleetwood, D., Draper, B., and Schwank, J. (1999). A non-volatile mosfet memory device based on mobile protons in sio2 thin films. *Journal of non-crystalline solids*, 254(1-3):1–10.

[254] Vanheusden, K., Warren, W., Fleetwood, D., Schwank, J., Shaneyfelt, M., Draper, B., Winokur, P., Devine, R., Archer, L., Brown, G., et al. (1998b). Chemical kinetics of mobile-proton generation and annihilation in sio 2 thin films. *Applied physics letters*, 73(5):674–676.

[255] Vanka, S., Shin, H., Davidson, B. A., Liu, C., and Zou, K. (2022). Hydrogen atom doping—a versatile method for modulated interface resistive switching. *Advanced Electronic Materials*, 8(10):2200353.

[256] Vieluf, S., El Atrache, R., Cantley, S., Jackson, M., Clark, J., Sheehan, T., Bosl, W. J., Zhang, B., and Loddenkemper, T. (2022). Seizure-related differences in biosignal 24-h modulation patterns. *Scientific reports*, 12(1):1–9.

[257] Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103.

[258] Von Neumann, J. (1993). First draft of a report on the edvac. *IEEE Annals of the History of Computing*, 15(4):27–75.

[259] Wakiya, N., Tajiri, N., Kiguchi, T., Mizutani, N., Cross, J. S., and Shinozaki, K. (2006). Activation energy of oxygen vacancy diffusion of yttria-stabilized-zirconia thin film determined from dc current measurements below 150 c. *Japanese journal of applied physics*, 45(6L):L525.

[260] Wan, W., Kubendran, R., Gao, B., Joshi, S., Raina, P., Wu, H., Cauwenberghs, G., and Wong, H. P. (2020). A voltage-mode sensing scheme with differential-row weight mapping for energy-efficient rram-based in-memory computing. In *2020 IEEE Symposium on VLSI Technology*, pages 1–2. IEEE.

[261] Wang, J., Hu, S., Zhan, X., Yu, Q., Liu, Z., Chen, T. P., Yin, Y., Hosaka, S., and Liu, Y. (2018a). Handwritten-digit recognition by hybrid convolutional neural network based on hfo2 memristive spiking-neuron. *Scientific reports*, 8(1):12546.

[262] Wang, J. and Trolier-McKinstry, S. (2006). Oxygen vacancy motion in er-doped barium strontium titanate thin films. *Applied physics letters*, 89(17).

[263] Wang, X., Lin, X., and Dang, X. (2020a). Supervised learning in spiking neural networks: A review of algorithms and evaluations. *Neural Networks*, 125:258–280.

[264] Wang, X.-Y., Dong, C.-T., Zhou, P.-F., Nandi, S. K., Nath, S. K., Elliman, R. G., Iu, H. H.-C., Kang, S.-M., and Eshraghian, J. K. (2022). Low-variance memristor-based multi-level ternary combinational logic. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 69(6):2423–2434.

[265] Wang, X.-Y., Zhou, P.-F., Eshraghian, J. K., Lin, C.-Y., Iu, H. H.-C., Chang, T.-C., and Kang, S.-M. (2020b). High-density memristor-cmos ternary logic family. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 68(1):264–274.

[266] Wang, Y., Chen, W. P., Cheng, K. C., Chan, H. L. W., and Choy, C. L. (2003). Optical degradation of indium tin oxide thin films induced by hydrogen-related room temperature reduction. *Japanese journal of applied physics*, 42(5B):L546.

[267] Wang, Y., Wu, S., Tian, L., and Shi, L. (2020c). Ssm: a high-performance scheme for in situ training of imprecise memristor neural networks. *Neurocomputing*, 407:270–280.

[268] Wang, Z., Joshi, S., Savel'Ev, S., Song, W., Midya, R., Li, Y., Rao, M., Yan, P., Asapu, S., Zhuo, Y., et al. (2018b). Fully memristive neural networks for pattern classification with unsupervised learning. *Nature Electronics*, 1(2):137–145.

[269] Wang, Z., Joshi, S., Savel'ev, S. E., Jiang, H., Midya, R., Lin, P., Hu, M., Ge, N., Strachan, J. P., Li, Z., et al. (2017). Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing. *Nature materials*, 16(1):101–108.

[270] Warren, W., Fleetwood, D., Schwank, J., Shaneyfelt, M., Draper, B., Winokur, P., Knoll, M., Vanheusden, K., Devine, R., Archer, L., et al. (1997). Protonic nonvolatile field effect transistor memories in si/sio/sub 2//si structures. *IEEE Transactions on Nuclear Science*, 44(6):1789–1798.

[271] Waser, R., Baiatu, T., and Härdtl, K.-H. (1990a). Dc electrical degradation of perovskite-type titanates: I, ceramics. *Journal of the American Ceramic Society*, 73(6):1645–1653.

[272] Waser, R., Baiatu, T., and Härdtl, K.-H. (1990b). Dc electrical degradation of perovskite-type titanates: Ii, single crystals. *Journal of the American Ceramic Society*, 73(6):1654–1662.

[273] Wei, Z., Kanzawa, Y., Arita, K., Katoh, Y., Kawai, K., Muraoka, S., Mitani, S., Fujii, S., Katayama, K., Iijima, M., et al. (2008). Highly reliable taox reram and direct evidence of redox reaction mechanism. In *2008 IEEE International Electron Devices Meeting*, pages 1–4. IEEE.

[274] Wen, S. and Zeng, Z. (2012). Dynamics analysis of a class of memristor-based recurrent networks with time-varying delays in the presence of strong external stimuli. *Neural processing letters*, 35(1):47–59.

[275] Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560.

[276] Wijesinghe, P., Ankit, A., Sengupta, A., and Roy, K. (2018). An all-memristor deep spiking neural computing system: A step toward realizing the low-power stochastic brain. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(5):345–358.

[277] Winokur, P., Boesch, H., McGarrity, J., and McLean, F. (1977). Field-and time-dependent radiation effects at the sio2/si interface of hardened mos capacitors. *IEEE Transactions on Nuclear Science*, 24(6):2113–2118.

[278] Woo, J., Moon, K., Song, J., Lee, S., Kwak, M., Park, J., and Hwang, H. (2016). Improved synaptic behavior under identical pulses using alo x/hfo 2 bilayer rram array for neuromorphic systems. *IEEE Electron Device Letters*, 37(8):994–997.

[279] Wronski, C. and Carlson, D. (1977). Surface states and barrier heights of metal-amorphous silicon schottky barriers. *Solid State Communications*, 23(7):421–424.

[280] Wu, W., Wu, H., Gao, B., Yao, P., Zhang, X., Peng, X., Yu, S., and Qian, H. (2018). A methodology to improve linearity of analog rram for neuromorphic computing. In *2018 IEEE symposium on VLSI technology*, pages 103–104. IEEE.

[281] Wu, X., Mei, S., Bosman, M., Raghavan, N., Zhang, X., Cha, D., Li, K., and Pey, K. L. (2015). Evolution of filament formation in ni/hfo2/siox/si-based rram devices. *Advanced Electronic Materials*, 1(11):1500130.

[282] Wu, X. and Saxena, V. (2017). Enabling bio-plausible multi-level stdp using cmos neurons with dendrites and bistable rrams. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 3522–3526. IEEE.

[283] Wu, X. and Saxena, V. (2018). Dendritic-inspired processing enables bio-plausible stdp in compound binary synapses. *IEEE Transactions on Nanotechnology*, 18:149–159.

[284] Wunderlich, T., Kungl, A. F., Müller, E., Hartel, A., Stradmann, Y., Aamir, S. A., Grübl, A., Heimbrecht, A., Schreiber, K., Stöckel, D., et al. (2019). Demonstrating advantages of neuromorphic computation: a pilot study. *Frontiers in neuroscience*, 13:260.

[285] Xia, L., Gu, P., Li, B., Tang, T., Yin, X., Huangfu, W., Yu, S., Cao, Y., Wang, Y., and Yang, H. (2016). Technological exploration of rram crossbar array for matrix-vector multiplication. *Journal of Computer Science and Technology*, 31(1):3–19.

[286] Xia, L., Huangfu, W., Tang, T., Yin, X., Chakrabarty, K., Xie, Y., Wang, Y., and Yang, H. (2017). Stuck-at fault tolerance in rram computing systems. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 8(1):102–115.

[287] Xia, Q. and Yang, J. J. (2019). Memristive crossbar arrays for brain-inspired computing. *Nature materials*, 18(4):309–323.

[288] Yakopcic, C., Taha, T. M., Subramanyam, G., and Pino, R. E. (2013). Generalized memristive device spice model and its application in circuit design. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 32(8):1201–1214.

[289] Yang, J. J., Strukov, D. B., and Stewart, D. R. (2013). Memristive devices for computing. *Nature nanotechnology*, 8(1):13–24.

[290] Ye, N., Cao, L., Yang, L., Zhang, Z., Fang, Z., Gu, Q., and Yang, G.-Z. (2023). Improving the robustness of analog deep neural networks through a bayes-optimized noise injection approach. *Communications Engineering*, 2(1):25.

[291] Yi, W., Savel'Ev, S. E., Medeiros-Ribeiro, G., Miao, F., Zhang, M.-X., Yang, J. J., Bratkovsky, A. M., and Williams, R. S. (2016). Quantized conductance coincides with state instability and excess noise in tantalum oxide memristors. *Nature communications*, 7(1):11142.

[292] Yon, E., Ko, W., and Kuper, A. (1966). Sodium distribution in thermal oxide on silicon by radiochemical and mos analysis. *IEEE Transactions on Electron Devices*, 1(2):276–280.

[293] Yu, S., Guan, X., and Wong, H.-S. P. (2012). On the switching parameter variation of metal oxide rram—part ii: Model corroboration and device design strategy. *IEEE Transactions on Electron Devices*, 59(4):1183–1188.

[294] Yu, S., Wu, Y., Jeyasingh, R., Kuzum, D., and Wong, H.-S. P. (2011). An electronic synapse device based on metal oxide resistive switching memory for neuromorphic computation. *IEEE Transactions on Electron Devices*, 58(8):2729–2737.

[295] Yun, J.-B., Kim, S., Seo, S., Lee, M.-J., Kim, D.-C., Ahn, S.-E., Park, Y., Kim, J., and Shin, H. (2007). Random and localized resistive switching observation in pt/nio/pt. *physica status solidi (RRL)–Rapid Research Letters*, 1(6):280–282.

[296] Zafar, S., Jagannathan, H., Edge, L. F., and Gupta, D. (2011). Measurement of oxygen diffusion in nanometer scale hfo2 gate dielectric films. *Applied Physics Letters*, 98(15).

[297] Zafar, S., Jones, R. E., Jiang, B., White, B., Chu, P., Taylor, D., and Gillespie, S. (1998). Oxygen vacancy mobility determined from current measurements in thin ba 0.5 sr 0.5 tio 3 films. *Applied physics letters*, 73(2):175–177.

[298] Zamarreño-Ramos, C., Camuñas-Mesa, L. A., Pérez-Carrasco, J. A., Masquelier, T., Serrano-Gotarredona, T., and Linares-Barranco, B. (2011). On spike-timing-dependent-plasticity, memristive devices, and building a self-learning visual cortex. *Frontiers in neuroscience*, 5:26.

[299] Zha, Y. and Li, J. (2017). Imec: A fully morphable in-memory computing fabric enabled by resistive crossbar. *IEEE Computer Architecture Letters*, 16(2):123–126.

[300] Zhang, X., Liu, S., Zhao, X., Wu, F., Wu, Q., Wang, W., Cao, R., Fang, Y., Lv, H., Long, S., et al. (2017). Emulating short-term and long-term plasticity of bio-synapse based on cu/a-si/pt memristor. *IEEE Electron Device Letters*, 38(9):1208–1211.

[301] Zhang, Y., Wang, Z., Zhu, J., Yang, Y., Rao, M., Song, W., Zhuo, Y., Zhang, X., Cui, M., Shen, L., et al. (2020). Brain-inspired computing with memristors: Challenges in devices, circuits, and systems. *Applied Physics Reviews*, 7(1).

[302] Zheng, N. and Mazumder, P. (2018). Learning in memristor crossbar-based spiking neural networks through modulation of weight-dependent spike-timing-dependent plasticity. *IEEE Transactions on Nanotechnology*, 17(3):520–532.

[303] Zhong, N., Shima, H., and Akinaga, H. (2010). Transient current study on pt/tio2-x/pt capacitor. *Japanese Journal of Applied Physics*, 49(4S):04DJ15.

[304] Zhou, P. (2022). *Memristive Spiking Neural Network for Neuromorphic Computing*. University of California, Santa Cruz.

[305] Zhou, P. and Abbaszadeh, S. (2020). Towards real-time machine learning for anomaly detection. In *2020 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, pages 1–3. IEEE.

[306] Zhou, P., Choi, D.-U., Eshraghian, J. K., and Kang, S.-M. (2022). A fully memristive spiking neural network with unsupervised learning. In *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 634–638. IEEE.

[307] Zhu, R., Loeffler, A., Hochstetter, J., Diaz-Alvarez, A., Nakayama, T., Stieg, A., Gimzewski, J., Lizier, J., and Kuncic, Z. (2021). Mnist classification using neuromorphic nanowire networks. In *International Conference on Neuromorphic Systems 2021*, pages 1–4.

[308] Zhu, X., Wang, Q., and Lu, W. D. (2020a). Memristor networks for real-time neural activity analysis. *Nature communications*, 11(1):2439.

[309] Zhu, Y., Zhang, G. L., Wang, T., Li, B., Shi, Y., Ho, T.-Y., and Schlichtmann, U. (2020b). Statistical training for neuromorphic computing using memristor-based crossbars considering process variations and noise. In *2020 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 1590–1593. IEEE.

[310] Zhuge, F., Fu, B., and Cao, H. (2013). Advances in resistive switching memories based on graphene oxide. *New Prog. Graphene Res*, 7:185–206.

[311] Zucker, R. S. and Regehr, W. G. (2002). Short-term synaptic plasticity. *Annual review of physiology*, 64(1):355–405.