Spiking neural networks based on resistive memory technologies for neural data analysis

Thilo Werner

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTE UNIVERSITE GRENOBLE ALPES

Spécialité : **Nano-Electronique et Nano-Technologies**

Arrêté ministériel : 25 mai 2016

Présentée par

Thilo WERNER

Thèse dirigée par **Dr. Blaise YVERT**
codirigée par **Dr. Barbara DE SALVO**

préparée au sein du **CEA-Leti** et **Braintech lab Inserm & UGA U1205**
dans l'**Ecole Doctorale Ingénierie pour la Santé, la Cognition et l'Environnement**

Réseaux de neurones impulsionnels basés sur les mémoires résistives pour l'analyse de données neuronales

Thèse soutenue publiquement le **« 10 juillet 2017 »**,
devant le jury composé de :

M Damien QUERLIOZ
DR, CNRS, Rapporteur
M Abdelkader SOUIFI
Prof., INSA, Rapporteur
M Thierry BARON
DR, CNRS, Président
M Jean-Michel PORTAL
Prof., CNRS, Examinateur
M Blaise YVERT
DR, Inserm, Directeur de thèse
Mme Barbara DE SALVO
DR, CEA-Leti, Co-directeur de thèse
Mme Elisa VIANELLO
Dr. Ing., CEA-Leti, Encadrant de thèse

Abstract

Title:
Spiking Neural Networks based on Resistive Memory Technologies for Neural data analysis

The central nervous system of humankind is an astonishing information processing system in terms of its capabilities, versatility, adaptability and low energy consumption. Its complex structure consists of billions of neurons interconnected by trillions of synapses forming specialized clusters. Recently, mimicking those paradigms has been attracting a strongly growing interest, triggered by the need for advanced computing approaches to tackle challenges related to the generation of massive amounts of complex data in the Internet of Things (IoT) era. This has led to a new research field, known as cognitive computing or neuromorphic engineering, which relies on so-called non-von-Neumann architectures (brain-inspired) in contrary to von-Neumann architectures (conventional computers). In this thesis, we explore the use of resistive memory technologies such as oxide vacancy based random access memory (OxRAM) and conductive bridge RAM (CBRAM) for the design of artificial synapses that are a basic building block for neuromorphic networks. Moreover, we develop an artificial spiking neural network (SNN) based on OxRAM synapses dedicated to the analysis of spiking data recorded from the human brain with the goal of using the output of the SNN in a brain-computer interface (BCI) for the treatment of neurological disorders. The impact of reliability issues characteristic to OxRAM technology on the system performance is studied in detail and potential ways to mitigate penalties related to single device uncertainties are demonstrated. Besides the already well-known spike-timing-dependent plasticity (STDP) implementation with OxRAM and CBRAM which constitutes a form of long term plasticity (LTP), OxRAM devices were also used to mimic short term plasticity (STP). The fundamentally different functionalities of LTP and STP are put in evidence.

Résumé en francais

Titre:
Réseaux de neurones impulsionnels basés sur les mémoires résistives pour l'analyse de données neuronales

Le système nerveux central humain est un système de traitement de l'information stupéfiant en termes de capacités, de polyvalence, d'adaptabilité et de faible consommation d'énergie. Sa structure complexe se compose de milliards de neurones, interconnectés par plusieurs trillions de synapses, formant des grappes spécialisées. Récemment, l'imitation de ces paradigmes a suscité un intérêt croissant en raison de la nécessité d'approches informatiques avancées pour s'attaquer aux défis liés à la génération de quantités massives de données complexes dans l'ère de l'Internet des Objets (IoT). Ceci a mené à un nouveau domaine de recherche, connu sous le nom d'informatique cognitive ou d'ingénierie neuromorphique, qui repose sur les architectures dites non-von-Neumann (inspirées du cerveau) en opposition aux architectures von-Neumann (ordinateurs classiques). Dans cette thèse, nous examinons l'utilisation des technologies de mémoire résistive telles que les mémoires à accès aléatoires à base de lacunes d'oxygène (OxRAM) et les mémoires à pont conducteur (CBRAM) pour la conception de synapses artificielles, composants de base indispensables des réseaux neuromorphiques. De plus, nous développons un réseau de neurones impulsionnels artificiel (SNN), utilisant des synapses OxRAM, pour l'analyse de données impulsionnelles provenant du cerveau humain en vue du traitement de troubles neurologiques, en connectant la sortie du SNN à une interface cerveau-ordinateur (BCI). L'impact des problèmes de fiabilité, caractéristiques des OxRAMs, sur les performances du système est étudié en détail et les moyens possibles pour atténuer les pénalités liées aux incertitudes des dispositifs seuls sont démontrés. En plus de l'implémentation avec des OxRAMs et CBRAMs de la bien connue plasticité fonction du temps d'occurrence des impulsions (STDP), qui constitue une forme de plasticité à long terme (LTP), les dispositifs OxRAM ont également été utilisés pour imiter la plasticité à court terme (STP). Les fonctionnalités fondamentalement différentes de la LTP et STP sont mises en évidence.

ACKNOWLEDGEMENTS

Certainly, I would not have been able to achieve the results presented in this PhD thesis without the help of numerous people. Therefore, it is my desire to express my deep gratitude to all those who have supported me during the last three years, both professionally as well as personally.

First, I would like to thank my PhD director Blaise Yvert for directing me through this 'exotic' project bridging neuroscience, electrical engineering and machine learning by consistently questioning the status quo, proposing several new ideas and also for his optimistic advices when I struggled with from time to time. Moreover, I am grateful to Barbara de Salvo for the co-direction of this PhD project. Her constant very professional and strategic support as well as fruitful technical discussions helped a lot to steer the work in a successful direction. My special thanks go to Elisa Vianello, my daily advisor, for always being there when I needed assistance, for constantly pushing me to give my best and for her patience with me and my perfectionism. Our countless and sometimes late discussions on neuromorphic subjects brought up a 1000 ideas (more or less) and I will never forget her valuable reminder that 'Research is frustrating!' when I was unsatisfied with the progress of my work. I also want to thank Olivier Bichler from CEA-LIST who assisted me a lot in debugging the code of their special purpose neural network simulator software that I was using (and messing up occasionally).

I am glad I have been able to meet, work and share break times with the people in our lab such as Brigitte, Cathy, Christelle, Eric, Etienne, Gabriel, Gabriele, Khalil, Laurent, Luca, Sabine, Remi, Sophie and Veronique. Of course there are also several people from other labs who helped me with the electrical characterization setup or supported me with expert advice in domains where I did not have enough expertise myself. Among them, I would like to thank especially: Alain Lopez, Carlo Cagli, Denis Blachier and Niccolo Castellani from LCTE as well as Elisabeth Delevoye from DSYS.

Then, there is a great number of interns, PhD students and post-docs that I am thankful to have met and who helped a lot to keep up the social life and reduce the stress level with coffee breaks and nice activities outside of work. In the naive hope not to forget anyone, I want to give my gratitude to the following people: Adam, Amine, Angelica, Annalisa, Anthonin, Blend, Boubacar, Cecile, Daesok, Daniele, Denys, Florent, Giuseppe, Jeremy, Julia, Julien, Loic, Luc, Maria, Marie, Marinela, Marios, Martin, Mouhamad, Mourad, Natalija, Paul, Rana, Selina, Thanasis, Thomas and Vincent. Just as important as people at work to share some creative breaks with are of course 'non-involved' people from the 'outside-of-work' world and I definitely owe finishing this thesis partly to the following people: Alan, Alexia, Daniel, Felipe, Felix, Joana, Johana, Lucy, Ning, Oscar, Pierre, Sebastien, Sholpan and Stephane.

Finally, I want to thank my family for always supporting me in following my ideas and achieving my goals as well as for motivating me when I was doubtful. Continuing my thesis is among the previously mentioned people strongly due to my girlfriend Julia who accompanied me

through this journey of the last three years which was perhaps not quite easy at times but she ever managed to bring me back into equilibrium when it was necessary. Last but definitely not least, I need to mention Marios. He persistently supported me personally when I was close to giving up and he kept listening to my complaints over and over. Therefore, I will be forever in those people's debt and I want to thank them greatly.

AUTHOR'S DECLARATION

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ..... DATE: 08/10/2017 .....

I walk slowly but I never walk backward.
Abraham Lincoln

TABLE OF CONTENTS

List of Tables

Table **Page**

List of Figures

Figure **Page**

The topic of this thesis is embedded within the three main research fields, being (i) biomedical engineering, (ii) memory technology and (iii) neuromorphic engineering. A short description of each of these fields is given before explaining the objective of this dissertation in more detail.

Biomedical engineering: Recording and stimulating the brain

Numerous people suffer from paralysis, e.g. after spinal cord injury (SCI), or neurodegenerative diseases such as Parkinson, Alzheimer, and Huntington's disease due to traumas and population aging (WHO, 2014). Official numbers are not available, however, estimations state that approximately 2.5% and 1.9% of US citizens are affected by neurodegenerative diseases and paralysis, respectively. This accounts to a total number of approx. 13 million people ($7.4M$ + $5.6M$) only in the USA. Consequences of this are a major loss of life quality for individual patients and enormous cost for healthcare, since patients are mostly dependent on around the clock assistance. Innovative therapies are needed in order to cure patients or re-establish their independence. Modern healthcare approaches do not only rely on molecular and pharmacological products but embrace more and more technological approaches for rehabilitation including brain-computer interfaces (BCIs) and neural prostheses (NPs). BCIs are paradigms designed to extract and decode neural signals to control an external device to restore motor commands or communication, while NPs use electrical stimulation of the Central Nervous System (CNS) to restore lost or missing functions as for instance audition with cochlear implants or vision with retinal implants. Therefore, probing motor cortical activity has recently received increased attention for the exploitation of human brain signals within BCI systems. It was shown that BCI's offer promising rehabilitation possibilities to improve life quality of patients suffering from neurodegenerative diseases or paralysis [3],[4], i.e. numerous signals have to be stored and decoded resulting in vast data rates and computational efforts. This requires the ability to precisely collect and analyse brain signals, e.g. triggered when a person intends to perform movements. The effectiveness and accuracy of BCI systems scale with the number of simultaneously recorded populations of neurons [26],[27]. To this end, advanced microelectrode array (MEA) technologies [28] are unique and increasingly powerful tools exploring the central nervous system in detail. Nowadays, they consist of hundreds or thousands of microelectrodes that allow recording the activity of

large neural ensembles and especially spikes (action potentials) generated by the surrounding single neuron cells. These technologies generate massive data due to sampling rates of typically $20-40\ kHz$ that have to be processed for further use and/or wireless transmission [29]. Spike sorting is a key technique to drastically reduce the amount of data by extracting relevant information as how many cells are active and the different instants at which they fire [30]. This allows to understand the neural code (e.g. of language or motor commands) and thus to develop revolutionary rehabilitation treatments.

Information storage: From early memories to advanced technologies for the Internet of Things/Big data era

The advent of portable electronics such as smartphones, tablets etc. has led to an ever increasing need of high capacity memory technologies operating on low energy budgets. The storage function was typically satisfied by Flash technologies for many years, more exactly by NOR and NAND for embedded and stand-alone products. These technologies find various applications in fields spanning everyday life products, such as micro-controllers, phones, cameras and automotive applications. Nowadays, novel technologies are emerging from research, appearing to be competitive for the replacement of Flash memories which faces several problems in the continuation of scaling according to Moores law. Among the new technologies are: phase change memory (PCM, PCRAM or PRAM), resistive memory (RRAM or ReRAM), ferroelectric memory (FeRAM) and magnetic memory (MRAM). All of those technologies are non-volatile (like Flash), e.g. do not require a power supply for data storage. They offer various advantages over Flash technology such as higher speed and endurance and are achieving storage capacities even higher with respect to Flash as shown in figure 0.1. The rapid emerging of non-volatile memory technologies, most importantly PRAM and RRAM, can also be attributed to their excellent suitability for 3-dimensional integration, hence, increasing massively the memory density on chip. Note that, due to their different characteristics and operation mechanisms, each emerging technology is believed to fulfil a different task in future memory applications.

Von-Neumann computers vs. Brain-Inspired architectures

The Von-Neumann computer architecture [31], based on the separation of processing (CPU, GPU) and data storage (memory), has enabled the rapid progress of human development throughout the late 20th and beginning of 21st century. In parallel, the Internet of Things (IoT) has evolved with billions of connected devices requiring fast and low energy data exchange. As the vast amount of data in the IoT era are rather complex and unstructured, the development of efficient approaches to extract information from them is becoming more and more challenging. Since the von-Neumann architecture is based on a deterministic approach which requires dedicated

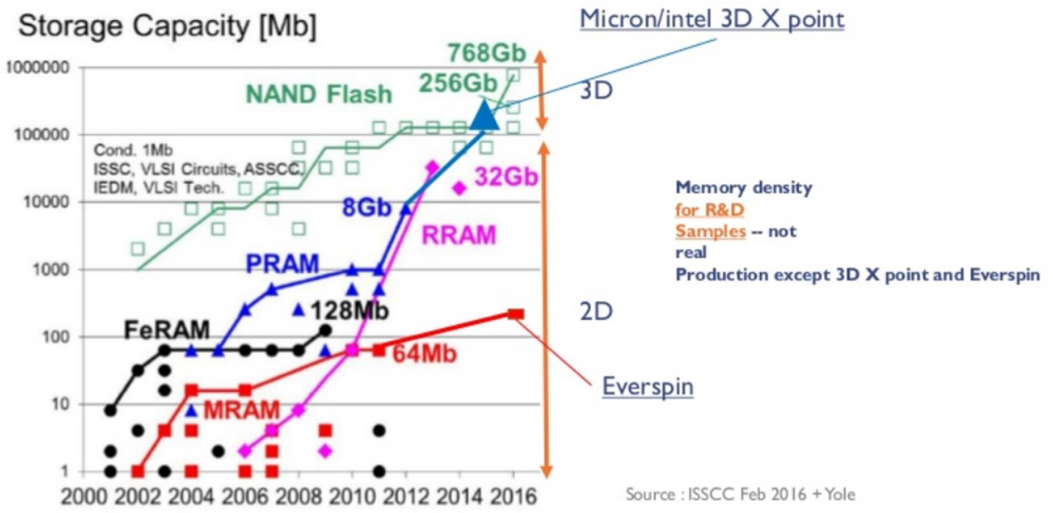


Figure 0.1: Evolution of storage capacity of several memory technologies: NAND Flash, phase change memory (PRAM), resistive memory (RRAM), ferroelectric memory (FeRAM) and magnetic memory (MRAM).

and sophisticated software, parallel processing of unstructured data becomes critical. An alternative could be brain-inspired (also known as non-von-Neumann or neuromorphic) computing approaches. This fundamentally different class of architectures is projected to exploit several new device technologies and change the computing paradigm with respect to available von-Neumann technologies, as shown in figure 0.2. Several studies among the last decade have demonstrated the potential of non-von-Neumann computing paradigms, in particular for (parallel) processing of vast amounts of complex data [32]. Indeed, the so-called brain-inspired networks are designed specifically for certain applications, often related to pattern recognition problems. Among the different approaches mimicking essential functions of biological neural systems are the so-called artificial neural networks (ANN). Currently, ANN are simulated using conventional computers which poses a number of problems such as high energy consumption and slow computation speed when highly parallel computation is needed. These bottlenecks could be overcome by customized physical implementations of ANN, which promise to achieve higher energy efficiency thus fitting the requirements of the Internet of Things (IoT) era. In this context, synapse implementations with novel memory technologies will play a key role as the synapses typically outnumber the neurons in the ANN by orders of magnitude. Properties as high integration density, CMOS process compatibility, low power consumption and long lifetime (high cycling numbers) make resistive RAM (RRAM) one of the main candidates for hardware synapse design.

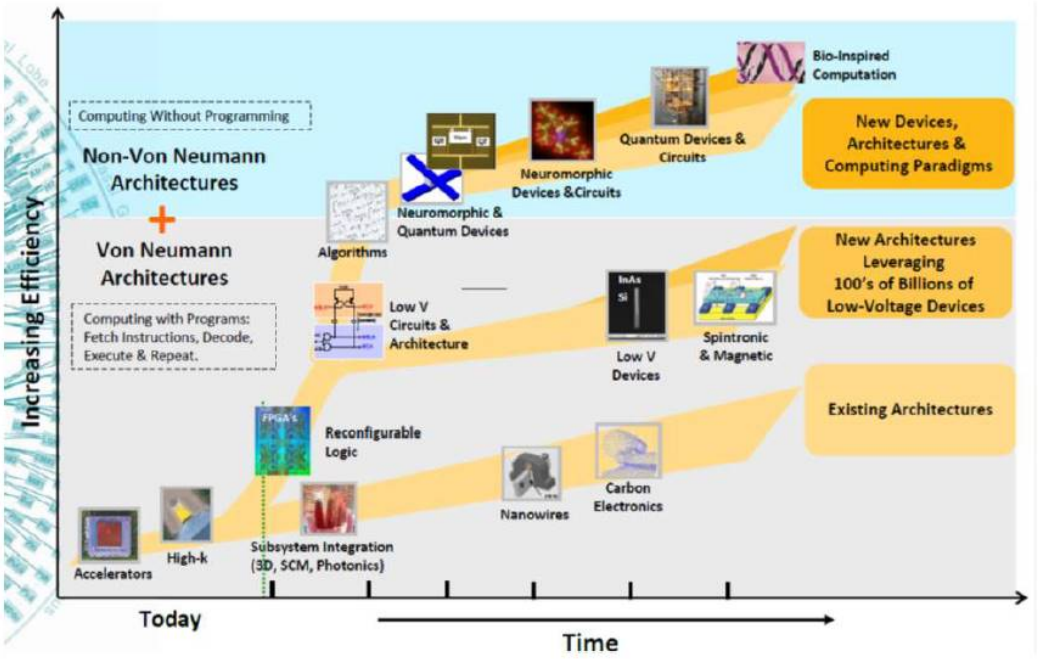


Figure 0.2: Prospected computing and technology roadmap based on von-Neumann and non-von-Neumann architectures. Source: IBM.

CHAPTER 1

BASICS FOR NEUROMORPHIC DEVICES

This chapter aims to provide the basics for the design of neuromorphic systems and moreover to understand the results in chapters 3 to 6. It is divided according to the pillars of this dissertation. First, the essentials of neuroscience and rehabilitation approaches for neurological disorders are described in section 1.1. Second, a survey of state-of-the-art non-volatile memory technologies as well as emerging memory technologies is given with an emphasis on resistive memories in section 1.2. Third, the role and potential applications of RRAM technologies in neuromorphic systems, in particular synapses, is introduced in section 1.3. Finally, the most sophisticated neuromorphic design concepts demonstrated previously are reviewed in section 1.4.

1.1 Neuroscience

This section introduces the physiological fundamentals of neuroscience in order to understand basic functional mechanisms governing nervous systems. Therefore, the role of neurons and synapses are discussed in detail. Furthermore, a comprehensive overview of state-of-the-art technologies used for sensing and/or stimulating neural activity is given. Finally, this section focusses on the challenge of recording/detecting single neuron activities with the ultimate goal to decode and understand the brain's activity for potential application of such techniques into brain-computer interfaces.

1.1.1 The central nervous system

The central nervous system (CNS) and the peripheral nervous system (PNS) constitute the two pillars for sensing, acting and processing information of a human individual [1]. The PNS spans the whole body and its purpose is to collect and distribute sensory and motor signals, respectively,

which are send to or received from the CNS. Figure 1.1 shows schematically the location of CNS and PNS in a human. The CNS consists of the brain and spinal cord and forms the basis for all higher level information processing. It is based on a large number of neurons communicating with one another whereas it is known that the brain is organized into specialized functional areas such as the sensory-motor cortex for motion control [33]. Other areas are, for instance, responsible for vision or language related tasks. Neurons in such areas may either be hard-wired or linked temporally by concerted oscillation [34] [35]. Common for all areas in the human brain is the basic structure, namely a large number of neurons being interconnected by synapses. The latter are used by neurons in order to communicate with each other. The entire brain is believed to consist of approximately 10^{11} neurons connected by some 10^{14} synapses. A detailed description of the structural and biochemical basics of neurons and synapses and there crucial role in neural network dynamics is given in the following sections.

1.1.2 Neurons

Neurons, also called nerve cells, are the principal building block of the human CNS. A neuron is a complex cell that features generally three cell areas dedicated to certain functions, namely (i) the soma where the information is processed, (ii) the dendrites as input and (iii) the axon as output terminals as shown schematically in figure 1.2. The soma is also called cell body and features the main signal processing of the neuron by receiving inputs from the dendrites and sending output signals through its axon. The dendritic arbour consists of relatively short branches of thousands of fine dendrites which establish numerous connections with axons of other neurons. These electro-chemical connections between individual axons and dendrites are known as the so-called synapses which are explained in detail in section 1.1.3. It is estimated that each neuron connects to about 10^4 other neurons constituting a massive connectivity. The computation and memory of the brain is based on the massively parallel communication of neurons with each other that is effected by the propagation of so-called action potentials (AP, also known as spikes) between neurons. An AP is a sharp electrical impulse with a rather uniform voltage amplitude of around 100 mV and exhibits a typical duration of about 1 ms. The AP is travelling from the soma of a pre-synaptic neuron along its axon and it is transmitted via a synapse to the dendrite of a post-synaptic neuron and finally arrives to this one's soma.

The electric potential of the inner neuron soma with respect to the area outside the cell body is governed by concentration gradients of various ion species such as potassium (K^+), sodium (Na^+), chlorine (Cl^-) or calcium (Ca^{2+}). These concentration gradients are controlled by the properties of so-called active transporters (made of proteins) which pump ions through the cell membrane to build up high concentration gradients and ion channels which allow ions to travel through the cell membrane of a neuron. Typically, a high ion concentration of K^+ dominates inside the cell whereas Na^+ exhibits a high concentration outside. Due to a high K^+ and low Na^+ permeability of the ion channels in the resting state of a neuron, the neurons have an intracellular

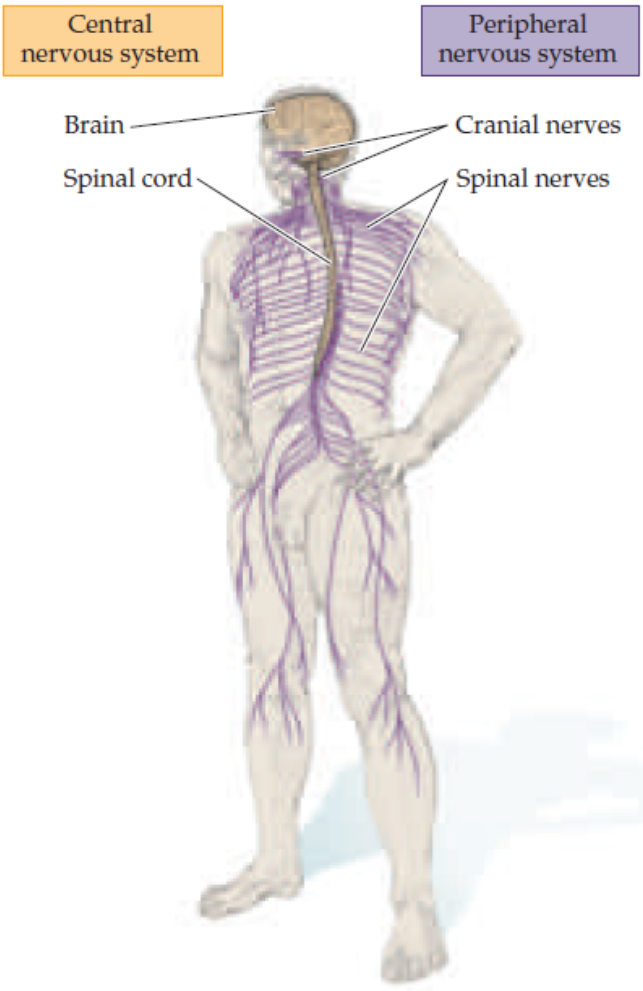


Figure 1.1: Schematic illustration of location of central nervous system (CNS) and peripheral nervous system (PNS) in human body. Note that the CNS consists of both brain and spinal cord. Figure taken from [1].

resting potential of around $-65\ mV$. In the event of an action potential, the ion specific channel permeabilities are inversed resulting in a positive polarization. By the enormous number of synaptic connections with other neurons, the soma receives numerous input information in the form of synaptic currents. These currents lead to a change of the intracellular potential.

Neurons exhibit an internal threshold at about $-40\ mV$, below which is called the sub-threshold region. If the potential rises to this threshold, the neuron emits an AP which is generated by the rapid depolarization up to $+40\ mV$, followed by a quick recovery towards the resting potential of $-65\ mV$. 1.2. This behaviour is caused by ion channels which open shortly and hence allow ions to flow in or out massively in order to create a concentration equilibrium. After this rapid process, the active ion transporters take control again and establish the resting potential inside the soma. During the transmission of the AP from the axon of a pre-synaptic

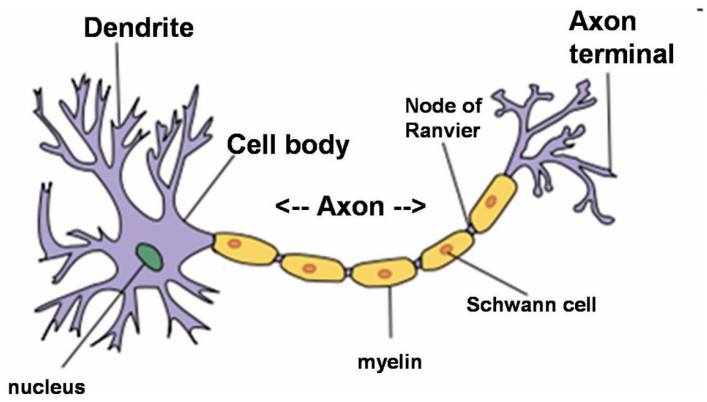


Figure 1.2: Structural illustration of a cortical neuron consisting of three main parts: the dendrites, soma and axon.

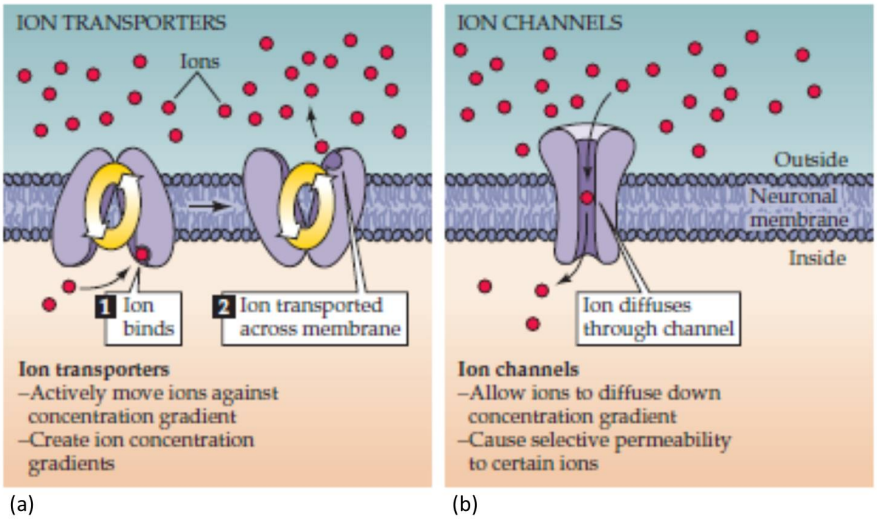


Figure 1.3: Illustration of a neuronal cell membrane. (a) Active ion transporters move certain species of ions in one preferable direction resulting in concentration gradients. (b) Ion channels permeable for certain species allow to equalize concentration gradients. Source: [1]

neuron to the dendrite of a post-synaptic neuron, the AP is converted into a current whose amplitude is characteristic for the individual synaptic strength.

1.1.3 Synapses

A synapse or synaptic cleft is the interface between two neurons and can be classified into chemical and electrical synapses. Chemical synapses are the majority and hold their name because they convert the electrical signal into a current of a specific chemical species. In a chemical synapse, the AP coming from a pre-synaptic neuron activates voltage-gated ion channels which triggers the release of neurotransmitters that bind to their corresponding receptors. Those receptors

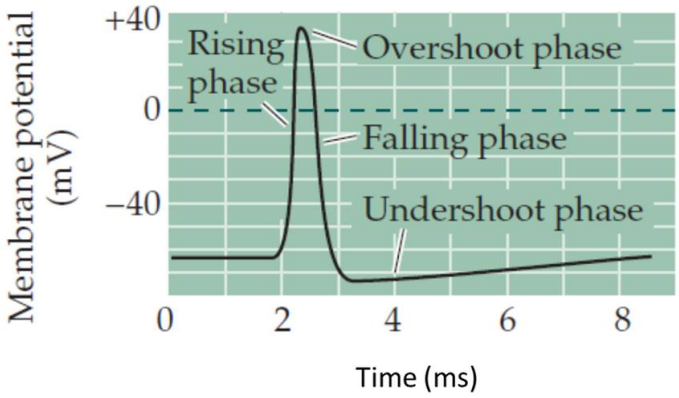


Figure 1.4: Schematic internal voltage of a neuron during the emission of an action potential (spike). Normally, a neuron lies at an internal resting potential of around $-65\ mV$ which can be changed by synaptic inputs. If a threshold at around $-40\ mV$ is reached, a rapid depolarization reaches to approximately $+40\ mV$ followed by the rapid re-polarization to the resting potential. Source: [1]

are located in the membrane of the post-synaptic neuron. When a neurotransmitter arrives to the post-synaptic neuron, it initiates an electrical response or it activates a second messenger. These two scenarios will either excite or inhibit the post-synaptic neuron. On the other hand, electrical synapses are rather rare and are constructed by well aligned gap junctions which are paired channels in the neuron membranes. They are typically of very short distance and allow for a direct ionic current flow between individual cells enabling a very fast information transfer minimizing delays. Due to the domination of chemical over electrical synapses, we will hereafter focus on chemical synapses. When an AP is transmitted over a synapse, a number of glutamate neurotransmitters from pre-synaptic vesicles and Na^+ via the ion channels is released into the synaptic cleft. The Na^+ causes a current flow which can either be an Excitatory Post-Synaptic Current (EPSC) or an Inhibitory Post-Synaptic Current (IPSC) and the amplitude depends on the individual strength of the synapse which is determined by the number of so-called AMPA ($\alpha-amino-3-hydroxyl-5-methyl-4-isoxazole-propionate$) receptors. An EPSC creates an Excitatory Post-Synaptic Potential (EPSP) which leads to the increase of the intracellular potential of a neuron and therefore increases its likelihood to emit an AP. Accordingly, an IPSC generates an Inhibitory Post-Synaptic Potential (IPSP) which decreases the intracellular potential and thus reduces the likelihood of a neuron to emit an AP. It was demonstrated that EPSPs are affected by fluctuations that might be linked to changing probabilities of neurotransmitter release [36] or due to the specfic axonal structure [37].

The specialization of the CNS throughout life occurs through learning and the creation of memory which are not identical, i.e. learning is a rather quick process to gain some capability and memory is the long term consolidation of the learned feature [38]. It is widely understood that learning and memory are due to synaptic pruning and synaptic plasticity. The former is the

elimination of synaptic connections which takes place mainly during early childhood and is a key mechanism for the beginning specialization of the CNS [39]. Synaptic plasticity is the ability of synapses to change their strength of connectivity according to their history of activations, environment and neural processes or even to form new synaptic connections [40]. Effects of synaptic plasticity can be observed on various time scales, ranging from milliseconds to months or years with characteristic effects. Several different kinds of plasticity can be observed in a single synapse making it an incredibly complex structure [41]. Typically, synaptic plasticity is distinguished between short term and long term effects [42]. While the former tend to exhibit a rather fast transient behaviour, the latter lead to stable modifications. Known short-term plasticity effects are: (i) Short-term facilitation (STF): This is the transient increase of synaptic strength which occurs upon the arrival of two or more APs within a short time at the pre-synaptic terminal. Its effect is an increased emission of neurotransmitters in the event of an AP. (ii) Post-tetanic potentiation: This occurs due to high-frequency bursts of pre-synaptic spikes, known as tetanus, and can be observed with some delay after the burst. Its effect is also an enhanced neurotransmitter release, however it is of longer duration than the STF. (iii) Short-term depression (STD): This effect can be regarded as the opposite of STF because repeated pre-synaptic APs result in a synaptic weakening due to the depletion of synaptic vesicles supplying neurotransmitters. Typically, a synapse can either feature STF or STD, depending on the initial state. If the initial probability for neurotransmitter release is high (low), the synapse tends to perform STD (STF) upon the application of several activations [41].

Known long-term plasticity effects are: (i) Long-term potentiation (LTP) is the persistent strengthening of a synaptic connection, usually referred to as synaptic efficacy or weight. It occurs when a weak pre-synaptic stimulus (low frequency) results in the release of glutamate from the pre-synaptic terminal. The glutamate binds to both NMDA (N-methyl-D-aspartate) and AMPA receptors while the former are blocked by magnesium (Mg^{2+}) in the resting state and the latter are permeable for Na^{+}. When a pre-synaptic spike activates the synapse, the Mg^{2+} is removed from NMDA making them permeable for Ca^{2+}. Hence, the Ca^{2+} can enter the post-synaptic neuron where its concentration is strongly increases which eventually triggers the phosphorylation leading to the incorporation of additional AMPA receptors, as shown in figure 1.5 (a). Thus, the sensitivity to pre-synaptic Na^{+} release is increased, i.e. more Na^{+} can be induced by a pre-synaptic spike. (ii) Long-term depression (LTD) is the opposite of LTP, hence, the persistent weakening of a synaptic connection, usually referred to as synaptic efficacy or weight. It is caused by rather long low-frequency (around 1 Hz) pre-synaptic stimuli which lead to a slight increase of the Ca^{2+} in the post-synaptic terminal. This results in the activation of phosphatases which cleave phosphate groups and essentially remove AMPA receptors from the post-synaptic terminal, as illustrated in figure 1.5 (b). Accordingly, the sensitivity to pre-synaptic Na^{+} release is decreased, i.e. less Na^{+} can be induced by a pre-synaptic spike. Note that the cellular mechanisms behind LTP and LTD are described for hippocampal synapses here and can

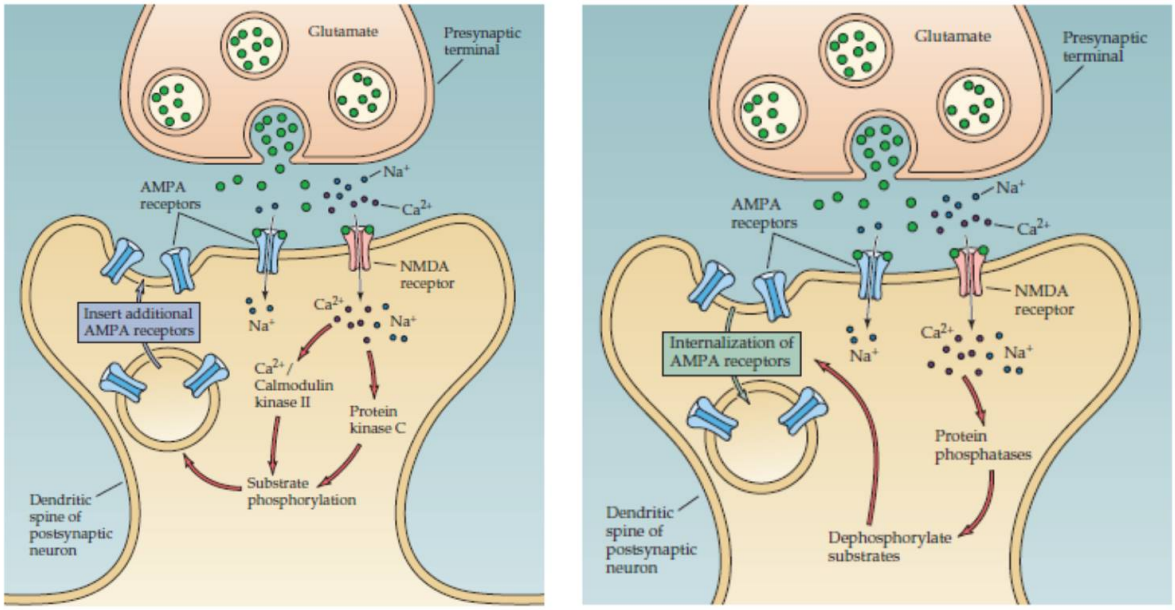


Figure 1.5: Schematic illustration of (a) Long-term potentiation (LTP) and (b) Long-term depression (LTD). The synaptic efficacy (strength) depends on the density of so-called AMPA receptors. Accordingly, LTP or LTD are due to an increased or decreased density of AMPA receptors. Source: [1]

differ for other CNS regions.

Spike-timing dependent plasticity (STDP) is a so-called learning rule which combines LTP and LTD based on the correlation of pre- and post-synaptic APs, evidenced by Tsodyks and Markram [43] as well as Bi and Poo [2]. When a post-synaptic spike occurs shortly after (before) a pre-synaptic spike, LTP (LTD) is performed on the synapse. Figure 1.6 shows the experimentally observed STDP. It is described as the change of the EPSC that occurs as a function of the relative timing of pre- and post-synaptic APs. A strong increase (decrease) was observed if the timing difference $\Delta t = t_{post} - t_{pre}$ was in the range $0\ ms < \Delta t < 40\ ms$ ($-40\ ms < \Delta t < 0\ ms$). This effect is also related to Hebbian learning proposed by Donald Hebb. In his theory, he stated that "The general idea is an old one, that any two cells or systems of cells that are repeatedly active at the same time will tend to become 'associated', so that activity in one facilitates activity in the other."

Other kinds of plasticity do certainly exist but are not yet fully understood, e.g. by controlling the membrane excitability, the CNS can actively adapt its sensitivity to sensory input activities [44]. It is believed that this threshold adaptation may as well enable or inhibit synaptic plasticity [45]. Furthermore, oscillation of neuronal ensembles may enhance synaptic plasticity temporally [34].

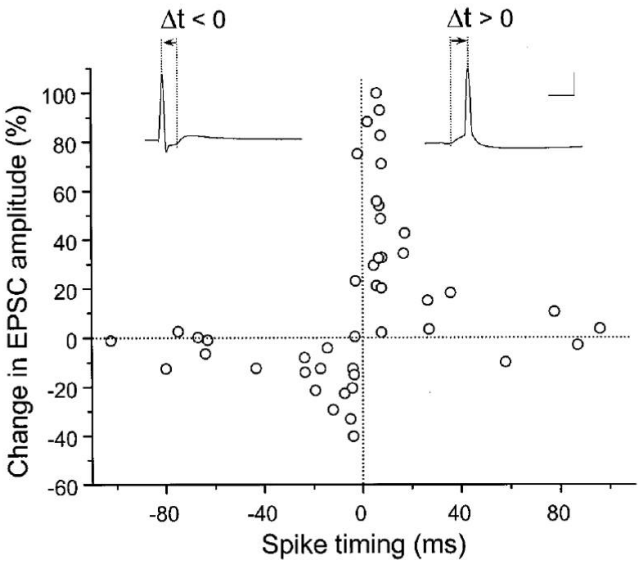


Figure 1.6: Experimental spike-timing-dependent plasticity found by Bi and Poo [2]. If a post-synaptic neuron spikes after (before) a pre-synaptic neuron, so that $\Delta T > 0$ ($\Delta T < 0$), their synapse is potentiated (depressed) in an LTP or LTD manner. This change of synaptic strength is expressed by the changed excitatory post-synaptic current (EPSC). A positive change, i.e. an increased EPSC means the synapse has become stronger whereas a negative change, i.e. a decreased EPSC indicates a weakened synaptic strength.

1.1.4 Electrophysiological techniques to record the brain activity

Neurodegenerative diseases or injuries can affect the physical structure of the brain as well as alter mechanisms that are critical for the proper functionality of the CNS. In order to find potential treatments or curing strategies, it is crucial to understand the functional organization and the basic principles of neural computation. This can be studied by using a variety of techniques that aim to record the brain's activity by means of various approaches:

- Non-invasive:
  - Electroencephalography (EEG)
  - Magnetoencephalography (MEG)
- Invasive by using implanted macroelectrodes:
  - Electrocorticography (ECoG)
  - Stereotactic electroencephalography (SEEG)
- Invasive by using implanted microelectrodes:
  - Microwires

– Microelectrode arrays (MEA)

The focus here is on electrophysiological techniques and the most important ones are summarized in figure 1.7. In EEG, one can simply attach a grid of electrodes onto the external surface of the scalp in order to record electric potentials. However, EEG lacks spatial resolution (usually in the range of a few cm) due to the large distance between the electrodes and the neurons and the electrical insulation of the skull (low conductivity of the bone). The electrode grid is positioned in contact with the brain below the skull (epidural or subdural) in the case of ECoG. By doing this, the spatial resolution can be increased by an order of magnitude to a few mm and noise usually decreases due to the absence of the bone. Typically, EEG and ECoG, do rather record large neuronal ensemble activities, however recent studies have proposed that even single neuron activities may be recorded by ECoG [46].

Implanted microwires and MEA's [28] exhibit greater potential for Brain-computer interfaces (BCI) and neural prostheses (NP) approaches than EEG/ECoG since they allow to record much more detailed signals from local electric potentials reflecting the activity of a single or a few cells [3] [4]. Two types of signals can be extracted from intra-cortical extracellular recordings: local field potentials (LFP's) and action potentials (APs) [47]. LFP's result from a number of active neurons around the electrode. LFP's mainly reflect synaptic activity and are mainly present in a frequency range below 200 Hz. With respect to LFP's, AP's originate from neurons on a shorter range with respect to the electrode and exhibit higher frequencies in a range between 200 Hz and 5000 Hz. Note that the AP's recorded in the signal correspond to the neural activity of different neurons and therefore may carry the encoded cortical message for a certain function, e.g. a motor command. To improve BCI performances, the current tendency is to record from several hundreds of microelectrodes simultaneously [26] [48] [27]. This is typically done at sampling rates ranging between $20-40$ kHz per channel, which generates huge amounts of data. This data needs to be processed in real time and sometimes transmitted wirelessly [29] [49].

Another technique which is not illustrated in figure 1.7 but should be mentioned is the

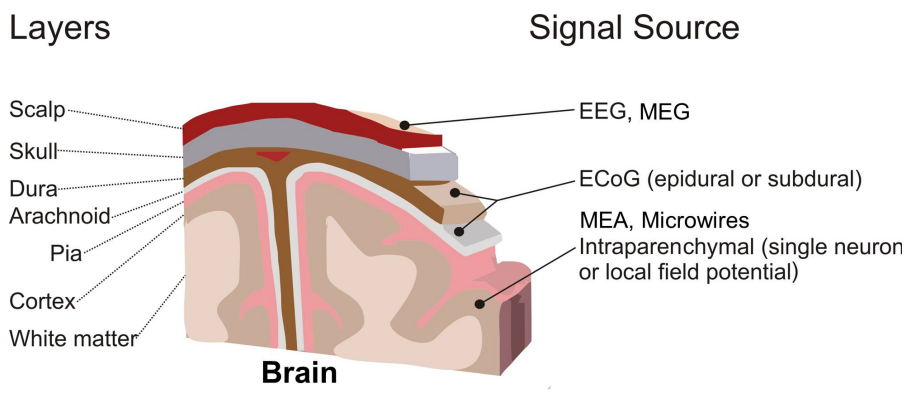


Figure 1.7: Overview of brain metrology techniques and their characteristic localization.

patch-clamp technique, developed around 1980 [50]. This technique uses μm-sized micropipettes which contacts the cell membrane and thereby encloses a small area of only a few ion channels. A second electrode is placed nearby the cell, but without contact to it. The setup allows to record the neuronal behaviour even in the absence of spikes, i.e. in the sub-threshold region. Since only a few or even just a single ion channel would be involved in the contact area of the micropipette, this experiment allows a significant improvement for the study of the behaviour of such channels. Alternative electrodes for neural recording were proposed but remained mostly at research level for niche applications [51].

1.1.5 Recording of single neurons

It was mentioned in section 1.1.4 that microwires and MEA's can be used to record the extracellular activity of single neurons, i.e. recorded by an electrode outside the neuron cell body, in the form of action potentials (or spikes). By doing this, a single electrode may record several cells in its close vicinity and the resulting signal is composed of a series of slightly different AP waveforms. Figure 1.8 shows an illustration of a multi-unit spike train (blue) which contains two characteristic AP waveforms. The shape of an AP strongly depends on the individual cell morphology, i.e. the geometry of the neuron, the location of the membrane channels involved in action potential generation and also importantly on the location of the electrode with respect to the neuron arborization. Typically, those AP waveforms have a rather stable shape that differs

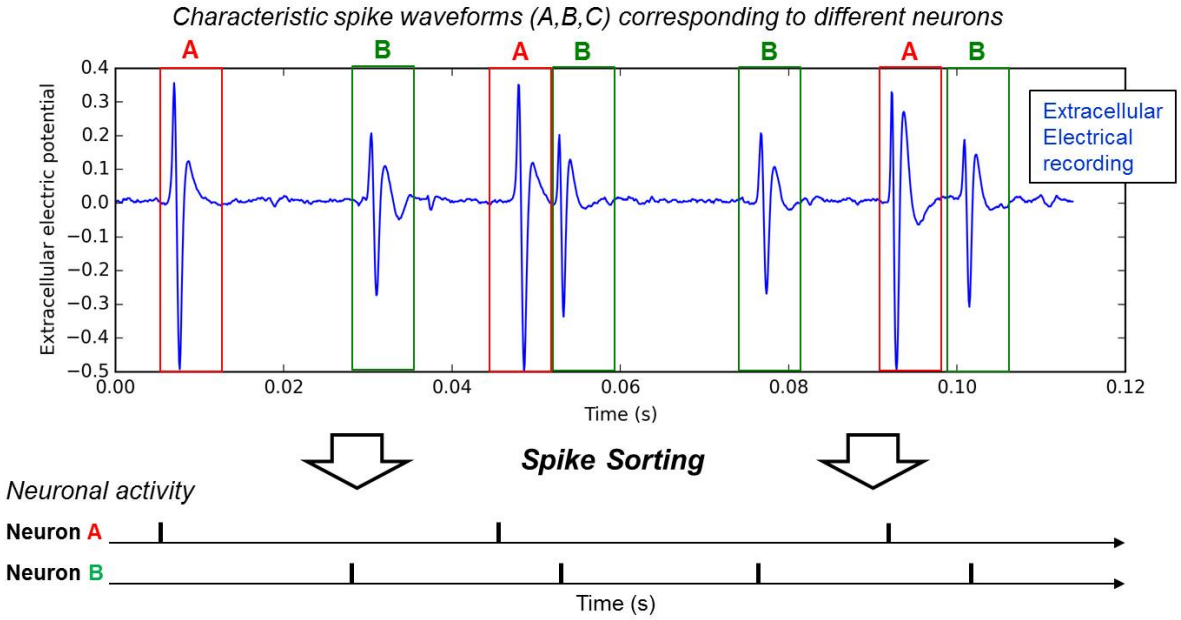


Figure 1.8: Schematic illustration of spike sorting from an extra-cellular electrical signal. Different spike waveforms are identified and associated to individual neurons. This allows to extract the spiking activity of single neurons from a multiplexed recording.

slightly from one cell to another [52]. Due to various distances between the recording electrode and individual neurons, the amplitudes (and delay times) differ for the detected APs [24]. Hence, it is possible to separate the activities in an extracellular signal corresponding to individual cells by separating the APs by their characteristic shape. This approach is called spike sorting [30] [53]. Spike sorting allows to follow the dynamics of several individual cells and thus offers great possibilities for the study of brain dynamics or the exploitation of the extracted information for brain-computer interfaces. Many spike sorting approaches have been proposed in the literature, but a current challenge is to implement real-time reliable methods into highly compact hardware at the level of a single microelectrode or array. Before describing the invention that makes a step toward this goal, the classical approaches for spike sorting are explained in the following. Note that the time at which a spike is detected in an extra-cellular signal, is in most cases not equal to the intracellular spike and it was even shown that extra-cellular spikes may be used to predict intra-cellular spike events [24].

1.1.6 State-of-the-art spike sorting techniques

The separation of single unit activities from a multiplexed electrical recording of several units (neurons) is conventionally done by following a 3-step procedure [54] [55]:

- **Detection:** In the first step, APs need to be discriminated from background noise in the recorded electrical signal. This is mostly done by defining a threshold level on the signal [56] [57]. Whenever the signal amplitude crosses this pre-defined threshold level, a spike is detected. Choosing a reasonable value for this threshold is crucial but not trivial since too low values can result in the identification of noise as spikes (False Positive) whereas too high values would miss small spikes for the analysis (False Negative). Automated spike detection directly implemented into a MEA was proposed in [58] for significant data bandwidth reduction.
- **Waveform analysis:** For step 2, the most commonly used techniques are based on template matching or feature analysis of the spike waveform.
  In *Template matching*, the detected spikes are extracted from the signal and matched to a certain template or a new spike shape template is created if the spike can not be associated with the already existing templates. This is done by temporally aligning the detected spike shape and computing an average waveform. Whether a detected spike matches a template or a new template is created, depends on tolerance levels specified in advance. It is possible to use either the entire data set or interpolating the spike shape templates based on sub-sampling the data. Note that this approach usually requires supervision by the user.
  Feature analysis is based on the characterization of the detected spikes by typical waveform features extracted from the data. The simplest way of discriminating between different

spikes is by comparison of their amplitudes (peak-to-peak), which is very fast but can lead to wrong results when different spikes have similar or equal amplitudes [56]. Hence, other features such as the local extrema, temporal width, integral, etc. can also be considered. A combination of n features is typically used for a precise spike shape description, which results in the representation of each spike by a point in a n-dimensional space, the axis of which are the features. The number and kind of features for extraction have to be chosen properly for sufficient spike distinction. Hereby, it is necessary to trade off precision versus computational effort since the extraction of a lot of features may increase the former but degrades the latter. For this reason, available spike sorting techniques extract many features in a first step and in a second step, the features that do not allow to distinguish different spike classes are eliminated. This results in a multidimensional space of less than n dimensions. Typical approaches to perform this dimensionality reduction are Principal or Independent Component Analysis, PCA or ICA, respectively [59] [60] [61].

- **Classification:** Once features are extracted, spikes are represented as points in a multidimensional space, the so-called feature space. Similar spike shapes typically result in points with similar location in this space. The definition of different spike classes is typically done by the so-called clustering which groups agglomerations of single points in the feature space, i.e. spikes with similar feature values, together. Thus, achieving spike sorting requires separating the clusters of points corresponding to the same cell signals. By taking n features into account, a n-dimensional cluster is created for each spike shape i.e. every cluster corresponding to one neuron. For this purpose, clustering algorithms such as k-means are typically used, which are usually very time and energy consuming. Typically it is useful or even necessary that the number of expected cells is known in advance. Current spike sorting approaches feature a critical bottleneck of being able to distinguish a maximum number of spike shapes which may be lower than the actual number of spike classes [55].

Strengths and limitations of standard spike sorting techniques and future requirements

Automatic unsupervised spike sorting approaches have been developed based on Bayesian clustering but generally lead to high computation loads that are difficult to implement in low-power consumption hardware [62] [63] [64] [65] [66]. Several studies focussed on low-power hardware implementations of spike sorting algorithms [65] [67] [68] [69] [68] [70] [71]. Alternative, more or less automatic, spike sorting approaches were proposed based on various kinds of feature extraction [54], neural network classifiers [72], iterative algorithms [73] [74], self-organizing maps [75] [76] [77], sparse data representation [78], wavelets [79] and frequency domain analysis [80]. However, a key problem is to find fully unsupervised methods allowing not only feature extraction but also fully automatic classification. Toward this goal, artificial neural networks incorporating learning capabilities (plasticity) can be considered. In a recent study, a STDP

neural network has been implemented, aimed at unsupervised spike sorting [81]. The approach uses a neural network of two layers which classifies snapshots of spikes (32 subsequent $8bits$ samples), however, it also relies on the threshold technique for spike detection. New approaches have to be developed for real time spike sorting of multi-electrode data recordings which can be integrated directly at the recording level of a micro-electrode array device [82]. Note that typically the number and characteristics of spikes is not known a priori.

1.1.7 Brain-computer interfaces

Brain-computer interfaces (BCI) are systems which measure the activity of the CNS and extract and decode specific information via signal processing. The extracted data is then used to control other technical devices such as prosthetic limbs [83]. BCI's can be based on both LFP and spiking data. The data can be recorded by a variety of technologies recording LFP as well as those which record spiking activities (see section 1.7 for an overview of techniques). As discussed in section 1.1.6, spiking data can be used in order to identify single neuron activities necessitating a method for spike sorting. Several studies have shown that best performances are achieved using spiking information rather than LFP's [84]. In the particular case of hand-control BCI's, it was shown that LFP (ECoG) data allows to achieve 7 different hand movements [85] compared to 10 based on spiking data [86]. Moreover, decoding performance is further increased when spiking activity has been sorted [87]. Finally, BCI performances increase when increasing the number of recorded cells [88] [26] [89] and spike sorting is an efficient way to reduce the amount of data to be transmitted when considering wireless systems.

Figure 1.9 shows two BCI studies by Hochberg et al. [3] [4] where MEA's were implanted into the brain of two patients who were tetraplegic, i.e. suffered from full paralysis of both legs and arms. The recorded signals were treated by sophisticated procedures using a standard computer whereas the operation was supervised by a scientist. Thanks to the extracted single neuron activities, the patient in figure 1.9 (a) is able to move a cursor on a computer screen while the other patient in figure 1.9 (b) was capable of controlling a robotic arm to reach and grasp a bottle and drink from it. Another study, based on the implantation of a tetraplegic subject with a 96-unit MEA, demonstrated a BCI that is able to control a prosthetic hand with 10 degrees of freedom [86] [90]. These are remarkable achievements towards the rehabilitation of paralysed humans. However, a number of critical bottlenecks follows from the used setup. First, a supervisor is needed to monitor (and adjust) the algorithm. This is a problem because ultimately BCIs should re-establish complete independence of patients. Second, the need of rather powerful hardware, i.e. a PC, in order to run the real-time signal processing prevents this BCI solution to become easily portable. Third, the MEA is connected to the outside via a cable instead of a wireless connection. This is not possible here because the data is processed externally by the PC which means that a huge amount of data needs to be streamed from the MEA to the PC. The problem of the wired connection is the significantly increased risk of infections at the location of the cable. For those

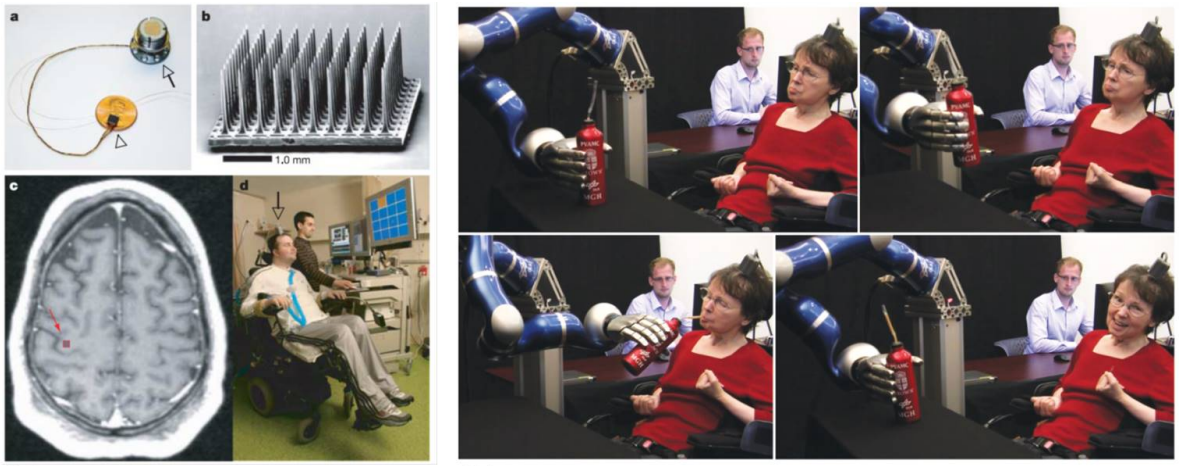


Figure 1.9: Selected brain-computer interface cased studies. Two tetraplegic patients were cortically implanted with multi-electrode arrays (left) and a dedicated setup was used for the signal treatment. (a) The subject is able to move a cursor thanks to harnessing his brain signals. (b) The extracted information allowed this subject control a robotic arm to grasp a bottle and drink independently. Source: [3] [4]

reasons, finding efficient automatic spike sorting methods that can be implemented in highly compact and real-time hardware is thus of crucial importance for the advance of BCIs and neural interfaces.

1.1.8 Neural prostheses

Neural prostheses (NP) aim to restore sensory or motor functions that were lost due to neurodegenerative diseases or injuries. Contrary to BCI, the approach of NPs is to stimulate parts of the human nervous system such as the retina, the cochlea or even the brain by means of implanted electrodes [91]. Hence, this approach can be used for hearing aids, tremor control, restoring vision or communication interaction [92]. Ultimately, it would be desired that BCIs and NPs are merged in a way that BCI are used to extract neural data which can then be used in a NP to control body functions. For example, the motor neural activity in a paralysed may be recorded and decoded in a BCI and the decoded information serves then in order to control a NP that stimulates the muscles of the limbs. Those applications could potentially enable patients to gain independence again by bypassing the damaged neural connection and using the own body instead of the need for additional prosthetic devices.

1.2 Memory technologies for data storage

Storing input and output data is a key function for information processing systems such as computers. Conventional computers are organized in the so-called von-Neumann computer architecture [31] where data is computed in a central processing unit (CPU) and stored in dedicated memory blocks. Those elements are physically separated and connected by a data bus which serves the information exchange. However, modern applications tend to be very data intensive which requires high data transfer rates between the CPU and memory. Moreover, modern CPUs have typical latencies in the range of nanoseconds while state-of-the-art memory technologies are in the range of microseconds or even higher. Thus, the total power consumption and speed is often limited by the memory access time and energy. These two issues are well known as the von-Neumann bottleneck or memory wall. In order to overcome these obstacles, several memory technologies are available on the market and a new generation of memory technologies is being researched, referred to as emerging non-volatile memory technologies.

1.2.1 Overview

Different memory technologies are available for the data storage, varying from their basic physical mechanism to their functional properties. Figure 1.10 shows an overview of the most important memory technologies which can be classified into volatile and non-volatile memories [93]. While volatile technologies lose the stored information within a short time after their power supply is cut off, non-volatile memories retain the data permanently. The established memory technologies (SRAM, DRAM, Flash) are all based on charge storage but exhibit very different characteristics which are exploited to fulfill specific roles in the memory hierarchy of von-Neumann architectures as shown in figure 1.11. The volatile technologies such as Static Random Access Memory (SRAM) and Dynamic Random Access Memory (DRAM) are used for time-critical tasks close to the processing core such as the cache and main memory, respectively. While SRAM is used as embedded memory close to the CPU, DRAM is typically used in stand-alone devices in a plug-and-play mode as main memory for computers. These technologies offer the lowest latencies (fastest read and write speed) which are necessary to exploit the ultra-low latency of nowadays processors. However, SRAM and DRAM consume a relatively large chip area (with SRAM being larger than DRAM) resulting in increased cost and need to be refreshed

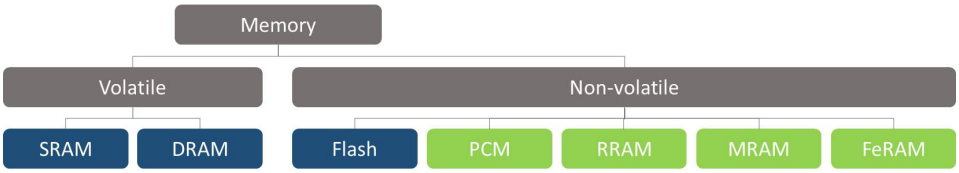


Figure 1.10: Overview of state-of-the-art (blue) and emerging (green) memory technologies.

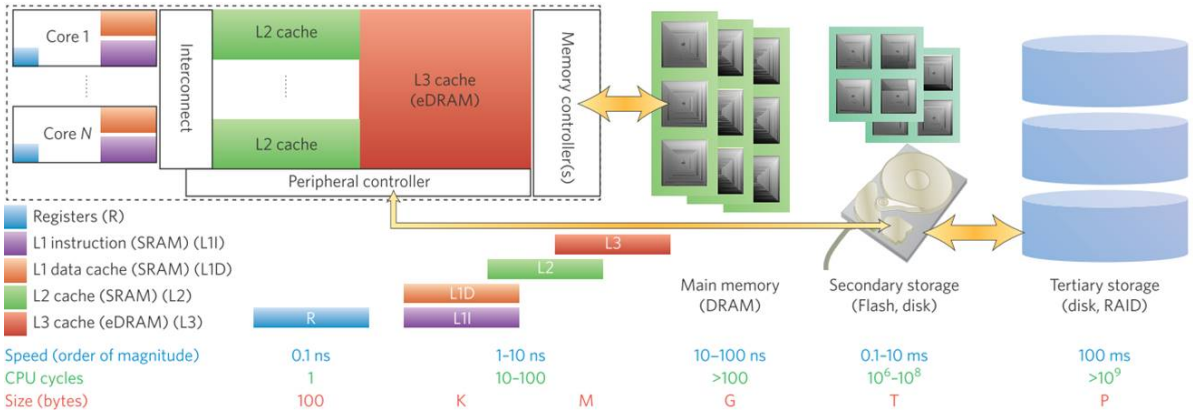


Figure 1.11: Memory hierarchy of conventional von-Neumann architectures constructed according the speed of the different technologies. Source: [5]

continuously to prevent data loss. This increases the energy consumption massively. On the other hand, non-volatile memory (NVM) technologies are generally slower but also occupy less chip area and therefore allow much higher bit densities. Moreover, they do not consume any power in standby. The main NVM technology is Flash while the NOR architecture (parallel organization of single memory cells) is rather used for embedded applications (micro-controllers) and the NAND (serial organization of single memory cells) is the standard technology for stand-alone products (USB drives, flash cards, solid state drives). Finally, the highest memory density at lowest cost and lowest speed is achieved by hard disk drives which are used for bulk data storage.

1.2.2 Emerging Non-Volatile Memory technologies

For several decades, Flash was scaled down leading to an ever increasing density of data storage [5]. However, Flash technology is facing fundamental scaling difficulties related to its basic physical mechanisms for switching and storing data. Moreover, its energy consumption for programming is rather large. Hence, a gap in the memory hierarchy is prospected to evolve in the future posing a critical challenge for the advancement of the von-Neumann computing architecture. This gap is projected to be filled by so-called storage-class memory (SCM) solutions which may be realized by some emerging memory technology (see figure 1.10). Research on those technologies has largely accelerated during the last decade and first products start to be commercialized. Unlike SRAM, DRAM or Flash, those emerging technologies are no longer based on charge storage but deploy some physical mechanism to change their electrical resistance, thus embodying so-called memristors, introduced by Leon Chua in 1971 [94] and discovered in 2008 [95]. As memristors are non-volatile, this new class of memories is typically referred to as emerging non-volatile memory (NVM) technologies. The major emerging NVM technologies are phase change memory (PCM), resistive random access memory (OxRAM), magnetic RAM (MRAM) and ferroelectric RAM (FeRAM). An increasing interest in the technologies can be observed not

only for classical NVM application but also for Internet of Things (IoT) related applications and the implementation of synapses for brain-inspired (so-called non-von-Neumann) computing architectures. Those applications require large memory densities, low power consumption and low cost. The most important NVM technologies with a high potential for future implementations of artificial synapses are introduced in the following.

1.2.2.1 Phase change memory

Phase change memory (PCM, PRAM or PCRAM) devices are based on the transition of a phase change material between a crystalline and amorphous phase. Those materials are known as chalcogenides and the most widely used material composition is $Ge_2Sb_2Te_5$, commonly referred to as GST. The two states exhibit a large difference in the electrical resistance as shown in figure 1.12 (a). The figure shows the experimental resistance for different chalcogenide materials which are in a high resistance state (HRS) as-deposited, i.e. $R > 1\Omega cm$. When the temperature is slowly increased while the resistance is monitored, one can observe a gradual decrease of the resistance. Furthermore, a sudden drop of the resistance occurs at the crystallization temperature that is characteristic for each material. This drop is the transition from the amorphous to the crystalline phase and leads to the low resistance state (LRS).

This resistance difference of several orders of magnitude between LRS and HRS can be exploited for memory applications, i.e. featuring distinct On ('1') and Off ('0') states. Therefore, the material is integrated in a lateral stack with a heater in series and metallic top and bottom electrodes which is also known as the mushroom structure, shown in figure 1.12 (b). The function of the heater is to apply thermal pulses that trigger the temperature induced phase changes in the chalcogenide. Figure 1.12 (c) illustrates schematically the applied pulses. In order to

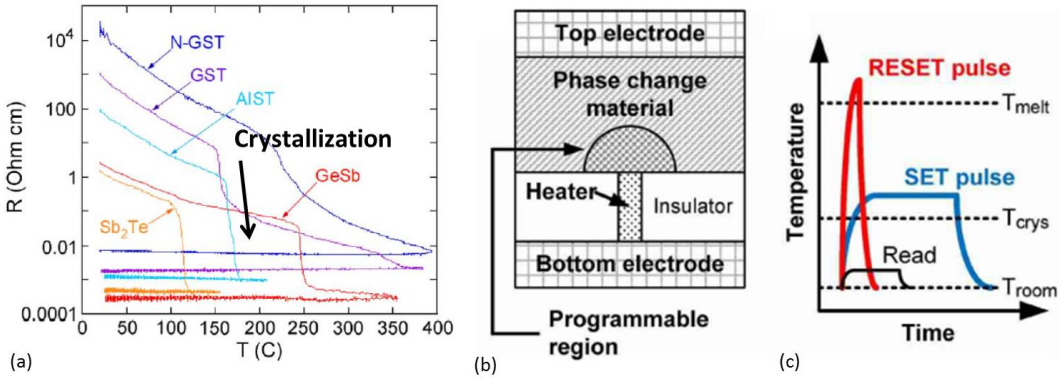


Figure 1.12: (a) Resistance as a function of temperature for major phase change materials. When a material in amorphous state is heated up, a crystalline phase will form at some temperature resulting in a significantly lower electrical resistance. (b) Typical integration of a phase change material in a mushroom cell for memory application. (c) Characteristic programming pulses to trigger conversion into crystalline (Set) or amorphous phase (Reset). Source: [6]

switch a PCM device to LRS, i.e. the Set operation, a rather long pulse of medium temperature ($T_{crys} < T_{pulse} < T_{melt}$) has to be applied to provide enough time for a complete crystallization of the active zone. To switch the PCM device back to the HRS, i.e. the Reset operation, a short pulse creating a high temperature ($T_{pulse} > T_{melt}$) is used to melt the material locally and quench it subsequently resulting in an amorphous region. The read operation is simply a low voltage IV measurement which should avoid to raise the temperature above T_{crys} to prevent altering the memory state (i.e. 'read disturb'). All device operations can be achieved using the same bias polarity, hence, PCM devices are called unipolar or nonpolar.

PCM characteristic drawbacks are the high power consumption for programming, in particular for the Reset operation and the drift of the resistance state, mainly in HRS. The drift is a problem especially for the design of so-called multi-level cell (MLC) architectures which aim at implementing more than 1 bit per PCM cell, i.e. instead of only LRS and HRS, intermediate resistance states are programmed with well defined programming conditions. This allows to increase the storage density. However, the drift can lead to erratic bits that are linked to single cells that change their state spontaneously due to the drift phenomenon. A possible solution to mitigate the drift effect is the projected phase-change memory concept where programming and reading affect different physical zones of the PCM device and thus drift, resistance-temperature dependence and thermal current noise can be significantly reduced [96]. Alternatively, a small additional pulse can be applied immediately after Reset to accelerate the drift effect [97]. Another trend in PCM research is to reduce the programming current for a lower energy consumption by means of inter-grain regions [98] or shrinking the device size in so-called confined structures [99]. Strong scalability down to $20nm^2$ and high integration density at $4F^2$ were demonstrated for the confined PCM. Hence, these concepts may allow to drastically reduce the operation currents and thus the overall energy consumption.

1.2.2.2 Resistive Random Access Memory

Resistive Random Access Memory (RRAM or ReRAM) is a generally broad term which describes a number of slightly different memory types. As for PCM, the principle of RRAM is the resistance modulation in order to store '0' and '1'. The basic structure of a RRAM device is very simple, namely a thin layer of one or multiple metal oxides sandwiched between two metallic Top (TE) and Bottom electrodes (BE), respectively, in a so-called metal-insulator-metal (MIM) structure, see figure 1.13 (a). It was reported in the literature that thin oxides exhibit a sudden switching phenomenon upon the application of a critical electric field between TE and BE that results in a drop of the electrical resistance of the oxide [100] leading to the so-called Low Resistance State (LRS). This resistance change is commonly attributed to the formation of the so-called Conductive Filament (CF). The oxide transformation is partly reversible by breaking the CF which is thought to create a tunnel barrier between the remaining part of the CF and the electrode thus blocking the current conduction and leading to the High Resistance State (HRS).

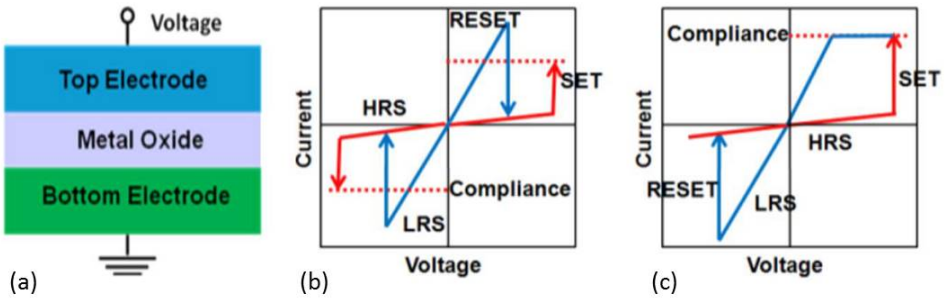


Figure 1.13: (a) Basic structure of a RRAM cell. (b) Unipolar and (c) Bipolar device operation. Source: [7]

The resistance levels of LRS and HRS depend on the applied programming condition. While LRS depends on the current compliance (I_{CC}) level during the set operation, the HRS level is determined by the applied Reset voltage (V_R) and also shows some dependence on I_{CC} [7]. The LRS dependence on I_{CC} is believed to be due to the CF geometry because a higher I_{CC} lead to a bigger diameter of the CF (d_{CF}), hence a lower LRS. On the other hand, the higher V_R the longer the tunnel gap, i.e. the higher the HRS. Moreover, the density of defects induced in the oxide during the Set operation depends on the I_{CC} and thus indirectly affects the HRS. This is because only a part of the defects is removed during the Reset operation.

RRAM can be classified according to (i) the IV switching characteristics and (ii) the physical mechanism that dominates the switching effect. Regarding the IV characteristic, one can discriminate between unipolar and bipolar devices. If the materials used for the TE and BE are both inert, the set and reset operations can be performed independently of the bias voltage polarity as it is illustrated in figure 1.13 (b). These devices are called unipolar. If one of the electrodes is replaced by an oxidizable material, a bipolar device is obtained, i.e. opposite polarities are necessary to switch the memory as figure 1.13 (c) shows. In some cases, both unipolar and bipolar switching characteristics could be observed in the same devices [101].

Furthermore, the choice of the materials for the metal electrodes (TE and BE) plays an important role for the dominating switching mechanism. If non-reactive metals such as platinum (Pt) or titanium (Ti) are used, the CF is created by anion migration while reactive metal electrodes such as silver (Ag) or copper (Cu) enable cation migration forming either an oxygen vacancy or metal ion based CF. Accordingly, it is useful to distinguish between oxide vacancy based RAM (OxRAM) and Conductive Bridge RAM (CBRAM). Both OxRAM and CBRAM technologies suffer from cycle-to-cylce as well as device-to-device variability in both LRS and HRS. This is a major concern for standard non-volatile memory applications.

Oxide vacancy based Random Access Memory

OxRAM technology relies on a functional oxide, typically transition metal oxides such as hafnia (HfO_2), alumina (Al_2O_3), titanium oxide (TiO_2) or tantalum oxide (TaO_x). The electric field during the Set operation causes the diffusion of oxygen ions (O^{2-}) towards the electrode interface which leaves oxygen vacancies (V_O^{2+}) behind while the path of highest V_O^{2+} density forms the CF [7]. The oxide transformation is partly reversible in the Reset process by breaking the CF which occurs when O^{2-} diffuse back (from the reservoir at the electrode interfaces) and recombine with V_O^{2+} of the CF. This is thought to create a tunnel barrier between the remaining part of the CF and the electrode thus preventing ohmic current conduction and leading to the High Resistance State (HRS) [102]. It is expected that Joule heating is involved in the Reset process by thermally activated O^{2-} diffusion [103].

Ultra-low (sub-μA) operation currents [104] [105] [106] [107] [108] [109] were demonstrated as well as the excellent scalability of OxRAM by $10x10\ nm^2$ functional devices [110]. Endurance of 10^{11} programming cycles in combination with good retention characteristics were experimentally demonstrated in [111]. Multi-level capability by tuning the reset voltage was demonstrated in [112]. The requirement of a forming voltage with a relatively high voltage poses a problem for potential applications, therefore it was shown in [113] that doping HfO_2 with silicon (Si) can provide forming free OxRAM devices.

On the other hand, variability poses a serious obstacle for the industrialization of OxRAM. HRS variability is widely attributed to the variation of the tunnel gap length between the CF

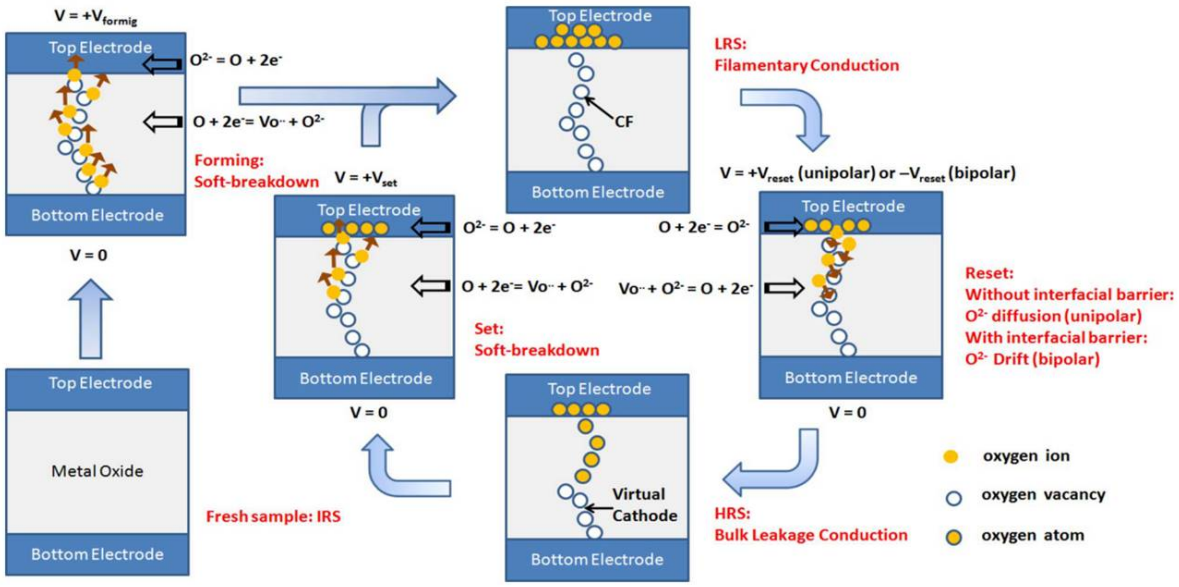


Figure 1.14: Illustration of basic physical mechanism involved in switching oxide vacancy based Random Access Memory. Source: [7]

and the electrode and the defect density/distribution in the dielectric [114] [115] [116]. HRS tail bits at low resistance were linked to new traps generated at the end of the reset process [114]. Relaxation effects were hypothesized to cause a drift of both LRS and HRS within a few microseconds, effectively reducing the resistance margin [117] [117].

Conductive Bridge Random Access Memory

CBRAM is very simlar to OxRAM, however, its resistance modulation is driven by metallic cation (Ag^+ or Cu^+) [118] migration originating from the reactive metal electrode (also known as active electrode). As shown in figure 1.15, metal ions diffuse into the oxide where they form a metallic CF during the Set operation. During Reset, the CF is broken which means that the TE and BE are disconnected resulting in the HRS [8]. The mechanism of the CF growth is not yet fully understood and in fact there are some controversial theories whether the growth starts at the interface of the inert or active electrode, mainly between the groups of Waser et al. [119] and Celano et al. [120]. Using a specialized 3-dimensional tomography technique based on conductive AFM, it was shown by Celano et al. [121] that the CF may have a conical shape with its constriction at the electrode interface. This gives rises to the assumption that the filament growth starts at the active electrode propagating towards the inert electrode and that it eventually becomes limited by the cation migration at the CF constriction. It was also shown that the HRS may be formed by breaking the filament in an abrupt process or it can be thinned down in a gradual Reset [122]. CBRAM is a promising memory technology since it allows to achieve a large margin between LRS and HRS, known as memory window, even if low programming

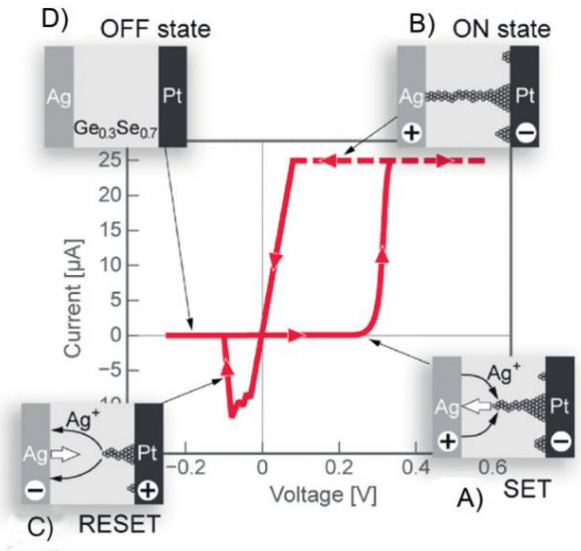


Figure 1.15: Illustration of basic physical mechanism involved in switching conductive bridge Random Access Memory. Source: [8]

currents (in the range of a few μA) are used [123] [124] [125].

1.2.2.3 Magnetic Random Access Memory

Magnetic Random Access Memory (MRAM) is one of the emerging NVM technologies that bears high potential to become a 'universal' memory, a technology that is thoroughly sought after. It provides some very promising advantages such as non-volatility, high speed, very long lifetime due to no stress effects and CMOS process compatibility. In MRAM, the information, i.e. $0's$ and $1's$, is stored in the magnetization of ferromagnetic materials. The principle origins back to the giant magnetoresistance effect (GMR) which was discovered three decades ago. It occurs when electrons pass a stack of two ferromagnetic layers separated by a non-magnetic layer. If the two layers are magnetized in parallel, the electrons with the spin parallel to that magnetization, called up electrons, can easily pass from one layer to the other. On the other hand, the neurons with the spin anti-parallel to the layer magnetization, called down neurons, undergo a strong scattering effect. Thus, the electrical resistivity for the up electrons is low while it is high for the down electrons. If one of the layers is magnetized anti-parallel to the other one, both up and down electrons experience a high electrical resistivity. Hence, a GMR ratio of $\Delta R = (R_{AP} - R_P)/R_P$ can be observed where R_{AP} and R_P are the electrical resistances in anti-parallel and parallel configuration. This effect is typically used in hard disk drives by keeping the magnetization of ferromagnetic layer constant while switching the magnetization of the other layer, known as free layer (FL) [126]. However, the GMR ratio is typically quite low (only a few %). Therefore, the non-magnetic layer between the two ferromagnetic layers is replaced with a thin dielectric barrier, the so-called magnetic tunnel junction (MTJ), as shown in figure 1.16. In case of a MTJ made of Al_2O_3 and MgO, an effect similar to the GMR, the tunnel magnetoresistance (TMR), can be observed but with a much higher TMR ratio of around $150 - 200\%$. Based on the MTJ, there is a number of different MRAM technologies as shown in figure 1.17. The toggle technology (figure 1.17 (a)) uses electric currents through the word lines (WL) to generate magnetic fields for switching the MTJ. The drawback is that two WL and a bit line and a high write current are needed, making this design poorly scalable. Figure 1.17 (b) is the thermally-assisted (TA-MRAM) scheme where programming is achieved by exploiting the thermal effect of the MTJ current with a coincident magnetic field. This technology offers some interesting properties for niche applications

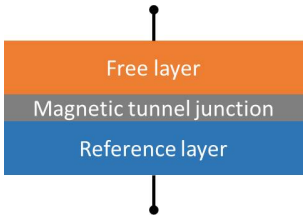


Figure 1.16: Basic structure of a MRAM cell.

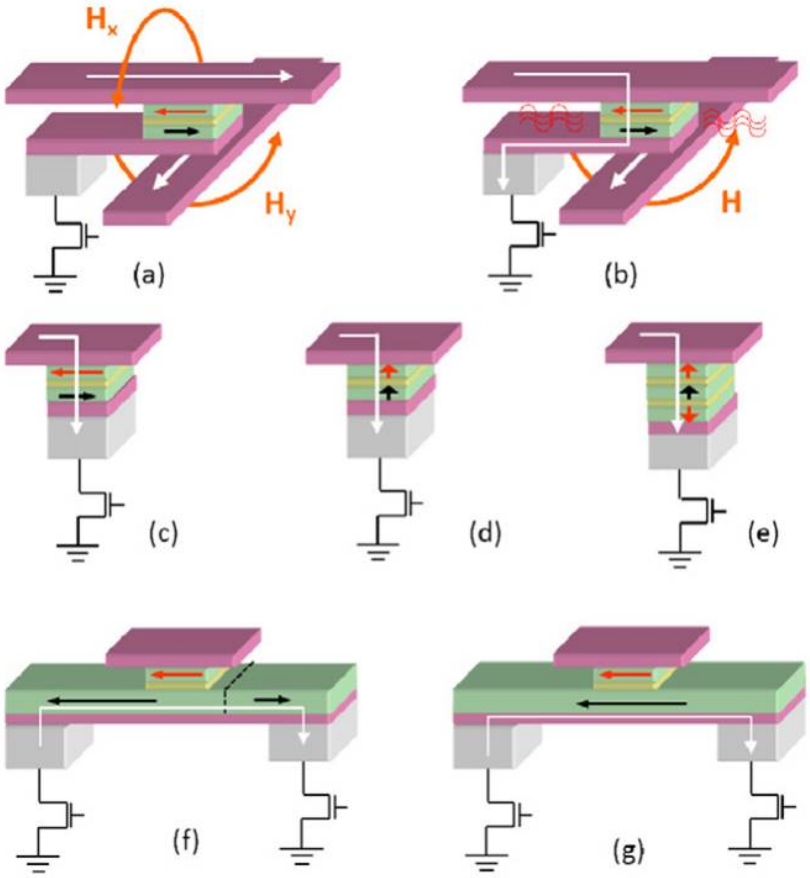


Figure 1.17: Overview of current MRAM technologies: (a) Toggle, (b) Thermally assisted MRAM, (c) in-plane STT-RAM, (d) and (e) perpendicular STT-RAM with single and double reference layers (f) domain wall propagation MRAM (g) Spin orbit torque MRAM. Source: [9]

but still suffer from scalability constrictions. The discovery of the spin-transfer torque (STT) effect enabled much lower currents and thus a better scalability for MRAM techniques. As shown in figures 1.17 (c) - (e), STT-MRAM cells exhibit rather simple architectures because the FL magnetization can be switched directly by means of the MTJ current. This is because the current becomes spin-polarized when it flows through the pinned layer and it can transfer their spin (or momentum) to the free layer and thus switch its magnetization [127]. The minimum write current of approximately $15\mu A$ of in-plane STT-MRAM (figure 1.17 (c)) can be further reduced by the perpendicular STT-MRAM architecture (figure 1.17 (d)) while even achieving better retention and scalability. Two anti-parallel polarized layers can be used in order to increase the STT efficiency as shown in figure 1.17 (e). Alternative three-terminal devices such as the domain-wall propagation (figure 1.17 (f)) or spin orbit torque (figure 1.17 (g)) may be used to design non-volatile logic elements thanks to the separation of the read and write paths [9].

1.2.2.4 Ferroelectric Random Access Memory

Ferroelectric Random Access Memory (FRAM or FeRAM) is, like MRAM, a promising technology for various future applications since it offers non-volatile data storage, fast write and read, low energy consumption and good retention [128]. It relies on ferroelectric materials, typically $PbZr_xTi_{1-x}O_3$ (PZT), $SrBi_2Ta_2O_9$ (SBT), $(Bi,La)_4Ti_3O_{12}$ (BLT) [129] and ZrO_2. Recently, HfO_x was found to exhibit ferroelectric properties if doped with e.g. Yttrium (Y), Aluminum (Al), Gadolinium (Gd), Strontium (Sr) or Lanthanum (La) [11] [10]. This offers enormous potential since HfO_x is among the most widely used materials in standard CMOS technology. FRAM relies on the spontaneous electric polarization of the mentioned materials while the the polarization can be inversed by means of the application of an opposite electric field. When the electric field is removed, the FRAM material keeps its polarization. This leads to a hysteresis of the polarization as a function of the applied external electric field, as shown in figure 1.18. At $E = 0$, the two polarizations provide the two memory states, 0 and 1, respectively. The polarizations can be measured and exploited in several FRAM device concepts such as $1T-1C$, $2T-2C$ or $1T$ which provide different performances [10].

1.2.3 Three-dimensional integration concepts

In order to improve the integration density of 2-terminal devices such as PCM and RRAM, two main concepts were proposed. Both of them aim to leverage the simple structure and back-end-

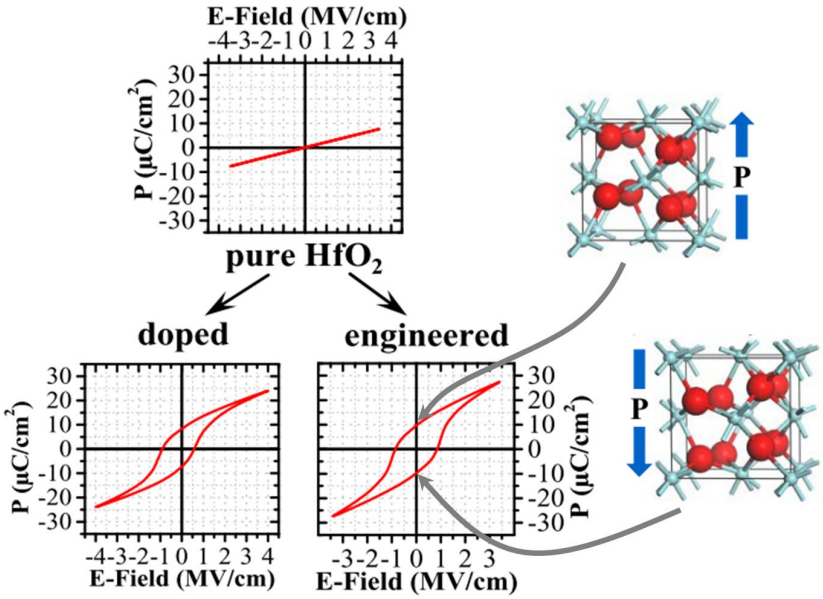


Figure 1.18: Principal mechanism of ferroelectric materials. Depending on the orientation of the crystal structure (note the red atoms), the material exhibits a spontaneous electrical polarization with positive or negative polarity. Source: [10] [11]

of-line (BEOL) process compatibility of such emerging NVM. The 3D cross-point architecture (shown in figure 1.19 (a)) relies on the idea to stack several layers of planar devices on top of each other [130]. Bit lines and word lines are perpendicular to each other with a memory cell at each of their intersections. This design allows to use every word and bit line for two layer of memory devices and thus reduce the number of metal layers by a factor of $2x$. Note that for this reason the polarity of two cells of layer n and $n+1$ is therefore inverted, i.e. the process flow for depositing the MIM layers has to be alternated. Due to the very high integration density, this architecture may be affected by thermal crosstalk of adjacent cross-points due to large reset currents [130]. Crossbar Inc., working with this 3D integration scheme, has started the production of 'silver-over-amorphous-silicon' based RRAM in 2016 [131]. Furthermore, in 2015, Intel and Micron announced to be working on the commercialization (expected for 2017) of their so-called $3DXPoint$ technology which is most likely a three-dimensional PCM design [132]. The second alternative integration concept is called vertical RRAM (VRRAM), illustrated in figure 1.19 (b) [12]. Instead of a planar device structure, the MIM layers are integrated vertically in a pillar reducing the number of masks, thus reducing the cost. The innermost material (bulk of the pillar) is the so-called vertical electrode whose side wall is covered by the resistive switching layer oxide. Another large scale horizontal metal sheet forms the second electrode. The memory element is located where the horizontal electrode surrounds the vertical one. High bit density at very low operating current ($< \mu A$) was demonstrated in [133].

1.2.4 Comparison

Table 1.20 provides a brief comparison of the current state of main emerging NVM technologies. Major challenges will be the process integration in the front-end-of-line (FEOL) of FeRAM and in the back-end-of-line (BEOL) of MRAM to replace Flash. For PCM and RRAM, reliability and power consumption issues are still restraining the industrial integration into reliable products.

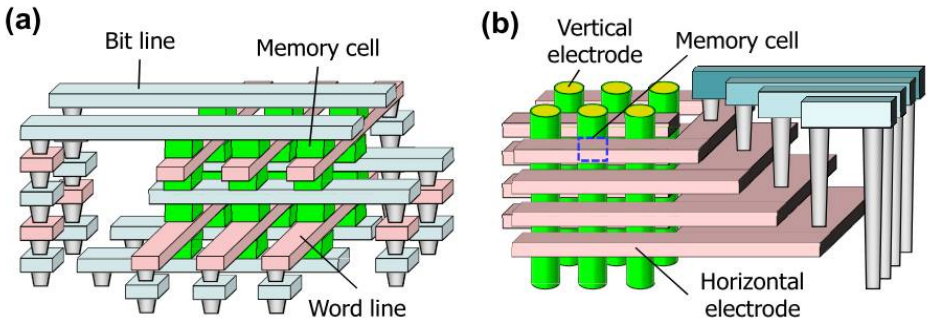


Figure 1.19: (a) 3D cross-point and (b) vertical RRAM (VRRAM) integration architectures. Source: [12]

	Main advantages	Key challenges
FeFET	• 1T cell structure • Low-power field-driven • High performance • Ferroelectric doped HfO_x	• Material and processing • FEOL integration • Reliability and parasitic effects (*e.g.*, charge trapping)
PCM	• Maturity • Proven performance	• Reliability • Disturbance • High switching power
STTRAM	• High performance • Well-understood physics • Novel mechanisms (*e.g.*, SHE, VCMA) to extend capabilities	• Reducing I_c/Δ (power-stability tradeoff) • MTJ patterning and etching • BEOL thermal budget
RRAM	• Simplicity and low cost • High density • Versatile materials, structures, and behaviors	• Reliability and failures • Stochastic mechanism and intrinsic variability • Forming

Figure 1.20: Comparison of most important emerging non-volatile memory technologies. Source: [13]

1.3 Emerging NVM in Neuromorphic Systems

After Leon Chua suggested the existence of a fourth basic electronic circuit element, the memristor in 1971 [94], he proposed this device class to be used for so-called memristive systems [134], where essentially the memristors function as synaptic weight elements. Finally, a structure behaving in the way Chua had postulated, was found in 2008 by HP labs [95]. The same group then proposed to use such nano-scale memristors as synapse with intrinsic properties that resemble biological STDP [135] to build neuromorphic systems which are inspired by the human brain. The discovery of memristors a decade ago, in combination with the development of emerging non-volatile memory technologies, has sparked a lot of interest and accelerated the worldwide research effort in the neuromorphic field. Here, computation and memory are co-localized and distributed in the form of numerous neurons which are interconnected by synapses (making them fundamentally different from von-Neumann architectures). Since software algorithms of artificial neural networks (ANN) are very time and energy consuming to run on conventional von-Neumann computers due to their intrinsic parallelism, it is necessary to develop approaches for the hardware implementation of ANN. One critical building block are synapses since they typically outnumber neurons if a connectivity comparable to the brain (around 10^4) shall be achieved. Synapses of neural systems as the CNS are essentially variable resistors that change their conductance according to certain rules. Moreover, they can be considered two terminal devices which makes them inherently similar to the structure of artificial synapses based on resistive memory technologies. Some emerging NVM such as RRAM and PCM offer excellent properties to mimic biological synapse features like low power consumption, high integration

density, long endurance and other synapse-like properties. In order to develop sophisticated hardware of plastic synapses and learning rules, it may be beneficial to use test platforms as an intermediate solution [136] which feature event-based simulation that is faster and more precise than von-Neumann computing [137]. It is of crucial importance to develop realistic models for the different types of memristive devices [138].

Both neuron and synapse implementations based on emerging NVM have been proposed. According to the scope of this dissertation the focus of the following sections will be on concepts of artificial synapses while neurons will not be explained here. Interested readers are referred to [139] [140] [141].

1.3.1 Application of NVM in synapses

The input to a neuron with more than one input synapse is essentially the so-called multiply-accumulate (MAC) function. That is, every synapse performs a multiplication of the activity impulse with its weight and propagates this value to the neuron. The neuron sums up all input from its synapses. Multiplication, however, is a computationally expensive process on a von-Neumann processor. Recent advancements in the field of emerging non-volatile memories have attracted a great interest by the brain-inspired computing community due to properties that are very promising for the implementation of artificial neural networks (ANN) in hardware. Their capability to alter the electrical conductance (or resistance) resembles to some extent the behaviour of synapses (e.g. in the human nervous system) which can change their density of ion channels, thus increasing or decreasing their response strength, known as synaptic efficacy. Hence, it derives naturally to use memory devices (i.e. memristors) as synapses because every memory cell performs the multiplication intrinsically physically due to Ohms law ($i = R * v$). Among the most promising technologies are resistive random access memory (RRAM) and phase-change memory (PCM). It was demonstrated that RRAM based synaptic arrays perform better in terms of area and leakage consumption with respect to SRAM based synaptic arrays but may be worse in latency [142]. Several emerging NVM technologies were explored for synaptic designs [143] and most often used in hybrid architectures, NVM based synapses and CMOS based neurons [144]. As biological synapses are complex structures with some characteristic features, it is however not trivial to mimic their exact behaviour by means of of one specific memory technology.

1.3.1.1 PCM synapses

It was found that PCM exhibits a progressive Set operation (crystallization) [145] [14] for the application of short (a few tens of nanoseconds) identical programming pulses as shown in figure 1.21 (a). This is due to the step-by-step crystallization of the active material and indeed a very promising result for the implementation of synaptic potentiation, i.e. the increase of a synaptic weight [14]. However, as figure 1.21 (b) shows, the reset process (amorphization) of PCM is abrupt,

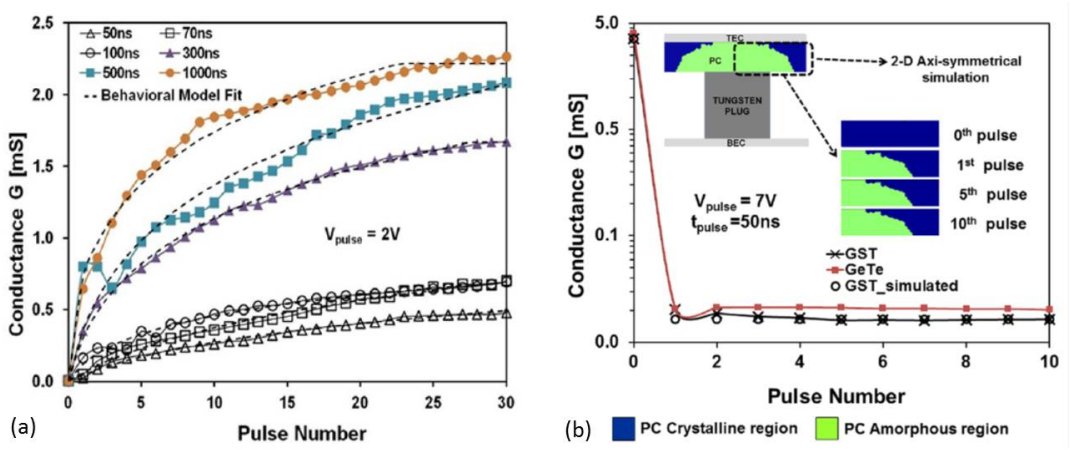


Figure 1.21: (a) Gradual crystallization of PCM cell upon application of identical set pulses. (b) Abrupt amorphization of PCM cell upon application of identical reset pulses. Source: [14]

i.e. a device is switched from HRS to LRS within one pulse and the resistance can no more be altered by the application of a pulse with the same parameters. This bottleneck was overcome with the introduction of the so-called 2-PCM synapse whose concept is the implementation of a synapse with two PCM devices, shown in figure 1.22 (a). Here, the long-term potentiation (LTP) and long-term depression (LTD) are performed by dedicated PCM devices. The synaptic weight is then measured as the differential current of LTP and LTD devices. It is possible that one device reaches its maximum conductance so that the overall conductance can not be altered any more by further LTP or LTD operations. Therefore, a refresh mechanism was used (see figure 1.22 (b)) after a fixed time interval or when a certain number of synaptic activations is reached [15]. In this case, both devices are reset to the HRS and the previous synaptic weight is restored by

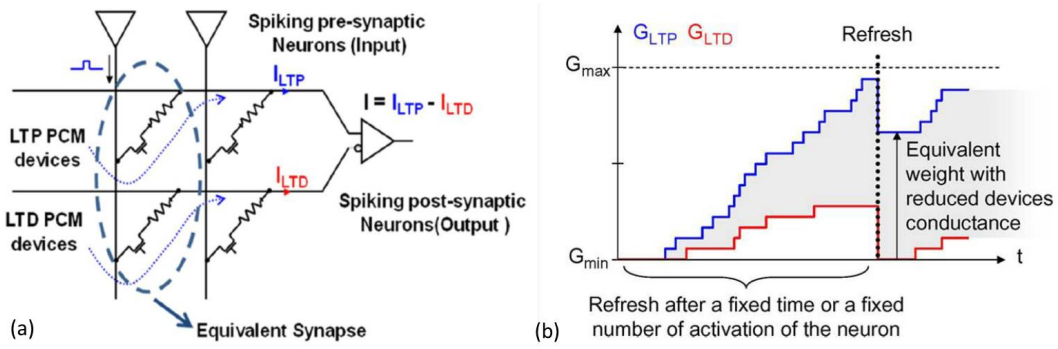


Figure 1.22: (a) The 2-PCM synapse design. Both LTP and LTD device use gradual crystallization in order to achieve a progressive potentiation and depression of the synaptic weight which corresponds to I. (b) Refresh algorithm to prevent saturation of synaptic weights. LTP and LTD devices are reset and the previous synaptic weight is restored by gradually programming the device which was stronger before the refresh. Source: [15]

gradually programming the PCM device which was stronger before the refresh to the desired level. It was demonstrated that this 2-PCM synapse device can be used to build spiking neural networks (SNN) [146] [147] [15] [148] and multi-layer perceptrons [149] [150] [151]. Inserting an additional HfO_2 layer between the heater and the plug allows to reduce the energy consumption in read and program mode and thus for the ANN [152] [153]. Since the 2-PCM synapse exploits only crystallization, it is inherently unaffected by drift mechanisms [154]. By tuning the Set and Reset pulse amplitudes, both potentiation and depression can be performed gradually in single devices [155] [156]. It was also shown that PCM operated in a binary mode, by application of long enough programming pulses, can be used as binary synapses in SNN [157]. Another approach was demonstrated in [158] where synapses were implemented with single PCM devices featuring gradual LTP and abrupt LTD being sufficient for the specific application.

1.3.1.2 OxRAM synapses

Section 1.2.2.2 explained that the programmed resistance in LRS or HRS of an OxRAM cell depends on the used current compliance (I_{CC}) or reset voltage (V_R), respectively. Figures 1.23 (a) and (b) show several set and reset operations while the I_{CC} and V_R are gradually increased from cycle to cycle. As the IV curves indicate, this leads to a gradual decrease or increase of the OxRAM device resistance. Thus, it is possible to use on OxRAM device as an analogue synapse by carefully tuning of the programming conditions in order to gradually tune its conductance, as shown in figure 1.23 (c) [16]. This concept of synaptic plasticity was used based on single devices [159] [160] [16] [161] [162] and OxRAM arrays [163] [164] [165] [166] [167]. However, tuning the programming pulse conditions at every programming step requires to read the synaptic state first and then adapt the driver circuit accordingly to induce a small conductance change. This increases significantly the circuit complexity and energy consumption. In order to avoid this, two invariant pulse conditions should be used, one for potentiation and another one for depression. This will not allow to achieve an analogue but a binary behaviour of the OxRAM devices. While a

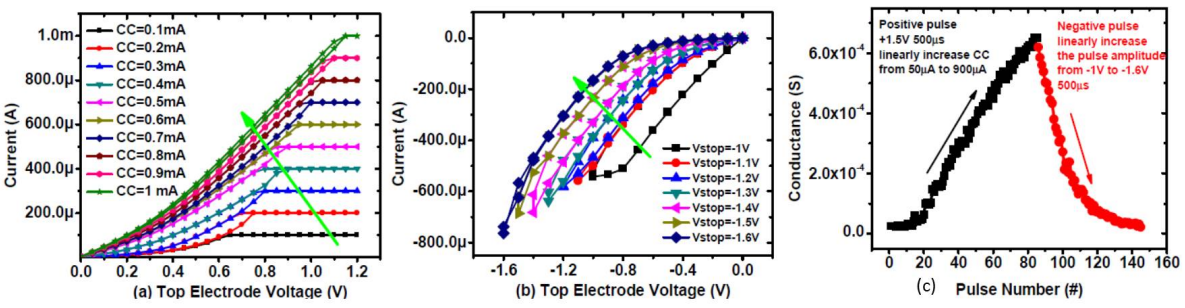


Figure 1.23: Gradual programming of OxRAM cell by increasing the (a) current compliance (CC) and (b) reset voltage (Vstop). (c) Gradual potentiation and depression achieved by tuning the programing conditions. Source: [16]

binary synaptic weight may be enough for certain applications, other applications require the possibility to use analogue weights. To achieve multi-level synapses based on binary devices, one can use multiple cells organized in a parallel circuit [168]. By doing this, the conductance of a synapse is equal to the sum of all device conductances, i.e. ranges between all devices in HRS or LRS. Hence, n devices allow to achieve $n+1$ levels of conductance (synaptic weight). The probabilistic programming of this compound synapse bears intrinsic similarity to the biological STDP [168] such as self adaptation, independently from the initial synaptic state [169], [170]. This strategy was used to implement OxRAM based synapses in Convolutional Neural Networks (CNN) [171].

It was argued that it is possible to exploit the gradual reset of OxRAM to implement gradual depression [172] or that it may be even possible to gradually change the resistance of a single OxRAM cell by applying identical sub-switching voltage pulses [173]. Several groups have reported the coexistence of short and long term plasticity in single OxRAM devices, i.e. volatile and non-volatile states with the possibility of a short-to-long transition upon repeated programming [174] [175] [176] [177]. Moreover, OxRAM synapses allow to implement spike-rate-dependent plasticity (SRDP) besides STDP [178]. Other designs include $1T1R$ and $2T1R$ synapses with STDP features [179] [180], recurrent OxRAM array based SNN [181] and complementary OxRAM based synapses which reduce the need for selectors and mitigate sneak paths currents [182].

Another strategy to elude the constriction of binary devices without compromising integration density may be the vertical RRAM (VRRAM) integration approach (see section 1.2.3). The number of synaptic levels can be tailored by the number of horizontal electrodes. It was recently shown how VRRAM based synapses may be used for synaptic plasticity [183] and to implement large scale CNNs [184].

1.3.1.3 CBRAM synapses

CBRAM shows the same principal switching behaviour like OxRAM when operated with relatively strong amplitude programming pulses, i.e. binary set and reset. Thus, gradual potentiation and depression can be achieved in single devices by tuning the programming pulses [185]. Otherwise, the multiple cell synapse architecture [168] and/or probabilistic programming can be used to achieve multiple-level synaptic weights [17] [186]. Therefore, the principle of tuning the set and reset probabilities by tuning the pulse voltages is illustrated in figure 1.24. On the other hand, the application of short programming pulses on single CBRAM devices allows to control the CF structure resulting in a short-term to long-term conductance modulation [187]. The authors claimed that their concept may be used to implement artificial synapses based on single CBRAM cells.

Sub-μA programming currents can be used to achieve gradual potentiation and depression in so-called programmable metallization cells (PMC) [188]. Otherwise, carefully tuning the programming pulses allows to achieve gradual LTP and LTD and thus STDP by means of a

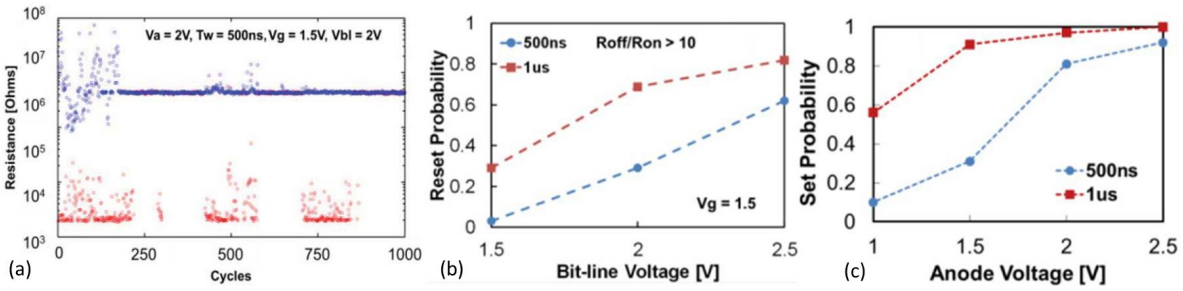


Figure 1.24: (a) Probabilistic programming is shown for a CBRAM cell. For a number of cycles, the Set pulses fail to switch the device into LRS. This phenomenon can be used to extract (b) Reset and (c) Set probabilities. Source: [17]

PMC [189] [190] [191] [192]. A PMC with several kinds of plasticity (STP, STDP, SRDP) was also proposed [193].

1.3.1.4 MRAM synapses

Switching a STT-MRAM MTJ from anti-parallel (AP) to parallel (P) configuration or vice versa, requires a certain time of the switching pulse to be applied. As a rule of thumb, the time is reduced as the applied switching current is higher [194], as shown in figure 1.25 (a). Since the exact switching time varies from device to device and cycle to cycle due to the fundamental physical STT mechanism, a switching variability can be derived and tuned by modulating the pulse voltage and current, shown in figure 1.25 (b) [18]. According to these results, binary synapses were implemented for STDP network [195]. Moreover, it was shown that the delay between two programming pulses has an inverse impact on the switching variability and can be used to balance synaptic modifications between STF and STDP, as demonstrated in figure 1.25 (c) [19]. A MTJ based synapse featuring a relatively large resistance range of 10% – 100% was shown in [196].

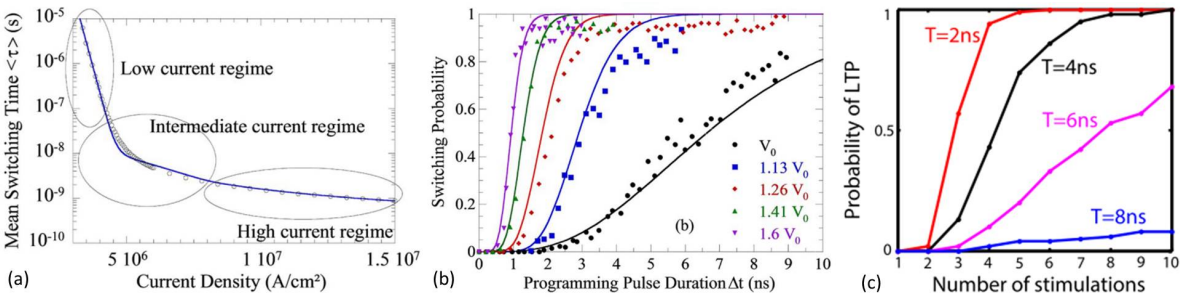


Figure 1.25: (a) Switching time as a function of the applied current density in a MRAM cell. Switching probability as a function of (b) applied programming time and (c) delay between programming pulses. Source: [18] [19]

1.3.1.5 FeRAM synapses

Ferroelectric PZT integrated in thin film capacitors (1C) was used in [20] to implement analogue synapses. Here, an analogue scheme was possible due to the existence of ferroelectric domains which can have differing polarizations compared to each other. An analogue weight characteristic can be achieved by tuning the amplitude of the applied sine wave used to switch the synaptic FRAM device, as shown in figure 1.26 (a). Panasonic has proposed an individual synapse based on a $ZnO(60nm)/Pb(Zr,Ti)O_3(250nm)/SrRuO_3(10nm)/Pt(30nm)$ stack where the PZT polarization is controlled by the gate voltage in order to implement analogue synaptic weights, shown in figure 1.26 (b) [21]. The authors of [22] have used ferroelectric tunnel junctions (FTJs) based on supertetragonal $BiFeO_3$ (BFO) tunnel barriers combined with $(Ca,Ce)MnO_3$ (CCMO) bottom and Co top electrodes. As they demonstrated, the ferroelectric domains can be polarized individually one-by-one, thus inducing a gradual resistance shift, depending on the number of pulses and applied pulse voltage as shown in figure 1.26 (c).

1.3.2 Neuromorphic concepts based on NVM synapses

Neuromorphic systems can be classified according to the applied strategy for synaptic weight specialization into the two fundamentally different approaches :

- **Supervised artificial neural networks:** This kind of ANN's are trained by means of large data sets and error back-propagation in order to master a specific classification task [197] [198]. Supervised ANN's typically operate on a synchronous time scale and use floating point numbers instead of spikes for the communication between neurons. The supervised approach is commonly chosen for deep neural networks (DNN) [199], convolutional neural networks (CNN), deep belief networks (DBN) and multilayer perceptrons.
- **Unsupervised artificial neural networks:** These ANN's are not trained per se but rather train themselves by following certain learning rules, e.g. STDP, which are inspired

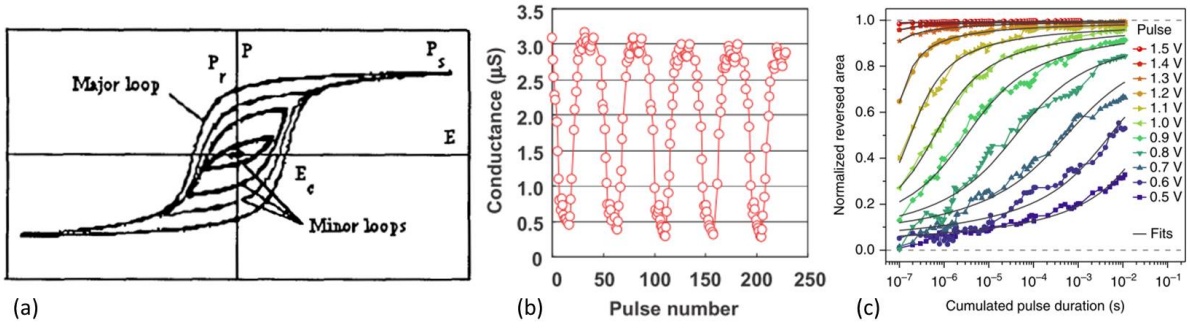


Figure 1.26: (a) Hysteresis loop of ferroelectric polarization indicating gradual polarization change. (b) Gradual conductance change observed in a FeFET. (c) Ratio of switched polarization area as function of the cumulated pulse time showing gradual changes. Source: [20] [21] [22]

by the mechanism of the human brain. This allows them to perform adaptation to a given task independently from a supervisor. Moreover, it was repeatedly argued that STDP is a key function for the detection of repeating patterns in the input data [200] [201] [202] [203] allowing to extract essential information from rather chaotic data. In particular, this feature was emphasized for using an adress event representation (AER) retina [204] [205] [206] [207] and an AER ear [208]. Unsupervised ANN's are also known as spiking neural networks (SNN) since their neurons communicate via instant action potentials (spikes).

It should be noted that software implemented CNN's are currently used for very powerful classification tasks such as face or object recognition and achieve state-of-the art performances [199]. On the other hand, SNN's are still more commonly used by the research community due to a lack of understanding how to effectively train large scale SNN.

1.3.2.1 Spiking Neural Networks based on emerging NVM

Several STDP featuring SNN's were demonstrated for two-dimensional visual pattern recognition based on OxRAM [161] [209] [173] [179] [180], CBRAM [17] and PCM [15] [157]. Furthermore, SNN's featuring STDP are able to process auditory signals [186] [164]. The gradual, self-limiting behaviour of OxRAM synapses was pointed out by means of a WTA network [210]. A perceptron with spike based plasticity in OxRAM memristors was demonstrated by [211]. Recently, binary synapses based on single STT-MTJ were used in a spiking neural network with dedicated pulse schemes adapted to the feature STDP achieving recognition rates up to 97% in a AER sensor visual pattern recognition application [195]. Panasonic has integrated a $9x9$ Hopfield neural network based on three-terminal ferroelectric memory ($3T-FeMEM$) cells [21] in a hardware chip. The network features an unsupervised learning rule inspired by biological STDP which allows them to achieve successful unsupervised learning and recognition of visual patterns. Crossbar arrays featuring analogue synapses based on single ferroelectric tunnel junctions were used for the implementation of a STDP network and it was demonstrated that it allows to learn different patters of two-dimensional visual inputs achieving a recognition rate of almost 100%, even in presence of noise [22].

It is interesting to note that SNN's were demonstrated to be extremely robust towards corrupted data [155] and that their performance is only slightly reduced in presence of various kinds of synaptic variability [212] [213]. Moreover, it was shown that homeostasis of neurons (i.e. threshold adaptation) can mitigate degradation of the SNN performance due to threshold variability.

1.3.2.2 Formal Neural Networks based on emerging NVM

This strategy was shown to be successful for the implementation of synapses in Convolutional Neural Networks (CNN) where synapses made of $n > 14$ allow to achieve very high recognition

rates of above 98% [214] [171]. Furthermore, it was demonstrated that even if OxRAM based synapses are affected by resistance variability, CNNs exhibit good robustness and still obtain high recognition rates [215]. Note that this application may be a major driver for future OxRAM technology application perspectives. The energy improvement of CNNs implemented with RRAM was investigated in [216]. A $1R$ RRAM crossbar was trained by the Manhattan update rule which is a variant of the common supervised training approach and successfully applied to a 3x3 pattern classification [166]. VRRAM based synapses were prospected for synaptic plasticity [183] and to implement large scale CNNs [184]. Moreover it was demonstrated that the variability of VRRAM exhibits a short ranged dependence of the resistance as a function of the recent device history in the order of a few tens of cycles which reduces effectively the overall resistance variability. This can be used to reduce the number of VRRAM cells per synapse without degrading the performance as demonstrated in [217]. Note that this is only possible for back-propagation training approaches where equal or similar cycling numbers of individual devices appear.

1.4 Applications of Neuromorphic Networks

First, it is important to understand that layers of neurons in a neuromorphic networks can be connected in fully-connected or convolutional scheme, as illustrated in figure 1.27. In a fully-connected architecture, each neuron of the layer m is connected to every neuron of layer $m+1$ by a synapse whereas in a convolutional architecture, a neuron of layer $m+1$ receives only synapses from a receptive field, i.e. a small number of close neurons in layer m. If all layers are fully-connected, one typically calls the network a fully-connected neural network (FCNN) whereas a convolutional neural network (CNN) comprises at least two layers of neurons connected in a convolutional way.

CNN's are usually used to extract rather abstract features by means of the receptive kernels of the single neurons. Those features may even be invariant from the given object type, i.e. a cat picture may be constructed by the same basic features that are needed to construct a bicycle picture [218]. Note that this property of feature extraction bears great potential for generalised classification approaches.

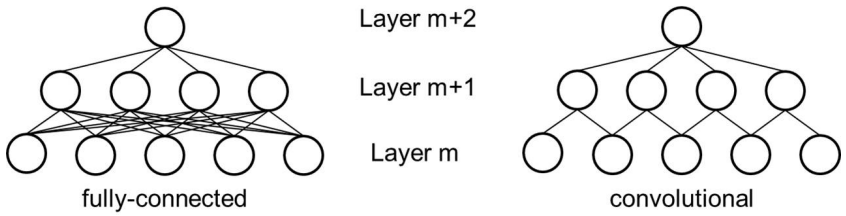


Figure 1.27: Schematic illustration of fully-connected neural network (FCNN) and convolutional neural network (CNN).

The applications that can be realized with neuromorphic systems are manifold and essentially all rely on pattern recognition/detection. They cover:

- Visual pattern analysis: character recognition, object detection, pattern completion
- Auditory pattern analysis: speech synthesizing, sound detection
- Data compression and reconstruction: Images, Video, Sound
- Healthcare: diagnostic applications [219], brain-machine-interfaces based on LFPs [220], brain-machine-interfaces based on action potentials [221]
- Optimization: Travelling salesman problem, Hamming distance
- Arithmetic computing
- Prediction: weather, stock market

CHAPTER 2

GOAL OF THIS WORK

A major challenge for the rehabilitation of the steadily growing number of patients affected by neurodegenerative diseases and injuries is the development of a practical spike sorting system, i.e. a simple system able to separate single unit activities from a multiplexed electrical recording of several units (neurons). To this aim, a number of critical tasks needed to be solved:

- Discrimination between noise and spikes
- Recognition of non-stationary spikes (bursts, dislocation of electrodes over time due to tissue relaxation, etc.)
- Ability to distinguish between simultaneous neuronal activity (overlapping spike trains)
- Computational effort to be minimal

Recently, brain-inspired computing imitations by means of neuromorphic network architectures have demonstrated to be superior candidates for the detection and prediction of patterns occuring in complex data with respect to conventional von-Neumann architectures [222], [223], [166]. For this reason, the key goal of this Ph.D. thesis is the exploration of neuromorphic systems targeting to perform real-time spike sorting with nanowatt-level power consumption and reasonable spike sorting performances. Therefore, the approach for the development of an innovative spike sorting system is to use an artificial neural network that is learns in an unsupervised manner from the data is put through its topology. Unsupervised learning ANN's require a learning rule which is typically inspired by nature such as spike-timing-dependent plasticity (STDP) applied for tuning the synaptic weights of the spiking neural network (SNN). A SNN is specifically designed and optimized to perform spike sorting in real-time. In order to do this, a real-time processing step of the neural signal is necessary before the processed signal

can be used by the SNN. The spike sorting application requires a quick adaptation to altering signals, i.e. the SNN needs to be able to learn fast and tune it's weight rapidly. Emerging resistive RAM (RRAM) memories are therefore a good candidate for the implementation of synapses becuase they offer the possibility to build complex brain-like (cognitive) computing hardware systems that operate with low latencies, are compact and consume low power. While several concepts for synaptic implementations based on RRAM have been proposed [224], [16], oxide based RRAM (OxRAM) technology is among the most promising candidates thanks to its low (sub-μA) operation currents [104], highly scalable lateral dimensions [110], low cost production and back-end-of-line (BEOL) process compatibility. While OxRAM in typical NVM applications is operated using switching currents higher than 50 μA for reliability reasons, we have analyzed the OxRAM device behavior in this paper for switching currents as low as 1 μA in order to understand the ultra-low energy operation. Switching and conduction properties of these new emerging technologies will be investigated in the perspective of implementation into potential artificial synapses for neuromorphic systems. It is well known that RRAM technology is subject to some issues related to reproducibility in the operation of single devices. The effect of those characteristic device features on SNN should be therefore thoroughly investigated. Furthermore, it is necessary that the SNN exhibits a high robustness to noise in the neural signals due to the nature of these recordings. While sometimes proposed in the literature, that ANN are inherently noise tolerant, this property should be studied and eventually improved by certain ANN functionalities.

In the following chapters, the results obtained in the framework of this dissertation will be demonstrated. Chapter 3 describes the electrical characterization of RRAM (OxRAM and CBRAM) device technologies and the implementation of synapses with RRAM devices. Chapter 4 explains the design of the Spiking Neural Network which was developed to perform real-time spike sorting based on RRAM synapses as well the real-time signal pre-processing. Chapter 5 studies the impact of RRAM related variability on the system level performances of spiking neural networks. Chapter 6 proposed a novel plasticity implementation based on RRAM synapses in order to mimic biological short term plasticity (STP). Finally, chapter 7 concludes the major thesis results and gives a small perspective to potential future challenges.

CHAPTER 3

SYNAPSE BASED ON RRAM

A critical challenge for the implementation of artificial neural networks (ANN) into a hardware electronic chip remains in the design of electronic synapses which mimic their biological counterparts appropriately. This chapter deals with the concept of designing artificial synapses based on resistive memory technologies (RRAM) such as OxRAM and CBRAM (see chapter 1). RRAM devices feature a low latency ($< 1\mu s$), high integration density ($< 1\mu m^2$) as well as a low energy consumption ($< 75pJ$). An original methodology to use Oxide based RRAM (OxRAM) as easy to program and low energy synapses is demonstrated.

First, the most critical requirements for the implementation of bio-inspired hardware synapses are reviewed in section 3.1. Then, section 3.2 presents a thorough investigation of RRAM, more precisely Oxide based RRAM (OxRAM) and Conductive Bridge RAM (CBRAM). Section 3.3 describes the utilization of OxRAM and CBRAM in an artificial synapse design overcoming the technology specific bottlenecks. Finally, section 3.4 concludes the findings from the electrical analysis and synapse concept.

3.1 Requirements to mimic biological synapses

The physiological structure and functionality of neurons and synapses as well as the biological basis for synaptic plasticity is described in detail in chapter 1. Accordingly, artificial synapses should fulfil a number of conditions:

- Structural
  - Two terminals: In analogy to a biological synapse, hardware implementations of artificial synapses should resemble a basic structure consisting of an input and an

output terminal separated by what is in the following called called synapse emulator (SE).

- Functional
  - Plasticity: The SE has to be capable to mimic basic synaptic features such as plasticity which means that the conductance has to be tunable according to a learning rule.
  - Analog character: The conductances which can be achieved for a synapse should be on a continuous scale, i.e. feature multi-level states.
  - Progressive programming: The tuning of the synaptic conductance shall occur in a progressive (also called cumulative manner). This means that the strength of modification has to be a function of the number of programming events.
  - Non-volatility: Typically, biological plasticity occurs on different time scales whereas the most important plasticity rule responsible for learning and memory are long-lasting, i.e. non-volatile.

3.2 Electrical analysis of Resistive RAM

Resistive RAM (RRAM or ReRAM) technology typically includes Oxide vacancy based RAM (OxRAM) and Conductive Bridge RAM (CBRAM). The resistance modificaition is attributed to the formation of a conductive filament (CF) due to mainly oxygen vacancy migration (OxRAM) or metal ion migration (CBRAM). For more details on the physical mechanisms of these technologies refer to chapter 1. OxRAM and CBRAM devices were co-integrated with n-type metal oxide semiconductor (NMOS) access devices in a standard 65 nm CMOS technology [225] into so-called $1T1R$ structures, consisting of a transistor (T) and a resistor (R). The transistor (T) is used as an access device and to precisely control the current compliance. For OxRAM, the resistive switching layer (R) is sandwiched between 5 nm or 10 nm Ti and 35 nm TiN electrodes, see figure 3.1 (a). The resistive switching layers of the CBRAM devices are sandwiched between optimized metals which are not disclosed for confidentiality reasons (see figure 3.1 (b)). As illustrated for OxRAM, three oxide compositions deposited by Atomic Layer Deposition (ALD) were studied: (i) $5nm\ HfO_2$, (ii) $1nm\ Al_2O_3/3nm\ HfO_2$ and (iii) $5nm\ HfO_2/4nm\ TaO_x$. For CBRAM, an undoped MO_x and a 20% doped MO_x, both 53Å in thickness were used.

The $1T1R$ OxRAM and CBRAM structures were characterized by applying DC voltage sweeps as shown in figure 3.2. As-fabricated devices exhibit a very high initial resistance ($> 10^{10}\Omega$), also known as the pristine resistance state (PRS). This is because the ALD grown oxide is virtually free of defects such as vacancies, interstitials or lattice defects. The first time a positive bias voltage sweep is applied (Forming), oxygen ions (O^{2-}) drift towards the electrode interface and leave oxygen vacancies (V_O^{2+}) behind. Once a critical number of V_O^{2+} generated by a certain electrical field, the resistance of the OxRAM cell changes abruptly up by several orders of magnitude

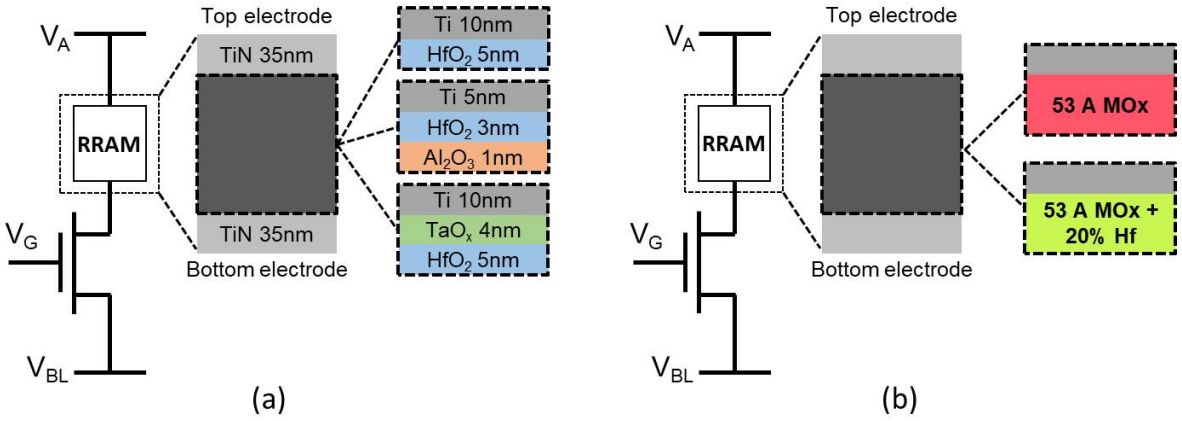


Figure 3.1: Schematic of 1-Transistor-1-Resistor ($1T1R$) co-integration. Overview of device structure and different material compositions analysed for this study for (a) OxRAM and (b) CBRAM.

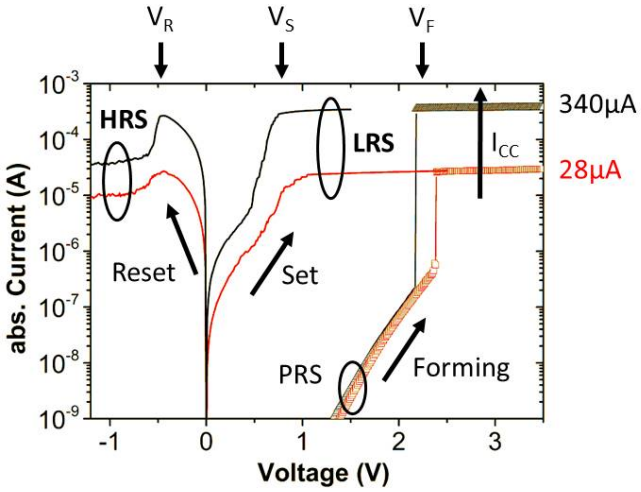


Figure 3.2: Schematic switching of 1-Transistor-1-Resistor ($1T1R$) co-integrated RRAM devices. RRAM IV characteristics (here for Al_2O_3/HfO_2 OxRAM device) for Forming (symbols) and Set/Reset (solid lines, 30 cycles averaged). Operation is shown for different programming currents (PC)

which can be noted by the significantly higher current flow. This change is due to the formation of the conductive filament (CF) which brings the cell into the Low Resistive State (LRS). The voltage at which this shift occurs is termed as the forming voltage (V_F) and depends typically on the oxide thickness, since the forming occurs at a critical field which depends inversely on the oxide thickness as $E = V/d$. Figure 3.2 shows the IV characteristic for two identical OxRAM devices, using different current compliances (I_{CC}) by varying the gate voltage of the access transistor. The I_{CC} can typically be used to control the diameter of the CF (d_{CF}), i.e. the higher the I_{CC} the higher d_{CF}. Hence, I_{CC} has a direct impact on the value of the LRS which will be

explained below. By applying a negative voltage sweep (Reset), O^{2-} drift back from the reservoir at the top electrode into the oxide and occupy a fraction of the V_O^{2+}, thus rupturing part of the conductive filament. A rather abrupt drop of the electrical current can be observed at a voltage which is termed the reset voltage V_R. This is where the CF is broken. This much lower current indicates that the CF was broken and the device was switched to the so-called the High Resistive State (HRS). Beyond V_R, the current still decreases slightly, i.e. the resistance can be further lowered by some extent. Since the V_O^{2+} generated during the forming operation are not fully recovered during a Reset operation, the level of HRS is significantly higher than PRS. Moreover the OxRAM device can be switched from HRS to LRS (Set) by applying a positive voltage while the set voltage (V_S) where the switch occurs is smaller than V_F. While OxRAM in typical NVM applications is operated using switching currents higher than 50 μA for reliability reasons, the electrical behaviour of OxRAM was analysed for switching currents as low as 1 μA. Switching and conduction properties are investigated in the perspective of implementation into potential artificial synapses for neuromorphic systems. Several OxRAM devices were therefore tested both by voltage sweeping (DC) and voltage pulses (AC). All resistance readings of the single OxRAM devices were performed using a bias voltage $V_A = 0.1\ V$ while reading the static current.

The CBRAM operation is similar to the previously described OxRAM operation. The main difference is that additionally to to the oxygen anion migration, a metallic cation migration occurs from the reactive metal electrode and the CF is of metallic nature.

3.2.1 Static IV analysis

Figure 3.3 (a) and (b) show typical IV sweep curves for I_{CC} ranging from 1.5 μA to 340 μA for Forming (first Set operation), Set and Reset operations for OxRAM and 4.5 μA to 200 μA for CBRAM. Note that it was not possible to switch the undoped MO_x devices using less than approx. 50μA whereas the I_{CC} could be reduced to as low as 4.5μA for the Hf-doped MO_x. During Forming or Set operations, a positive bias anode voltage (V_A) is applied to the TE to switch the OxRAM/CBRAM devices from HRS to LRS. During Reset operations, a negative bias voltage is applied to TE switching from LRS to HRS. Although the forming voltages (i.e. voltage of abrupt current increase) are similar for all operation currents, the Set voltage increases when I_{CC} is reduced and the Set process appears to be more gradual. Furthermore, the reset current (I_{Reset}), defined as the maximum current during the reset process, is typically equal or slightly higher than the current compliance during the Set operation. This is true for $I_{CC} > 20\ \mu A$, however, if I_{CC} is reduced below 20 μA, I_{Reset} drops significantly below I_{Set} (figure 3.4). This applies regardless of the oxide material whereas the effect is the strongest for the HfO_2/TaO_x layer which is the oxide layer with the highest overall thickness of 9 nm tested in this work. This suggests that the electric conduction involves mainly tunneling transport phenomena.

As d_{CF} is a function of the I_{CC}, the value of the LRS is inversely proportional to the I_{CC}, i.e. the higher the I_{CC} the lower the LRS, shown in figure 3.5 (a). While the LRS seems to be

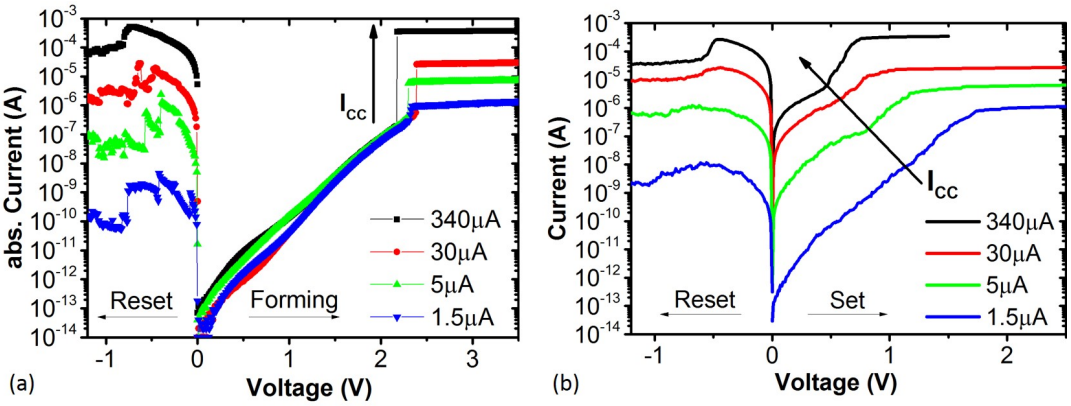


Figure 3.3: IV characteristics (shown for Al_2O_3/HfO_2 OxRAM) for (a) Forming/1^{st}Reset and (b) Set/Reset. Operation is shown for I_{CC} (i.e. current compliance) ranging from 1.5 μA to 340 μA. Note the shift of the Set IV curve towards higher voltages for reduced I_{CC}.

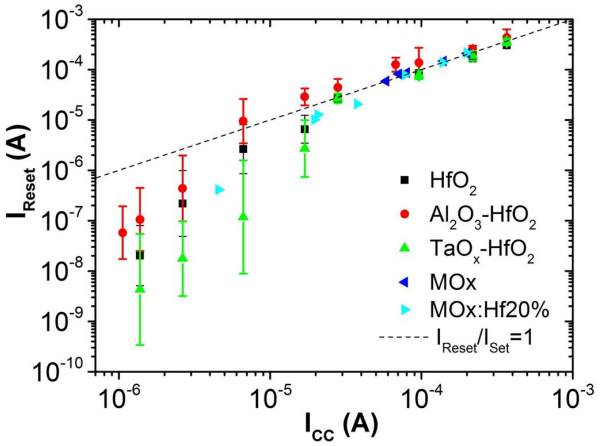


Figure 3.4: Reset current (I_{Reset}) as a function of I_{CC} for OxRAM and CBRAM material compositions.

independent from the oxide material for $I_{CC} > 20\ \mu A$ (in agreement with the literature [226]), the LRS value shows a strong dependence on the oxide material for $I_{CC} < 20\ \mu A$. Interestingly, the LRS increases as a function of the total oxide thickness such that the largest oxide layer (HfO_2/TaO_x) exhibits the highest LRS values, giving rise to a non-filamentary conduction mechanism in the $sub-20\mu A$ regime. Moreover, the LRS of all materials seems to depend strongly on I_{CC} in this low current range. On the other hand, the value of the HRS depends typically on the number and distribution of the V_O^{2+} in the oxide layer(s) which in turn is a function of the I_{CC} and the maximum voltage applied during Reset (V_R). The HRS is shown for various I_{CC} and different OxRAM materials in figure 3.5 (b). A similar trend for HRS depending on I_{CC} and the oxide thickness can be observed. The slight variation of the experimental results of this work and reference data from literature may be attributed to statistical variations or

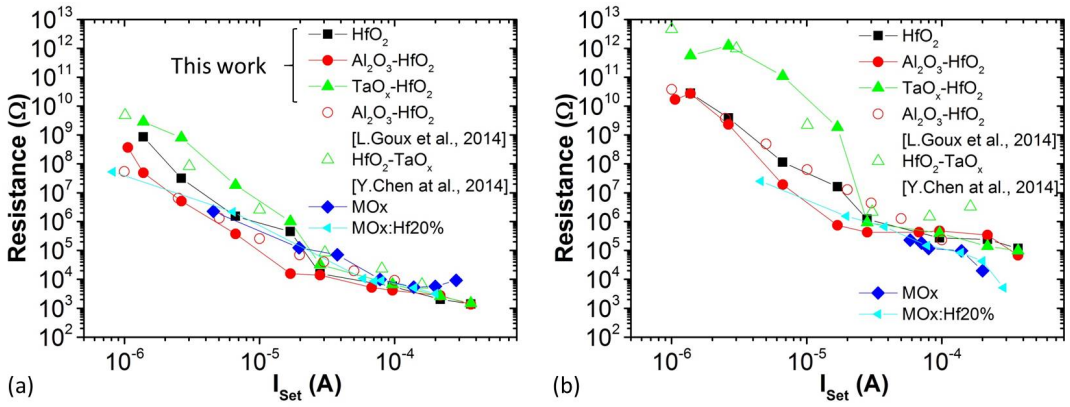


Figure 3.5: (a) LRS and (b) HRS as a function of I_{CC} for different oxide materials.

slightly different operation conditions such as the programming voltages.

Interestingly, the doped MO_x operated with a $I_{CC} = 4.5\mu A$ exhibits a very low cycle-to-cycle variability in HRS and still despite the low programming current still offers a relatively large resistance margin between LRS and HRS as shown in figure 3.6. The high variability of the LRS may be explained by the same theory as presented for the OxRAM technology using ultra-low operation currents that no filament is formed during Set but the (intrinsic) distribution of defects in the oxide material is changed optimizing tunneling paths. The reduced reset current also points towards a non-filamentary LRS which is rather unstable. Furthermore, the low variability of the HRS and its level equal to the pristine resistance give rise to the assumption that no additional defects are introduced into the oxide during application of programming pulses with such low currents.

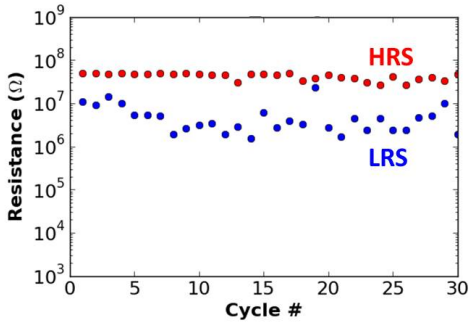


Figure 3.6: (a) IV characteristic of doped MO_x CBRAM operated using $I_{CC} = 4.5\mu A$ and (b) corresponding resistance values for Low and High Resistance States (LRS and HRS) for 30 switching cycles.

3.2.2 Resistance variability

Both the structure of the CF (in LRS) and the gap between the remaining filament and top electrode as well as the distribution of V_O^{2+} (in HRS) fluctuate from one switching cycle to another as well as among different devices. This behaviour is well known as variability whereas it can be differentiated between cycle-to-cycle and device-to-device variability. Figures 3.7 and figure 3.8 show the experimental distributions of Low Resistive States (LRS) and High Resistive States (HRS) varying the I_{CC} during the forming and set operations. Each set of lines (LRS and HRS for same I_{CC}, one color) represents the complete ensemble of tested RRAM cells cycled 30 times. A high I_{CC}, e.g. I_{CC} = 367μA, results in a sharp distribution of the LRS whereas the HRS distribution is rather broad spanning a good order of magnitude. As the I_{CC} is reduced, both LRS and HRS distributions widen while the entire distributions are shifted towards higher values. This means, that the variability is largely dependent on the I_{CC}.

The LRS and HRS distributions of each device (obtained through 30 cycles) were used to investigate the dependence of the variability on the OxRAM resistance. As proposed in [227] and [116], the mean resistance μ_R (median) and the resistance variability σ_R as the resistance range between 30% and 70% were extracted therefore, see figure 3.9 (a). Figure 3.9 represents the resistance variability σ_R of all tested oxide materials as a function of the mean resistance μ_R. As we previously stated in [171], the LRS and HRS variabilities form a continuous curve and are thus presented together for each material. As one can see, σ_R increases with μ_R, i.e. when I_{CC} is reduced. Indeed, the variability depends strongly on the resistance level but is identical for different oxide materials. The dependence of σ_R on μ_R is slightly reduced for $\mu_R > 10^6$ Ω.

3.2.3 Resistance margin

As described in section 3.2.1, both LRS and HRS depend inversely on the current compliance I_{CC}. Moreover, also the variability increases when I_{CC} is reduced which prevents to predict the exact resistance value upon programming. Relatively high I_{CC} of approximately 100μA and fixed programming voltages for Set and Reset result in changing the resistance of a OxRAM device between two rather distinct distributions, LRS and HRS, separated by a gap which is also known as memory window (MW) or resistance margin. However, as I_{CC} is reduced, the increased variability results in a shrinking MW which eventually disappears if I_{CC} is below a certain value, depicted in figure 3.10. The overlapping distributions of LRS and HRS are a critical problem for conventional memories where a clear separation between LRS and HRS is necessary in order to determine the memory state reliably in a read operation. The experimental distributions for LRS and HRS of several devices are used to extract the memory window (MW) at different confidence intervals, 0σ (median to median), ±1σ, ±2σ and ±3σ, as illustrated in figure 3.11 (a). Furthermore, the experimental data can be used to extract a dynamic range (DR) of a single RRAM device, depicted in figure 3.11 (b). The DR is somewhat the opposite of the MW and may be of more relevance for synapses in neural networks. The choice of the confidence interval (σ-range)

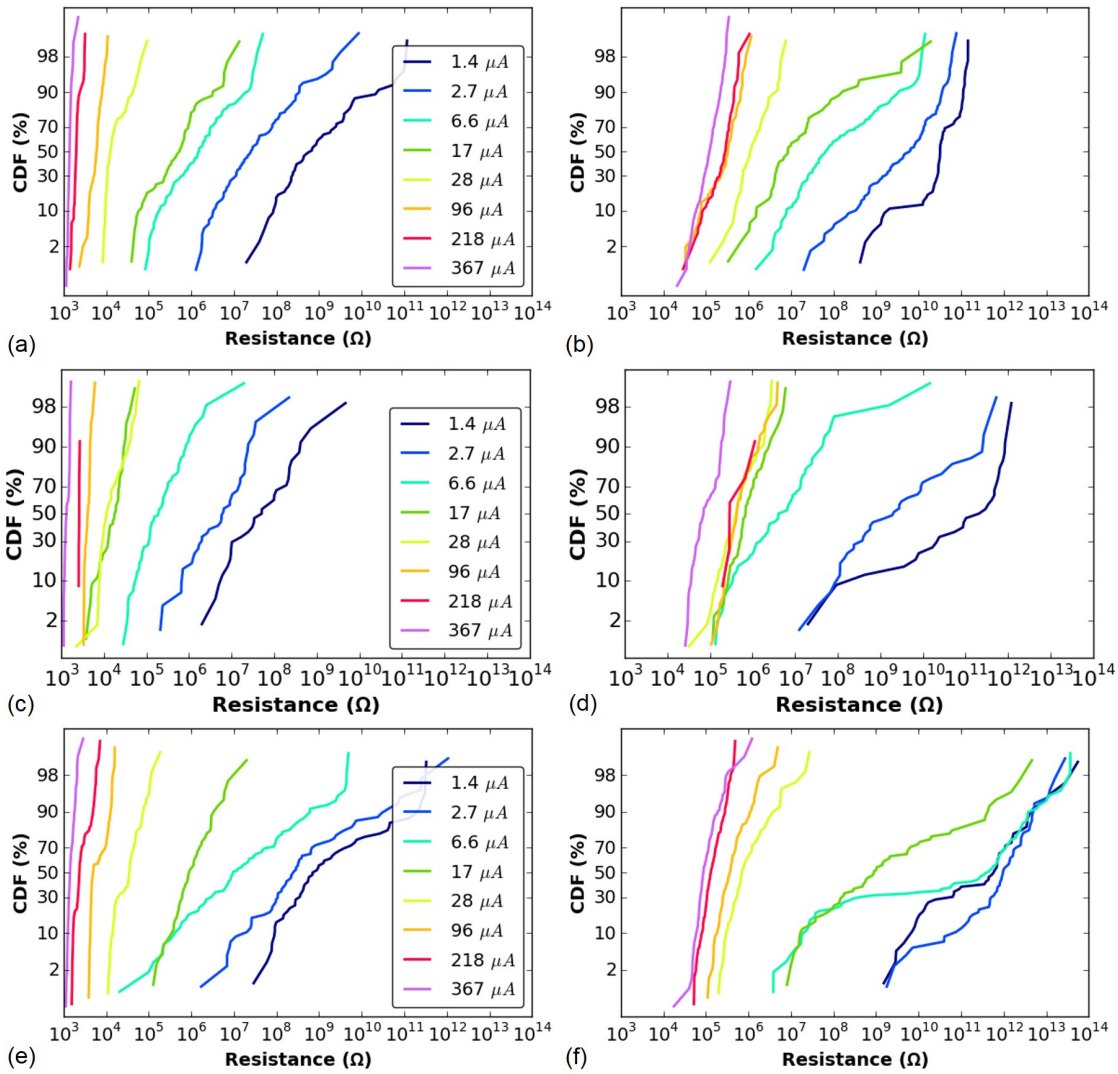


Figure 3.7: Cumulative distribution functions (CDF) of Low Resistive State (LRS) and High Resistive State (HRS) as function of the current compliance (I_{CC}) for different OxRAM materials: (a) LRS and (b) HRS for $5nm\ HfO_2$ (c) LRS and (d) HRS for $1nm\ Al_2O_3/3nm\ HfO_2$ (e) LRS and (f) HRS for $5nm\ HfO_2/4nm\ TaO_x$. Note the shift and widening of the CDF in both LRS and HRS for reduced I_{CC}.

depends on the number of of memory devices concerned in an array or more general in a memory based system. Hence, the more devices a system contains, the larger confidence interval has to be taken into account. Figures 3.12 and 3.13 show the extracted MW and DR as a function of the programming current (I_{CC}) for the tested OxRAM materials and CBRAM materials, respectively. While the memory window is critically reduced for small I_{CC}, the dynamic range of OxRAM devices is increasing by several orders of magnitude due to the large resistance variability. On the other hand, both the memory window and the dynamic range are reduced for small I_{CC} for

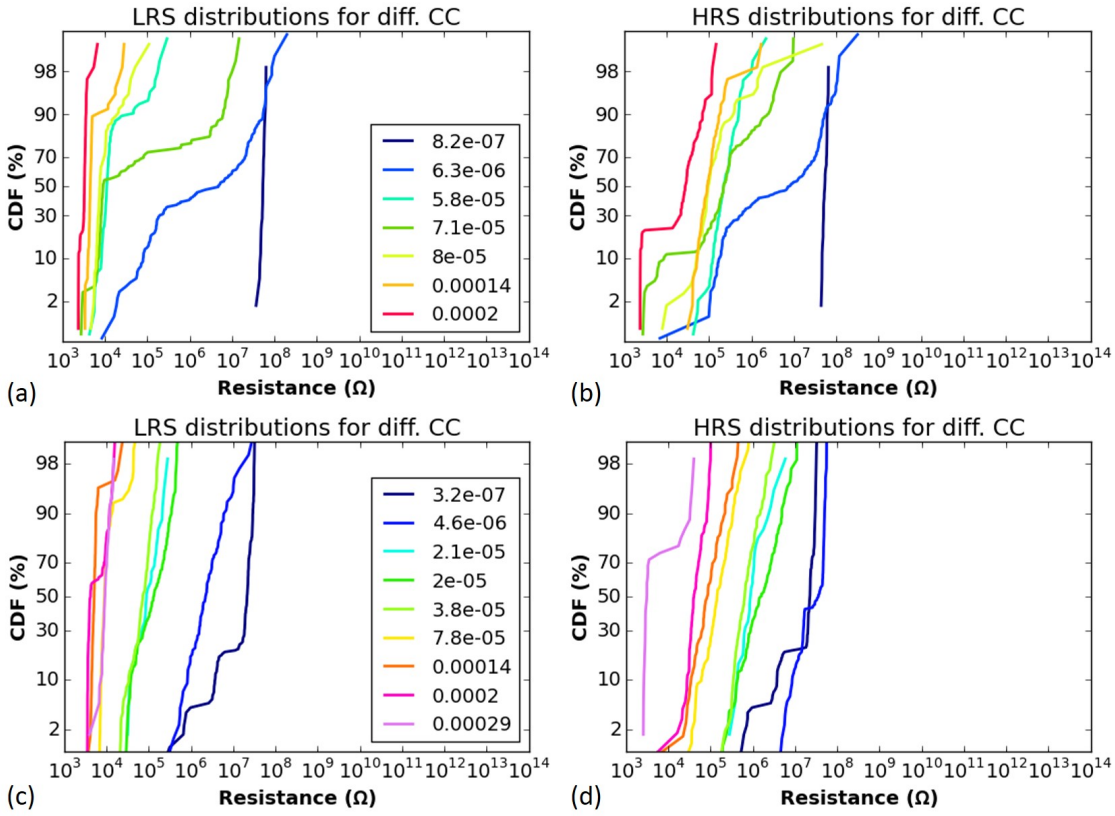


Figure 3.8: Cumulative distribution functions (CDF) of Low Resistive State (LRS) and High Resistive State (HRS) as function of the current compliance (I_{CC}) for different CBRAM materials: (a) LRS and (b) HRS for undoped MO_x (c) LRS and (d) HRS for $20\%Hf - MO_x$. Note the shift and widening of the CDF in both LRS and HRS for reduced I_{CC}.

the undoped MO_x CBRAM device while they are rather constant for the $20\%Hf - MO_x$ CBRAM devices.

3.2.4 Switching variability

The voltage at which an OxRAM cell switches from HRS to LRS during the Set process is affected by the variability and thus fluctuating from cycle to cycle as on can see in figure 3.2 (b). Since the chance that a certain LRS is reached, is affected by variability, a switching probability P_{Set} can be derived which is the cumulative density function of switched devices as a function of the applied voltage during the set process (V_S). As shown in figure 3.14, the higher the V_S the higher P_{Set}. The set switching probability (P_{Set}) can therefore be adjusted between 0 and 1 by tuning V_{Set} while those values are a function of the I_{CC} used during programming. Similarly to the resistance variability that increases upon reduction of I_{CC}, also the switching variability, i.e. the range of V_{Set} to achieve $0 < P_{Set} < 1$, increases. Note that P_{Set} for the very low I_{CC}, indicated by the shaded area, corresponds to both the probability to trigger a switch as well as its magnitude

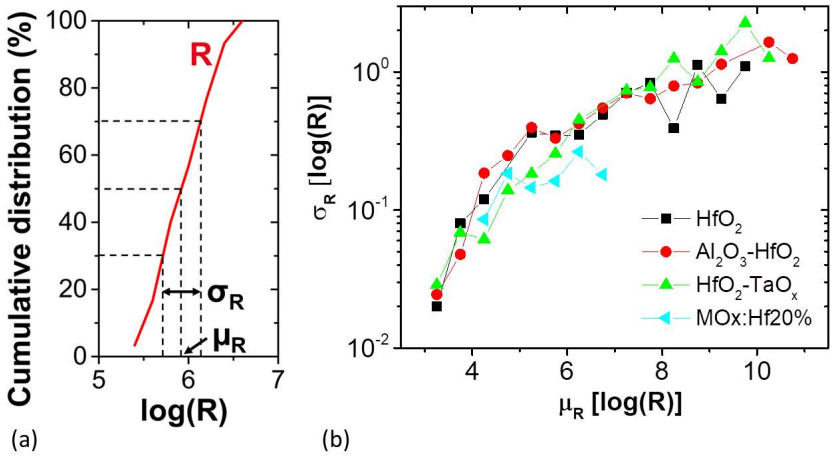


Figure 3.9: (a) Variability (σ_R) as a function of programmed mean resistance (μ_R). (b) μ_R and σ_R extraction methodology from experimental resistance distribution of 30 cycles for one device.

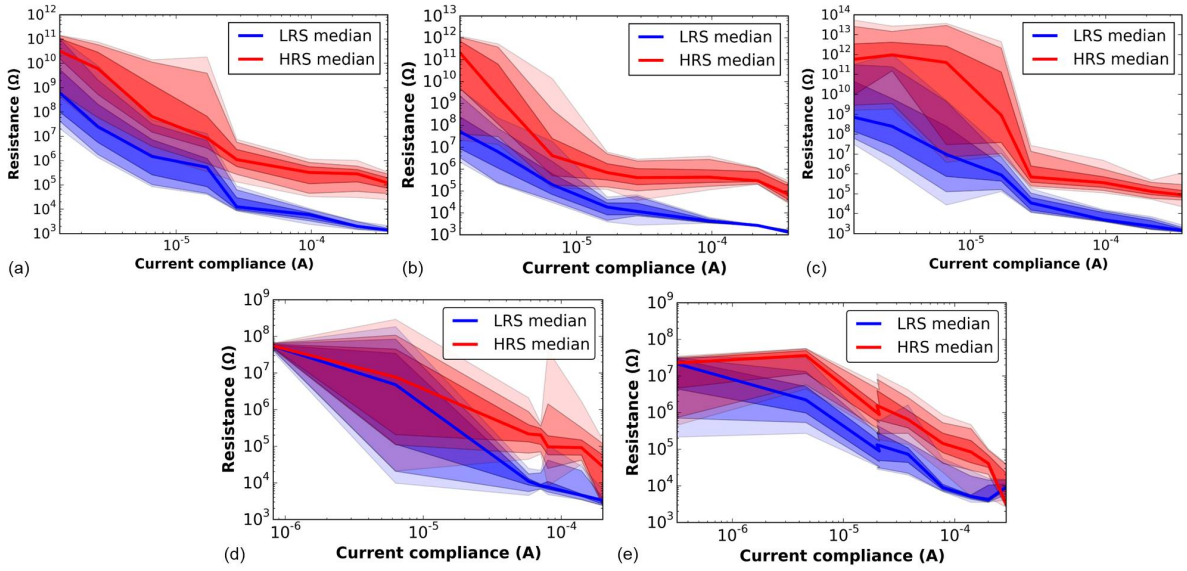


Figure 3.10: LRS and HRS as a function of the current compliance (I_{CC}) for (a) $5nm\ HfO_2$, (b) $1nm\ Al_2O_3/3nm\ HfO_2$ and (c) $5nm\ HfO_2/4nm\ TaO_x$. The bold lines show the geometrical mean values of LRS and HRS, the shaded areas represent different confidence intervals of the experimental sample, i.e. 1σ, 2σ and 3σ.

of resistance change, i.e. the higher V_{Set} the higher P_{Set} and ΔG.

3.2.5 Endurance

The lifetime of OxRAM devices was characterized by extended cycling, i.e. repeatedly switching between Set and Reset using short programming pulses. Different programming currents (I_{CC}) and reset voltage (V_{Reset}) were used to cycle several OxRAM devices 10^8 times, thus accounting

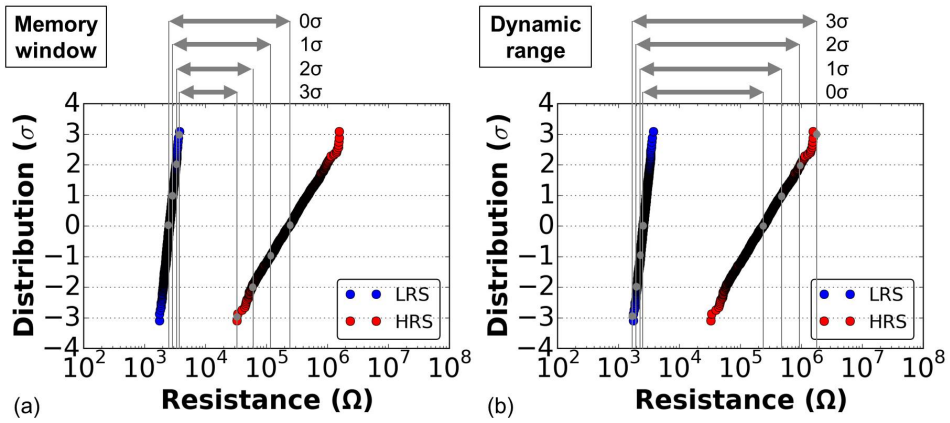


Figure 3.11: The extraction of (a) the memory window (MW) and (b) the dynamic range for different confidence intervals (σ) is schematically illustrated.

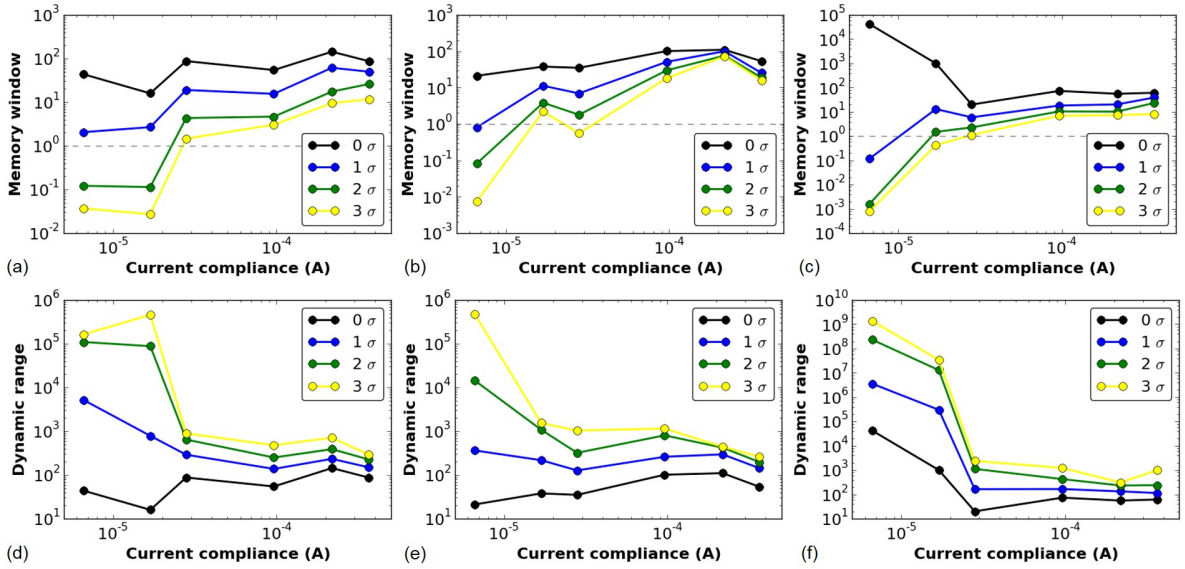


Figure 3.12: Memory window as function of I_{CC} for (a) $5nm\ HfO_2$, (b) $1nm\ Al_2O_3/3nm\ HfO_2$ and (c) $5nm\ HfO_2/4nm\ TaO_x$. The different lines correspond to different confidence intervals of the experimental sample, i.e. 0σ, 1σ, 2σ and 3σ. The dashed lines represent a MW of 1, i.e. LRS and HRS distributions blend into each other. Dynamic range as function of I_{CC} for (a) $5nm\ HfO_2$, (b) $1nm\ Al_2O_3/3nm\ HfO_2$ and (c) $5nm\ HfO_2/4nm\ TaO_x$. The different lines correspond to different confidence intervals of the experimental sample, i.e. 0σ, 1σ, 2σ and 3σ.

for both device-to-device and cycle-to-cycle variabilities. The set pulse voltage $V_{Set} = 2.5V$ was not varied as well as the pulse duration $t_{Set,Reset} = 1\mu s$. The electrical tests are shown in figure 3.15 where each device is represented by grey lines and the mean values for LRS and HRS are shown in blue and red, respectively. When $V_{Reset} = -1.2V$, functionality for 10^8 cycles could be achieved for all tested OxRAM materials and I_{CC} while device failure occurs around 10^6

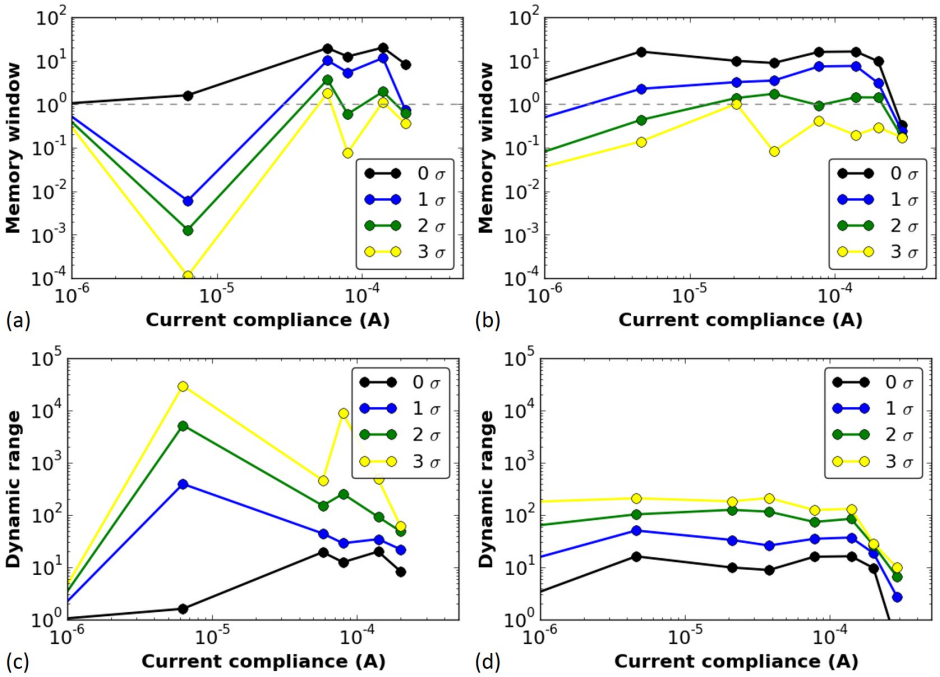


Figure 3.13: Memory window as function of I_{CC} for (a) undoped MO_x and (b) $20\%Hf - MO_x$. The different lines correspond to different confidence intervals of the experimental sample, i.e. 0σ, 1σ, 2σ and 3σ. The dashed lines represent a MW of 1, i.e. LRS and HRS distributions blend into each other. Dynamic range as function of I_{CC} for (a) undoped MO_x and (b) $20\%Hf : MO_x$. The different lines correspond to different confidence intervals of the experimental sample, i.e. 0σ, 1σ, 2σ and 3σ.

cycles for $V_{Reset} = -1.5V$. As expected, both LRS and HRS are increasing as I_{CC} is reduced from $135\mu A$ to $30\mu A$. Furthermore, the variability increases strongly, especially for the LRS. Upon high cycle numbers, some drift towards higher resistance can be observed both in LRS and HRS. In the case of $V_{Reset} = -1.5V$, this drift affected cycling period (starting from some 10^3 cycles) is followed by the device breakdown. For this reason, the drift effect might be an indicator for device degradation which could be used to predict device breakdown as well as to recondition the device when its drift is detected.

The experimental distributions for LRS and HRS of several devices are used to extract the memory window (MW) at different confidence intervals, 0σ (median to median), 1σ, 2σ and 3σ, and different instants of the device cycle number, illustrated in figure 3.11. As shown in figure 3.19 (c) and (f), using a $I_{CC} = 135\ \mu A$ results in two separated distributions for LRS and HRS enabling a clear memory window (MW). The MW shrinks for $I_{CC} = 85\ \mu A$ whereas using a slightly stronger reset condition of $V_{Reset} = -1.5V$ (see figure 3.19 (e)) can enhance the MW with respect to $V_{Reset} = -1.5V$ (see figure 3.19 (b)). When reducing the I_{CC} to 30 μA, both LRS and HRS distributions expand strongly and overlap each other, hence the MW vanishes (as shown in figure 3.15 (a) and (d)). Even if the strong reset condition is used, the separation between LRS

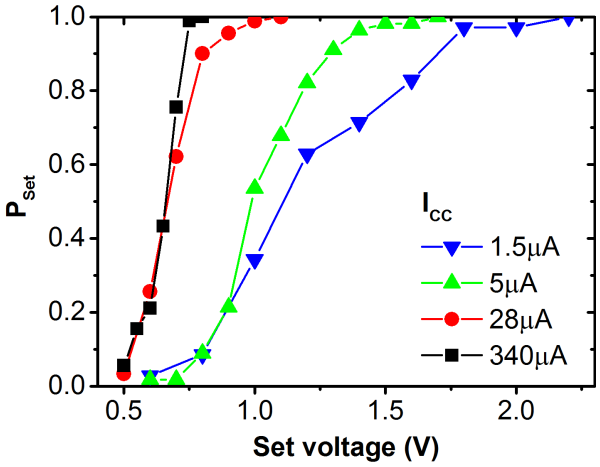


Figure 3.14: Probability to perform a Set operation (P_{Set}) as a function of the applied Set voltage (V_{Set}).

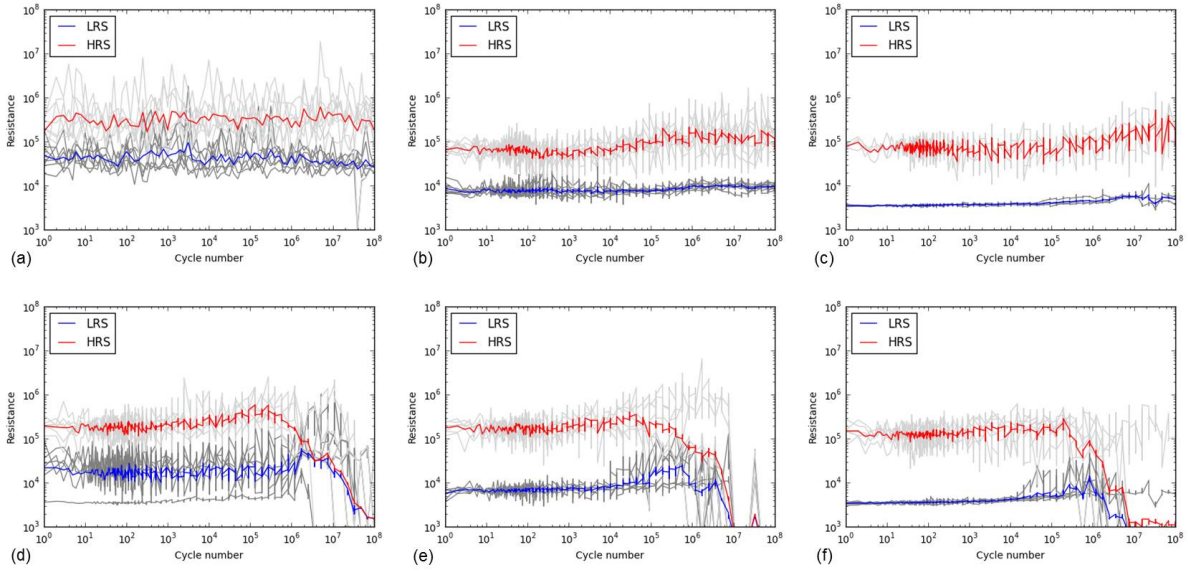


Figure 3.15: HfO_2 endurance test using pulsed programming with $V_S = 2.5V$ and $t_{Set,Reset} = 1\mu s$ and a variation of current compliance I_{CC} and reset voltage V_R: (a) $I_{CC} = 30\mu A$, $V_R = -1.2V$, (b) $I_{CC} = 85\mu A$, $V_R = -1.2V$, (c) $I_{CC} = 135\mu A$, $V_R = -1.2V$, (d) $I_{CC} = 30\mu A$, $V_R = -1.5V$, (e) $I_{CC} = 85\mu A$, $V_R = -1.5V$, (f) $I_{CC} = 135\mu A$, $V_R = -1.5V$. The single devices LRS and HRS are represented in grey lines while the mean LRS and HRS are shown in blue and red.

and HRS distribution is no longer sufficient.

Figure 3.16 shows the extracted MW for different σ ranges and a variation of programming conditions introduced in figure 3.15. Using a relatively low $I_{CC} = 30\mu A$ and $V_{Reset} = -1.2V$ does not exhibit a MW for $MW > 2\sigma$. The MW is significantly increased by increasing the I_{CC} and/or V_R. However, it seems preferable to increase I_{CC} rather than V_{Reset} in order to prevent the

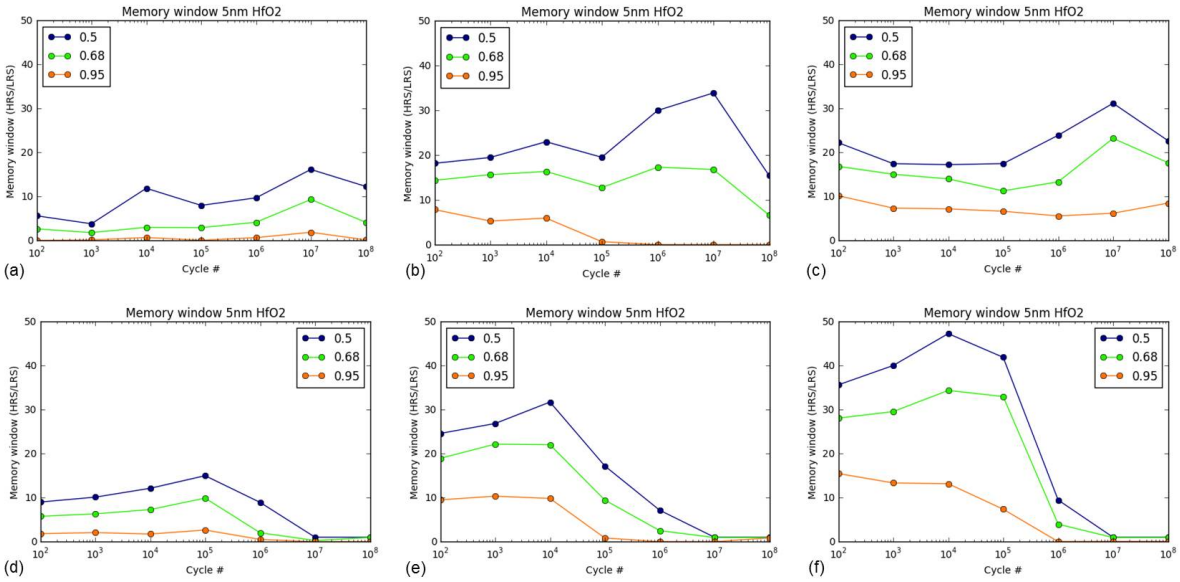


Figure 3.16: Memory window (MW) for different distribution intervals (σ) of HfO_2 endurance test using pulsed programming with $V_S = 2.5V$ and $t_{Set,Reset} = 1\mu s$ and a variation of current compliance I_{CC} and reset voltage V_R: (a) $I_{CC} = 30\mu A$, $V_R = -1.2V$, (b) $I_{CC} = 85\mu A$, $V_R = -1.2V$, (c) $I_{CC} = 135\mu A$, $V_R = -1.2V$, (d) $I_{CC} = 30\mu A$, $V_R = -1.5V$, (e) $I_{CC} = 85\mu A$, $V_R = -1.5V$, (f) $I_{CC} = 135\mu A$, $V_R = -1.5V$.

enhanced device degradation which is likely to be due to the higher electric field stress.

A number of devices was cycled beyond 10^8 programming cycles in order to investigate the device lifetime as function of the reset conditions. The device failure rate is extracted as the number of devices that break down at different cycle numbers during the endurance test and is shown in figure 3.17. Accordingly, it is clear that the device lifetime is linked to V_R, i.e. a higher

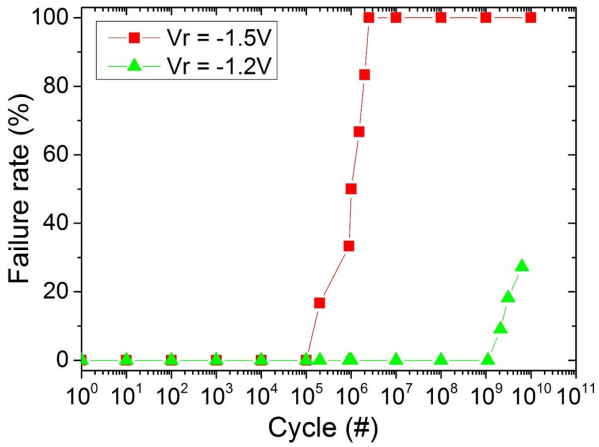


Figure 3.17: Endurance failure rate of RRAM as a function of the reset voltage V_R. Early HRS failure rate is induced by high V_R.

V_R leads to much higher failure rates with respect to lower V_R. Note that the breakdown occurred exclusively in LRS.

Figure 3.18 demonstrates for one OxRAM device that it is possible to achieve more than 10^9 switching cycles, simply by using optimized programming conditions, here $V_R = -1.2V$. Note that the impact of the set voltage was not studied here.

3.2.6 Filamentary vs. non-filamentary switching

In the previous sections it was explained that the OxRAM operation with very low current, i.e. $I_{CC} < 20\ \mu A$, results in a significantly smaller I_{Reset} with respect to I_{Set} (figure 3.4), a strong dependence of LRS and HRS values on I_{CC} (figure 3.10) and a similar variability for LRS and HRS (figure 3.9). These findings may be explained by bulk switching and conduction mechanisms rather than filamentary ones [105],[115] when very low I_{CC} ($< 20\ \mu A$) are used. This means, that during Set switching from HRS to LRS no filament is created but the spatial distribution of defects (mainly oxygen vacancies) is changed altering effectively the tunneling or hopping distances for charge carriers which results in a changed resistance. We believe that in this case the current conduction in the LRS is dominated by trap-assisted tunneling as it is the case for the HRS [7]. This assumption is supported by experimental results from pulsed cycling of the OxRAM devices in both current regimes shown in figure 3.19. Whereas $I_{CC} = 30\ \mu A$ is still sufficient to achieve a defined switching with a significant resistance margin between LRS and HRS (see figure 3.19 (b)), the LRS and HRS distributions for $I_{CC} = 5\ \mu A$ cover several orders of magnitude and are overlapping (i.e. no resistance window). In the case of $I_{CC} > 30\ \mu A$, the resistance window can be improved by increasing the I_{CC}.

As explained above, the switching and conduction mechanisms seem to change upon reducing

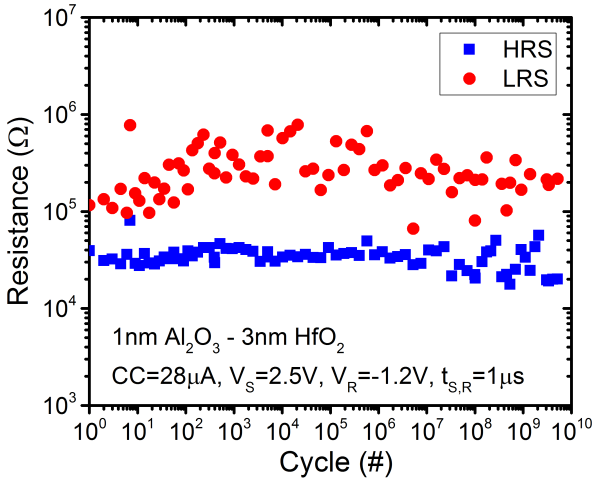


Figure 3.18: Endurance failure rate of RRAM as a function of the reset voltage V_R. Early HRS failure rate is induced by high V_R.

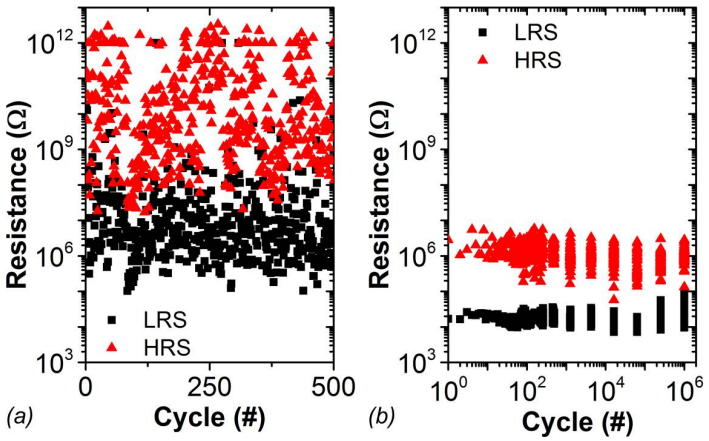


Figure 3.19: TaO_x/HfO_2 endurance for pulsed operation using (a) $I_{CC} = 5\ \mu A$, $V_{Set} = 3\ V$, $V_{Reset} = -1.5\ V$, $t_{Set/Reset} = 10\ \mu s$ (no resistance window) and (b) $I_{CC} = 30\ \mu A$, $V_{Set} = 2.5\ V$, $V_{Reset} = -1.5\ V$, $t_{Set/Reset} = 1\ \mu s$ (1 decade median-median resistance window).

I_{CC} beyond 20 μA. Therefore, two values for I_{CC} were chosen, $I_{CC} = 30\ \mu A$ and $I_{CC} = 5\ \mu A$. The switching process depending on the choice of I_{CC} was experimentally studied by applying a train of identical Set or Reset pulses on the OxRAM device in HRS or LRS, respectively. When a pulse with $I_{CC} = 30\ \mu A$ is repeatedly applied to a single OxRAM cell in HRS, the Set process occurs abruptly in a probabilistic manner after a number of pulses, see figure 3.20. Any subsequent pulses do not result in further changes of the achieved LRS. The Reset process on the other hand is not as abrupt as the Set process as shown in figure 3.20. When a reset pulse is applied, the OxRAM cell resistance is immediately reduced by a probabilistic amount achieving a HRS. The exact value of the HRS can still be tuned in a small range by the application of additional Reset pulses.

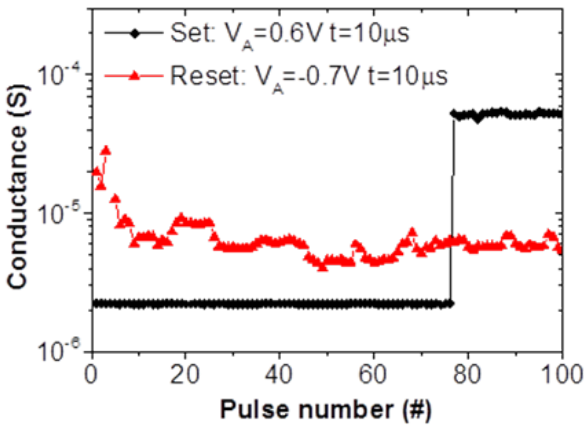


Figure 3.20: Abrupt Set of single TaO_x/HfO_2 device obtained by applying 100 identical Set pulses with$I_{CC} = 30\ \mu A$.

On the contrary, for pulses of $I_{CC} = 5\ \mu A$, the Set process is no longer abrupt but rather progressive and the achieved LRS depends on the number of applied Set pulses (see figure 3.21 (a)). A similar behavior can be observed for the Reset process as shown in figure 3.21 (b). Note that the conductance of single devices (grey lines) changes over several orders of magnitude orders with the pulse number while the different devices exhibit significant differences in conductance values, i.e. OxRAM cells exhibit a very strong device-to-device variability when ultra-low programming currents are used.

Due to the small number of tested devices here (around 10), it is ambiguous to extract the correct maximum range ΔG that can be achieved for the synaptic weight using one OxRAM device. Therefore, we have defined three possible ranges $\Delta G = [100, 300, 1000]$. Figure 3.22 shows the number of programming pulses as a function of the pulse duration that is needed to achieve a certain synaptic weight change ΔG. In order to achieve a certain ΔG of the OxRAM synapse, a certain number of set pulses is required whereas the longer the set pulse the stronger the change of the synaptic conductance and thus the less pulses are needed. Moreover, ΔG scales with the number of pulses, i.e. more pulses cause a larger ΔG.

3.2.7 Retention for ultra-low programming currents

It was stated before that the reset current (I_R) needed to switch an OxRAM device from LRS to HRS is typically similar to the current (I_{CC}) applied during the Set operation (HRS to LRS) unless $I_{CC} < 20\mu A$ (see figure 3.4). For lower currents, I_R drops significantly below I_{CC}. The retention, i.e. the capability of a memory device to remain in its programmed state over time, was analysed for 15 devices which were programmed using a very low $I_{CC} = 6.5\mu A$, shown in figure 3.23. Note that the measurements were taken at room temperature. A clear trend towards higher resistance values can be observed already after approximately 10^4 seconds which confirms a

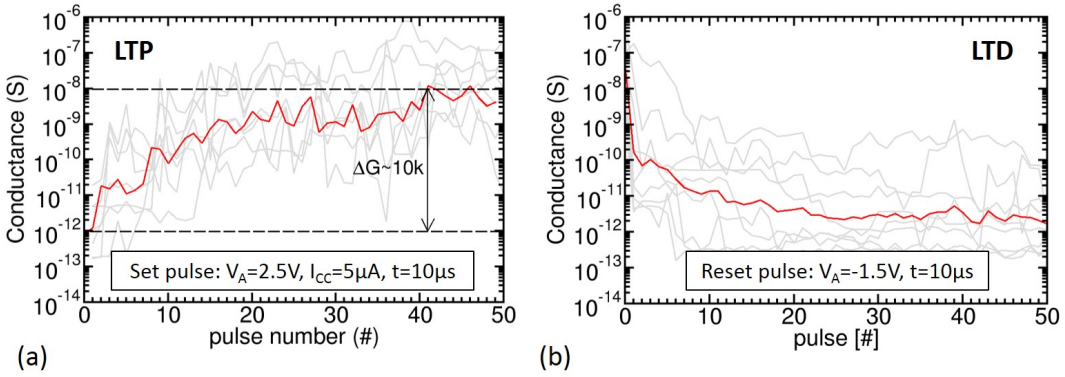


Figure 3.21: (a) Long Term Potentiation (LTP) and (b) Long Term Depression (LTD) of 10 TaO_x/HfO_2 devices (grey) obtained by application of 50 identical Set and Reset pulses with $I_{CC} = 5\ \mu A$. Geometric mean over all devices is also shown (red).

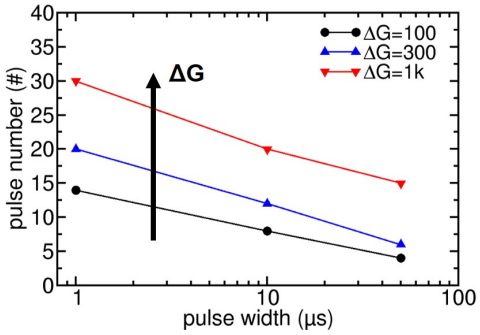


Figure 3.22: The pulse number required to increase the single OxRAM device conductance by a certain ratio ΔG is shown as a function of the pulse duration for $\Delta G = 100$, $\Delta G = 300$ and $\Delta G = 1000$. $I_{CC} = 5\ \mu A$.

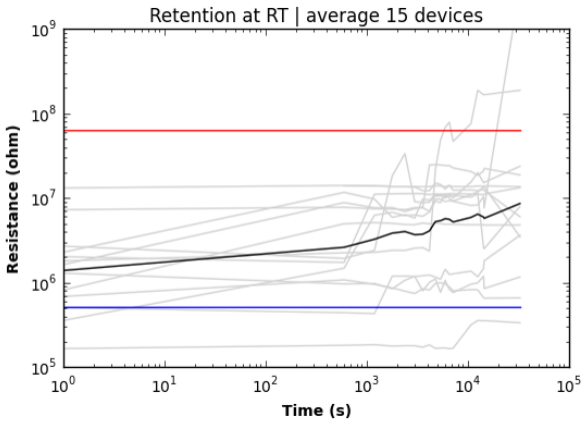


Figure 3.23: Data retention of $15x\ 1nm\ Al_2O_3/3nm\ HfO_2$ devices programmed into LRS using $I_{CC} = 6.5\mu A$. The test was performed at room temperature. Blue and red lines represent the average LRS or HRS levels. Grey lines show the single device behaviour and black the mean value of all devices.

relatively low retention capability of OxRAM cells that are programmed using ultra low currents.

3.3 Synapse design

OxRAM and CBRAM technologies were described in sections 3.2. It was shown that both technologies exhibit a binary switching behaviour if predefined programming conditions and typical programming currents are used for Set and Reset, i.e. these memory cells feature two distinct states rather than an analogue range of (resistance) states. Moreover, these technologies normally do not exhibit a cumulative behaviour, i.e. the incremental change of the resistance state upon application of low-magnitude programming events. However, if the programming current (I_{CC}) is reduced below a certain value which was found to be between $10\mu A$ and $30\mu A$, an analogue, cumulative switching behaviour could be observed when a sequence of identical Set or Reset

pulses was applied. The latter (analogue operation, non-filamentary) may be used directly to implement one synapse with one resistive memory device. On the other hand, the former (binary operation, filamentary) requires a dedicated circuit designs and/or programming strategies in order to compensate the intrinsic device shortcomings for the implementation of sophisticated synapse models. In the following, both solutions to this end are introduced and a biology inspired learning rule using the synapse concept is explained.

3.3.1 Synapse based on filamentary RRAM

Filamentary switching OxRAM exhibits two critical obstacles for the implementation of synaptic models inspired by biology. First, although the resistance of one device can be tuned throughout a continuous range spanning several orders of magnitude, this tuning requires a careful design of the programming conditions ($V_{S,R}$, I_{CC}) for each programming event which necessitates a rather complex circuitry (increasing chip area). Moreover, tailoring the pulse conditions for each device requires to access them sequentially one by one preventing parallel programming of several devices and thus causing major drawbacks in terms of speed. Finally, using this approach makes it necessary to know the current resistance in order to define the weight to be programmed and set the corresponding programming conditions. This is a major concern for memory size and/or parallelism. For this reason, two dedicated programming conditions shall be defined for Set and Reset, respectively. This can simplify the circuit design of the driver circuit (for programming) significantly, hence improving the integration density. The second problem evolves from the application of those invariant programming conditions because this allows typically to switch the RRAM cell in a binary fashion between two distinct states, the Low Resistive State (LRS) and High Resistive State (HRS) by only one programming pulse. Further programming pulses with the same voltage and current do not alter the device resistance as it was shown in figure 3.20. However, a synapse should feature multiple states and a cumulative behaviour, i.e. more programming pulses result in a stronger synaptic weight change.

In order to overcome the RRAM specific drawbacks, a synapse based on $n > 1$ RRAM devices can be used [171]. Several OxRAM devices (n) are combined in a parallel architecture, see figure 3.24 to build one synapse. This allows to achieve $n + 1$ states of synaptic weight and therefore the granularity (number of states) can be tailored according to the needs of any specific application. The devices are programmed using the driver circuit which applies Set or Reset pulses with $I_{CC} > 30\ \mu A$. This triggers the abrupt switching of single devices. In order to control the increase and decrease ratio, the programming pulse voltage has to be adjusted according to figure 3.14 or a pseudo random number generator (PRNG) can be used. Whereas the former allows to directly affect the device switching probability, the latter applies pulses with switching probabilities of 1 probabilistically to devices according to predefined Set and Reset probabilities, P_{Set} and P_{Reset}. Figure 3.25 demonstrates the functionality of the synapse concept featuring 20 devices per synapse and using the external PRNG with $P_{Set} = 0.071$ and $P_{Reset} = 0.047$. The individual

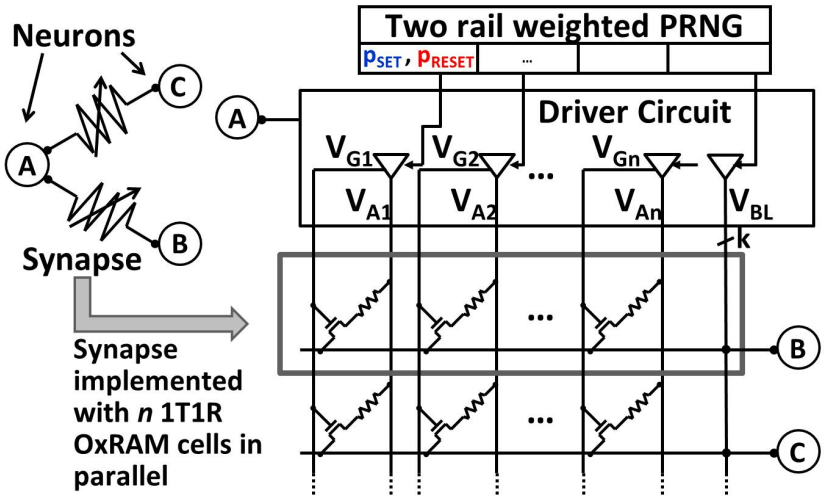


Figure 3.24: Multi-cell synapse concept. Each equivalent synapse consists of a series of 1T1R integrated RRAM devices, i.e. the corresponding synaptic weight is the sum of device conductances. A driver circuit including a pseudo random number generator (PRNG) is used to enable gradual tuning of the synaptic weight, thus overcoming the typical abrupt switching characteristic of RRAM shown in figure 3.20.

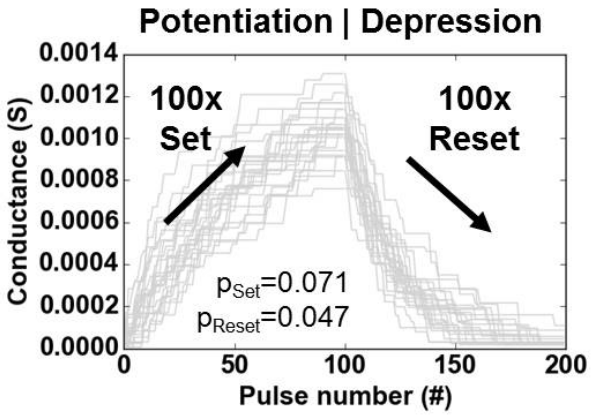


Figure 3.25: Potentiation and Depression for 20 synapses each based on 20 OxRAM devices using a pseudo random number generator (PRNG) for the application of Set and Reset programming pulses with p_{Set} and p_{Reset}. OxRAM devices are fitted using experimental data from figure 3.15 (a).

device resistances are achieved by using the experimental distributions of LRS and HRS of the 1T1R HfO_2/TaO_x structures operated by a programming current of $I_{CC} = 30\ \mu A$ (shown in figure 3.15 (a)). Upon the application of 100 identical Set or Reset pulses, the equivalent synaptic weight (conductance) can be gradually increased or decreased, respectively.

3.3.2 Synapse based on non-filamentary RRAM

The gradual resistance change of single OxRAM devices observed in the ultra-low current OxRAM operation (using $I_{CC} = 5\ \mu A$) seems very promising for the implementation of LTP and LTD with one device per synapse. Instead of a parallel circuit of several devices this may allow for very compact low power synaptic networks. Moreover, the cumulative resistance change could be exploited to significantly reduce the circuit complexity since no driver circuit (featuring the pseudo-random number generator (PRNG) for probabilistic potentiation and depression) would be needed. The conductance of a single device changes by $3-4$ orders of magnitude, thereby providing a very large dynamic range for the synaptic weight. However, the device-to-device variability poses a problem because it is in the same order of magnitude as the dynamic range of ΔG for single devices. This means, that a strong (potentiated) synapse may have a weight that is comparable or even lower than a weak (depressed) synapse or vice versa. This effect prevents the gradual switching OxRAM based synapse from straightforward integration into a neuromorphic network circuitry.

3.3.3 Probabilistic Spike-Timing-Dependent Plasticity for RRAM synapse

Spike-timing dependent plasticity (STDP) is a so-called learning rule which combines LTP and LTD based on the correlation of pre- and post-synaptic activities. When a post-synaptic spike occurs shortly after (before) a pre-synaptic spike, LTP (LTD) is performed on the synapse. The synapse implementation based on multiple filamentary switching OxRAM devices is adopted here. The relative weight change of a synapse according to biological Spike-Timing-Dependent Plasticity (STDP, see section 1.1.3) depends strongly on the exact timing of pre- and post-synaptic spikes to one another, i.e. if the two spikes occur within a narrow time frame, the weight change is relatively high and vice versa. In order to achieve this behaviour in an OxRAM based synapse (section 4.2.4), the pulse parameters for potentiation and depression would need to be modulated according to each timing difference Δt. This may cause a significant complication for the design of the driver circuit for the OxRAM programming which shall be avoided by using a simplified STDP rule for the OxRAM synapse. It was demonstrated that a simplified probabilistic approach for the STDP approximation is sufficient [228] to induce gradual Long Term Potentiation (LTP) and Depression (LTD). Here, it is only of importance whether the post-synaptic neuron spikes before or after the pre-synaptic neuron and if the timing difference lies within a certain range, as illustrated in figure 4.7. Within the LTP and LTD time regimes, the corresponding probabilities p_{Set} and p_{Reset} are constant.

The probabilistic STDP means that the synaptic weight changes when a post-synaptic spike occurs. If the pre-synaptic neuron was activated recently ($\Delta t < t_{LTP}$), LTP is performed on the synapse with a given Set probability p_{Set}, otherwise ($\Delta t > t_{LTP}$), LTD is performed with a Reset

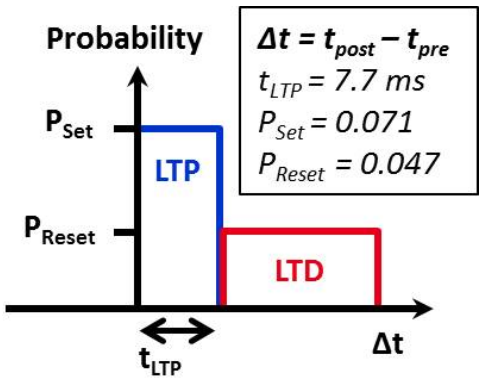


Figure 3.26: Probabilistic learning rule used for online learning in our SNN inspired by spike timing dependent plasticity (STDP). Set and Reset probabilities, p_{Set} and p_{Reset} as well as the LTP time window t_{LTP} are indicated.

probability p_{Reset}. The change of the synapse resistance follows accordingly

$$\Delta R = \begin{cases} P_{Set} \cdot dR_{Set} & \text{for } \Delta t < t_{LTP} \\ P_{Reset} \cdot dR_{Reset} & \text{for } \Delta t \geq t_{LTP} \end{cases} \tag{3.1}$$

The probabilistic STDP can be achieved by using the intrinsic switching probability (tuning Set and Reset voltages) or by using the extrinsic probability implemented with a PRNG.

3.3.4 From OxRAM variability to synaptic variability

A synapse is storing a specific weight which can either be potentiated (synaptic connection strengthened) or depressed (synaptic connection weakened). In order to perform this tuning precisely in an OxRAM based synapse, it is desired to be able to perform incremental well-defined changes of the conductance. However, if the single OxRAM devices are affected by variability, each potentiation or depression step induces uncertainty of the final weight to some extent as shown in figure 3.27. The figures show the evolution of the synaptic weight for 100 potentiation events and subsequently 100 depression events for single synapses (grey) and the average evolution of 20 synapses (red). Every synapse is based on 20 OxRAM devices in parallel whereas the Set/Reset probabilities are tuned by means of a Pseudo Random Number Generator (PRNG). As shown, a low variability (figure 3.27 (a)) allows to achieve a fine step-wise increase or decrease of the synaptic weight, i.e. conductance, whereas a high variability (figure 3.27 (b)) results in a strongly differing conductance change from step to step. Moreover, the high variability leads to a strong variation of the weights among different synapses as marked by the grey lines.

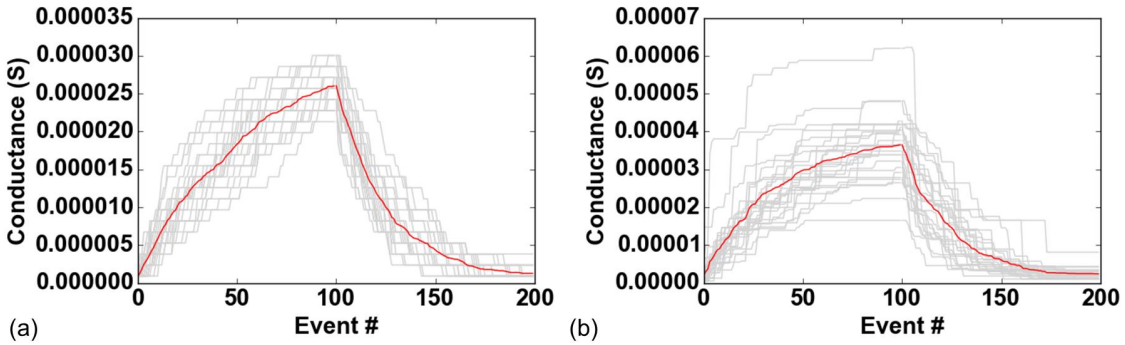


Figure 3.27: Representation of synaptic evolution for 100 events of potentiation and depression each for synapses based on (a) OxRAM operated at $I_{CC} = 340\mu A$ and (b) OxRAM operated at $I_{CC} = 30\mu A$. The OxRAM was based on a $1nm\ Al_2O_3/3nm\ HfO_2$ dielectric. Each grey line represents one synapse based on 20 OxRAM devices. The average synaptic weight evolution is shown in red.

3.4 Summary

This chapter has covered the three main aspects of implementing artificial synapses with solid-state electronic devices. First, the conditions that have to be fulfilled for a biology-like synaptic behaviour were explained. These are mainly a simple, or two-terminal, structure with an input and output terminal representing the dendrite and axon of two neurons. In addition, a synapse should typically feature multiple states of conductance, a progressive programming behaviour and it should be non-volatile. Second, the electrical characteristics of RRAM were studied in detail and described with respect to their principal mechanisms, reliability and in the perspective of exploiting this technology for the implementation of artificial hardware synapses. It was found that OxRAM features two operation regimes depending on the programming current (I_{CC}) that is used for the Set operation. For $I_{CC} > 20\mu A$, a filamentary switching was observed whereas for $I_{CC} < 20\mu A$, the resistance switching seems to be governed by a bulk process. An interesting effect was observed on the CBRAM technology based on a doped oxide material which offered a rather high resistance margin between LRS and HRS while operating at only $I_{CC} = 4.5\mu A$. Typically, those ultra-low currents do not allow to have a separation of LRS and HRS due to large resistance variability of both states. Third, the design of an artificial hardware synapse based on RRAM was explained and it was shown that the requirements to mimic a behaviour similar to biology can be obtained by using a probabilistic learning rule that was inspired by Spike-Timing-Dependent Plasticity.

CHAPTER 4

SPIKING NEURAL NETWORK FOR SPIKE SORTING

Ddespite enormous progress in technology, solutions for real-time processing of (human) neural signals for applications such as brain-computer interfaces (BCI) are still not available. In this chapter, an innovative approach to decode complex brain signals by so-called Spike Sorting is introduced. The approach is based on a compact spiking neural network (SNN) which is able to identify spike shapes in neural signals. The synaptic weights of the SNN are trained in-situ by an on-line learning strategy inspired by biological Spike Timing Dependent Plasticity (STDP). This concept offers promising advantages to conventional spike sorting techniques for BCI applications because the SNN architecture is potentially suitable for hardware implementation by using resistive random access memory (RRAM) technology for the design of the synapses while relying on standard CMOS technology for the integration of neurons. For the particular application of spike sorting, the excellent RRAM properties described in the chapter 3 may enable real-time functionality, integration of large SNN and wireless implantable devices for rehabilitation purposes.

This chapter is structured as follows. First, the motivation to design a spike sorting system is reviewed briefly in section 4.1. Section 4.2 describes approach of the spike sorting system, which is based on the two main components data encoding by frequency filtering and detection/classification by a Spiking Neural Network. Section 4.2.3 describes the architecture of the Spiking Neural Network. Section 4.3 introduces the biological data used in this work. Section 4.4 presents the performance of the spike sorting application and finally, section 4.5 summarizes the strengths and weaknesses of the presented approach.

4.1 Motivation for spike sorting

Spike sorting is the process of separating contributions in the electrical signal according to its single units, i.e. neurons (see chapter 1). Algorithms based on the conventional three steps of spike detection, feature extraction and classification are able to provide powerful analysis tools, however, they present several limitations, as they often require (i) *user supervision* (manual tuning of the threshold parameters, choice of features to be extracted, choice of the number of clusters). Most of the available spike sorting approaches are performed via (ii) *off-line processing* which is not practical because it does not allow for real-time processing in closed-loop applications (e.g. in BCI) or real-time data compression prior to wireless transmission. Moreover, the off-line processing using conventional computers or powerful GPU's is (iii) *computationally expensive* posing a problem for the design of low-power portable BCI solutions. Therefore, new spike sorting approaches are required to address the needs for future healthcare applications.

4.2 Spike Sorting system

To tackle the shortcomings of state-of-the-art spike sorting techniques, an extremely promising approach may be brain-inspired computing by means of artificial neural networks (ANN) which have demonstrated to be superior candidates for the detection and prediction of patterns occurring in complex data [222], [223], [166]. Moreover, they offer several advantages over conventional von-Neumann based computing paradigms such as: (i) *Unsupervised operation* of an ANN can be achieved by using (feed-forward) learning rules, e.g. Spike-Timing-Dependent Plasticity (STDP). (ii) *On-line functionality* is achieved by ensuring that the ANN operation cycle time meets the requirements of the application. In the case of spike sorting, single spike events are in the order of $1ms$ which means that the ANN response time should be $1ms$ or lower. Synapses based on RRAM technology can be used for this purpose since their latency is typically in the range of microseconds. (iii) *Low-power consumption* may be realized by an efficient network structure and low-power building blocks (neurons and synapses). Therefore, CMOS based neurons and RRAM based synapses are excellent candidates to facilitate ultra-low energy ANN's.

4.2.1 General approach

Figure 4.1 shows the schematic view of our system designed to perform real-time spike sorting of spiking cortical signals, i.e. to extract, learn and recognize different spike shapes. The heart of the system is a Spiking Neural Network (SNN) with key features such as synaptic plasticity and a bio-inspired learning rule similar to Spike-Timing-Dependent Plasticity (STDP). Extracellular cortical signals are recorded by fine electrodes which are implanted in-vitro or in-vivo in neural tissue. Before the SNN can be used to detect and classify spikes the recorded electric potentials, the raw signal has to be pre-processed by some means into different signal components that

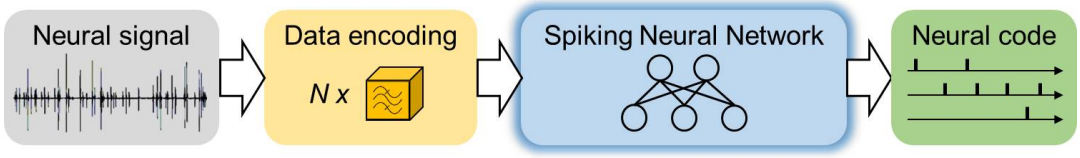


Figure 4.1: Overall schematic of spike sorting approach based on data encoding by N band-pass filters and a spiking neural network. The approach aims to extract the neural code from electrical neural signals.

allow to analyse different features of the signal. Therefore, we chose the approach of separating the signal according to its frequency components by using a set of band-pass filters (BPF) which enable the real-time pre-processing of the continuous electrical signal. This approach is described in detail in section 4.2.2. The SNN is then using the BPF signals to perform the actual spike sorting task. The components of this system were tailored to the needs of Spike Sorting and its synapses were implemented with OxRAM devices (see section 4.2.3). The spike sorting functionality was tested by using our event-driven simulator 'Xnet' [148].

4.2.2 Input data encoding

Spiking neural data has characteristic properties, namely:

- A priori unknown number of classes (i.e. spike waveforms)
- Characteristic spike times around 1-2ms
- Short inter-spike intervals in range of spike time are possible
- Non-stationary spike waveforms (i.e. characteristic spike shape of a given class might change over time)
- Amplitudes of electrically measured spikes variable and dependent on technology used to record neuronal activity
- Spiking signals typically exhibit frequency ranges between 300-3000Hz

Figure 4.2 shows a biological neural signal that was recorded with a micro-electrode located next to a nerve fibre. Since this electrode was placed outside of neurons, the signal is referred to as extra-cellular signal (ES). It shows the electrical potential evolution over time where two spikes can be observed between $2-4.5\ ms$ and $5.5-8\ ms$. These two spikes are accordingly labelled as 'Spike A' and 'Spike B'. In the right hand side of figure 4.2, the corresponding spectrogram is shown which represents the energy variation of the frequency spectrum over time. Low energies (blue) are observed in a wide frequency range in the absence of spikes (only background noise due to Local Field Potentials) whereas high energies (red) can be observed over a wide range of frequencies whenever a spike is present in the data. Furthermore, the two different spike shapes

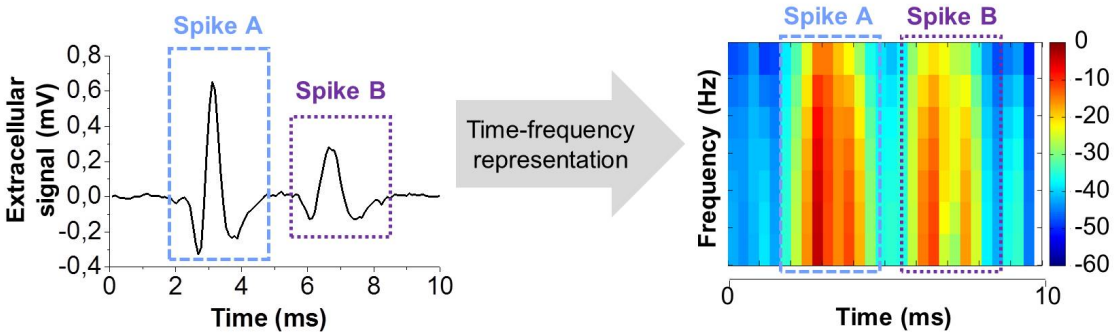


Figure 4.2: Signal encoding for spike sorting paradigm based on continuous time-frequency decomposition of the analog extracellular signal (ES). Different spike shapes (here Spike A and B) exhibit distinct patterns in the spectrogram. These 'finger prints' are used to distinguish between different spike shapes.

shown in this example differ in their appearance in the spectrogram, i.e. a larger spike produces both higher frequencies and higher energies. This gives rise to the assumption that any given spike waveform exhibits a characteristic representation in the time-frequency domain (if the resolution in terms of time and frequency is sufficient). This may be used to recognize a specific waveform among others and therefore a means of distinction between different waveforms. By using those 'finger-print' spectra of different waveforms, it may be possible to trace the activity of single neurons.

In order to perform the spectral analysis of a recorded signal in the time-frequency domain, the signal is filtered by a series of band-pass filters. The role of this filter bank is to distribute the energy of the spikes over an optimized number of channels providing the crucial possibility to observe even small differences in the spike shapes thanks to differences in their spectra. Hence, each filter should feature (i) a high frequency resolution to analyse only a small frequency range of the signal and (ii) a high temporal resolution and minimum response delay to allow the distinction of consecutive spikes (within a few tens of milliseconds) and low-latency applications. As time and frequency resolution are inversely related, a sufficient trade-off between frequency and temporal resolution has to be found. To address this, a low filter order (≤ 3) is required to have fast filter responses of less than a few tens of ms. The filter bandwidths (B) should be narrow to achieve a reasonable frequency resolution, however, the temporal resolution degrades (longer filter response) as the filter bandwidth is reduced. It is known that the typical frequency spectrum of neural spikes is invariant and does not exceed 2000 Hz [82] [46] and that those spikes have a characteristic duration of $1-2$ ms. Bandwidths of around $60Hz$ and $2nd$ order Butterworth filters offer a good compromise for this application. Here, we defined a minimum frequency of 100 Hz for the spectral analysis to exclude low frequent background signals ($\ll$100 Hz). For this reason, 32 filters with $B\approx$60 Hz for each filter are used to cover the frequency range between 100 Hz and 2000 Hz. The center frequency (f_0) of filter n is shifted by $\Delta f = 60Hz$ for filter

$n+1$. Hence, the filter bank introduces a (partly redundant) encoding of the frequency intensity since adjacent filters start to overlap at -3 dB filter gain which may even enhance the ability to distinguish the spectral content of spikes whose waveforms are relatively similar to one another. The filter characteristics for the described set of BPF is shown in figure 4.3. Note that increasing the number of filters, i.e. reducing the frequency spacing between two BPF, may not be beneficial because it results in excessive filter redundancy.

Figure 4.4 shows an example for the signal encoding approach based on the band-pass filters. The signal used here is a short sequence of an in-vitro Crayfish recording which features two different spike shapes, i.e. Spike A at $2-4.5$ ms and Spike B at $5.5-8$ ms (see also figures 4.2) and 4.3). First, the signal is fed to all 32 filters which generate continuous filter response signals ('Raw'), shown in figure 4.4 (a). As the water fall diagram shows, those filter responses fluctuate between positive and negative values. However, for the SNN it is necessary to have non-negative input signals in order to avoid the auto-cancellation of signals of opposite polarity. This will be explained in more detail in section 4.2.3. For this reason, the filter signals are full-wave rectified (FWR) which yields a signal representation as shown in figure 4.4 (b). These positive valued signals can then be used as input signals for the first layer neurons of the SNN.

It is worth noting that the filter bank may be used for any spiking neural data given that the frequency range of spikes does not exceed 2000 Hz and that the temporal duration of single spike events is around 1 ms. This is important because a key challenge for spike sorting is that one can not foresee the specific properties of the signal to be recorded, e.g. low-frequent drift, signal-noise-ratio or the frequency of spike events. The spike sorting system is therefore tested on other data sets as well and the results are presented in section 4.4.4.

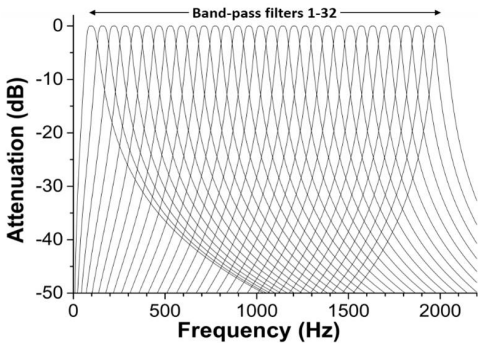


Figure 4.3: Band-pass filter characteristics for 32 order 2 Butterworth filters equally distributed between 100 Hz and 2000 Hz. The bandwidth for each filter is $B = 60$ Hz. This filter set is used to pre-process biological spiking data.

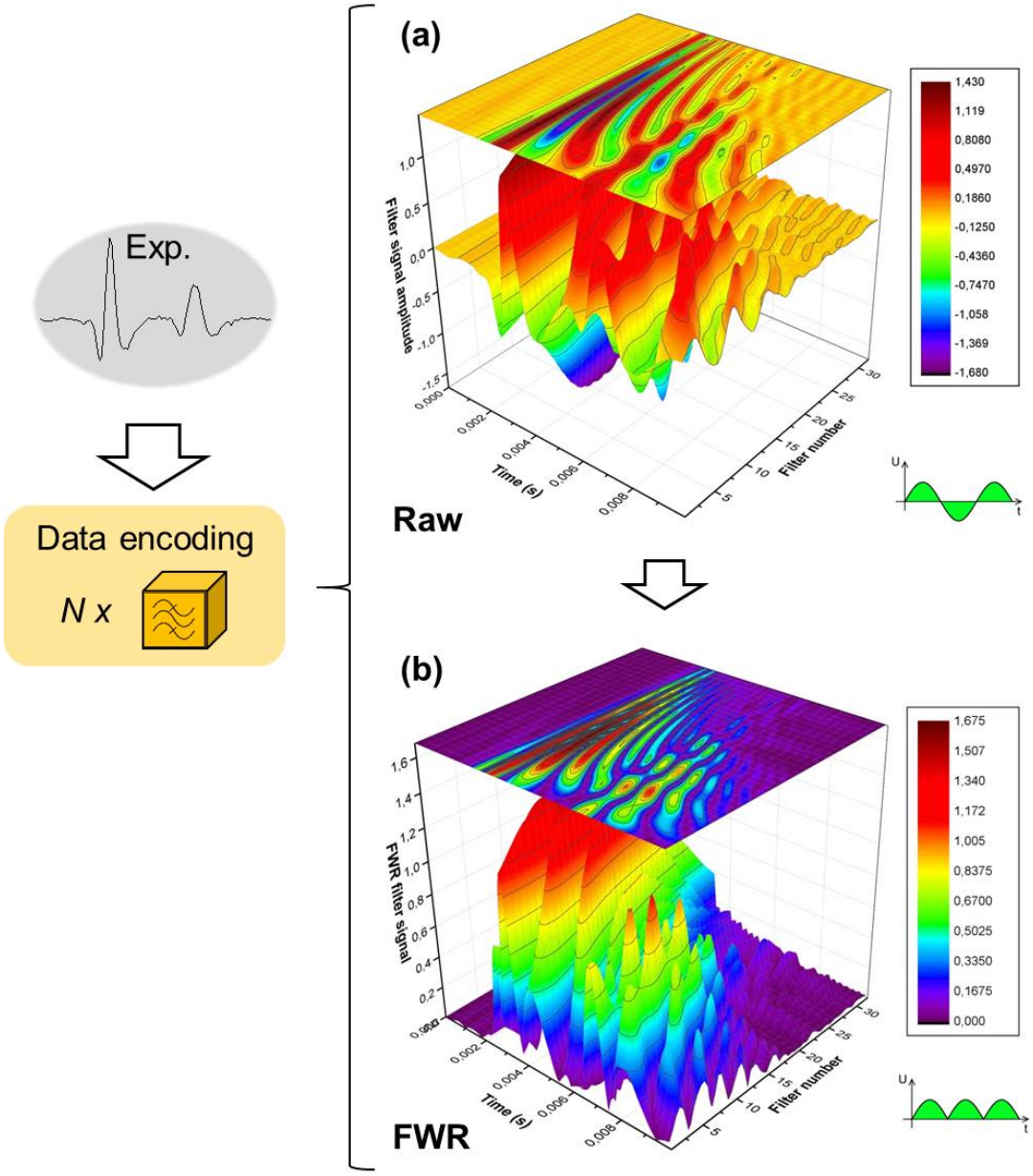


Figure 4.4: Band-pass filter output signals from 32 filters applied to a 10ms long signal (see figure 4.2). (a) The raw continuous filter responses are shown as a function of the time which are then full-wave rectified resulting in signals shown in (b).

4.2.3 Spiking neural network architecture

The Spiking Neural Network (SNN) uses the rectified BPF output signals (see section 4.2.2) to detect spikes and by repetitive occurrences of characteristic spectra becomes selective to those, which allows to perform spike sorting. Figure 4.5 shows schematically how the BPF and the SNN are connected to each other. Since the spike sorting task is a dynamic classification task

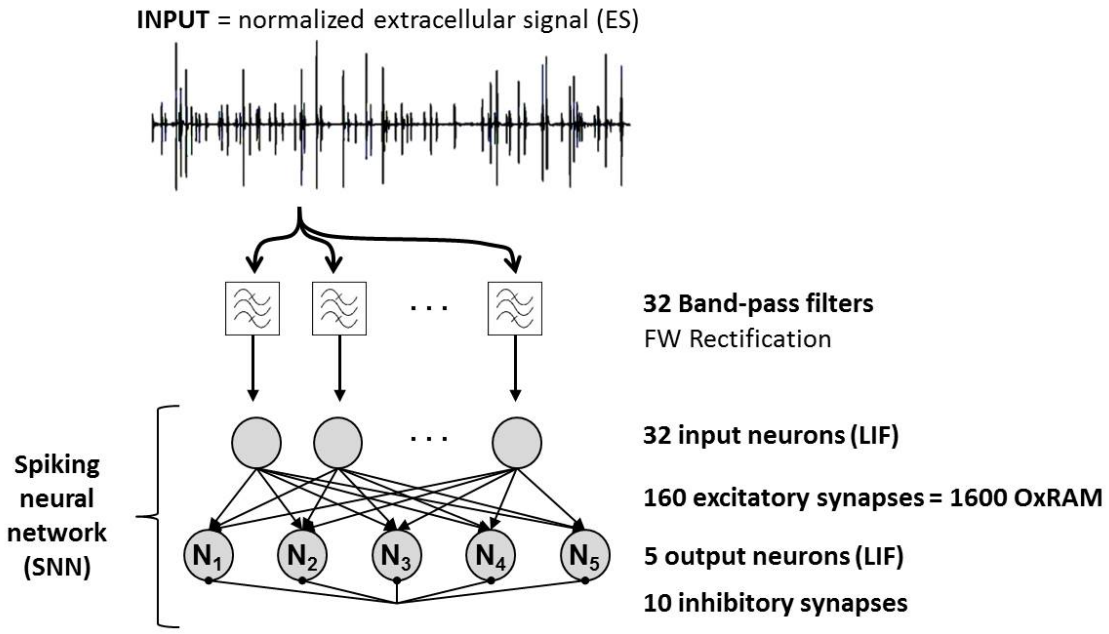


Figure 4.5: Functional schematic of spike sorting system based on a Spiking Neural Network. The extracellular signal (ES) is fed through 32 frequency band-pass filters which are connected one-to-one to the input layer of the SNN. Synapses are based on OxRAM devices. Output neurons are interconnected by inhibitory synapses to feature the winner-take-all principle which allows them to become selective to different input spike shapes.

in the real-time signal processing domain, spike based signal coding has been chosen instead of formal coding (used in most artificial neural networks such as deep neural networks). Spike coding allows to reduce the energy consumption due to the asynchronous, sparse coding nature of information. Furthermore, it allows to implement temporal features such as the detection of correlation of neuronal activities in time. Finally, SNNs are able to incorporate permanent learning based on interference of spiking activities, e.g. by Spike-Timing-Dependent Plasticity.

The neurons of both input and output layer are described by the Leaky Integrate Fire (LIF) model which is a simplified spiking neuron model with respect to the biologically precise Hodgkin-Huxley model to facilitate computational efficiency [229]. It is described in detail in [230]. The LIF is a popular spiking neuron model which considers the neuron as a parallel circuit of a resistor R and capacitor C. The input current to the neuron $I(t)$ is divided into the two components I_R and I_C, respectively,

$$I(t) = I_R + I_C = \frac{u(t)}{R} + C\,\frac{du}{dt} \tag{4.1}$$

whereas R is the membrane resistance, C the membrane capacitance and $u(t)$ the membrane potential. By introducing a membrane leak time constant $\tau_m = RC$, this leads to

$$\frac{du}{dt} = \frac{1}{\tau_m}[-u(t) + R\ I(t)] \tag{4.2}$$

A neuron emits an action potential (or spike) when $u(t)$ reaches a pre-defined membrane integration threshold I_{thres}. The membrane potential is reset to $u(t) = 0$ and the integration is deactivated for a refractory period t_{refrac}. This results in the case-wise equation for the time-dependent membrane potential

$$\frac{du}{dt} = \begin{cases} \frac{1}{\tau_m}[-u(t) + R\ I(t)] & \text{for } t < t_{refrac} \\ 0, & \text{for } t > t_{refrac} \end{cases} \tag{4.3}$$

If an input spike arrives to a neuron within t_{refrac}, $u(t)$ is not increased by the synaptic current. Otherwise, the synaptic current is integrated raising $u(t)$. The parameters of the LIF input and output neurons are given in table 4.1.

The input and output neuron layers of the SNN are fully connected (by $32x5$) excitatory synapses, i.e. every input neuron has a synaptic connection with every output neuron. These synapses are emulated by 10 OxRAM devices per synapse, described in section 4.2.4. A biologically inspired learning rule, STDP, is used to tune those synaptic weights in an unsupervised way. The goal is that every spike shape will be learned and recognized by one of the output neurons whereas non-selective neurons remain silent, i.e. the number of spiking output neurons indicates the number of spike classes. Classification redundancy has to be avoided, i.e. for each spike shape occurring in the data, only one neuron shall spike. For this reason, lateral inhibition is implemented with recurrent inhibitory synapses across the output layer to prevent the neurons from simultaneous spiking (i.e. winner-takes-all principle).

4.2.3.1 Input layer of SNN

As shown in figure 4.5, the number of input neurons corresponds to the number of filters, i.e. 32 neurons. When the 32 filtered full-wave rectified signals are presented to the input neurons, the analogue continuous signals are converted into spikes according to the LIF neuron model (equation 4.3). The corresponding LIF parameters were manually tuned using the two spike waveforms of the in-vitro Crayfish dataset described in section 4.3 in such a way that the output of the neuron's activity represents the spectral magnitude of the signal throughout the tested frequency range ($100\ Hz - 2000\ Hz$), i.e. the stronger the energy in a specific frequency band the more input spikes are generated. Thus, the input neurons create characteristic patterns

Table 4.1: Leaky Integrate Fire (LIF) neuron parameters of the 2-layer spiking neural network used for spike sorting of extracellular spiking data.

Symbol	Parameter	Layer 1	Layer 2
I_{thres}	Integration threshold	0.1 (a.u.)	0.58 (a.u.)
T_{leak}	Leak time constant	0.2 ms	5.1 ms
$T_{refractory}$	Refractory period	4 ms	46.1 ms

for different spike waveforms as figure 4.6 shows. Spike A generates a much higher activity throughout all frequency channels with respect to Spike B. Moreover, the duration of the input activity seems to represent the energy in the individual frequency bands since input spikes can be observed up to 25 ms after Spike A compared to 20 ms for Spike B. There seem to be waves of input activity, i.e. two input spike are roughly 5 ms apart from each other. Note that these delays are most likely imposed by the refractory period of the input neurons.

4.2.3.2 Output layer of SNN

The number of output neurons determines the maximum number of spike classes that can be classified by the SNN. A sufficiently high number of output neurons has to be chosen so that every spike shape contained in the extracellular data can be assigned to one output neuron, i.e. the number of output neurons has to be at least as high as the number of spike shapes in the extracellular signal. However, this number is typically not known a priori. The dataset that was used for the calibration (section 4.3) contains two spike classes which need to be classified. Therefore, we used deliberately a higher number of output neurons, namely five, to verify that the network is able to detect the number of classes independently.

The spikes of the input layer neurons are propagated along the excitatory synapses to the 5 SNN output layer neurons. Hence, in the event of a spike, a synapse performs a multiply-function according to Ohms law, i.e. it converts the unitary spike signal (V_{spike}) into a current I that is a function of the individual synaptic weight G.

$$I = V_{spike} \times G \tag{4.4}$$

Each output neuron receives the pulsed input currents from 32 synapses which lead to an increase of the neurons membrane potential $u(t)$ following equation 4.3. The firing event of an output neuron means that the input currents were sufficient to reach the threshold for spike emission

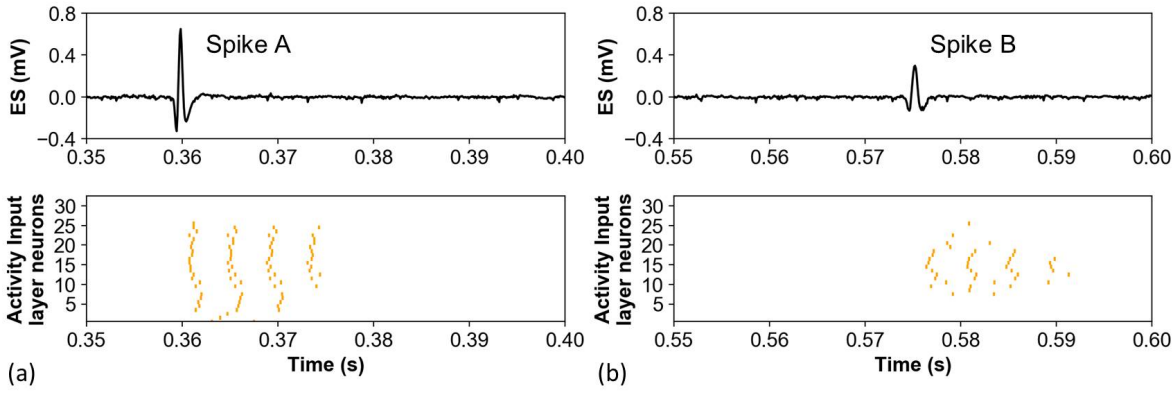


Figure 4.6: Recorded extracellular (ES) signals (black) and representation of frequency bands by input neurons (orange) for (a) Spike A and (b) Spike B.

and the learning rule is activated (see section 4.2.5). For a well trained network, i.e. specialized synaptic weights, an output spike further indicates that the spike inspected in the input signal (Spike A or B in the example) belongs to the specific class corresponding to this output neuron.

The parameters (I_{thres}, T_{leak}, $I_{refractory}$) of the output neurons were tuned manually and then optimized by using a genetic algorithm to make the system capable to detect and sort spikes. A genetic algorithm performs simulations of a system while each simulation varies the parameters slightly. A number of different parameter sets (specimen) is simulated in one generation, amongst which a number of winners is chosen to create a next generation which parameters are again varied. This process is repeated for a number of generations. Here, we randomly varied the parameters (by maximum 20 %) within one generation and evaluated the classification rate. The number of specimen was 32 and the number of generations was 8. Based on the simulation results of each generation, four winners were chosen for further parameter variation. The level of variation was decreased as the classification rate saturated.

4.2.4 Synapse design

The synapse implementation based on multiple OxRAM devices is adopted in this work. This concept features multiple states for the synaptic weight and allows to overcome the abrupt Set switching limitation of single OxRAM devices inducing gradual/progressive Long Term Potentiation (LTP) and Long Term Depression (LTD), see section 3.3. 1T1R OxRAM structures have been fully characterized using a programming current I_{CC} = 30 μA (described in detail in section 3.2) and the experimental LRS and HRS distributions have been used to model the OxRAM based SNN architecture presented in figure 4.5. Based on the electrical tests of OxRAM, the HfO_2/TaO_x resistive layer was chosen here to implement the synapses since it has the highest resistance values compared to the other tested materials and thus consumes the lowest power in read mode. Ten OxRAM devices were used per synapse resulting in a total number of 1600 OxRAM devices required for the SNN. In this work, a pseudo random number generator (PRNG) is used as part of a driver circuit for the application of the Set and Reset electrical pulses with the corresponding probabilities (p_{Set} and p_{Reset}).

The individual weights of the synapses are the key to recognition and distinction of different input patterns. Those weights are achieved by the application of an on-line learning rule, described in section 4.2.5. The output neurons are connected all-to-all with inhibitory synapses which are not implemented with RRAM here but simply result in a reset of the internal potential of the neurons of layer if another one emits a spike. Thus, the output layer features lateral inhibition to prevent output neurons from simultaneous spiking (i.e. winner-takes-all) which would lead to spiking of multiple neurons for the same input event and hence learning the same features. This is a critical feature in order to allow spike sorting.

4.2.5 Unsupervised learning by Spike-Timing-Dependent Plasticity

One of the key challenges for spike sorting algorithms is the real-time functionality for a priori unknown data. This requires an online learning algorithm, i.e. the fast adaptation of the spike sorting system to new data (new spike shapes in the ES, changing number of classes) and specifically for SNN a synaptic latency that is lower than the duration of biological spikes (approx. 1 ms). Spike-timing dependent plasticity (STDP) is used to meet the first requirement whereas the latter is accomplished thanks to the fast switching synapses ($< 1\ \mu s$), in our case the OxRAM devices. Note that a fast switching time of the SNN synapses is required since the online learning is permanently active which necessitates in-situ modifications of individual OxRAM cell resistances. Without online learning, classification does not typically require fast switching synapses.

As previously explained, our synapses are composed of multiple binary-state devices (figure 4.7 (a)) in order to achieve multi-level synaptic weights [168]. The relative weight change of a synapse caused by biological STDP (chapter 1) depends strongly on the exact timing of pre- and post-synaptic spikes to one another, i.e. if the two spikes occur within a narrow time frame, the weight change is relatively high and vice versa. In order to achieve this behaviour in an OxRAM based synapse (section 4.2.4), the pulse parameters for potentiation and depression would need to be modulated according to each timing difference Δt. This may cause a significant complication for the design of the driver circuit for the OxRAM programming which shall be avoided by using a simplified STDP rule for the OxRAM synapse. Here, it is only of importance whether the post-synaptic neuron spikes before or after the pre-synaptic neuron and if the timing difference lies within a certain range. The value of the relative change is equalized.

Figure 4.7 (a) illustrates the simplified probabilistic STDP [228] for online learning and figure 4.7 (b) shows the conductance behaviour of 20 synapses, each based on 20 OxRAM devices for the application of 100 Long Term Potentiation (LTP) and 100 Long Term Depression (LTD) operations. The synaptic weight changes when a post-synaptic spike occurs. If the pre-synaptic neuron was activated recently ($\Delta t < t_{LTP}$), LTP is performed on the synapse with a given Set probability p_{Set}, otherwise ($\Delta t > t_{LTP}$), LTD is performed with a Reset probability p_{Reset}. The probabilities as well as t_{LTP} were optimized by means of a genetic algorithm together with the parameters of the output neuron layer. Note that once all the parameters for the filters, SNN and probabilistic STDP are set, the spike sorting system may in principal be used on any spiking dataset without changing those parameters.

4.2.6 System level description

The operating principle of the developed spike sorting system is summarized in figure 4.8. The neural activity is recorded and streamed to the band-pass filters connected to the SNN with the sampling frequency $f_{sampling}$. Therefore, the data value at time t is read and fed to the filters. The filter outputs are depending on the preceding data values and the pass band frequency,

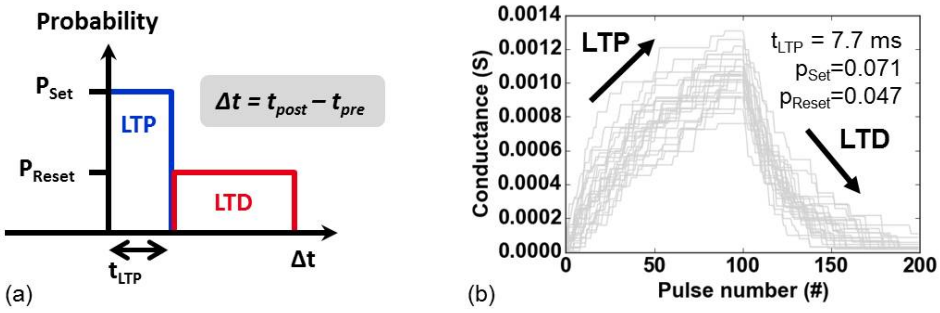


Figure 4.7: (a) Probabilistic learning rule used for online learning in our SNN inspired by spike timing dependent plasticity (STDP). Set and Reset probabilities, p_{Set} and p_{Reset} as well as the LTP time window t_{LTP} are indicated. (b) Long Term Potentiation (LTP) and Long Term Depression (LTD) for 20 synapses each based on 20 OxRAM devices using p_{Set} and p_{Reset}. OxRAM devices are fitted using experimental data from figure 3.19.

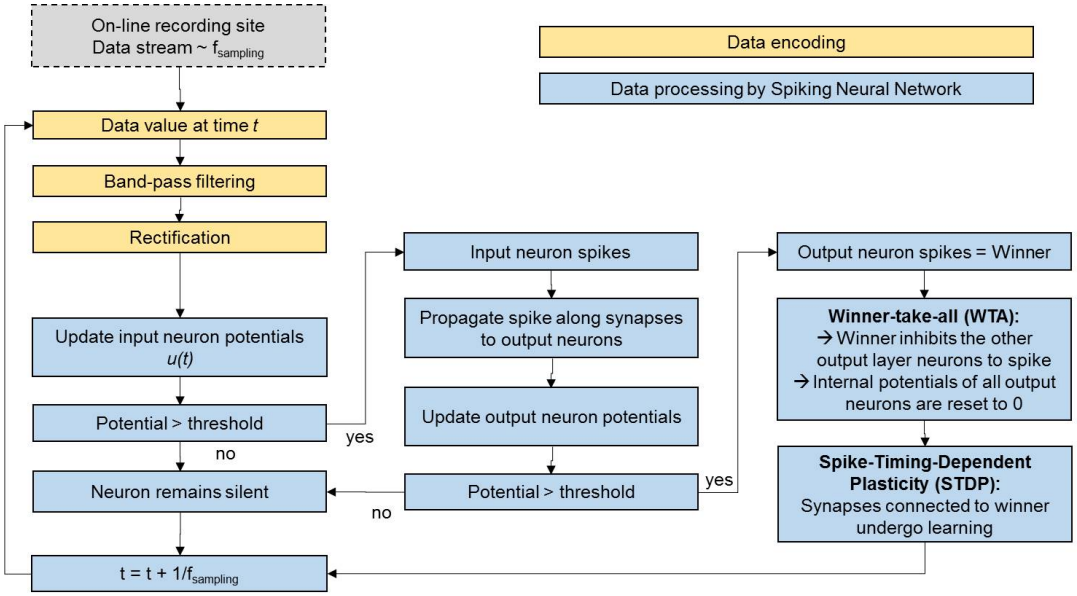


Figure 4.8: Schematic algorithm of the proposed spike sorting system.

respectively. The filter signals are rectified and used as input to the layer 1 of the SNN. By using the band-pass filter approach to encode spiking data, the SNN receives strong input signals if a spike is observed in the input data whereas rather low-frequency signals are not able to excite the network sufficiently. Thus, no dedicated method to remove low frequent noise is required and spike detection is inherently implemented. The potentials of the input neurons are updated according to the integration at this time step and the neuronal parameters. In the next step, it is checked whether neurons cross the pre-defined integration threshold. If this is not the case, the neuron does not produce an output activity. The time t is iterated to the next sample, restarting the loop. Otherwise, neurons crossing the threshold, emit a spike. This spike is sent to

all output neurons via the synapses. The neuron potentials of layer 2 are updated according to the received spikes whereas the magnitudes are modulated by the synaptic strengths. Again, it is checked whether neurons cross the pre-defined integration threshold. The neuron that crosses the threshold first, is identified as the winner following the winner-take-all principle. This neuron emits an output spike, inhibits the other output neurons from spiking and resets their membrane potential to $u(t) = 0$. Based on the spiking times, the simplified STDP rule is applied for learning.

It is possible to implement the presented SNN in a co-integrated circuit using complementary metal oxide semiconductor (CMOS) technology for the design of hardware neurons [231] as well as the band-pass filters and Oxide based resistive RAM (OxRAM) for the synapses [171]. The electrical conductance of OxRAM devices can be modified by means of voltage pulses which is exploited to tune the synaptic weights, described in chapter 3. The synapse design is explained in more detail in section 3.3. The validity of the proposed network and the OxRAM synapse model extracted from electrical data will be demonstrated in section4.4 by means of simulations using our special purpose event-driven simulator tool 'Xnet'.

4.3 Spiking biological data

In order to illustrate the validity of the proposed spike sorting methodology, real biological data recorded from crayfish and rats were used. When real data are recorded in neural tissue, it is difficult to retrieve the reference of the spiking activity, hereinafter called ground truth, because it requires to record single neurons which is typically done by so-called intracellular electrodes. An in-vitro preparation of a Crayfish nervous system was used for testing and calibrating the spike sorting system. Extracellular and intracellular activity were recorded simultaneously [23] [232]. Two electrodes were implanted in an in-vitro preparation, as illustrated in figure 4.9 (a). One electrode was inserted into a motor neuron of the T5 ganglion and is therefore referred to as the intracellular electrode. The other electrode was attached to a nerve fiber outside of the neurons, hence called extracellular electrode, that recorded action potentials of several cells simultaneously. In these data, the extracellular signal (ES) contains two different spike shapes, labelled as Spike A and Spike B in figure 4.9 (b), corresponding to two different neurons. The spikes simultaneously observed in the intracellular signal (IS) correlate with the activity of Spike A in the ES. Therefore, the IS activity can be used as the ground truth to assess the spike sorting capability of our system for the detection of Spike A in the ES data. The entire data set duration comprises 681 seconds and is called CF1 subsequently.

4.4 Performance

The complete spike sorting system consisting of band-pass filters and SNN was simulated with the 'Xnet' (event-driven) simulator for the treatment of the Crayfish data (CF1) introduced in section 4.3.

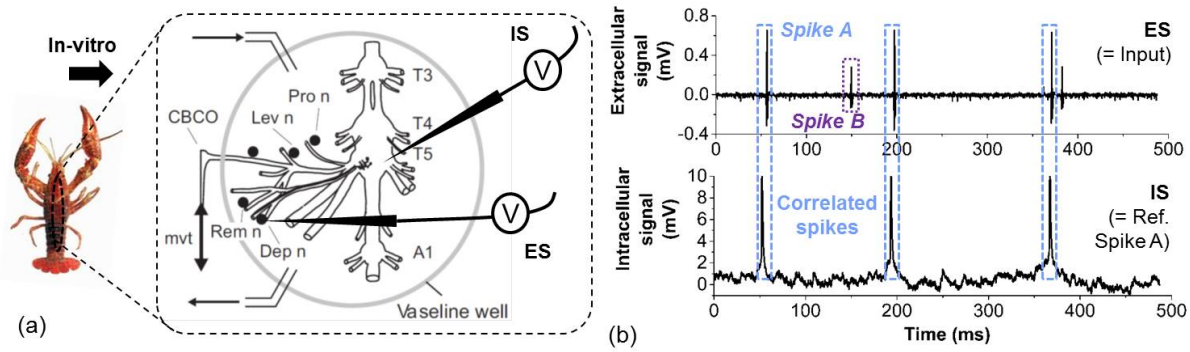


Figure 4.9: (a) Illustration of the experiment used to obtain real biological data. The crayfish is dissected and two electrodes are used in-vitro, one intracellular electrode inside a motor neuron in the T5 ganglion and one extracellular positioned against a depressor nerve ('Dep'). (b) The extracellular signal (ES, short sequence shown) contains two different spike shapes, labelled as Spike A and Spike B. The intracellular signal (IS) contains spiking events matching only Spike A of the ES.

4.4.1 Functionality

Figure 4.10 illustrates the unsupervised learning response of our SNN to the input signal (ES) described in section 4.3. The intracellular signal (IS) is shown at the top showing the reference signal used to quantify the recognition rate of spike A. Below, the extracellular signal (ES) is shown with two snapshot sequences and the corresponding representation of the input activity (orange, middle part of figure 4.10) and the output activity of the SNN (green, bottom-most part of the figure). Initially ($0\ s - 285\ s$), only Spike B is present in the extracellular signal. The SNN output, i.e. the firing patterns of the five output neurons $N_1 - N_5$ are completely random. Thanks to the introduced lateral inhibition, one output neuron, one output will become gradually selective to Spike B. Then ($285\ s - 545\ s$), also Spike A is observed in the input signal. In this period, another output neuron starts to spike predominantly when the Spike A appears, while the one that learned the previous pattern for Spike B continues to fire when Spike B appears. In the following, the neurons corresponding to the Spike A and Spike B are referred to as N_1 and N_2. However, it is important to note that it is unknown before the learning which neuron will become selective to a given pattern. The remaining output neurons N_3, N_4 and N_5 are rather silent. The output activity after learning (bottom right of figure 4.10) shows how two output neurons are selectively spiking when Spike A and Spike B occur in the data while no activity can be observed from the other neurons. At the end of the test case ($545\ s - 681\ s$) only Spike B is present. Therefore, only N_2 shall show a spiking activity whereas N_1, N_3, N_4 and N_5 should be inactive.

The activities of all output neurons N_1 to N_5 are shown in figure 4.11 whereas the activity is defined as the number of output spikes in time intervals of $10\ s$. As one can see, the N_1 activity is in good agreement with the intracellular reference, i.e. N_1 detects Spike A. The activity of N_2

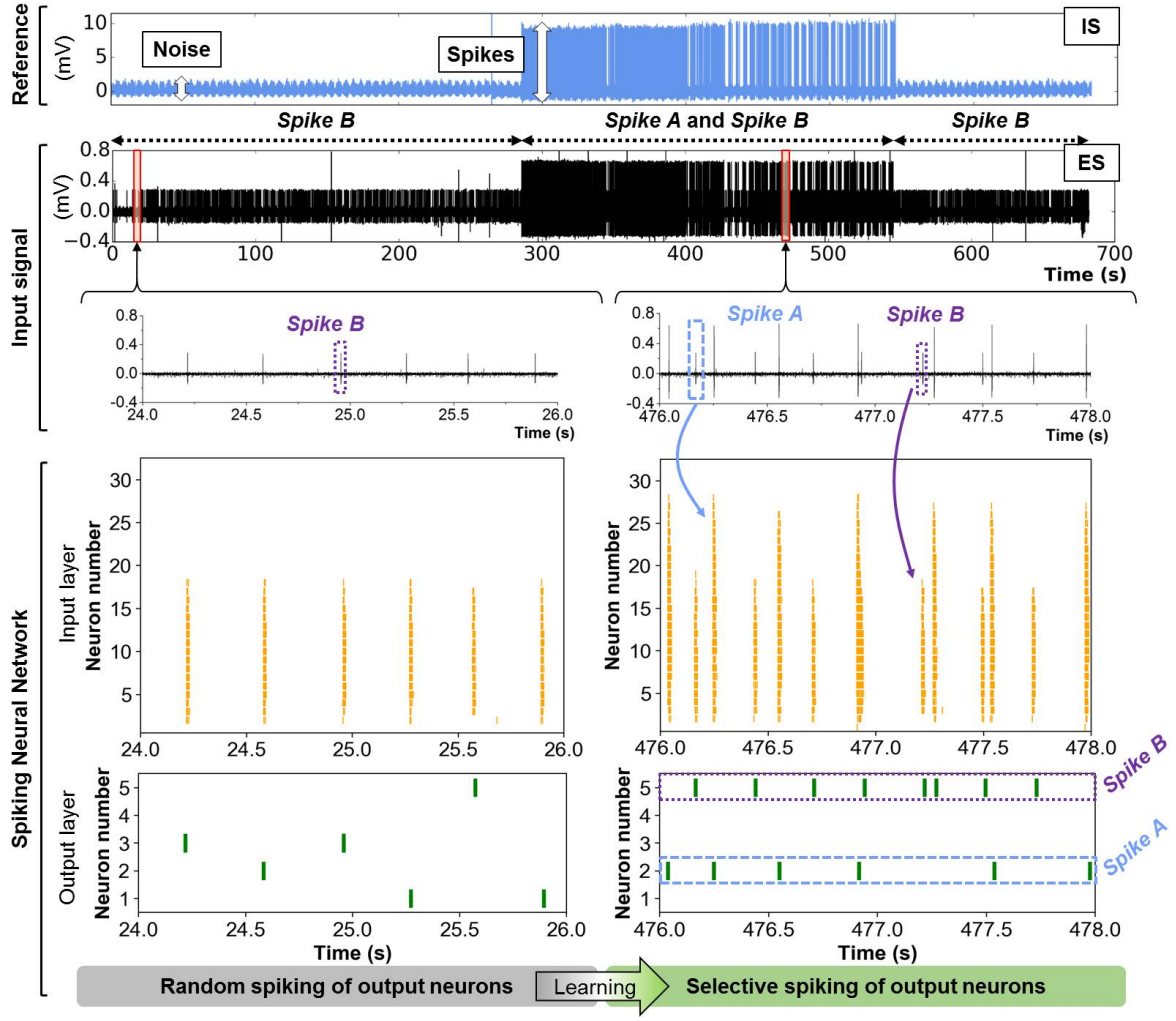


Figure 4.10: Schematic illustration of the learning phase for the SNN (see figure 4.5) applied on the biological data (see figure 4.9). Initially, the SNN is untrained for new input spikes (in the ES signal) and output neurons spike randomly. Due to online learning, different output neurons become gradually selective to certain input spike patterns.

is found to be correlated to Spike B, however, no ground truth (intracellular signal) is available for a reliable quantification of the recognition rate. N_3, N_4 and N_5 show very small activity meaning that they do not become selective to input spikes in the extracellular signal. These results prove the qualitative functionality of the proposed spike sorting algorithm. Note that, even if the frequency patterns of Spike A and Spike B are overlapping, two independent output neurons are assigned for the two different spikes. The delay between an observed biological spike and an output spike from the SNN is typically within $50ms$.

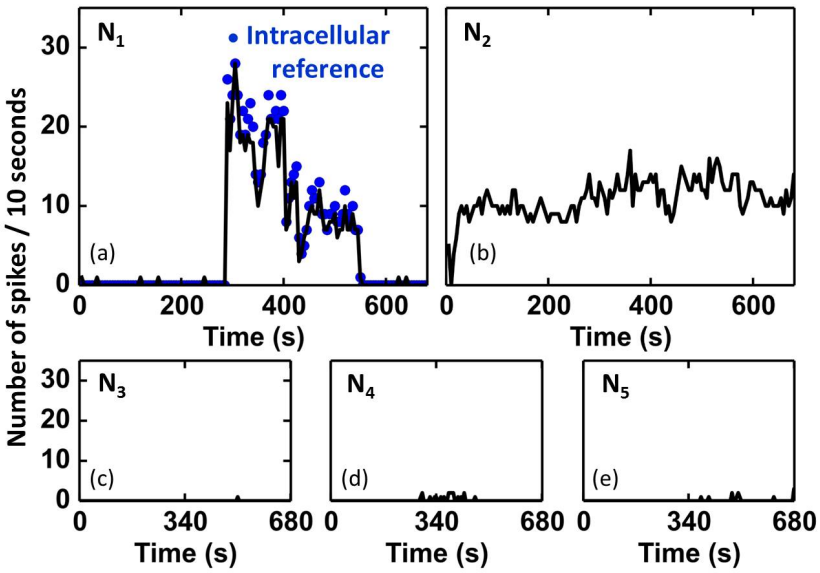


Figure 4.11: Activity of SNN output neurons during 681 s of continuous input signal. Activity is plotted as the number of spikes in time intervals of 10 seconds. N_1 activity matches well with the intracellular reference (blue dots), i.e. N_1 detects Spike A. N_2 seems to be selective to Spike B, however, no reference data is available for verification.

4.4.2 Reliability

In order to quantify the recognition rate of Spike A (figure 4.11), we correlated the activity of N_1 with the intracellular signal (IS in figure 4.10). A Spike A event is considered to be recognised if N_1 spikes within 50 ms after the Spike A event. The recognition rate was calculated as the ratio of recognized spikes to the total number of Spike A events (truth from IS data) in a given time interval (fixed to ten seconds). As shown in figure 4.12, recognition rate starts at 0 because Spike A only appears after 285s in the data. The network is associating one neuron quickly and reaches a mean spike recognition rate of up to 85.5 % after 15 seconds (corresponding to approximately 50 Spike A events), calculated starting from the first occurrence of Spike A in the ES signal at $t = 285$ s. Note that the recognition rate is fluctuating which means that the learning of the network is not completely stable. This may be due to the fact that the STDP learning rule is permanently active. However, this is crucial because otherwise the network would be prevented from learning spike shapes that occur in the data after STDP was disabled.

4.4.3 Power consumption

Table 6.1 summarizes the statistics of the SNN for the application on the ES data used here. The total duration of the signal is 681 s and the activity of all neuronal and synaptic events was recorded. Note that the average number of set and reset events per OxRAM device is very small, 17 and 37, respectively. This means that the SNN learning is fast and rather stable and

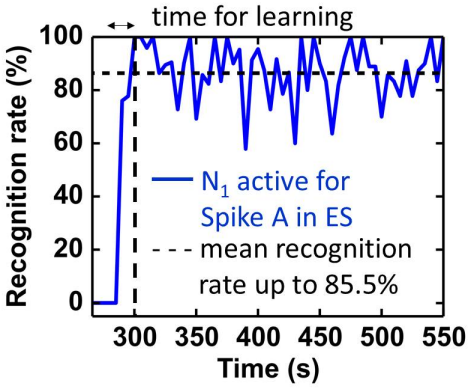


Figure 4.12: Temporal evolution of recognition rate of Spike A by N_1. A mean recognition rate of 86.4% (dashed line) is reached within 15 seconds starting from the first Spike A occurence.

OxRAM device degradation can be neglected. Furthermore, extrapolation of these statistics to an application time of 10 years, accounts to $8x10^6$ Set and $1.7x10^7$ Reset events per OxRAM device. These cycling requirements are satisfied by state-of-the-art OxRAM technologies [171].

We estimated the specific energy dissipation for a single synaptic event in our SNN by considering the pre-defined operation conditions for the OxRAM devices according to:

$$E_{mode} = V_{mode} \cdot I_{mode} \cdot t_{mode} \tag{4.5}$$

where the index $mode = [Set, Reset, Read]$ denotes the type of synaptic event, i.e. Set, Reset or Read operation. V_{mode}, I_{mode} and t_{mode} are the respective values for the voltage, current and time of the applied pulse. For Set and Reset, the pulse conditions reported in figure 3.19 (b) were used. For the Read operation, $V_{Read} = 0.1V$ and $t_{Read} = 1\mu s$ whereas I_{Read} is determined by the device resistance. Based on the statistics reported in table 6.1 and the event specific energies, the total energy dissipation and corresponding power consumption $P = E/t$ of the synaptic part of the

Table 4.2: Spiking Neural Network (SNN) statistics.

Input signal duration	681 s
Network statistics	
Number of synapses	160
Devices/synapse	10
Read events	16,235,500
Set events	27,467
Reset events	58,577
Number of spikes	329178

SNN are calculated following to

$$E_{total} = \sum_{mode} E_{mode} \cdot N_{mode} \tag{4.6}$$

whereas N_{mode} is the number of Set, Reset or Read events. The estimated energy consumptions of the synaptic part of the SNN are reported in table 4.3. The event specific energies in the low pJ range in combination with the relatively low number of switching events, result in extremely low synaptic power consumption of 8.1 nW. Considering a state-of-the-art analog neuron design in the 65nm technology node [231] with an energy per spike of 2 pJ may add 0.66 μJ (i.e. 5.6 %) to the total energy dissipation. Hence, the power consumption remains at a very low competitive level of 8.6 nW.

4.4.4 Versatility

We tested our spike sorting SNN with respect to its applicability on other neural spiking data. Therefore we used another dataset recorded in-vitro from Crayfish and a dataset recorded from anesthetized (in-vivo) rat hippocampus (publicly available online provided by the Buszaki lab [24][88]). Both datasets feature simultaneous recording of extra- and intracellular signals and are in the following referred to as CF2 and B1, respectively. As before in the case of CF1, we use the intracellular recording as a ground truth for the quantification of the recognition rate of the SNN output. CF2 is much more complex with respect to CF1 since it contains more different spike shapes and a higher overall spiking frequency which results in overlapping spikes. B1 comprises a strongly increased background noise level with respect to CF1. Snapshots of both datasets are shown in figure 4.13. The recognition rates for CF2 can be up to 74.2 % and 82.1 % for B1 after learning. However, as the recognition rate is defined as the ratio of detected spikes to true spikes in the data, it does not take into account false positive events, i.e. output spikes in the absence of the corresponding input event. Therefore, table 4.4 reports the recognition rate (RR), false negatives (FN), false positives (FP) and the corresponding FN and FP rates. Moreover, F1 being a general reliability metric was calculated by $F1 = 2 * TP/(2 * TP + FN + FP)$

Table 4.3: Spiking Neural Network (SNN) power metrics.

Energies per event	
Set event (E_{Set})	75 pJ
Reset event (E_{Reset})	45 pJ
Read event (E_{Read})	0.39 pJ
Total power estimation	
Energy dissipation	11 μJ
Power consumption	8.1 nW

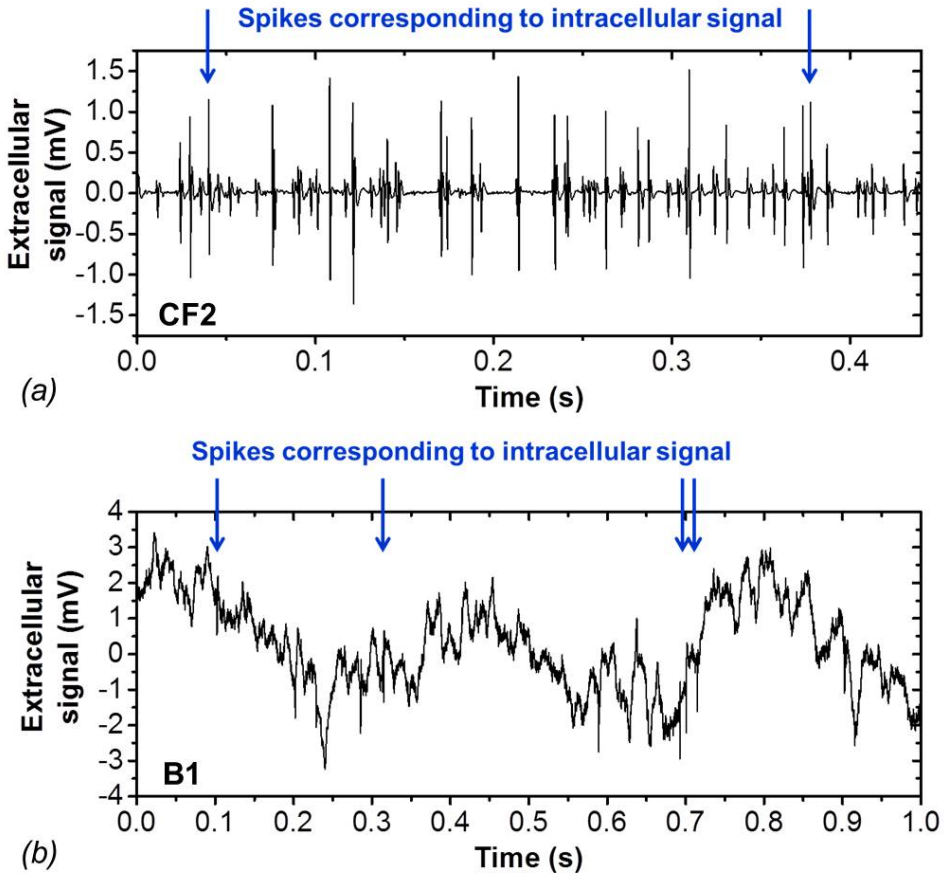


Figure 4.13: Sequences of real biological spiking data used for verification of Spike Sorting system, recorded in (a) in-vitro crayfish [23] and (b) in-vivo implanted rat hippocampus [24]. Intracellular recordings were simultaneously obtained and provide the ground truth for valid quantification of the spike recognition rate for the labeled spikes (blue arrows).

Table 4.4: Quantitative evaluation of spike sorting on different biological datasets.

	RR (%)	FN (#)	FP (#)	F1	FN (%)	FP (%)
CF1	80.5	142	50	0.86	19.5	6.9
CF2	50	331	1817	0.24	50	274
B1	64.8	299	519	0.58	35.3	61.2

and is reported in table 4.4. Note that all metrics are calculated based on the entire datasets, i.e. the learning period is included resulting in lower performance on average. It is clear, that the spike sorting network works rather well for CF1 with an F1 around 0.85. However, F1 drops significantly for both CF2 and B1 which is mainly a problem of the high number of FP. Thus, even if the RR may be considered acceptable for this first proof of concept, the true reliability based on F1 is much lower. These results show that the STDP learning rule allows some tolerance to other

datasets but the proposed network has to be significantly improved and/or fine tuned to suit other spiking data sets. This can be attributed to the data encoding approach by the band-pass filters. Apparently, the filtering allows to reduce the low-frequency fluctuation of the signal and thus improves the detection of spikes. On the other hand, CF2 contains many spikes with spike shapes (and amplitudes) that are quite similar from one to another and the intervals between two consecutive spikes are much shorter with respect to CF1 which was used to optimize the network parameters. The former property of CF2 may lead to an increased number of FN because of too long time constants for the refractory ($t_{refractory}$) and inhibitory ($T_{inbitory}$) periods of the neurons. The latter demonstrates that the data encoding is not precise enough to be able to distinguish between relatively similar spike waveforms.

State-of-the-art spike sorting algorithms based on spike detection, feature extraction and clustering (i.e. standard methodology) achieve recognition rates around 90 % on the dataset B1 [233][234] and therefore outperform the proposed approach in terms of reliability.

4.4.5 Qualitative comparison to standard spike sorting techniques

The alternative spike sorting approach based on a SNN has been qualitatively compared with the standard methodologies (template matching, principal component analysis) in table 4.5. The advantage of our approach is clearly the real-time functionality without the need for supervision as well as the computational efficiency which results in very low power consumption. These benefits may enable our approach to be suitable for rather simple hardware implementation for long-time, portable and low-power implants whereas standard spike sorting techniques do not meet these requirements. On the other hand, the spike sorting accuracy is considerably lower with respect to standard techniques. This issue should be addressed by a more sophisticated network (e.g. more neuron layers, better data encoding etc.).

Table 4.5: Qualitative comparison of Spike-Timing Depending Plasticity (STDP) based Spike sorting (this work) with standard approaches (template matching, PCA).

Criterion	STDP based (this work)	Standard techniques
Real-time functionality (permanent adaptation to spikes shapes)	+	-
Unsupervised operation	+	-
Computational efficiency	+	-
Energy efficiency	+	-
Reliability	-	+
Suitability for (long-term) hardware integration	+	-

4.5 Summary

An alternative approach towards unsupervised spike sorting of brain activity signals, relevant for the analysis of large-scale brain signals, was proposed in this chapter. Since the approach is compatible for low-power applications and hardware integration, it bears a high potential for embedded Spiking Neural Networks (SNN) used in spike sorting applications. It was shown that these systems allow for fast adaptation to new input data and completely unsupervised operation, independently from the number of spikes in the input signal, yielding a good reliability on rather easy signals. However, as the network was tested on complex sets of real biological spiking data without parameter tuning, the reliability of the spike sorting system could not be proven to be sufficiently high. Spike sorting performances are considerably lower with respect to conventional power-hungry spike sorting methodologies. In order to improve this critical issue, both the data encoding algorithm (by band-pass filtering) and the classification approach (by using a SNN) have to be drastically improved, i.e. complementary input information and/or more sophisticated SNN architectures have to be developed. Nevertheless, it should be emphasized that in contrast to standard spike sorting techniques, SNN based approaches offer several advantages, e.g. no power-consuming CPU or GPU are needed and no parameters (e.g. threshold level for spike detection) have to be optimized manually as a function of the input data. Hence, SNN's offer a powerful alternative to standard spike sorting methodologies. Moreover, we proposed OxRAM technology for the hardware implementation of synapses with ultra-low power consumption and fast operation times ($< 1\ \mu s$). This enables the system for real-time application to neural data in potential medical devices featuring high energy-efficiencies. Extended OxRAM cycling capabilities ($> 10^8$ switching cycles) ensure that the SNN retains its learning capability throughout the application lifetime and thus allows for long-term functional implants. We believe that compact hardware implementations of SNN's will enable spike sorting directly at the recording site within the brain thus solving the bottleneck of data storage and power consumption. Finally, data reduction rates of about 1000 (depending on the spiking frequency of the input data) open the path to wireless data streaming of the spike sorted data to an external receiver.

CHAPTER 5

SYNAPTIC VARIABILITY IN SPIKING NEURAL NETWORKS

Neurological research has demonstrated that variability in neural systems exist both for synapses and neurons whereas it is believed that synaptic variability dominates [37] and it was shown that synaptic variability may be beneficial for reliable spike firing [235]. Moreover, it was found that variability improves the performance of extreme learning machines [236]. An inherent property of RRAM technology is variability, both in terms of switching success as well as for the reproducibility of exact resistance values. However, the specific effects of these reliability issues of RRAM devices on Spiking Neural Network (SNN) are yet to be understood. Therefore, the impact of RRAM variability on Spiking Neural Networks (SNN) using on-line (unsupervised) learning, where the RRAM resistance status (synaptic weight) is tuned in-situ using probabilistic Spike-Timing-Dependent-Plasticity (STDP) [186], was studied. Two Spiking Neural Networks (SNN) using OxRAM based synapses were used for a systematic study of the impact of synaptic variability on the application reliability.

Section 5.1 explains the requirements on the OxRAM programming conditions for synapses used in SNN. Section 5.2 and 5.3 demonstrate the effects of synaptic variability on the performance of classification and detection tasks. The results are summarized in section 5.4.

5.1 Artificial synapse implementation with RRAM technology

RRAM is a promising technology for next generation non-volatile memories replacing Flash technology as described in chapter 3. Different implementation concepts have been explained in literature and in this work.

5.1.1 OxRAM operation for synapses

Neuromorphic hardware implementations impose a number of requirements on the electrical properties of its individual components, e.g. the synapses. The brain-inspired computing approach is inherently low-power with respect to the conventional von-Neumann architecture. In order to harness and further improve this advantage in non-von-Neumann hardware implementations, the energy used to switch a synapse from one synaptic weight to another (i.e. synaptic plasticity) has to be minimized. Furthermore, the current through a synapse triggered by the propagation of a pre-synaptic spike is proportional to its conductance and is integrated by the connected post-synaptic neuron which receives many more input synapses adding up their currents. This additive behaviour of a neuron leads to large input currents that have to be sustained. For this reason and the fact that most of the power of neural networks is typically consumed in read mode, the electrical resistance of a synapse shall be rather high in order to reduce power consumption and post-synaptic currents.

When using RRAM based synapses, both the program and read energies can be reduced by decreasing the I_{CC} as shown in figure 5.1. The tested electrical devices are described in detail in chapter 3. However, a drawback of using a very low I_{CC} is the increased variability which essentially prevents to predict the exact resistance value upon programming. The widening distributions for LRS and HRS result furthermore in a reduced resistance memory window (MW), defined as the separation gap between LRS and HRS, e.g. $LRS_{+3\sigma}$ and $HRS_{-3\sigma}$, depicted in figure 5.2. This is a critical problem for common memory applications where a clear separation between LRS and HRS is necessary in order to determine the memory state reliably in a read operation. In order to increase the MW, HRS can be slightly shifted up by using higher reset voltages (V_R), though this induces an enhanced degradation in the oxide due to the stronger electrical fields. This leads to much higher device failure rates with respect to lower V_R as

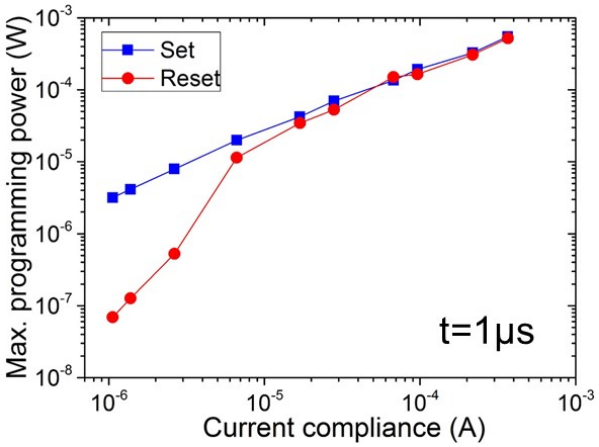


Figure 5.1: Estimation of maximum programming power of OxRAM ($1nm\ Al_2O_3/3nm\ HfO_2$) as a function of the current compliance.

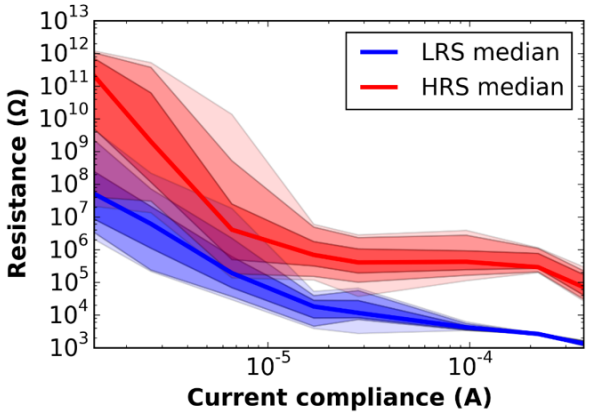


Figure 5.2: Experimental LRS and HRS distributions as a function of the current compliance (CC) for 1T1R OxRAM devices ($1nm\ Al_2O_3/3nm\ HfO_2$). Set and reset voltages were $2.5V$ and $-1.2V$ The bold lines mark $Median$ for both LRS and HRS, the shaded areas include 95% of the samples, i.e. reflect the distribution at 2σ.

described in section 3.2.5. As demonstrated, more than 10^9 cycles can be achieved using an optimized reset condition of $V_R = -1.2V$.

Good endurance is a fundamental requirement for networks using STDP learning, especially when the input data is not known a priori and therefore the network is in permanent learning mode which can rapidly increase the number of set and reset events (cycle number) in order to optimize the synaptic weights. Typical OxRAM device failures such as 'stuck-in-LRS' are detrimental and have to be avoided because those cells would generate very high synaptic weights which can disturb the network (learning) because they produce a permanent strong input current to other neurons.

5.2 Effects of synaptic variability on SNN in Classification tasks

The spiking neural network that was previously developed for spike sorting, was used to study the impact of synaptic variability on its performance, using the simple crayfish data. A detailed description of the encoding and the network can be found in chapter 4.

Pulsed endurance measurements were performed on three OxRAM stacks ($5nm\ HfO_2$, $1nm\ Al_2O_3/3nm\ HfO_2$, $5nm\ HfO_2/4nm\ TaO_x$, Ti and TiN top and bottom electrodes) with three different programming currents (30 μA, 85 μA, 135 μA). By applying pulse voltages of $2.5V$ and $-1.2V$ for Set and Reset, $10-15$ OxRAM devices were cycled up to 10^8 times, thus accounting for both device-to-device and cycle-to-cycle variabilities. Independently of the programming current and OxRAM material, a switching functionality for 10^8 cycles could be achieved. The experimental data are explained in more detail in chapter 3. As shown in figure

3.19 (a), the HfO_2 cycled using a $I_{CC} = 135\ \mu A$ exhibits two separated distributions for LRS and HRS enabling a clear memory window (MW). When reducing the I_{CC} to 30 μA, both LRS and HRS distributions expand and overlap each other, hence the MW vanishes (as shown in figure 5.3 (c) for Al_2O_3/HfO_2). For each I_{CC} and oxide material, the experimental distributions of LRS and HRS were used to extract the memory window (MW) as $MW = Median_{HRS}/Median_{LRS}$ as well as the variability $\sigma_{LRS,HRS} = \pm 1\sigma$ (see figures 5.3 (b) and (d)). Figures 5.4 (a) and (b) report the MW and $\sigma_{LRS,HRS}$ as functions of the programming current I_{CC}, respectively. When I_{CC} is increased, the MW can be enhanced and the variability $\sigma_{LRS,HRS}$ is drastically reduced at the cost of a higher power consumption of the OxRAM device. While the biggest memory window can be achieved with HfO_2, the variability is the smallest for Al_2O_3/HfO_2, in particular for low I_{CC} (see figure 5.4 (b)). Therefore, two potential conditions are chosen to study the impact of both MW and variability on the performance of the Spiking Sorting application, namely HfO_2 using $I_{CC} = 135\ \mu A$ and Al_2O_3/HfO_2 using $I_{CC} = 30\ \mu A$, hereafter called High Current OxRAM (HCO) and Low Current OxRAM (LCO) devices, respectively.

The LRS and HRS distributions for the two selected synapse devices, LCO and HCO, were normalized to equal mean values for LRS and HRS in order to study only the impact of the variability on the network performance. A third condition was artificially created by using the same mean values for LRS and HRS fixing both variabilities to $\sigma_{LRS,HRS} = 0$, i.e. a purely digital switching without any cycle-to-cycle and device-to-device variability. The three conditions are shown in figure 5.5 whereas $C1$ is the artificially created condition serving as a reference and C2 and C3 correspond to the experimental variabilities observed for the HCO and LCO.

5.2.1 Reliability

The number of devices-per-synapse (n) was varied from 1 to 100 for the three OxRAM conditions. The recognition rate (RR) was quantified both for Spike A and Spike B as the ratio of detected spikes to true spikes. The RR is presented in figure 5.6. Apparently, a very low n, for example using only $n = 1$ devices per synapse does not allow the SNN to achieve a good RR for the two spike classes. However, RR can be strongly improved up to around $N = 10$ while it increases only slightly for $N > 10$. Interestingly, for $N < 10$, $C1$ achieves the best Recognition Rate with respect to $C2$ and $C3$. This trend is inversed for $N > 10$. Since the variability of both LRS and HRS increases from $C1$ to $C2$ to $C3$, it seems that for the implementation of synapses affected by variability, two regimes of the network accuracy exist as a function of n which is separated by a critical number of devices per synapse n_{crit}: $n < n_{crit}$, where synaptic variability degrades the network performance and $n > n_{crit}$, where synaptic variability enhances the network performance.

In order to better understand the effect of variability in this application, the recognition rates for the two spike classes (Spike A and B) are plotted separately. As it is shown in figure 5.6 (a), the RR corresponding to the class of Spike A increases with n up to approximately 80%, independently from the OxRAM condition. On the other hand, figure 5.6 (b) shows that the RR for

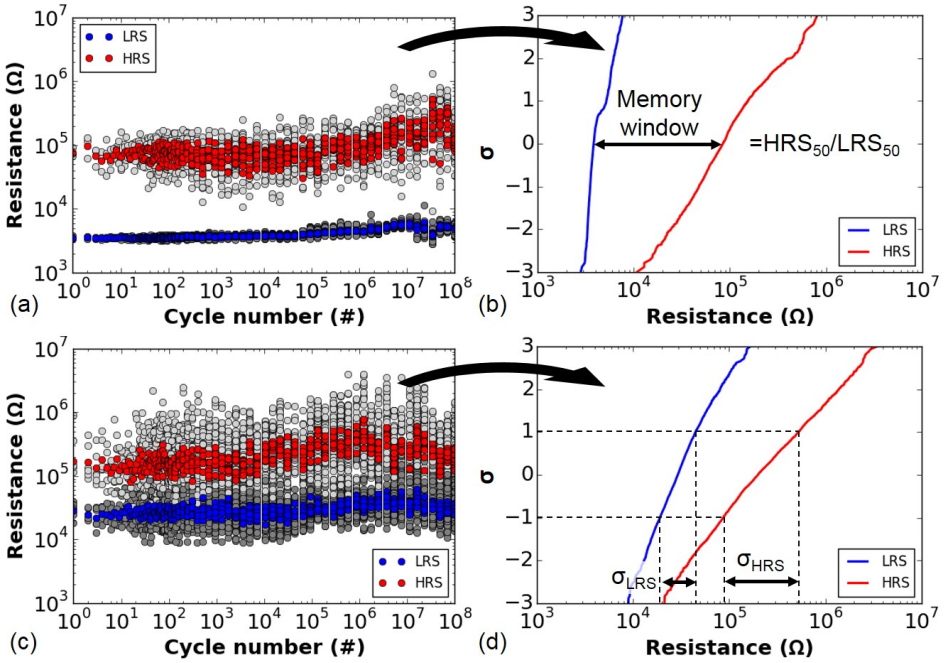


Figure 5.3: (a) HfO_2 endurance test using a $I_{CC} = 135\mu A$ and (b) extracted σ for LRS and HRS. (c) Al_2O_3/HfO_2 endurance test using a $I_{CC} = 30\mu A$ and (d) extracted σ for LRS and HRS. The extraction of the memory window (MW) and variabilities σ_{LRS} and σ_{HRS} is illustrated in (b) and (d).

Spike B improves only up to around 75% for $C1$ and $C2$ while $C3$ is able to reach a significantly higher RR of around 75%.

The contour plots in figure 5.7 represent the overall recognition rate as a function of LRS and HRS variability depending on the number of devices per synapse $n = [1,5,10,20,50,100]$. For $1 < n \leq 10$ (5.7) (a) - (c), the overall RR is rather low and decreases if the variability of LRS and/or HRS is increased. As n is increased, i.e. for $1 < n \leq 10$ (5.7) (d) - (f), the RR is globally increased as mentioned before and moreover, the RR increases as the LRS variability increases. On the other hand, the RR seems to decrease for an increased HRS variability.

Synapses implemented with binary RRAM devices, are usually dominated by the devices which are in LRS because the LRS is considerably higher than HRS. Hence, the synaptic weight depends mainly on the sum of conductances corresponding to LRS devices while the conductances of HRS devices differ by an order of magnitude typically and therefore play a minor role. The finding that variability introduced to the synaptic weight by single RRAM devices improves the performance of Spiking Neural Networks seems counter-intuitive and is therefore studied in more detail. In order to verify the results and shed more light on the effect of synaptic variability on the learning and functionality of Spiking Neural Networks in more detail, another application has been systematically studied under the impact of variability in section 5.3. This application aims to extract visual patterns from encoded video data with a much higher number of neurons

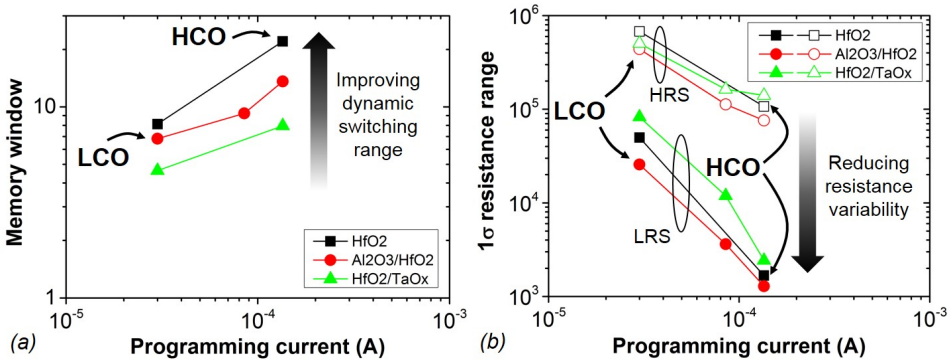


Figure 5.4: (a) Median-to-median memory window (MW) for the three tested OxRAM materials as a function on the I_{CC}. (b) Resistance variability $\sigma_{LRS,HRS}$ of the three tested OxRAM materials depending on the PC. Two device approaches are chosen as indicated in the graphs: Low Current OxRAM ('LCO', Al_2O_3/HfO_2, PC = 30μA) and High Current OxRAM ('HCO', HfO_2, PC = 135μA).

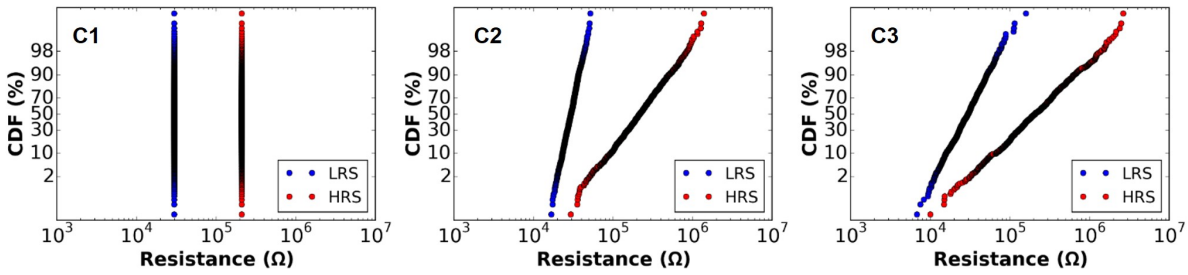


Figure 5.5: LRS and HRS distributions of test conditions for SNN of Fig.10.

and synapses and moreover 6 different classes has to be identified by the SNN [15]. Thus, this network allows to draw more general conclusions with respect to the very small SNN used for Spike Sorting.

5.3 Effects of synaptic variability on SNN in Detection tasks

The Internet-of-Things era bears numerous opportunities for smart systems which offer some autonomy, both in terms of functionality and energy consumption. In the application presented here, cars passing on a motorway have to be detected without any user intervention. The number and position of lanes thereby is not known a priori. This requires a means of autonomous lane detection and distinction as well as an adaptation of the system to detect single cars passing on these lanes. For this reason, a SNN was previously developed featuring Spike-Timing-Dependent Plasticity (STDP) which provides online, unsupervised learning, see figure 6.18 for a schematic illustration of the SNN based application. The concept was previously demonstrated in detail by Bichler et al. [15] exploiting multi-level Phase-Change Memory [147] and binary Conductive

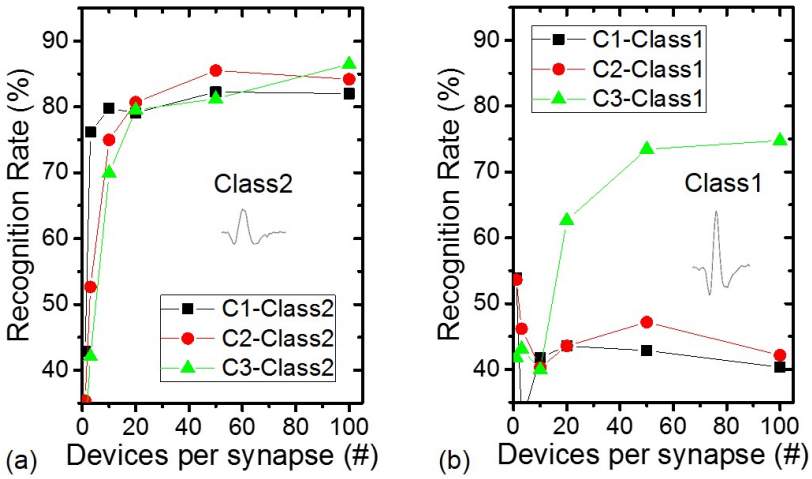


Figure 5.6: Overall Recognition Rate of spike sorting SNN as a function of number of devices per synapses and for different conditions $C1-C3$.

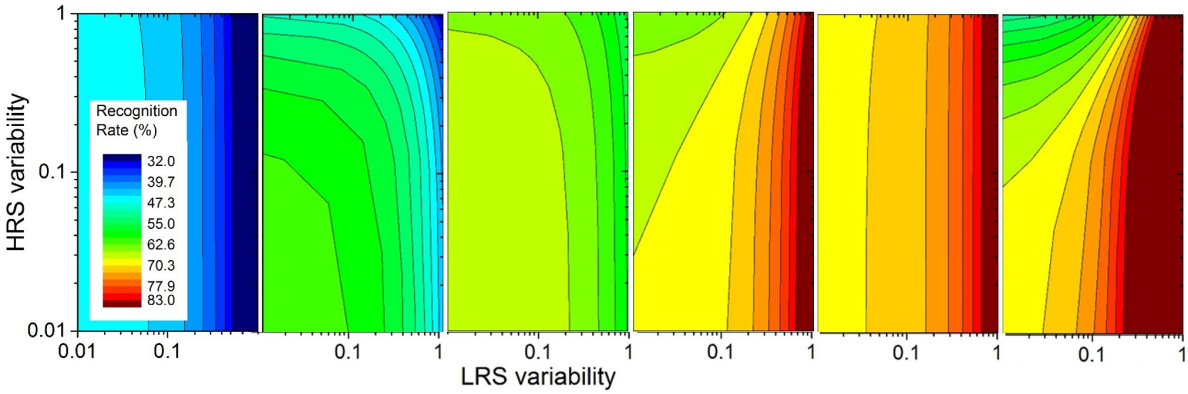


Figure 5.7: Recognition rate of SNN used for neural spike classification as a function of LRS and HRS variability. Synaptic redundancy accounts to (a) 1, (b) 5, (c) 10, (d) 20, (e) 50 and (f) 100. Note that these results were obtained for using the same set of parameters, i.e. neuron threshold etc.

Bridge RRAM synapses [17].

A video of cars passing on a six-lane freeway (Pasadena, CA) is recorded by a Dynamic Vision Sensor (DVS) with $128x128$ pixels. The DVS is a retina-inspired sensor where each pixel features two sub-pixels which detect a luminosity increase or decrease of a small section of the picture, respectively. The camera featuring the DVS records data in the Address Event Representation (AER) format [237]. The AER data is then presented to a two-layered FCNN consisting of $1.97M$ synaptic connections ($128*128*2*60$). The unsupervised learning thanks to STDP and the winner-take-all (WTA) principle enable the output neurons to become sensitive to different traffic lanes. In order to validate and quantify the SNN performance, the SNN output activity is compared to the manually labelled reference. Figure 5.9 illustrates the extraction of the number

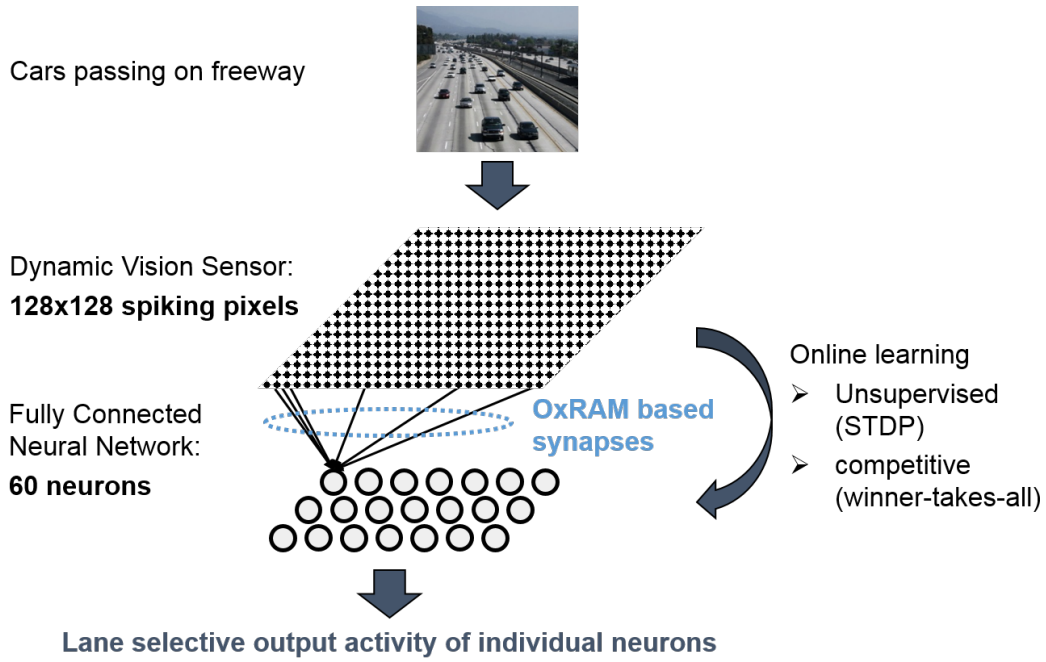


Figure 5.8: Two-layer Spiking Neural Network used for unsupervised detection of cars in different traffic lanes.

of True Positive (TP), False Negatives (FN) and False Positives (FP) from the reference activity (blue) and the output neuron activity of the corresponding traffic lane (red). A TP means that a car was accurately detected, while it was missed in the case of a FN. The FP means that a car was detected even if no car was present at the specific traffic lane. This event is likely due to background noise or other effects specific to this application, i.e. cars crossing lanes which may lead to double detection.

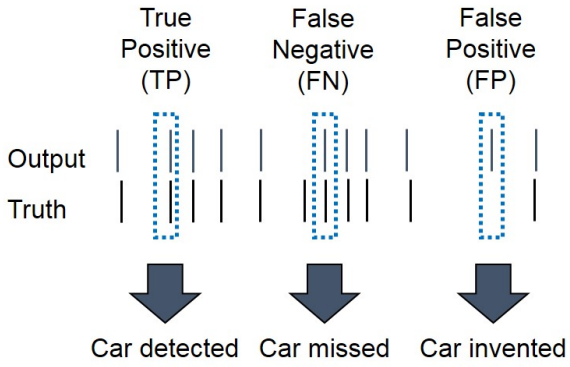


Figure 5.9: The reference activity (blue) is compared to the activity of the corresponding output neurons (red) to calculate the number of True Positives (TP), False Negatives (FN) and False Positives (FP).

To study the impact of the synaptic variations induced by the characteristic OxRAM variability of both LRS and HRS, the synapses were modelled assuming different OxRAM test cases. Therefore, a number of various combinations of LRS and HRS distributions were defined, summarized in table 5.1.

The SNN parameters, e.g. for the neurons and learning rule, were calibrated using a genetic algorithm similar to the one described in chapter 4 using the OxRAM test condition $C2$. This condition was experimentally obtained by cycling several HfO_2 devices using a current compliance of $I_{CC} = 200\mu A$. The remaining conditions ($C1$, $C3 - C9$) were simulated using the same network parameters while changing only the parameters of the synapses related to the OxRAM variability.

5.3.1 Reliability

The performance of the SNN simulations was quantified by extracting the number of False Negatives (FN), False Positives (FP) as well as the F1 score, described in section 5.3. FN, FP and $F1$ averaged over the six lanes are presented in figure 5.10 for the simulated OxRAM operations conditions $C1$ to $C9$. The FN seems to be reduced significantly by the conditions $C1$, $C4$ and $C7$ while the FP are slightly increased. F1, taking into account both FN and FP, is indeed high for all conditions but the highest F1 scores can be achieved by $C1$, $C4$ and $C7$. These conditions have

Table 5.1: LRS and HRS test conditions for OxRAM based synapses used to simulate Spiking Neural Network for visual signal processing.

		σ_{LRS}		
		-100	-2.14	-0.326
σ_{HRS}	-100	C3	C6	C7
	-2.14	C8	C5	C4
	-0.326	C9	C2	C1

the highest variability of the LRS in common.

In order to understand the differences in the performance depending on the OxRAM specific variability, it is necessary to investigate the network performance in detail for the single classes of events, i.e. the six single lanes. As shown in figure 5.11 (a), the conditions $C1$, $C4$ and $C7$ achieve a significantly lower number of False Negatives (FN) with respect to the other conditions on lanes 1 and 6. Lanes 2 to 5 are approximately the same for all conditions in terms of the FN. On the other hand, the number of False Positives (FP) is slightly higher for $C1$, $C4$ and $C7$ in comparison to the other conditions, especially on lane 3 and 4, as shown in figure 5.11 (b). This gives rise to the assumption that the increased LRS variability enhances the network to detect cars on the outermost lanes (1 and 6) while a few more cars are 'invented' throughout the central lanes. To draw a proper conclusion, considering both FN and FP, the F1 score was calculated and shown in figure 5.11 (c). It is clear that F1 is very similar for the lanes 2 to 5 whereas lane 1 and 6 are significantly increased leading to the overall higher F1 score in figure 5.11 (c).

In order to understand the influence of OxRAM variability in a synapse, the detection rate was plotted as function of the LRS and HRS variabilities in the contour graph in figure 5.12. Whereas the performance of the SNN seems to improve with an increased LRS variability, the HRS variability appears to decrease it.

Typically, the car detection rate of the SNN is reduced for the traffic lanes 1 and 6 [15] [17], since these are at the edge of the AER sensor, thus, cars appear smaller which results in less input activity from the 2-dimensional AER sensor. Here, it seems that especially this shortcoming can be compensated by introducing synaptic variability by means of LRS variability. This may be explained as following. A car is detected by the SNN when a neuron of the first layer emits a spike which happens whenever the integration threshold of a neuron is crossed. The threshold for a spike emission was constant for all neurons of the same layer in our simulations and moreover, the same threshold value was used for all OxRAM conditions $C1 - C9$. Hence, the lower input activity may prevent the neurons of the SNN associated with lane 1 and 6 to reach the threshold for spiking, i.e. they fail to detect a passing car. The conditions $C1$, $C4$ and $C7$ feature the highest

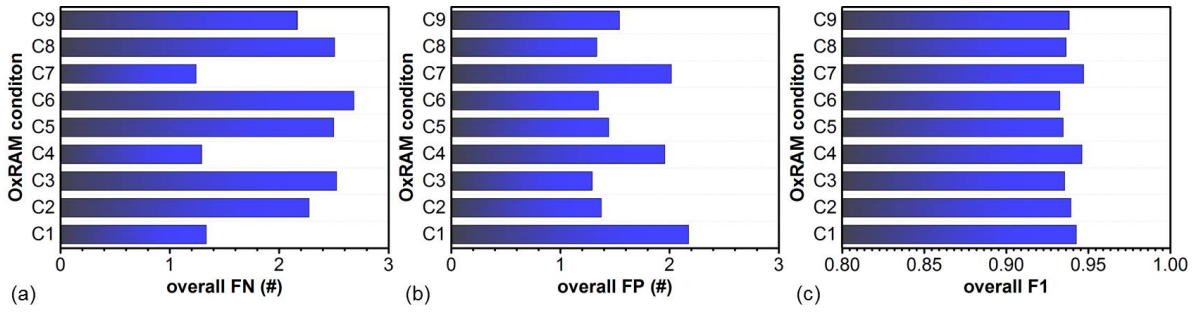


Figure 5.10: (a) False Negatives (FN), (b) False Positives (FP) and (c) F1 score for the different OxRAM conditions (see figure 5.10). All numbers are averaged over the six traffic lanes. Note that FN and FP shall be as low as possible while F1 has to be maximized (i.e. converge to 1).

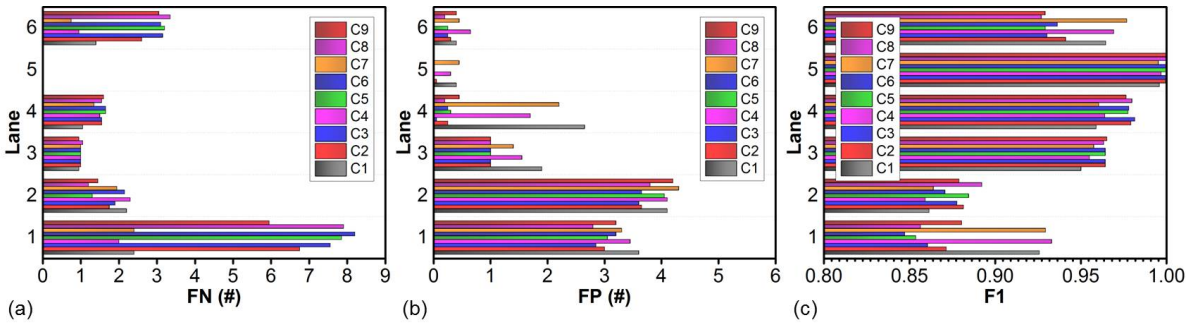


Figure 5.11: (a) False Negatives (FN), (b) False Positives (FP) and (c) F1 score for the different OxRAM conditions (see figure 5.10) for the six traffic lanes. Note that FN and FP shall be as low as possible while F1 has to be maximized (i.e. converge to 1).

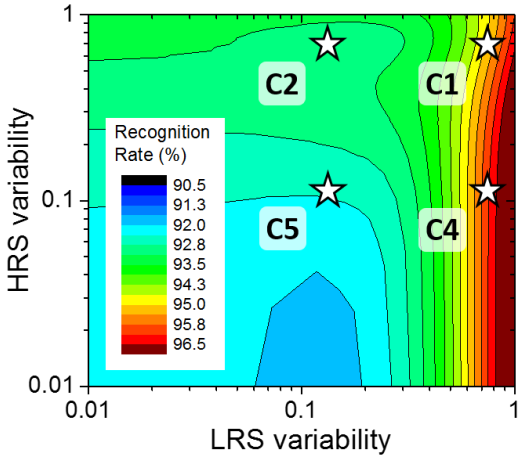


Figure 5.12: Recognition rate of car detection SNN (figure 5.8) as a function of LRS and HRS variability of the OxRAM devices used for the implementation of the SNN synapses.

variability in LRS, this allows to achieve relatively high device conductances (very low resistances at the distribution tails) with respect to the conditions of lower LRS variability. In our simulations, the mean values for the LRS and HRS distributions were constant across different conditions, hence, the higher the OxRAM device variability, the higher the final synaptic weights can be. The impact of the variability effect is shown in figure 5.13 for the OxRAM conditions $C4$ and $C5$. For those conditions, the neurons corresponding to lane 1 were identified and their integration potential was plotted as function of time for the same sequence of data. As it was previously explained, the number of non-detected cars (FN) is higher for $C5$ with respect to $C4$ on lane 1 and 6. The reason for this can be found in evaluating the integration trace over time. While the threshold for spiking is reached every time when a car is passing for $C4$, this threshold is missed two times for $C5$ (around $22s$ and $23s$).

Since the EPSP induced in a post-synaptic neuron (increasing the integration) depends on the individual synaptic weights which are typically dominated by the OxRAM cells of highest

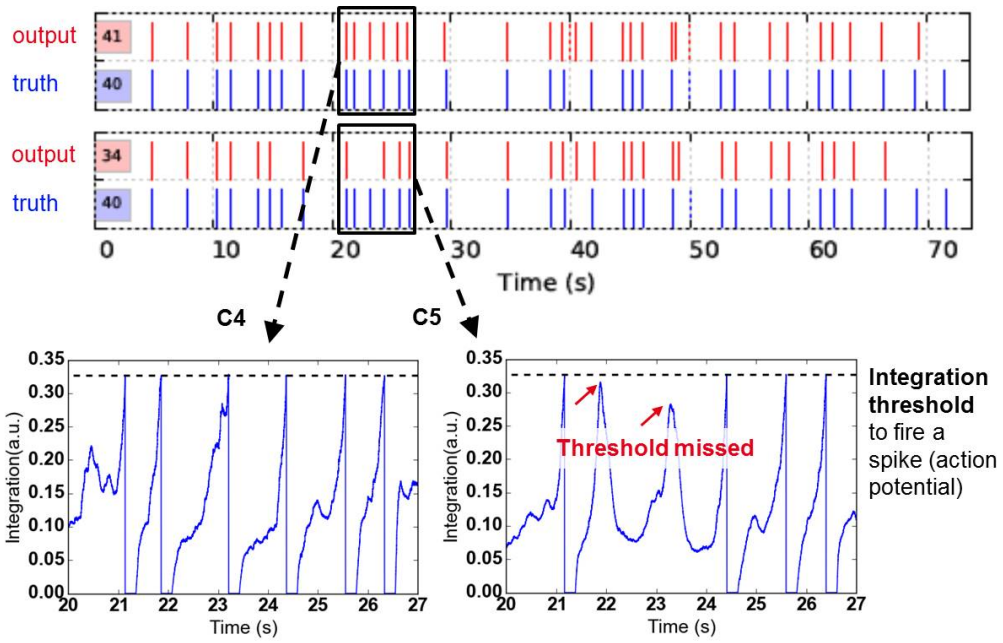


Figure 5.13: (top) The spike trains of the output neurons corresponding to lane 1 (red) of the SNN shown in figure 5.8 are compared with the reference (blue) for the OxRAM conditions C4 and C5. (bottom) The integrated membrane potential for the two neurons is shown. Every time a car passes (spike in the truth), the integration increases significantly and eventually reaches the threshold. This is true for all events using C4 but 2 events are missed for C5. Note that the increase of the integration is proportional to the input synaptic weights.

conductance, missing the threshold and thus not detecting a car can be certainly attributed to the lower synaptic weights when $C5$ is used instead of $C4$. The distributions of synaptic weights after 8 learning passes ($5500s$) is plotted in figure 5.14 for all programming conditions. $C4$ features apparently a wider distribution of synapses corresponding to the LRS, i.e. a few synapses exhibit weights being almost an order of magnitude higher than $C5$. Those are the synapses that enable the SNN to provide enough input current to the neurons to reach the threshold and to fire a spike. Note that the two distributions for depressed and potentiated synapses are usually separated except for the programming condition $C1$.

5.3.2 Threshold dependence

A possibility to overcome the problem of missing cars because of a low input activity may be to reduce the threshold so that a neuron can more easily fire a spike. For this reason, the threshold value was varied on a wide range and the simulation results were quantified. Figure 5.15 shows the number of FN, FP and the F1 score for the condition C4 and C5 as a function of the threshold of the LIF neurons in layer 1 of the SNN. Indeed, the number of FN can be reduced by using a lower threshold for C5, thus compensating the lower synaptic weights. There are optimum

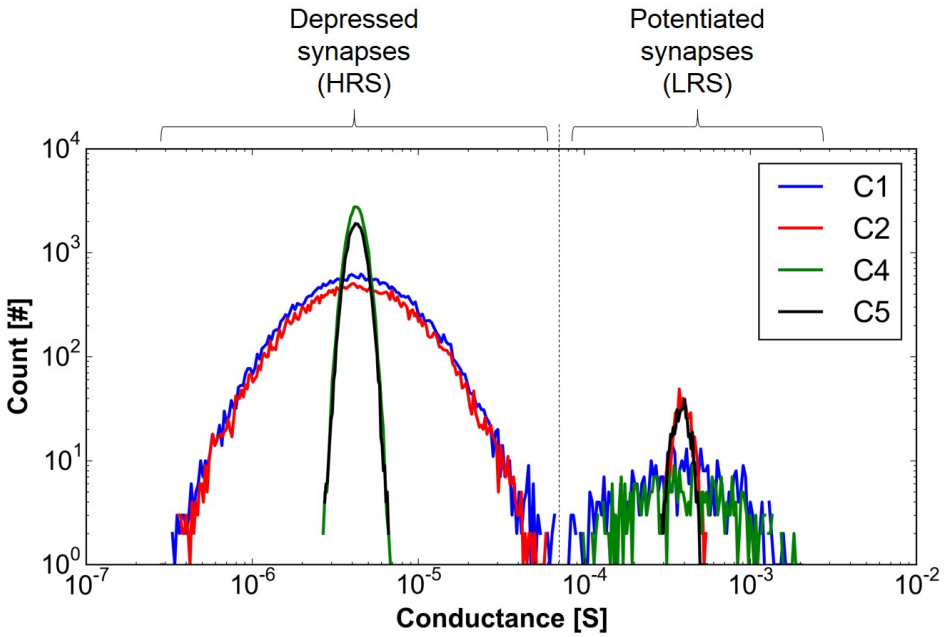


Figure 5.14: Synaptic weight distribution after sufficient learning period for C4 and C5. The majority of the OxRAM based synapses is depressed, i.e. in High Resistive State (HRS) while a small fraction is potentiated, i.e. in Low Resistive State (LRS). The synapses in LRS are the ones corresponding to relevant input information for a specific lane and allow to detect a passing car while the HRS synapses detect events outside the former lane. C4 bears a much wider distribution of LRS synapses with respect to a very sharp distribution for C5.

thresholds for both C4 and C5 in terms of the FN, while above this optimum, the number of FN increases strongly since neurons can no longer reach the threshold and on the other hand, below the threshold, noise starts to disturb the proper detection of cars. The FP increases also significantly when the threshold is reduced since the increased excitability of the neurons tends to lead to detection of noise rather than cars. Finally, F1 was computed and shown in figure 5.15 (c) which shows clearly that an optimum threshold exists for both conditions which maximizes the performance. It is worth noting that the performance parameters FN, FP and F1 show an asymmetric dependence on the threshold, i.e. using a lower threshold with respect to the optimum value causes a much stronger degradation of the performance than using a higher threshold. The absolute value of F1 is the same for different conditions which suggests that the synaptic variability does not critically affect the SNN precision. However, it is clear that the parameters of the SNN such as the threshold have to be precisely adjusted according to the synaptic device characteristics.

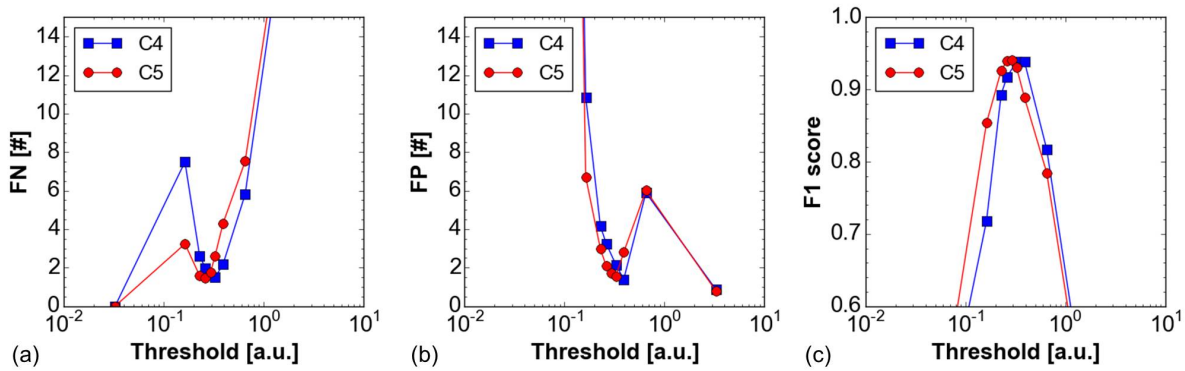


Figure 5.15: False Negative (FN), False Positives (FP) and F1 score as a function of the threshold of the LIF neurons for C4 and C5.

5.3.3 Memory window dependence

Variations of the LRS and HRS variability, introduced previously, change the resistance margin, also known as the memory window (MW). The MW characterizes the margin between the resistance distributions (LRS and HRS) and is accordingly calculated by

$$MW = \frac{R_{HRS,\,-3\sigma}}{R_{LRS,\,+3\sigma}} \tag{5.1}$$

Moreover, it seems to be important for the proper functionality of a Spiking Neural Network that potentiated synapses have a significantly higher weight than depressed synapses since the former are supposed to provide relevant input signals to a neuron while the latter usually provide input signals corresponding to uncorrelated or noisy events. For this reason, the impact of the MW on the reliability of the SNN application was studied. Since the simulated conditions featuring virtually no variability in one or both of the resistance states ($\sigma = -100$) are not possible to be

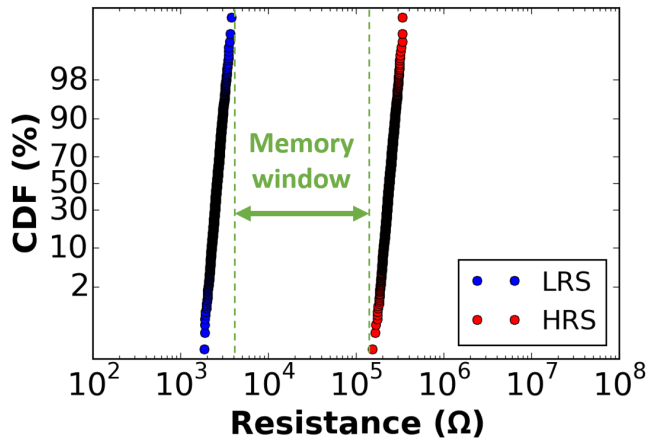


Figure 5.16: Schematic illustration on how to retrieve the MW from experimental distributions of LRS and HRS. The MW described the gap between those two distributions.

achieved by using OxRAM devices, only the conditions with significant variability are used in the following, i.e. C1, C2, C4 and C5. The MW was varied by changing either the mean value of the LRS or HRS distribution and the equivalent effect was verified by simulations, but not shown here.

As it was pointed out in section 5.3.2, the optimum neuron threshold depends strongly on the synaptic weights, thus on the programming conditions. If the MW is changed, the influence of synapses in HRS changes which can be observed in figure 5.17. Here, the threshold has to be increased for a shrinking MW. This is because lowering the MW results in HRS that are comparable to LRS while for high MW, the HRS is orders of magnitudes lower than LRS.

Figure 5.18 shows the quantified results for the FN, FP and F1 of the four conditions as a function of the MW. For high MW, the number of FN and FP are low resulting in a F1 score close to 1, i.e. a high accuracy of the SNN to detect single cars on specific lanes. On the other hand, both the number of FN and FP increase rapidly below a certain MW while this threshold memory varies from condition to condition. For this reason, the minimum MW providing a sufficient performance seems to depend on the LRS and HRS variabilities. It is interesting to note that the lowest MW can be used with C1 which features the highest variabilities among the tested conditions. This finding gives rise to the assumption that rather than the MW, the dynamic range of the synaptic weight (LRS to HRS), in the following called synaptic window (SW), plays a keyrole for the accuracy of a SNN.

This effect can be understood by considering the basic functionality of the neurons performing the Multiply-Accumulate-Function (MAC)

$$I = \sum N_i * G_i \tag{5.2}$$

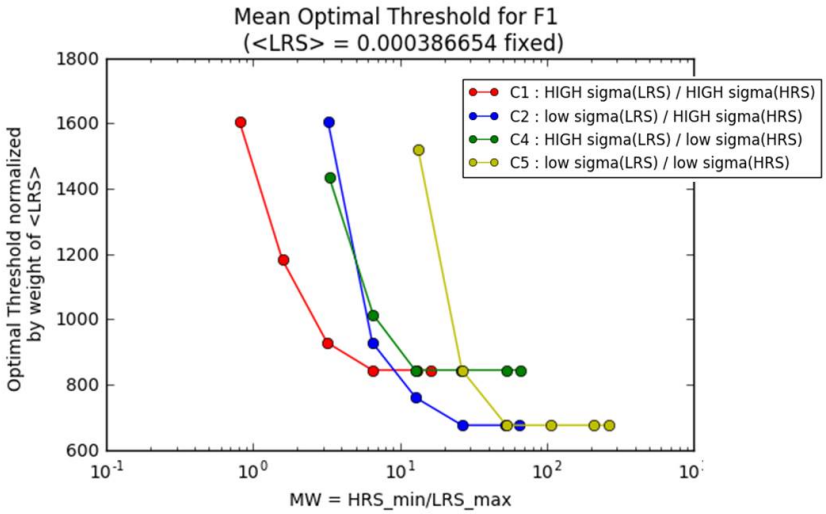


Figure 5.17: False Negative (FN), False Positives (FP) and F1 score as a function of the threshold of the LIF neurons for C4 and C5.

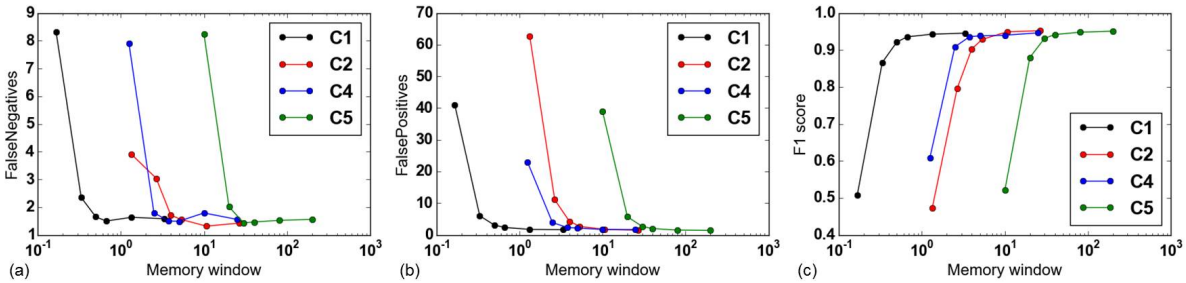


Figure 5.18: False Negative (FN), False Positives (FP) and F1 score as a function of the MW of the LIF neurons for C1, C2, C4 and C5.

where I is the neuron integration, N_i is the number of activations and G_i is the synaptic weight for a specific synapse i. One may separate I into I_P and I_D which are the two contributions from potentiated and depressed synapses (see figure 5.14), respectively, with

$$I_P = \sum N_P * G_P \; and \; I_D = \sum N_D * G_D \tag{5.3}$$

where N_P and N_D are the number of activations of potentiated and depressed synapses and G_P and G_D are the weights of the potentiated and depressed synapses. I_P refers to the integration that is caused by relevant input activity corresponding to the event to be detected/classified whereas I_D is the integrated potential caused by background signals. In order to achieve a good $F1$, or in other words a high selectivity to relevant inputs over background noise, I_P has to dominate the I_D, thus $I_P >> I_D$. Assume that in a short time interval all synapses transmit a spike and thus contribute to I_P or I_D, depending whether they are in LRS or HRS. The ratio of I_P/I_D depends both on the MW and the variability as shown in figure 5.19. In order to maximize SW, one can either increase the MW or tune the variability of LRS and/or HRS.

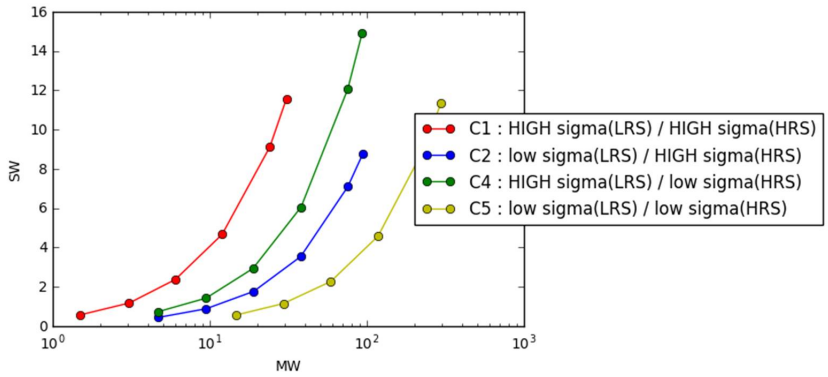


Figure 5.19: Synaptic window (SW) as a function of the statistical RRAM device memory window (MW) for different variabilities of low and high resistance states (LRS and HRS). Note that the populations of RRAM devices in both LRS and HRS are assumed to follow Gaussian distributions.

5.3.4 Synaptic granularity

Synaptic granularity is the capability of a synapse to emulate a certain number of states, i.e. synaptic weights. For example, if one binary device such as a RRAM cell is used to mimic a synapse, the granularity is very high because the synapse can only have two distinct states. The implementation of such a binary synapse may be sufficient for certain applications as shown previously for the visual pattern extraction application, see section 5.8. This application essentially detects events and classification is enabled inherently by the topology of the Spiking Neural Network. However, for other types of applications, binary synapses are not sophisticated enough and it may be necessary that the synapses are able to attain multiple different states to ensure the functionality of a given artificial neural network. This is the case for neural networks where the neural activity corresponding to different input event classes (e.g. different spike waveforms) is propagated along common network paths, i.e. synapses and neurons. Classifying different events based on those overlapping activity patterns results in a much higher demand on the synapse features. The need for synapses which resemble a rather analogue weight characteristic than a binary one was demonstrated for the Spiking Sorting application in section 5.2.

Synaptic redundancy, i.e. multiple binary switching devices, can be combined to build one synapse as explained in more detail in chapter 3. Using n devices per synapse, $n+1$ values of synaptic conductance can be achieved accordingly. Figure 5.20 represents the distribution of synaptic weights for the car detection application after learning using the OxRAM programming condition C5 for two different synaptic redundancies ($n = [1, 10]$). The peaks in the synaptic weight distribution are labelled according to the number of devices in Low Resistance State (LRS). In case of implementing a synapse with $n = 1$ OxRAM devices, the distribution exhibits only two peaks corresponding to the single device resistance in LRS or HRS, respectively. For $n = 10$ devices per synapse, a number of peaks can be observed. The peak at lowest conductance corresponds to synapses where all OxRAM devices are in HRS. For higher conductances, 10 more peaks can be observed which reflect the synaptic weights as a function of the number of OxRAM synapses in LRS, i.e. n between 1 and 10. Apparently there is a rather large gap between the peaks for $n = 0$ and $n = 1$ which means that the synaptic weight spectrum here does not allow to achieve synaptic weights in this gap. Moreover, it is worth noting that the ratio between minimum and maximum conductance is not enhanced for $n = 10$ with respect to $n = 1$. This means that the dynamic range depends solely on the programming condition of individual RRAM devices within the synapse structure.

5.3.5 Learning time

The previous sections have demonstrated that the overall accuracy of a Spiking Neural Network does not degrade with the variability that is induced by single synapses if parameters such as the threshold are optimized. An equivalent performance according to the F1 scores can be achieved,

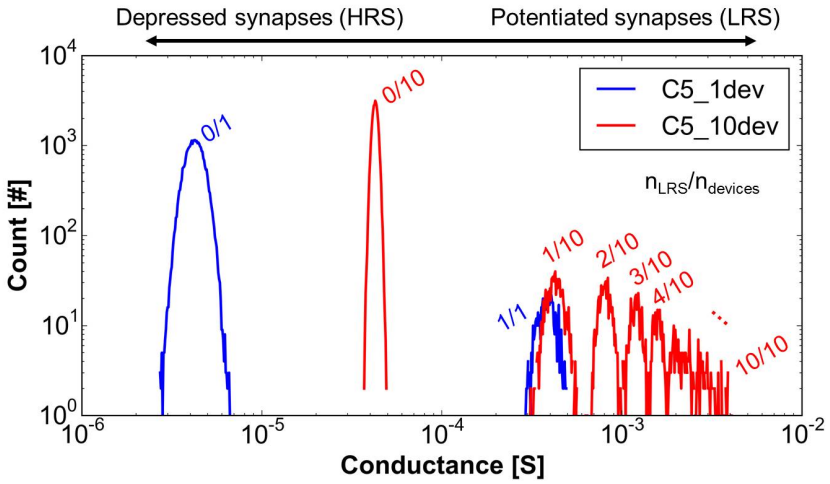


Figure 5.20: Synaptic weight distribution after sufficient learning period for synapses based on 1 C5 device (blue) and 10 C5 devices (red). The majority of the OxRAM based synapses is depressed, i.e. in High Resistive State (HRS) while a small fraction is potentiated, i.e. in Low Resistive State (LRS). The synapses in LRS are the ones corresponding to relevant input information for a specific lane and allow to detect a passing car while the HRS synapses detect events outside the former lane. Note that the absolute ratio between lowest and highest conductance synapses is equivalent for synapses based on 1 and 10 devices, i.e. the dynamic range can not be enhanced by increasing the number of devices per synapse. However, this approach can be used to achieve intermediate synaptic levels instead of only binary weights.

however, the learning speeds seems to be slightly lower for high variabilities, i.e. the network needs a longer time for the appropriate weight tuning that allows the optimum classification performance. This effect is shown in figure 5.21. While the detection rate of $C5$ rises abruptly to the maximum of, the learning curve for $C4$ starts slightly lower and its slope is weaker. The extended learning time needed by a network that is affected by synaptic variability may be

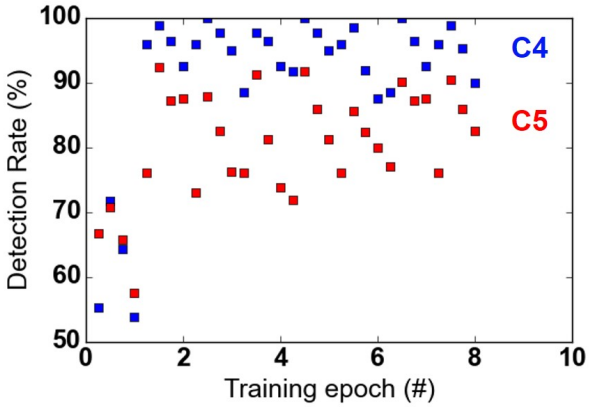


Figure 5.21: Detection Rate (DR) as a function of the number of training epochs.

due to neurons which initially learn a pattern. However, at a later stage during learning, the network might be able to identify other neurons that have a more suitable (probabilistic) weights distribution. Furthermore, the variability is likely to cause the network to change synaptic weights more often due to the uncertainty of the resistance programming introduced by synaptic variability. Note that due to the poor statistical evidence of this theory, depicted in the highly noisy learning curves in figure 5.21, this theory should be verified by repeated neural network simulations.

5.4 Summary

In this chapter, the impact of the synaptic variability on spiking neural networks was studied for two applications, one related to object detection and the other one to classification. When a synapse is implemented based on binary devices, whereas each of the two states is affected by variability, it was found that synaptic variability can have different effects, dependent on whether the variability is due to electrical device variability in low or high resistive state. First, it was found that the reliability could be improved by enlarging the memory window. Second, it was demonstrated that is necessary to carefully tune the network parameters such as the neuron threshold for spike emission as a function of the electrical characteristic of the synapse devices. This was shown to achieve a reliability that does not depend on the variability but only on the relative dynamic range of the synapses. This dynamic range depends on the average weight of synapses in potentiated state compared to the average weight of the ones in depressed state. Since the average is an arithmetic value, variability in LRS can increase the dynamic range of a synapse while HRS variability leads to a decrease, assuming that the mean values are constant. The findings of the impact of variability on the reliability of Spiking Neural Networks shall be used for the design of optimized operation conditions of RRAM (or other) technologies for the application in artificial synapses.

CHAPTER 6

SHORT-TERM PLASTICITY

Various concepts to mimic stable, i.e. Long Term Plasticity (LTP), effects in artificial neural networks (ANN) were proposed previously [141]. Experimental findings from the neuroscience community have shown however that synaptic changes are not only governed by LTP but also unstable, i.e. Short Term Plasticity (STP), effects can be observed. Those unstable effects have a different biochemical origin and serve different purposes with respect to the stable ones. In this chapter, it is demonstrated how OxRAM devices and their intrinsic switching probability can be exploited to emulate both STP and LTP inspired by biology. A new circuit concept is proposed to co-implement STP and LTP using non-volatile OxRAM devices and some rules for the design of the STP synapse are described. It is showcased by two realistic applications based on Fully Connected Neural Networks that LTP enables the networks to learn patterns in the data without any supervision while STP ensures a highly reliable signal detection even in presence of significant background noise in the input data.

This chapter is structured as follows. First, the features of a biological synapse are briefly reviewed section 6.1. Then, the model for Short Term Plasticity based on experimental findings by Tsodyks and Markram is described in section 6.2. Next, the approach to emulate STP using non-volatile OxRAM technology is explained in detail in section 6.3, followed by the concept of a compound synapse to merge STP with LTP in section 6.4. An example for the implementation of this synapse concept using OxRAM arrays is introduced in section 6.5. The STP related performance enhancement of Spiking Neural Networks exposed to highly noisy data is demonstrated in section 6.6 and compromises regarding integration complexity and energy consumption are discussed. Finally, the results of the previous sections are summarized concluding the relevance of short term plasticity for spiking neural networks.

6.1 Biological synapse review

Figure 6.1 shows an illustration of a biological synapse (i.e. the electro-chemical connection of an axon of a pre-synaptic neuron and a dendrite of a post-synaptic neuron) restricting itself on the essential features needed to understand the functionality of the synaptic features. The pre-synaptic terminal contains different species of neurotransmitters, each of a certain concentration. The post-synaptic terminal features a number of channels or receptors which are specific for a certain kind of neurotransmitter. When a pre-synaptic neuron emits an action potential, commonly known as spike, towards the post-synaptic neuron, a number of glutamate particles is released from the pre-synaptic terminal transported through the synaptic cleft and channelled into the post-synaptic terminal via the element specific channels, in this case the so-called AMPA channels (see chapter 1 for more details). This process induces an impulse into the post-synaptic terminal which results in a modulation of the internal cell voltage, termed as the Excitatory Post-Synaptic Potential (EPSP). The EPSP essentially increases probability of a neuron to emit a spike by approaching its internal voltage towards the threshold voltage. Both the number of neurotransmitters (N_T) in the pre-synaptic terminal and the number of receptors (N_R) in the post-synaptic terminal determine the amplitude of the EPSP induced due to a spike from the pre-synaptic neuron. As a rule of thumb, it is valid to assume the higher N_T and/or N_R the stronger the influx and thus the stronger the EPSP in the post-synaptic cell body.

$$EPSP \propto (N_T, N_R) \tag{6.1}$$

N_R determines what is usually referred to as the synaptic weight which can be understood as the sensitivity of the post-synaptic neuron to inputs from the pre-synaptic neuron. This synaptic weight is known to be subject to various kinds of plasticity (see chapter 1). Plasticity describes the modification of the synaptic weight, or efficacy, as a function of the activity of pre- and post-synaptic activations. One of the most well known types of synaptic plasticity is Spike-Timing-Dependent Plasticity (STDP) which affects the number of receptor channels in the post-synaptic terminal based on the relative timing of pre- and post-synaptic spikes, see figure 6.1. Changes due to STDP are long-lasting and are therefore responsible for Long Term Plasticity used in the process of learning and memory creation. N_R is increased or decreased in Long Term Potentiation or Long Term Depression, respectively. On the other hand, N_T changes in a rather dynamical manner based on the spiking activity of the pre-synaptic neuron. For example, two consecutive spikes do not evoke the equal EPSP in a post-synaptic neuron if the delay between those spikes is within a few milliseconds. This can be due to a lack of time to recover the reservoir of neurotransmitters to a sufficient population. Due to this temporary nature, this effect is corresponding to Short Term Plasticity.

For the time being, no implementation concepts that feature long and short term plasticity independently from each other were demonstrated but STP concepts which have been proposed

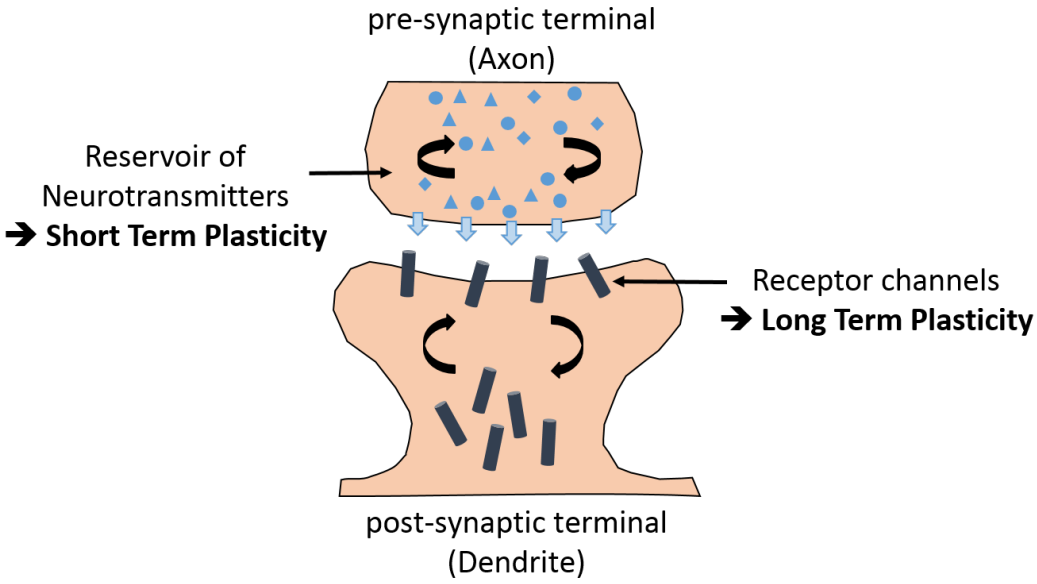


Figure 6.1: Schematic illustration of synaptic connection between a pre-synaptic axon and a post-synaptic dendrite. Both the number of neurotransmitters in the pre-synaptic terminal and the number of channels in the post-synaptic terminal determine the amplitude of the voltage in the post-synaptic neuron induced by a spike of the pre-synaptic neuron. Note that the number of neurotransmitters is changed dynamically and modifications decay in an exponential relaxation where the channel number modifications are permanent. Therefore, those two effects are affiliated with Short and Long Term Plasticity, respectively.

so far either rely on switched capacitors [238], conventional CMOS [239] or on the short to long term plasticity conversion of emerging NVM such as RRAM [177].

6.2 Tsodyks-Markram model

Besides the well-known STDP [43] [2] (described in chapter 1), it was found in biological synapses that the EPSP during a spike train beyond a certain frequency propagated along a synapse decreases proportionally to the pre-synaptic spiking frequency (f_{pre}) as shown in figure 6.2 [25]. The EPSP decrease is due to the corresponding reduction of pre-synaptic neurotransmitters. In fact, this particular effect is called Short Term Depression (STD) while the opposite effect of Short Term Facilitation (STF) can be found as well in biology but is not described here.

Indeed, these synaptic modifications depend only on the pre-synaptic activity, i.e. the history of the pre-synaptic spike events, in contrary to modifications by e.g. Spike Timing Dependent Plasticity, where the synaptic weights change based on correlation of the pre- and post-synaptic neuron activities and are stable over time. This is one of the main characteristic features of Short Term Plasticity (STP) along with the attribute that the synaptic weight recovers quickly towards its resting level in case of no pre-synaptic activity. A phenomenological model describing the

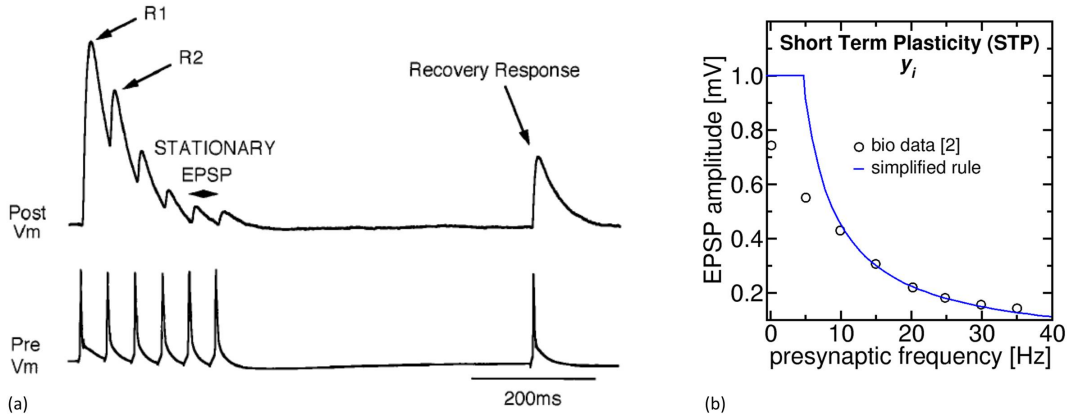


Figure 6.2: (a) Functional observation of Excitatory Post-Synaptic Potential (EPSP) evoked in post-synaptic neuron during a pre-synaptic spike. The amplitude of the EPSP reduces upon a rising number of input spikes, e.g. $R2 < R1$. Note that a stationary EPSP occurs for a spike train of constant frequency. (b) Biological data (symbols) and a simplified rule (line) for the stationary EPSP as a function of the pre-synaptic frequency. Figures reproduced from [25].

depression upon a pre-synaptic spike and subsequent transient relaxation has been developed by Tsodyks and Markram [25]. The weight of a synapse of a pre-synaptic neuron i associated to STP is expressed as y_i which evolves over time t according to

$$\frac{dy_i}{dt} = \frac{1-y_i}{\tau_D} - f_D \cdot y_i \cdot \delta(t-t_{pre}) \tag{6.2}$$

where $0 < f_D < 1$ controls the degree of depression when a pre-synaptic spike occurs at time t_{pre} and τ_D is the recovery time constant for the transient decay of $y_i(t)$ towards its resting level, here $y_{i,max}$. Every time a spike was propagated along the synapse, its weight is depressed while the absolute value of synaptic depression $\Delta\, y_i$ is a function of f_D and the momentary weight $y_i(t)$ at time t, hence

$$\Delta\, y_i = - f_D \cdot y_i \tag{6.3}$$

Figure 6.3 illustrates the STP weight characteristic for an arbitrary spike train assuming different values of the parameters f_D and τ_D. The spike train consists of 5 spikes with different inter-spike-intervals (ISI) ranging from $10s$ to $90s$. The plotted weights are normalized to the resting state (equal to 1). For fast relaxation times, i.e. $\tau_D < ISI$, $y_i(t)$ can recover quickly towards 1 so that the weight during consecutive spike is not influenced. On the other hand, for $\tau_D > ISI$, the time is not sufficient for a full recovery of y_i, i.e. the STP effect is additive for several consecutive spikes and y_i can become smaller than $y_{i,max} - f_D$.

6.3 Emulation of Short Term Plasticity using RRAM

It was previously demonstrated that HfO_2 based OxRAM cells are capable to emulate Long Term Plasticity (w_{ij}) in [214] [141] and in chapters 3 and 4. The challenge addressed in this work

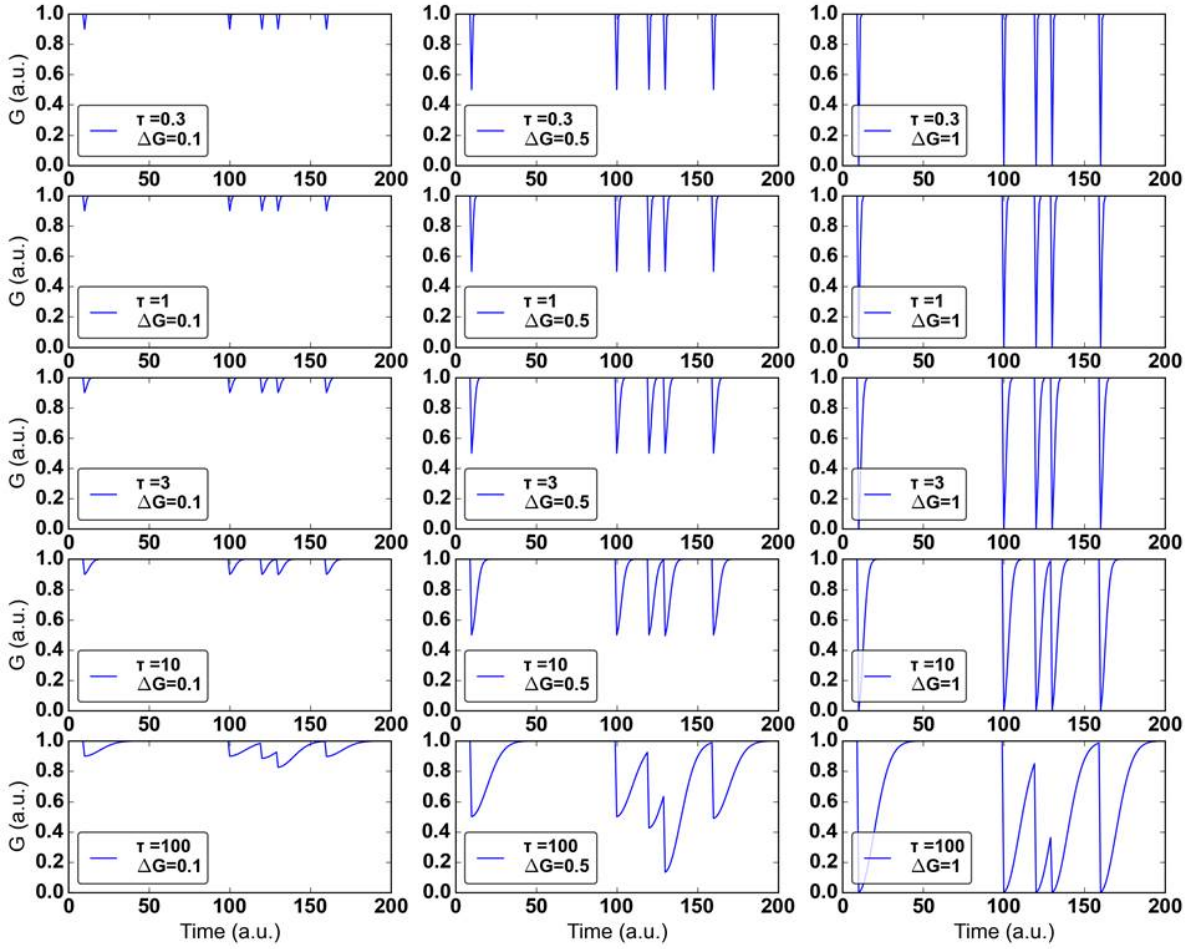


Figure 6.3: Schematic illustration of Short Term Plasticity model according to equation 6.2. Several traces are shown for the same spike train example using different STP parameters $\tau_D = [0.3, 1, 3, 10, 100]$ and $f_D = [0.1, 0.5, 1.0]$.

is the joint hardware integration of several synaptic features such as Long Term Plasticity (LTP) and Short Term Platsticity (STP). To this end, it is desirable to use compatible or even the same technology, in our example OxRAM technology. While the non-volatility of the resistance state can be directly exploited to feature the long-term remaining synaptic modifications and thus implement LTP, synaptic changes due to STP are dynamic and decay over time, as explained in section 6.2. This requires a volatile behaviour of the OxRAM based synapse to implement the STP. However, OxRAM does not intrinsically comply to this behaviour, i.e. this shortcoming must be overcome by a dedicated programming strategy.

In the following, it is demonstrated that synapses based on OxRAM technology are capable to reproduce both STP ($y_i(t)$) and LTP (w_{ij}). In order to reproduce the characteristic decay of the STP effect, the changed synaptic weight has to recover towards an initial state over time. Moreover, this recovering phase occurs typically progressively and therefore it seems useful to

rely on a synapse design featuring multiple levels. For this reason, a number of binary devices are employed to represent one STP synapse, as described in detail in section 3.3. Figure 6.4 shows the principle hardware implementation of one STP synapse $y_i(t)$ between one pre-synaptic neuron and several post-synaptic neurons based on multiple OxRAM cells. Note that it is possible to mutualize the STP weights of all post-synaptic neurons connected to the same pre-synaptic neuron by one single STP synapse since the latter depends only on the pre-synaptic spiking frequency, i.e. the synaptic connections between all post-synaptic neurons and one specific pre-synaptic neuron experience the same spike train. The pre-synaptic spikes are applied by the driver circuit to the top electrodes of all n devices and the post-synaptic current enters the post-synaptic neuron from the cell's bottom electrodes.

Figure 6.5 describes the programming strategy of the STP synapse (figure 6.4). The weight $y_i(t)$ is tuned using two invariant pulse conditions for Set and Reset, which can either increase or decrease the STP synaptic weight (conductance). A Set programming pulse (blue line in graph) is applied at a constant 'clock' rate ($1/\Delta T$). Therefore, in case of no pre-synaptic spiking activity, the synaptic weight tends to be at its resting level ($y_i(t)/y_{i,max} = 1$). When a spike occurs (black line on top of figure), an abrupt decrease of the normalized synaptic weight (red line in graph) is induced by applying a Reset programming pulse on the OxRAM based synapse. Due to the series of Set programming pulses (blue line in graph) applied after every timing delay ΔT, the characteristic STP relaxation is induced in the synapse resistance. This results in the gradual recovery towards the high (conductance) resting level of the synaptic weight. Note that a series of pre-synaptic spikes, as illustrated for a high spiking activity, can cause a strong depression of $y_i(t)$ strongly so that $y_i << y_{i,max} - f_D$, resembling the biological model of STP which is based on experimental results. The model parameters f_D and τ_D can be approximated by tuning the absolute change of $y_i(t)$ triggered by a Reset or Set pulse. Typically a Reset pulse shall result in a bigger absolute change compared to a Set pulse. The level of conductance increase or decrease

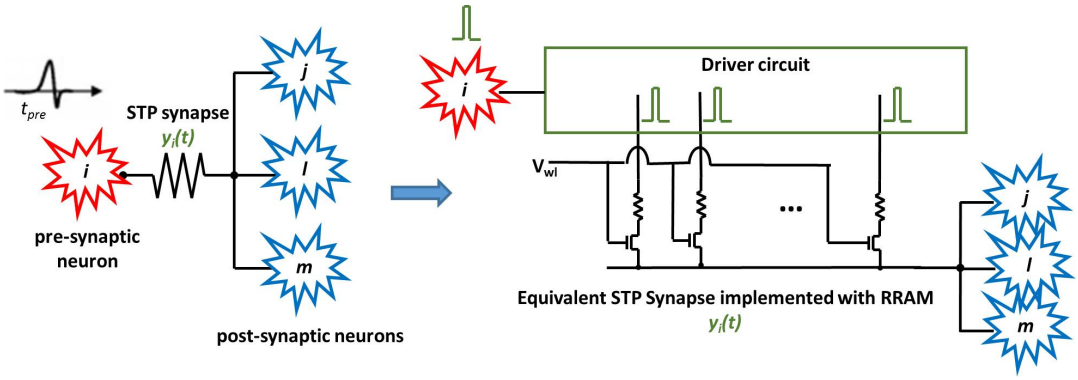


Figure 6.4: Schematic of proposed Short Term Plasticity synapse ($y_i(t)$) using 10 HfO_2 based OxRAM cells. Top electrode: Ti PVD 10 nm, resistive switching layer: HfO_2 ALD 5 nm, bottom electrode: TiN PVD, 130 nm node.

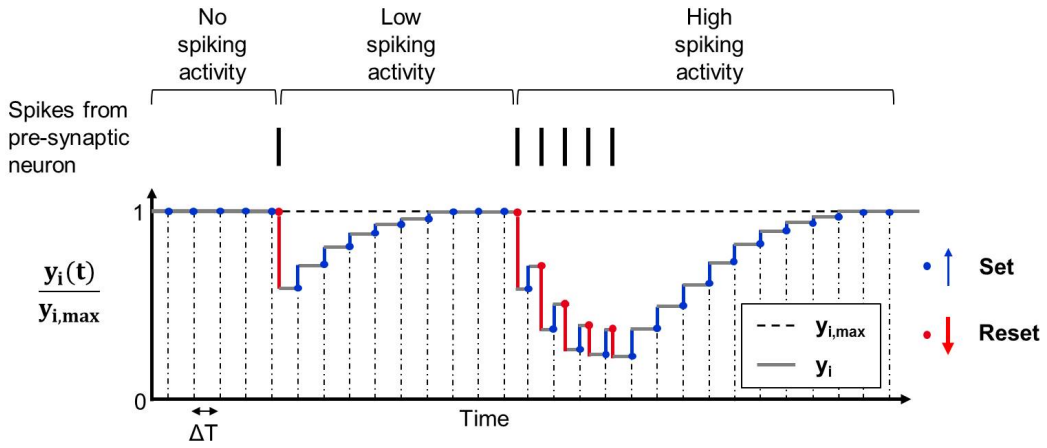


Figure 6.5: Programming strategy to reproduce Short Term Plasticity using the OxRAM based synapse (figure 6.4). The synaptic weight is decreased at each pre-synaptic spike and periodically increased (every ΔT) in absence of pre-synaptic spikes.

can be controlled by adjusting the Set and Reset pulse conditions which can be realized by either tuning the programming probabilities in the driver circuit or using the intrinsic switching probability of OxRAM devices. For a detailed description consult section 3.3. The Set probability p_{Set} in combination with the programming delay ΔT are used to fit τ_D while f_D is controlled by the Reset probability p_{Reset}.

Figure 6.6 demonstrates that the synaptic weight evolution $y_i(t)$ of the STP model can be reproduced using a synapse based on $n = 10$ OxRAM devices (see figure 6.4) by applying the programming strategy presented in figure 6.5. Figure 6.6 (a) shows the STP characteristic (green line) for $f_D = 1$ and $\tau_D = 1ms$ and the corresponding experimental approximation of the OxRAM synapse. The programming conditions of the STP OxRAM synapse were $p_{Set} = 0.1$ and $p_{Reset} = 1$. Accordingly, figure 6.6 (b) shows the fitted STP for another set of STP parameters, $f_D = 0.5$ and $\tau_D = 10ms$. In order to achieve a sufficient approximation, the OxRAM programming conditions have to be changed to $p_{Set} = 0.05$ and $p_{Reset} = 0.5$. From a qualitative point of view, this simple programming strategy seems to enable a promising emulation of the STP model by a relatively simple synapse of only 10 devices/synapse, i.e. 11 synaptic weight levels. Here, the parameters p_{Set} and p_{Reset} are adjusted by tuning the set and reset voltages, V_{Set} and V_{Reset}, respectively (explained below). The Set programming interval ΔT was fixed at τ_D/n. The results obtained with the phenomenological model of equation 6.2 (green lines) are reported for verification.

Note that for a pre-synaptic spike train of constant spiking frequency f_{pre}, the STP weight $y_i(t)$ reaches a stationary value, in the following referred to as the effective weight $y_i(f_{pre})$. The $y_i(f_{pre})$ is the weight that encodes the EPSP induced by a pre-synaptic spike. It expresses the equilibrium between synaptic depression (pre-synaptic spike) and synaptic relaxation (pre-synaptic inactivity). In other words, between two consecutive spikes, the synaptic weight can

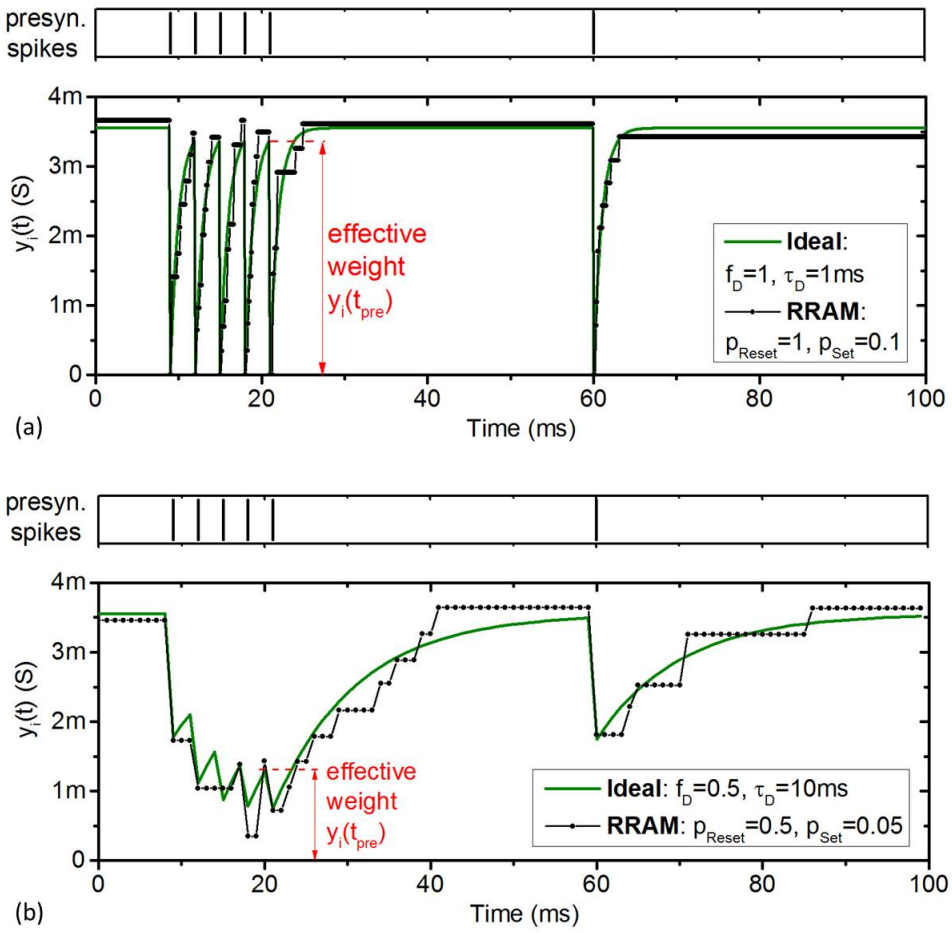


Figure 6.6: Short Term Plasticity synaptic weight evolution obtained using the OxRAM synapse structure ($n = 10$) and programming scheme presented in figures 6.4 and 6.5 (black symbols) and the Tsodyks and Markram model (green line). Different values for τ_D and f_D can be experimentally obtained by changing the set and reset probabilities, p_{Set} and p_{Reset}. The programming interval was set to $\Delta T = \tau_D/n$, hence (a) $\Delta T = 0.1ms$ and (b) $\Delta T = 1ms$.

relax just as much as it is depressed in the event of a spike. The value of $y_i(f_{pre})$ depends obviously on f_{pre} and furthermore on f_D and τ_D. The effective weight normalized to its resting (i.e. maximum) level $y_{i,max}$ is shown in figure 6.7 as a function of the pre-synaptic spiking frequency f_{pre} (constant relaxation time $\tau_D = 1\ ms$). If the spikes along the synapse are driven beyond a certain frequency, defined as the limiting frequency, thus $f_{pre} > f_{lim}$, the effective weight decreases proportional to f_{pre}. In this case, the time between two subsequent spikes ($1/f_{pre}$) is too short for the synaptic weight to recover to $y_{i,max}$. Otherwise, if $f_{pre} < f_{lim}$, the time for synaptic relaxation is sufficient and therefore the STP related synaptic changes do not affect the subsequent spiking activity. Moreover, it can be derived from the graph that $y_i(f_{pre})$ depends also on f_D, i.e the stronger f_D the smaller $y_i(f_{pre})$. Figure 6.7 (b) shows $y_i(f_{pre})$ for different τ_D and a constant $f_D = 0.5$. While the characteristic S-shaped curve remains the same, $y_i(f_{pre})$ is

shifted along f_{pre} as a function of τ_D. This means that f_{lim} depends solely on the value of τ_D. These results are in qualitative agreement with the biological STP behavior reported in figure 6.2.

As mentioned above, the STP model parameters τ_D and f_D can be converted into switching probabilities P_{Set} and P_{Reset} which are used in the driver circuit of the OxRAM based synapse (figure 6.4). In order to emulate a certain τ_D, the two parameters P_{Set} and ΔT can be tuned correspondingly. It is clear that the longer ΔT, i.e. the less set pulses inducing the relaxation effect, the higher P_{Set} has to be used in order to ensure the correct transient behaviour. The level of synaptic depression f_D on the other hand is simply adjusted by tuning P_{Reset}, as shown in figure 6.8 (b).

6.4 Compound synapse featuring Short and Long Term Plasticity

It was previously demonstrated how OxRAM can be used to emulate STP (section 6.3) as well as LTP (chapter 3). Another challenge is to merge those two functionalities into one synapse thus featuring both dynamic and static changes. The underlying principle for the interaction of STP and LTP is shown in figure 6.9. Whenever a pre-synaptic spike occurs, the short term synaptic weight $y_i(t)$ is depressed followed by the relaxation towards its resting level. In case of a post-synaptic spike shortly after a pre-synaptic spike, the long term synaptic weight w_{ij} is potentiated (LTP) or if the post-synaptic spike shortly before a pre-synaptic spike, w_{ij} is depressed (LTD). If there is no correlation between pre- and post-synaptic spikes, w_{ij} remains

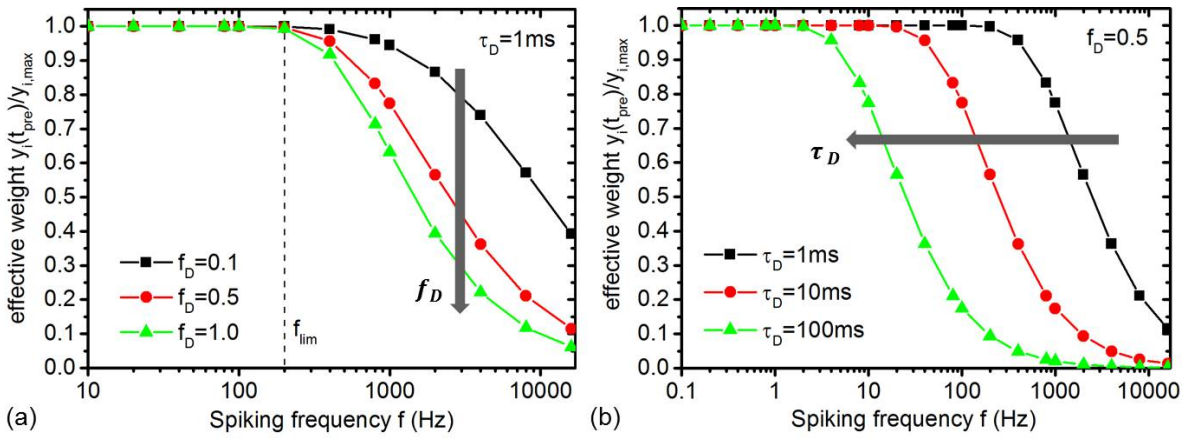


Figure 6.7: Stationary amplitude of $y_i(t)$ (Fig.4) reached during a train of spikes with a given pre-synaptic frequency, f_{pre}, for different (a) f_D and (b) τ_D values. The limiting frequency f_{lim} decreases as the τ_D and is independent of f_D.

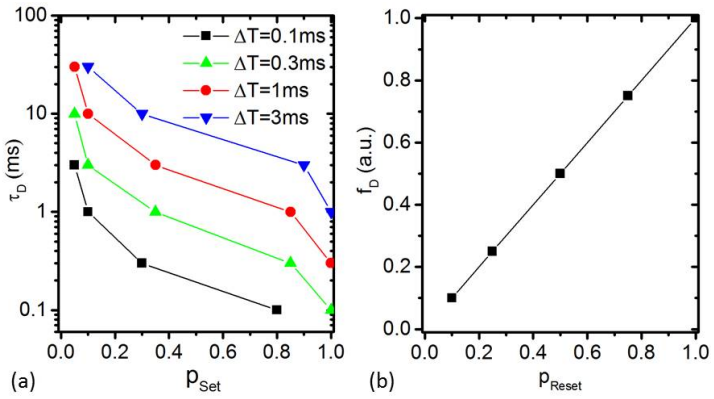


Figure 6.8: (a) R^2 for correlation of Short Term Plasticity based on RRAM and model as a function of $\Delta T/\tau_D$. Relationship between (b) τ_D and the set probability and (c) f_D and the reset probability. p_{Set} and p_{Reset} are modulated by the OxRAM programming voltages (Fig.13).

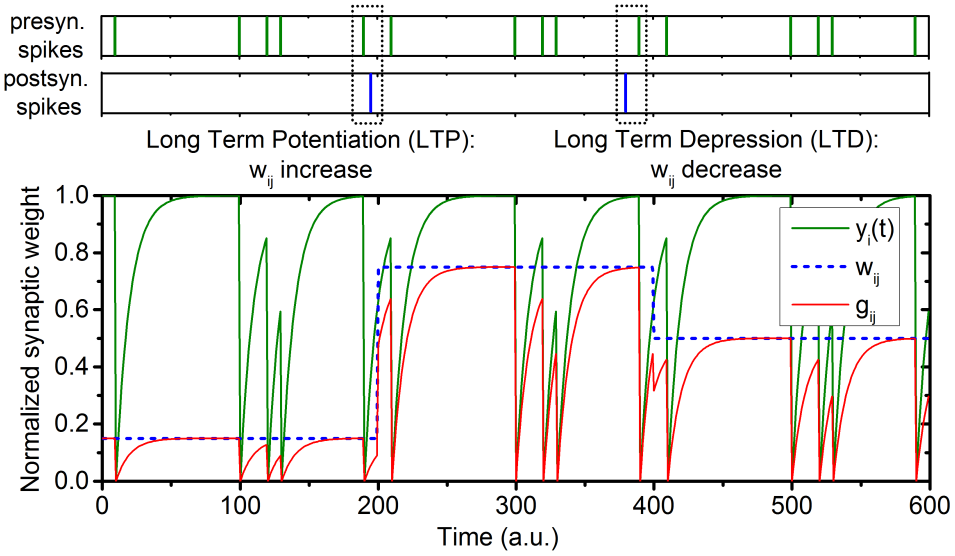


Figure 6.9: Schematic illustration for association of Short Term Plasticity weight ($y_i(t)$) with Long Term Plasticity weight (w_{ij}) to create total synaptic weight (g_{ij}).

unchanged. The overall synaptic weight g_{ij} follows a multiplicative rule according to

$$g_{ij} = y_i(t) \cdot w_{ij} \tag{6.4}$$

which means that g_{ij} is dominated by the smallest weight, either $y_i(t)$ or w_{ij}. It may always be assumed that $0 < g_{ij} < w_{ij}$.

Figure 6.10 shows the circuit proposed to (i) reproduce both the STP and LTP rules according to their biological models and to (ii) associate the STP and LTP weights which are stored in two separate OxRAM elements, $y_i(t)$ and w_{ij}, respectively. The conductance multiplication of $y_i(t)$ and w_{ij} during the reading operation is performed by means of a buffer which is essentially a

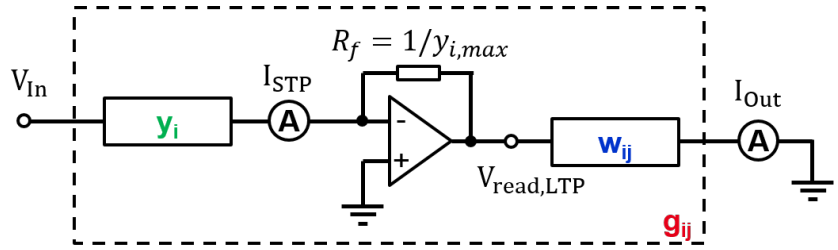


Figure 6.10: Principal circuit proposed to reproduce both the Short Term Plasticity and Long Term Plasticity rules using non volatile OxRAM cells. The conductance multiplication during the read operation is performed by means of a buffer which modulates the read voltage for the Long Term Plasticity synapse w_{ij}.

current-to-voltage converter (or transimpedance amplifier) where the feedback resistor has the value $1/y_{i,max}$. When a pre-synaptic neuron i emits a spike, the synapse receives an incoming event which generates a voltage pulse (V_{In}) that propagates through the STP synapse (y_i) and causes a current (I_{STP}) corresponding to its weight. The buffer modulates the read voltage applied to w_{ij} ($V_{read,LTP}$) as a function of I_{STP}, i.e. as a function of the conductance value $y_i(t)$.

$$V_{read,LTP} = -V_{In} \cdot y_i(t) \cdot \frac{1}{y_{max}} \tag{6.5}$$

Consequently, $V_{read,LTP}$ can be varied between a minimum and maximum voltage, corresponding to $y_{i,min}$ and $y_{i,max}$, respectively

$$y_i(t) = \begin{cases} y_{i,max} \Rightarrow V_{read,LTP}(t) = -V_{In} & \text{no STP impact} \\ y_{i,min} \Rightarrow V_{read,LTP}(t) = -V_{In} \cdot \frac{y_{i,min}}{y_{i,max}} & \text{highest STP impact} \end{cases} \tag{6.6}$$

Finally, the higher $V_{read,LTP}$, the higher the resulting output current I_{Out} which reflects the overall weight of the synapse g_{ij}.

$$I_{Out}(t) = V_{read,LTP} \cdot w_{ij} \tag{6.7}$$

6.5 Synapse implementation with OxRAM arrays

Figure 6.11 presents an array implementation of the compound synapse into a Fully Connected Neural Network (FCNN) topology. Each layer's (i) neurons drive the input signals for the next layer's (j) neurons through the individual synaptic weights g_{ij} (consisting of $y_i(t)$, the buffer and w_{ij}). Remember that the STP depends only on the pre-synaptic frequency and for this reason, the total number of STP synapses including the corresponding buffers is equal to the number of input neurons (N_i), i.e. each $y_i(t)$ synapse is shared by all its output neurons. On the other hand, the total number of LTP synapses w_{ij} is equal to the number of input neurons times the number of output neurons ($N_i x N_j$). Every $y_i(t)$ is implemented with the multiple OxRAM scheme described

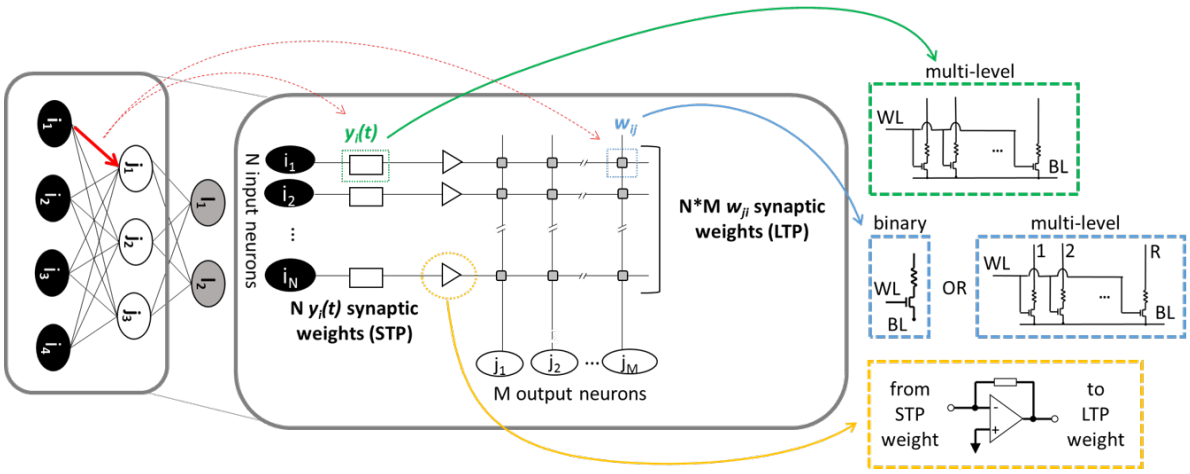


Figure 6.11: Integration concept for a Fully Connected Neural Network (FCNN) using $1T-1R$ OxRAM arrays. Each layer,Äôs neurons drive the next layer through weights $y_i(t)$ (Short Term Plasticity) and w_{ij} (Long Term Plasticity).

in section 6.3. The LTP weight (w_{ij}) can be implemented by means of the same approach or even one single OxRAM element per synapse, thus introducing binary long term synaptic weights.

We mimicked a FCNN on a 64 $kbit$ Ti/HfO_2-based OxRAM array manufactured in 130 nm CMOS node, see figure 6.12. As indicated, the OxRAM element is located between the metallization layers $M4$ and $M5$. The OxRAM devices are co-integrated in a so-called $1T1R$ structure, i.e. every resistive memory element ($'R'$) is in series with a transistor ($'T'$) which is used to both access and control the current compliance of the OxRAM memory cell. A wide range of pulse voltages and current compliances can be used for Set and Reset programming of the $1T1R$ devices resulting in different resistive switching behaviour. Figure 6.13 presents the experimental results for switching a $4-kbit$ OxRAM array between Low Resistive State (LRS) and High Resistive

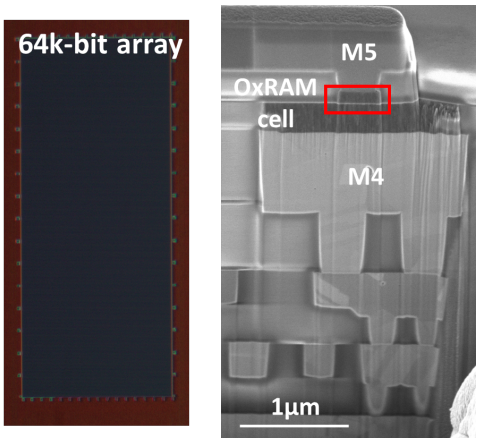


Figure 6.12: Photograph of 64 $kbit$ circuit demonstrator and SEM image of CMOS stack including the OxRAM cell between $M4$ and $M5$.

State (HRS) using two different values of current compliance, defined as strong ($I_{set} = 400\mu A$) and weak ($I_{set} = 40\mu A$) programming conditions. The plots represent the statistical distributions of LRS and HRS for one cycle of switching of all OxRAM cells. It is clear that in case of a strong programming, both LRS and HRS are well separated and only the tails of the distributions (beyond 2σ) start to deviate from the narrow distributions. This means that 95 % of the samples (within $\pm$ 2 σ) feature a low resistance variability and therefore offer a clear resistance window margin of more than one order of magnitude. On the other hand, for the weak programming current, the LRS distribution varies substantially with respect to the one of the strong programming which may be attributed to the filament geometry. The lower the programming current, the lower the filament diameter, hence, the fluctuations in the exact filament structure can affect the final LRS value strongly. This results in a rather wide LRS distribution spanning almost three orders of magnitude and partly overlapping with the HRS. As a consequence, the weak condition does no longer offer a resistance margin. The HRS distribution remains rather narrow because of the sneak path current effect that occurs in resistive memory arrays. This effect limits the maximum resistance that can be measured to about $5{\cdot}10^6$ Ω, which limits the distribution accordingly.

Figure 6.14 shows the resistance distributions after Set ($HRStoLRS$) and Reset ($LRStoHRS$) pulsed programming for various pulse amplitude biases. As V_{Set} is increased, the LRS distribution shifts to lower resistances while the HRS distribution shifts towards high resistance as V_{Reset} is increased. Both for Set and Reset, one can extract the percentage of switched cells as a function of the applied bias voltage.

Therefore, one can consider that the higher the pulse amplitude, the higher the percentage of cells switching to the respective state. For this reason, Set and Reset voltages can be used to control the probability to switch between the memory states (figure 6.15). The Set probability p_{Set} of this OxRAM implementation can be tuned between 0 and 1 applying voltages between

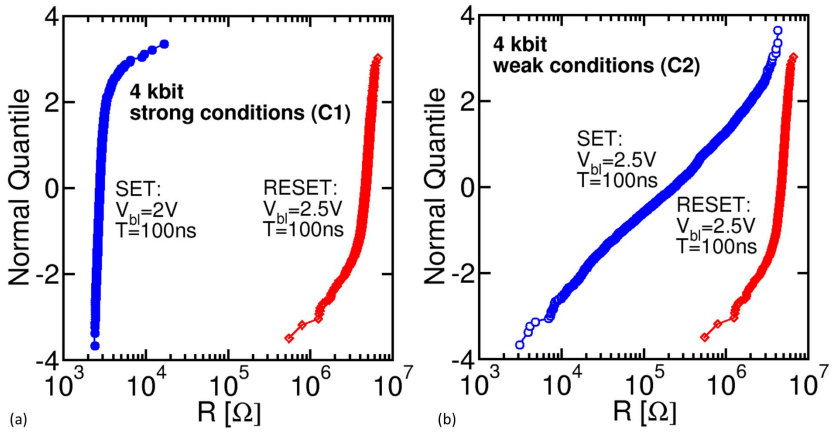


Figure 6.13: $4-kbit$ resistance distributions for (a) strong (I_{Set} = 400 μA) and (b) weak (I_{Set} = 40 μA) programming conditions. 5 $M\Omega$ is the resistance measurement limit.

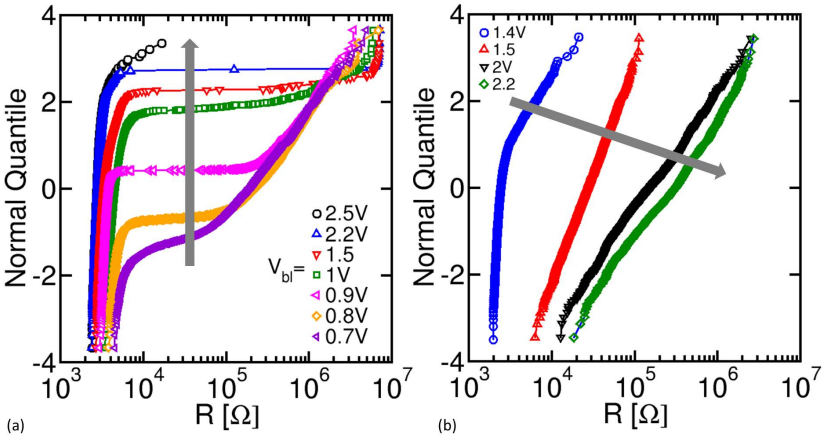


Figure 6.14: $4-kbit$ resistance distributions for different (a) set and (b) reset voltages for strong programming condition ($I_{Set} = 400\ \mu A$) .

0.6 V and 1.7 V. For the Reset process, the switching probability p_{Reset} can be tuned between 0 and 1 applying voltages between 1.4 V and 2.2 V. Finally, p_{Set} and p_{Reset} are used to calibrate the STP parameters τ_D and f_D (see section 6.1).

Both experimental OxRAM array operation conditions (strong, weak) introduced in figure 6.13 were tested for the implementation of the STP synapse in order to identify an optimized condition. Therefore, the geometrical means and variabilities (2σ range) were extracted from the statistical distributions of LRS and HRS in order to calibrate the RRAM model for the synapse. The synapse was implemented with $n = 10$ OxRAM devices each. Using the calibrated synapse model for the two conditions, several combinations of p_{Set} and p_{Reset} were simulated and compared to different set of τ_D and f_D. Each simulation was performed 100 times since the synapses exhibit fluctuations due to OxRAM device variability. As shown in figure 6.16 (a), the

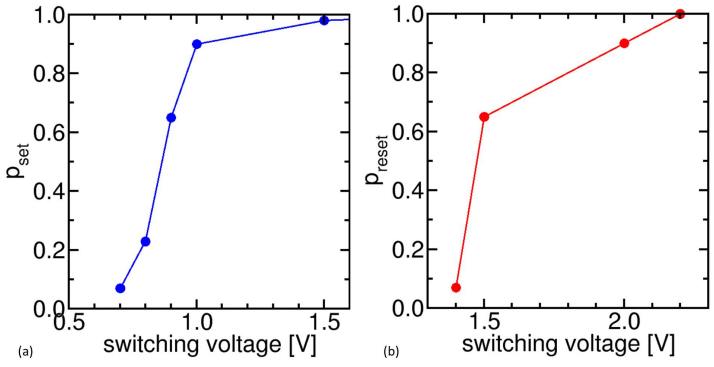


Figure 6.15: (a) Set and (b) reset switching probabilities extracted from the $4-kbit$ array resistance distributions of figure 6.14 and used to tune the Short Term Plasticity conditions, τ_D and f_D.

synapses operated with the strong condition allow to achieve a good qualitative accordance with the model while using the weak condition leads to a significantly different STP characteristic (see figure 6.16 (b)). The discrepancy between experimental RRAM STP and modelled STP for the two operation conditions can be mainly attributed to the difference in the LRS variability between the two conditions. The high variability in LRS leads to a few OxRAM devices with a very high conductance with respect to the mean value, thus, the average STP synapse (red) exceeds the model STP weight (blue) by approximately a factor of 2. It is worth noting that a few STP synapses can feature conductances which are considerably higher than the one according to the model, i.e. $10x$ higher. This is very critical for the proposed concept of the co-integration of STP and LTP in figure 6.10 because according to this concept, the read current of the STP synapse, which scales directly with the conductance, is converted into a pulse voltage for the LTP synapse during read mode. That means that a very high conductance would trigger a high voltage pulse on the LTP synapse which can potentially cause read disturbs if the voltage lies in the range of the Set programming voltage (see figure 6.15). A read disturb is the unintentional switching of an OxRAM cell from LRS to HRS or vice versa while its current state is supposed to be determined.

The quantitative accuracy of the OxRAM based STP approximation with the STP model was evaluated by calculating the Pearson correlation coefficient r^2 using simulated synaptic weights of figure 6.16. Figure 6.17 reports the correlation as a function of ΔT. The optimum is reached around $\Delta T = 0.1 \cdot \tau_D$ which is interesting to note because it gives rise to the assumption that ΔT is related to the number of devices n used to implement one synapse, here $n = 10$. For this reason, it can be concluded that generally $\Delta T/\tau_D = n$ Obviously, the strong condition achieves a much higher r^2 than the weak condition. For this reason and to avoid read disturbs, the strong condition is subsequently used for the implementation of the STP synapses while the weak condition is assumed for the LTP synapses.

6.6 Short Term Plasticity in Spiking Neural Networks

The effect of synaptic Short Term Plasticity (STP) in Spiking Neural Networks (SNN) is demonstrated on two independent applications. Both applications rely on dedicated two-layer Fully Connected Neural Network (FCNN) topologies where each synaptic connection features separate components to provide STP and LTP functionality, see figure 6.18. The SNN's are modelled by means of an in-house developed Neural Network simulator ('Xnet' [148]) where all parameters are matched to the specific problem and stochasticity of the OxRAM devices encountered in the electrical characterization. Moreover, the performance in presence of significant background noise in the input data is studied.

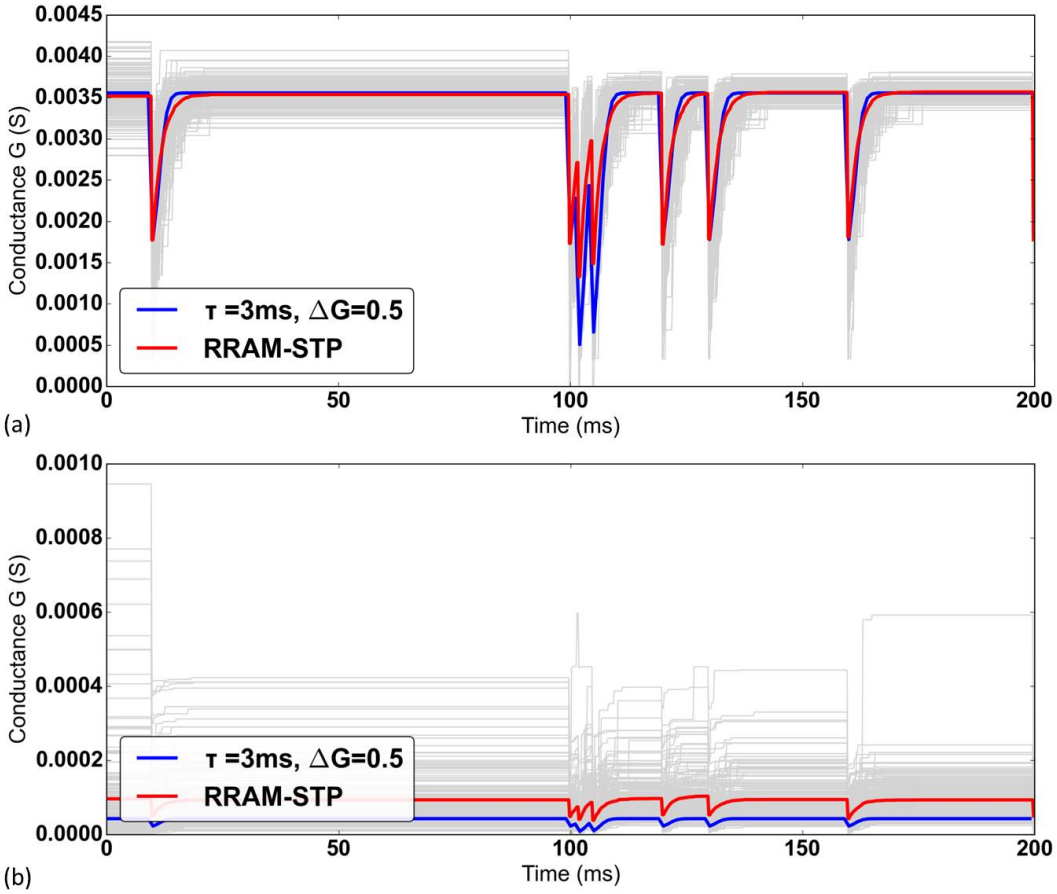


Figure 6.16: Probabilistic STP synapses (grey) based on 10 OxRAM cells in parallel architecture and mean value (red) for (a) $I_{Set} = 400\mu A$ and (b) $I_{Set} = 40\mu A$. The ideal STP trace according to the Tsodyks-Markram model is shown for comparison (blue).

6.6.1 Visual processing with highly noisy input data

This application based on a dedicated SNN was described in detail in chapter 5 and by Bichler et al. [206]. An SNN is used to process temporally encoded video data, recorded directly from an artificial silicon retina [237]. A video of cars passing on a freeway recorded in Address Event Representation (AER) format is presented to a two-layered SNN. In each layer, every input is connected to every output by a single RRAM synapse [157]. The quantification procedure, i.e. extracting the number of True Positives (TP), False Positives (FP) and False Negatives (FN) is performed analogue to chapter 5.

Figure 6.19 reports the number of FP and the DR obtained using the strong and weak programming conditions for the OxRAM array (see figure 6.13). The synapses in these simulations featured only Long Term Plasticity (LTP) to study the impact of the number of OxRAM cells per synapse, i.e. the number of resistance levels that can be achieved for a synaptic weight. As one can see, the detection performances do not improve significantly by increasing the number of cells

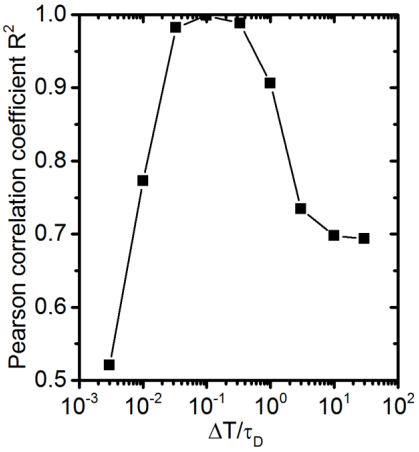


Figure 6.17: Pearson correlation coefficient r^2 for correlation of OxRAM based STP and STP-model as a function $\Delta T/\tau_D$.

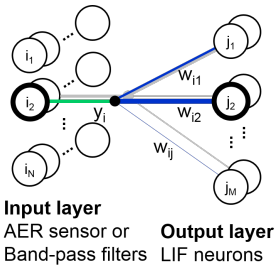


Figure 6.18: Two-layer Spiking Neural Network (SNN) used for car ($N = 16384$, $M = 60$) or spike detection ($N = 32$, $M = 5$).

per synapse. Therefore, we adopt one OxRAM per LTP synapse in the following since this strategy may as well simplify the integration circuitry. The weak programming condition (blue) achieves a lower FN rate (> 3 %) but shows a slightly higher FP rate (< 2.5 %). The weak condition is subsequently used for the LTP synapses because it allows to reduce the power consumption by 26 % with respect to the strong condition.

For the following study, the previous synapse implementation featuring only LTP (STDP) was replaced by the novel synapse design featuring both STP and LTP (see section 6.4 for description). Each synapse implicates two OxRAM based synaptic weights, one for STP and another one for LTP. For this reason, $32k$ STP synapses ($128 * 128 * 2$) are required additionally to the $1.97M$ LTP synapses to implement the SNN. By doing this, the False Positive Rate (FPR) could be reduced without degrading the DR (black crosses in figures 6.19 (a) and (b)).

Furthermore, additional noise was artificially introduced in the SNN input signal by adding random spiking activity in the (already noisy) AER data as illustrated in figure 6.20. This scenario mimics natural phenomena which could affect this kind of application, e.g. light reflections, fog or rain. As one can see, the contours of the cars become very blurry for increased noise levels

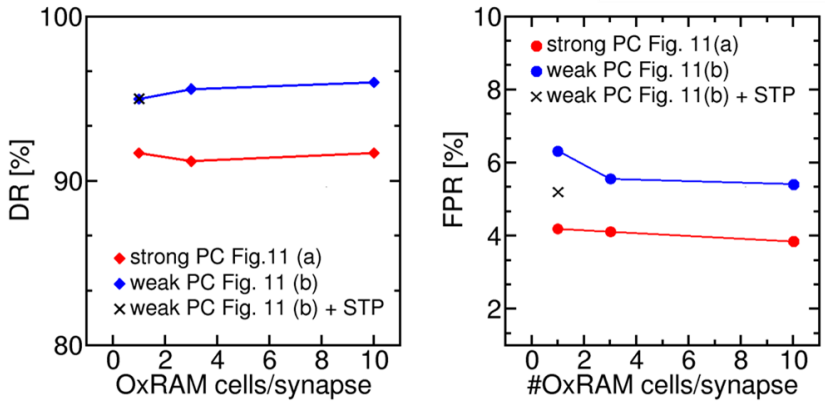


Figure 6.19: Detection Rate (DR) (a) and False Positive Rate (FPR) (b) as a function of the number of cells per Long Term Plasticity synapse. Only Long Term Plasticity is considered. Results have been obtained on a bench of 20 simulations (no added noise).

because of the high background spiking activity. This significantly higher number of background spikes may lead to disturbs in the learning of the SNN and hence degrade the reliable detection of cars.

Simulation results based on the input data with 30% additional noise (i.e. 30% of the total amount of spikes in the sequence are entirely random) are presented in figure 6.21. Without STP (black dashed line), the DR decreases around 15% with respect to the reference case (0% noise) and the FPR increases from a few percent to about 30%. The coloured plots show the simulated performance for STP conditions varying τ_D and f_D. Indeed, both DR and FPR are strongly improved thanks to the introduction of the STP. Higher τ_D values decrease the limiting frequency f_{lim}, (figure 6.7) making the Short Term Plasticity more effective, thus lowering FPR. However, for very long τ_D such as $\tau_D = 30ms$, the DR degrades since the STP affects also the detection of relevant data (i.e. true input spikes from passing cars). It seems to be useful to use high f_D (around 1) in combination with τ_D of around $10ms$ for this specific application.

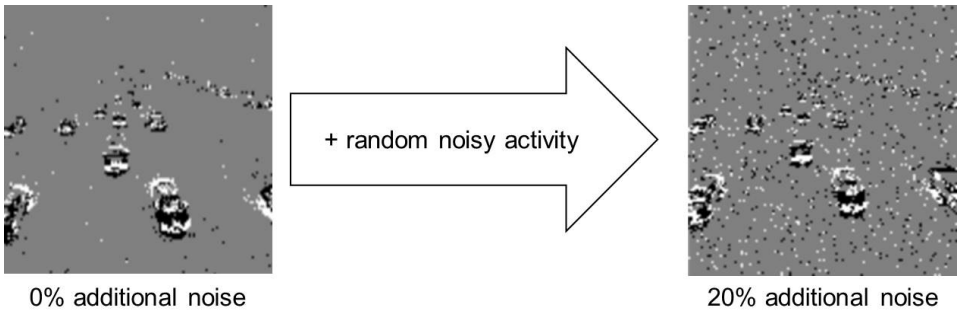


Figure 6.20: Input representation of AER signal while recording cars passing on a freeway. Random noise is added in the right-hand side presentation.

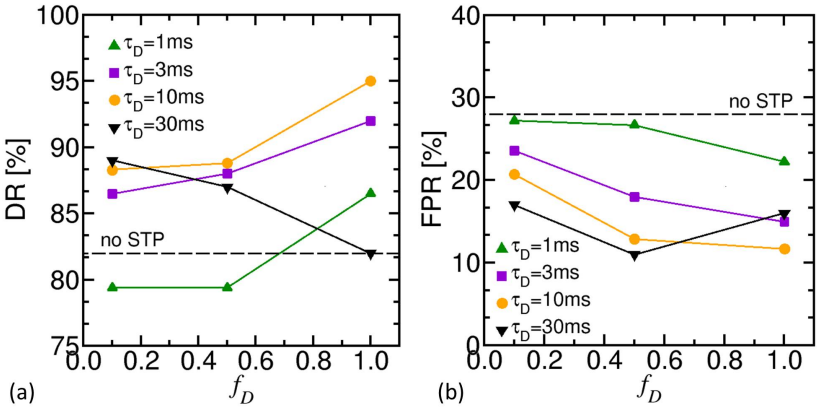
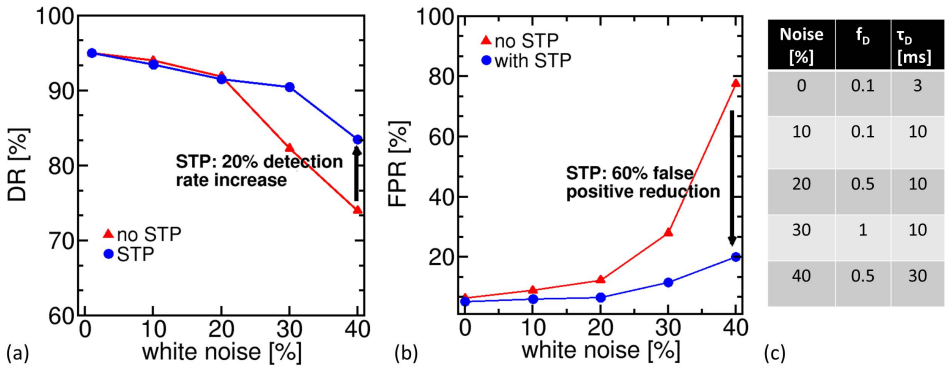


Figure 6.21: Detection Rate (DR) and False Positive Rate (FPR) as a function of f_D and for different τ_D. 30 % of random noise is artificially introduced in the input data. Both Detection Rate (DR) and False Positive Rate (FPR) can be increased by additional Short Term Plasticity with respect to a network featuring only Long Term Plasticity.

Figure 6.22 compares the car detection SNN rates for FN and FP with and without STP as a function of the noise level contained in the input data. Without STP, the DR reduces significantly to approximately 73% and the FPR is heavily increased to about 79% (for 40% noise). On the contrary, by using optimized STP parameters (f_D, τ_D, table in figure 6.22), it is possible to enhance the DR by 10% while reducing the FPR by 60%. Essentially, these results mean that the SNN featuring only LTP can no longer provide reliable car detection results in noisy environments while adding STP to the SNN enables its functionality in such scenarios. This promising result demonstrates that the STP technique is fundamental to maintaining functionality of the network for elevated noise levels.



Noise [%]	f_D	τ_D [ms]
0	0.1	3
10	0.1	10
20	0.5	10
30	1	10
40	0.5	30

Figure 6.22: Detection Rate (DR) and False Positive Rate (FPR) as a funcion of the noise level artificially introduced in the Adress Event Representation (AER) input data. The Short Term Plasticity parameters for each noise level are reported in the right table. Short Term Plasticity maintains the functionality of the Spiking Neural Network for high noise levels.

6.6.2 Spike detection in noisy brain signals

In order to test the spike sorting system described extensively in chapter 4 towards its applicability for noisy spiking data, different artificial data sets were used throughout which the signal-noise-ratio (SNR) has been changed largely, as illustrated in figure 6.23. The data was generated by an algorithm similar to the one described in [240]. The snapshot of the spiking data of $SNR = 80$ shown in figure 6.23 allows easily to identify the spikes since their amplitude exceeds many times over the one of the background noise. Therefore, even simple spike detection based on applying a signal threshold to distinguish between a spike and noise would be sufficient. However, in a realistic signal this is not the case because signals are typically much more noisy and moreover non-stationary, i.e. showing an overall low-frequent drift over time. As for the lowest $SNR = 3$, the spikes are widely hidden by the 'white' noise and it is no longer possible to identify discriminate a spike from noise 'by hand'. Hence. this prevents such simple methods in order to detect spikes in practical applications.

The number of True Positives (TP), False Negatives (FN) and False Positives (FP) are extracted according to figure 5.9. Since the datasets are artificial data sets, a ground truth of the spiking activity is available which can be used for reliable quantification of the spike detection by the spiking neural network. Based on these numbers, the detection rate (DR) and false positive rate (FPR) are calculated according to

$$DR = \frac{N_{TP}}{N_{spikes}} \tag{6.8}$$

where N_{TP} is the number of TP and N_{spikes} is the number of spikes and analogue

$$FPR = \frac{N_{FP}}{N_{spikes}} \tag{6.9}$$

where N_{FP} is the number of FP.

If the SNR is high (80), the network achieves very good Detection Rates (DR) and False Positive Rates (FPR) around 97 % and 1.6 %, even without STP (see figure 6.24). Using additionally STP achieves very similar results of 96 % and 1.4 % for DR and FPR, accordingly. Here, it is worth noting that introducing STP does not degrade the performance in terms of accuracy. However, for a slightly increased noise level ($SNR = 27$), the FPR increases significantly to about 70%. The DR remains at a high 80%, however, the spike detection is no longer reliable due to the very high

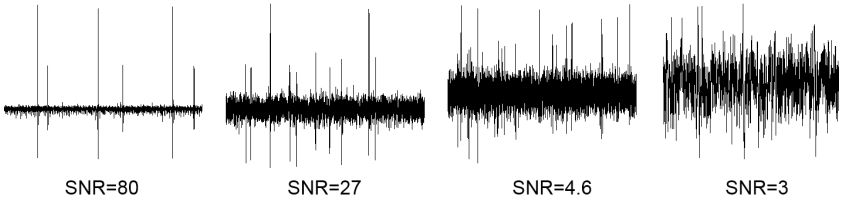


Figure 6.23: Snapshots of spiking data featuring different signal-noise-ratios (SNR).

FPR. For this reason, the introduction of STP is mandatory since it allows to decrease the FPR by 40 % while the DR decreases only 3 %.

The simulated performances of the spike detecting SNN for data with a $SNR = 27$ are shown in figure 6.25 as a function of the STP parameters τ_D and f_D. A small f_D seems to be required for this type of application (to avoid degradation of the DR) whereas long τ_D seem beneficial to reduce the FPR.

Figure 6.26 summarizes the results for the DR and FPR as a function of the SNR. It shows that the DR and FPR are always > 88 % and < 35 %, respectively, over a wide range of the SNR as low as $SNR = 3$ thanks to STP. This allows to use the described system for the detection of neural spikes (or other kinds of signals) superimposed by strong background noise while a SNN without STP would be strongly disturbed by the noise and hence detect much more spikes than actually existent in the data. For this reason, STP seems crucial to ensure the reliable detection of biological spikes from non-relevant background noise.

6.6.3 Implications due to STP

One of the striking advantages of brain-inspired computing approaches (e.g. by Spiking Neural Networks) over conventional von-Neumann architectures is the ultra-low power consumption. This makes SNN especially useful for mobile applications which require parallel computing on a low energy budget. Those applications are very likely to be subject to noisy signals due to the environment which are constantly changing their properties. Thus, SNN's need to be able to adapt independently and in a dynamic manner to the signal, for example provided by STP.

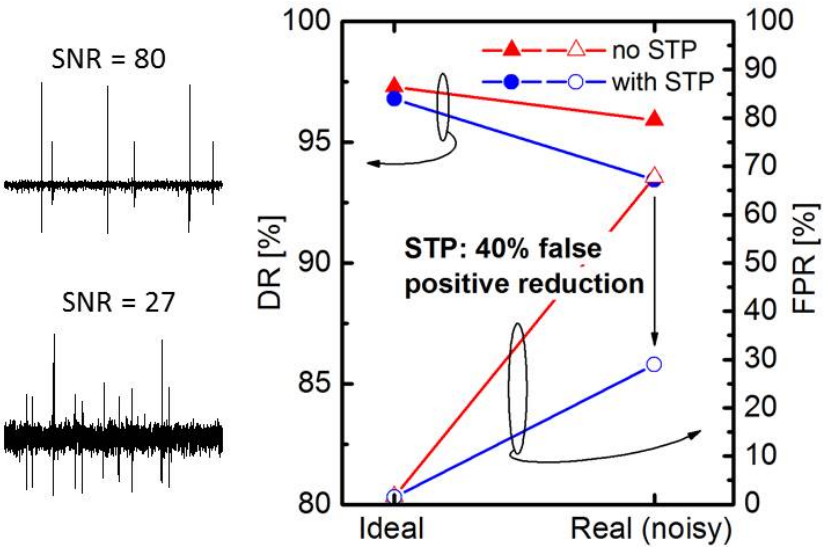


Figure 6.24: Detection Rate (DR) and False Positive Rate (FPR) for the cases of ideal ($SNR = 80$) and real ($SNR < 27$) biological data. Short Term Plasticity is mandatory to reduce the False Positive Rate (FPR) for reliable spike detection in real data.

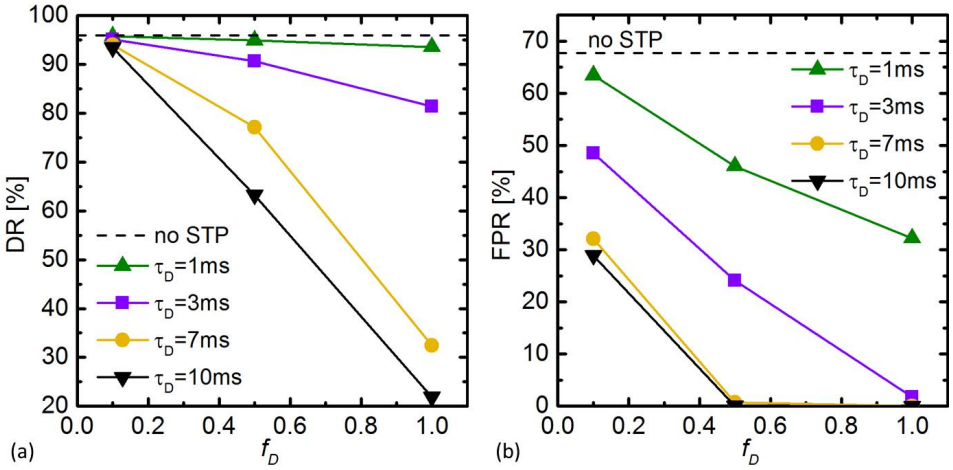


Figure 6.25: (a) Detection Rate (DR) and (b) False Positive Rate (FPR) as a funcion of the degree of depression, f_D, for different recovery times τ_D. Short Term Plasticity allows to reduce the False Positive Rate (FPR) significantly while maintaining a high DR. DR decreases for $f_D > 0.1$ since the Short Term Plasticity disturbs detection of relevant data in this case.

However, the introduction of STP in the synapses of a SNN besides LTP implies a number of obstacles which are briefly explained in the following.

Increased circuit complexity

Considering the concept presented in section 6.3 (based on OxRAM technology) for the implementation of synapses, one layer of LTP synapses has to be integrated and another one needs to be added for the STP synapses. Furthermore, the buffer needs to be integrated to associate LTP and STP. This adds a certain number of process steps (lithography, deposition, metallization) to deposit the two layers of OxRAM devices as well as the buffer. It may be assumed that the processing steps concerning the OxRAM deposition can be used for both LTP and STP fabrication, thus reducing process complexity and saving cost.

Increased energy consumption

STP increases the number of synaptic programming steps since the STP weight is updated after every pre-synaptic spike event. The statistics of the SNN applications for car and spike detection were extracted from the Xnet simulations and are shown in table 6.1. As one can see, the number of LTP programming events can be slightly reduced by around 30% (cars). This is because the STP reduces the synaptic weights which results in reduced probabilities of the neuron to reach their threshold and thus to perform and STDP operation. On the other hand, an enormous number of STP switching events is observed due to the high input activity caused by noise. The number of read events per synapse (STP and LTP component) does not depend on whether STP is used or

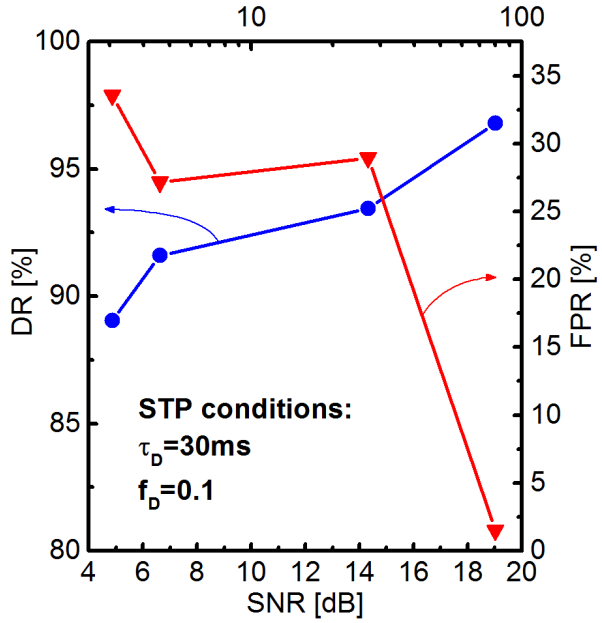


Figure 6.26: Detection Rate (DR) and False Positive Rate (FPR) as a function of the Signal-Noise-Ratio (SNR).

Table 6.1: Statistics of SNN based applications for unsupervised car or spike detection for the highest noise levels: Cars: 40%, Spikes: SNR=2.7.

	Car detection		Spike detection	
	w/o STP	w/ STP	w/o STP	w/ STP
Duration of video (s)	681 s	681 s	200 s	200 s
LTP synapses (#)	2 M	2 M	160	160
STP synapses (#)	-	32 k	-	32
Devices/LTP synapse (#)	1	1	50	50
Devices/STP synapse (#)	-	10	-	10
LTP programming events (#)	56 $x10^6$	39 $x10^6$		470 $x10^3$
STP programming events (#)	-	1.16 $x10^{10}$		275 $x10^6$
Total Read events (#)	4.14 $x10^9$	4.14 $x10^9$		269 $x10^6$

not.

The statistics were used to extract the estimated energy consumptions of the SNN for the application on the noisy datasets, reported in table 6.2. As expected, the significantly higher number of switching events due to the STP synapses results in an increase of the energy consumption by a factor of 175x. Despite the much higher energy consumption, it should be clarified that without using STP, no more functionality is given for SNN applications when the input data is affected by high noise levels.

Table 6.2: Energy estimation of SNN based applications for unsupervised car or spike detection for the highest noise levels: Cars: 40%, Spikes: SNR=2.7.

	Car detection		Spike detection	
	w/o STP	w/ STP	w/o STP	w/ STP
Energy (J)	0.0063	1.1		0.036
Power (mW)	0.0093	1.7		0.18

6.7 Summary

In this chapter, the capability of OxRAM to implement synapses reproducing both Long Term Plasticity (LTP) and Short Term Plasticity (STP) was studied. Therefore, a programming strategy was developed in order to impose a bio-inspired STP behaviour on the intrinsically non-volatile OxRAM devices. A synapse concept merging STP with LTP was proposed and demonstrated by spiking neural network simulations for two applications. It was shown that STP enhances the unsupervised learning of Spiking Neural Networks in case of highly noisy input data and thus enables reliable signal detection. Thanks to the STP implementation, the demonstrated applications exhibit very strong performances, while systems without STP would no longer deliver valid results for signals hidden by 'white' noise. The detection rate of the visual pattern extraction could be increased by 20 % ($\sim$ 83 %) whereas the false positive rate was decreased by 60 % ($\sim$ 20 %) for up to 40 % added random spikes in the input data. Furthermore, it was shown that the STP allows to decrease the false positives by 40 % ($<$ 35 %) while maintaining a detection rate of 89 % for a neural detection application at an ultra low $SNR = 3.7$. It can be concluded, that STP is a critical feature of SNN's which are likely to be exposed to significant noise in target applications. As shown, the performance is not degraded by STP in absence of noise while it is strongly enhanced for high noise levels. Hence, combining LTP with STP results seems a very promising approach towards significantly higher classification reliability. Moreover, the STP adds some autonomy to the network since it behaves as a kind of gain control without any supervision. This is extremely interesting for the implementation of versatile sensing and classification applications into low-power analogue devices.

CHAPTER 7

CONCLUSIONS AND PERSPECTIVES

The scope of this thesis was to gain knowledge and assess the potentialities of applying spiking neural networks to the decoding of neural signals in a spike sorting application. Moreover, we argued that spiking neural network could be realized in low-power hardware implementations in the near future. To this aim, the design of not only of the high level network structure but also of the specific building blocks, namely the neurons and synapses, should be customized.

In the framework of this thesis, we have focused solely on the design of synapses since they typically outnumber the neurons and consume the major part of the total energy of spiking neural networks. We first identified the general conditions that are known to be necessary in order to fulfil a biology-like synaptic behavior, such as analogue weight range and progressive behavior. In order to design our artificial synapses based on resistive memory technologies such as OxRAM and CBRAM, we analyzed these technologies in depth with a focus on the properties that are important to mimic synaptic behavior. It was shown that OxRAM can be either used in a binary or analogue mode, depending on the programming current (I_{CC}) used for the Set operation. For $I_{CC} > 20\mu A$, a filamentary switching was observed whereas for $I_{CC} < 20\mu A$,the resistance switching seemed to be governed by a bulk process. While the binary mode can be used in the so-called compound synapse [168], the analogue operation regime will allow for the use of a single OxRAM device, thus reducing the chip area, circuit level and power consumption due to ultra-low operation currents and high resistances. In order to do so, the very large device-to-device variability should be addressed in the future, to allow for a better understanding of its origin and for the device optimization. Concerning CBRAM, we showed that doped oxide material offers a rather high resistance margin (memory window) between LRS and HRS while operating only at a very low programming current of $I_{CC} = 4.5\mu A$. Typically, those ultra-low

currents do not allow to have a separation of LRS and HRS due to large resistance variability of both states. Hence, doping the oxide of CBRAM (and potentially even OxRAM) devices is a very promising approach. Finally, the design of artificial hardware synapses exploiting binary resistive memories as synapses coupled with a probabilistic learning rule (inspired by biological Spike-Timing-Dependent Plasticity) was addressed. An original spike sorting approach based on a compact spiking neural network (SNN) whose synapses are implemented with OxRAM technologies was proposed. It was shown that this approach allows for sorting simple biological spiking data without supervision, i.e. the network can identify, recognize and distinguish a small number of different spike shapes in the input signal. This autonomy of the network was achieved by using an on-line learning strategy inspired by biological Spike Timing Dependent Plasticity (STDP) which results in the programming of the OxRAM based synaptic weights as a function of the activity of single neurons. It was shown that this concept offers promising advantages compared to conventional spike sorting techniques for brain-computer interfaces applications. In fact, the SNN architecture is suitable for hardware implementation due to promising OxRAM device features, such as low latency ($< 1\mu s$), high integration density ($< 1\mu m^2$) as well as a low energy consumption ($< 75pJ$). For the particular application of spike sorting, these properties may enable real-time operation, opening the way for integration of large SNN and wireless implantable devices for rehabilitation purposes, not yet possible with conventional spike sorting techniques. However, it was also shown that this approach, using frequency filtering to encode neural data, poses a very complex problem of classification since input patterns are overlapping, i.e. the same group of input channels (band-pass filters and input neurons) is used to encode several spiking waveforms. Hence, the spike sorting reliability was not efficient with noisy data containing many different spike shapes. In order to address this critical issue, the data encoding has to be drastically improved, for example by introducing additional input signals such as the phase information instead of using only the amplitude of the filter signals. Also, the temporal and frequency resolution of the filtering should be studied in more detail in order to optimize the performance of the encoding. Moreover, the network topology has probably to be changed from the simple fully-connected 2-layers network towards a convolutional architecture, in order to be able to extract finer sections of the frequency spectra. Eventually, in this kind of applications, the use of band-pass filters for the frequency analysis will be probably replaced by other signal processing techniques such as wavelets, discrete Fourier transforms or others.

The impact of the synaptic variability on spiking neural networks was studied for two applications, one related to object detection and the other one to classification. In both cases, the synapses were implemented using binary (filamentary) OxRAM technology, i.e. a memristive technology featuring two distinct resistance states, low and high resistance state, which are typically separated from each other by some margin (memory window) and both affected by variability. First, the reliability in terms of the detection/classification rate can be improved by enlarging the memory window. Second, independently from the OxRAM device variability, the

same reliability could be achieved assuming that the threshold integration parameter controlling the emission of a neural spike is adapted to this variability. Third, it was found that synaptic variability can have different effects, depending on whether the variability is due to the electrical device variability in the low or in the high resistive state. We concluded that the reliability depends on the relative dynamic range of the network of synapses, i.e. the ratio of the sum of conductances in potentiated state and the ones in depressed state. Hence, it was shown that this dynamic range can be expressed as the average weight of synapses in these two states, defined as the synaptic window. Since the average weights of a synaptic conductance distribution is an arithmetic metric, variability in LRS can increase the dynamic range of a synapse while HRS variability leads to a decrease, assuming that the geometric values of LRS and HRS are constant. Thus, it was proposed that resistance variability in LRS has a positive effect on the performance of a spiking neural network while variability in the HRS has a negative impact. This result is encouraging with respect to RRAM operations because higher variabilities are generally achieved by using lower programming currents and therefore reduce the energy consumption of SNNs as a side effect. These results may be used for the optimization of the operation conditions of RRAM used to design artificial synapses and moreover, the overall understanding of variability effects can be applied for the adaptation of other technologies.

Finally, the capability of OxRAM to implement synapses reproducing both Long Term Plasticity (LTP) and Short Term Plasticity (STP) was studied. Therefore, a programming strategy was developed in order to impose a bio-inspired STP behavior on the intrinsically non-volatile OxRAM devices. A synapse concept merging STP with LTP was proposed and demonstrated by spiking neural network simulations for two applications. It was shown that STP enhances the unsupervised learning of Spiking Neural Networks in case of highly noisy input data and thus enables reliable signal detection. Thanks to the STP implementation, the demonstrated applications exhibit very strong performances, while systems without STP would no longer deliver valid results for signals hidden by white noise. It was shown that this benefit is due to the significant reduction of false positive errors that originate from the noisy input signals. It can be concluded, that STP is a critical feature of SNNs since real-world applications based on SNNs are likely to be exposed to significant (and changing) noise levels. As shown, the performance is not degraded by STP in absence of noise while it is strongly enhanced for high noise levels. Hence, combining LTP with STP seems a very promising approach towards significantly higher robustness of the classification reliability. Moreover, the STP adds some autonomy to the network since it behaves as a kind of gain control without any supervision. This is extremely interesting for the implementation of versatile sensing and classification applications into low-power analogue devices. It should be noted though that the STP increases the energy consumption of SNN's drastically and leads to higher numbers of programming and reading cycles which has to be taken into account in the design of the artificial synapses.

Future work should cover mainly the following subjects:

- Development of a specialized real-time signal processing approach for encoding spiking biological data in the frequency domain allowing to extract specific features from the frequency spectrum.
- Continue the study of variability effects of several network parameters in order to identify systematically the aspects where variability of synaptic or neuronal properties leads to improved or degraded SNN performance.
- Implementation of several types of synaptic (and neuronal) plasticity for certain network functionalities such as noise resilience, flexibility to input data variations and to reduce the requirements of SNN's to precise parameter tuning.

APPENDIX A

APPENDIX A: BAND-PASS FILTERING

A signal can be spectrally analysed in real-time by means of band-pass filters (BPF) which pass signals with frequencies within a specific range (pass band) and attenuate signal components of frequencies outside this range. The BPF generate continuous analog output signals comparable to a Frequency Domain Spectrogram (also commonly known as waterfall, i.e. the evolution of the Fourier spectrum over time). As shown in figure A.1, a BPF is characterised by a low and high cut-off frequency, f_L and f_H, respectively, where the signal attenuation amounts to $-3\ dB$ ($-50\ \%$) relatively to the peak ($-0\ dB$). The range between f_L and f_H is referred to as the bandwidth B. Equally distanced between f_L and f_H lies the so-called center frequency f_0. The filter characteristic in the pass-band (flatness) and the shape factor of the slopes (i.e. frequency selectivity) are determined by the filter order (≥ 1). The higher the order, the steeper the slopes of a BPF as illustrated in figure A.2. A high order enables an increased

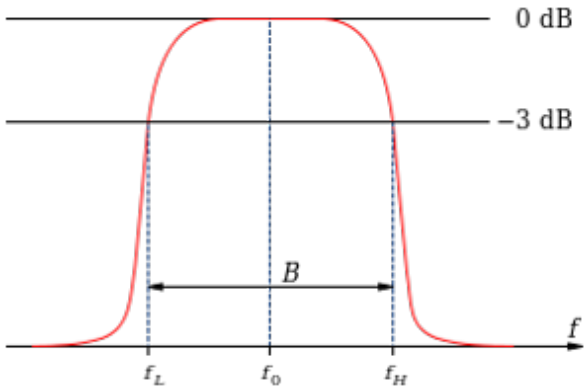


Figure A.1: Band-pass filter characteristic.

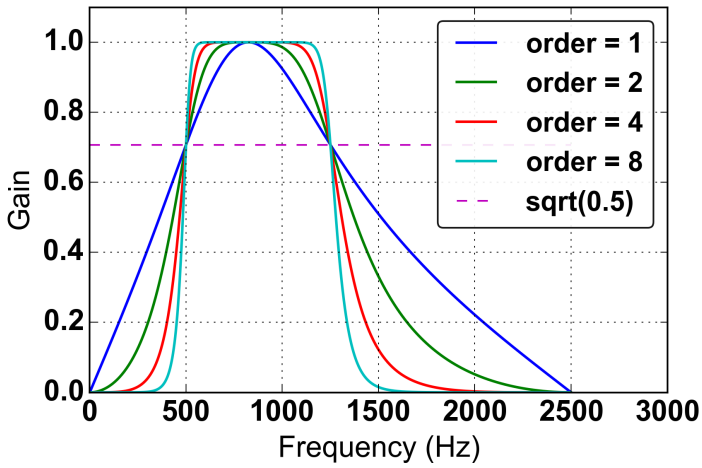


Figure A.2: Band-pass filter characteristics for Butterworth filters of order 1, 2, 4 and 8. The horizontal dashed line indicates the cut-off level of $-3\ dB$.

frequency resolution, however, reduces the temporal resolution. This time-frequency dilemma is a common concern to real-time signal processing.

APPENDIX A

APPENDIX B: PUBLICATIONS

- E. Vianello, **T. Werner**, A. Grossi, E. Nowak, B. De Salvo, L. Perniola, O. Bichler, B. Yvert, 'Bioinspired Programming of Resistive Memory Devices for Implementing Spiking Neural Networks', *Proc. of GLSVLSI*, 2017.
- E. Vianello, **T. Werner**, G. Piccolboni, D. Garbin, O. Bichler, G. Molas, J.M. Portal, B. Yvert, B. De Salvo, L. Perniola, 'Binary OxRAM/CBRAM Memories for Efficient Implementations of Embedded Neuromorphic Circuits' (book chapter), Neuro-inspired Computing Using Resistive Synaptic Devices, pp.253-269, January 2017.
- **T. Werner**, E. Vianello, O. Bichler, A. Grossi, E. Nowak, J.-F. Nodin, B. Yvert, B. De Salvo, L. Perniola, 'Experimental Demonstration of Short and Long Term Synaptic Plasticity Using OxRAM Multi k-bit Arrays for Reliable Detection in Highly Noisy Input Data', *Proc. of IEEE IEDM*, 2016.
- **T. Werner**, E. Vianello, O. Bichler, D. Garbin, D. Cattaert, B. Yvert, B. De Salvo and L. Perniola, 'Spiking Neural Networks Based on OxRAM Synapses for Real-Time Unsupervised Spike Sorting', *Frontiers in Neuroscience*, 2016.
- **T. Werner**, E. Vianello, O. Bichler, B. Yvert, B. De Salvo, L. Perniola, 'Exploitation of RRAM variability to improve on-line unsupervised learning in small-scale Spiking Neural Networks', *Proc. of SSDM*, 2016.
- G. Piccolboni, G. Molas, D. Garbin, **T. Werner**, E. Vianello, B. De Salvo, G. Ghibaudo, and L. Perniola, 'Investigation of variability in Vertical Resistive RAM (VRRAM): Physical Model', *Proc. of SSDM*, 2016.
- **T. Werner**, D. Garbin, E. Vianello, O. Bichler, D. Cattaert, B. Yvert, B. De Salvo, L. Perniola, 'Real-time decoding of brain activity by embedded Spiking Neural Networks using OxRAM

synapses', *Proc. of IEEE ISCAS*, 2016.

- **T. Werner**, E. Vianello, B. Yvert, B. De Salvo, and L. Perniola, 'Low power OxRAM devices for the design of artificial synapses', *Proc. of IEEE EMBS NER*, 2015.

BIBLIOGRAPHY

[1] D. Purves, G. J. Augustine, D. Fitzpatrick, W. C. Hall, A.-S. LaMantia, J. O. McNamara, S. M. Williams, M. Bear, B. Connors, and M. Paradiso, *Neuroscience*.
3 ed., 2007.

[2] G. Q. Bi and M. M. Poo, "Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type.," *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 18, pp. 10464–72, dec 1998.

[3] L. R. Hochberg, M. D. Serruya, G. M. Friehs, J. a. Mukand, M. Saleh, A. H. Caplan, A. Branner, D. Chen, R. D. Penn, and J. P. Donoghue, "Neuronal ensemble control of prosthetic devices by a human with tetraplegia.," *Nature*, vol. 442, pp. 164–71, jul 2006.

[4] L. R. Hochberg, D. Bacher, B. Jarosiewicz, N. Y. Masse, J. D. Simeral, J. Vogel, S. Haddadin, J. Liu, S. S. Cash, P. van der Smagt, and J. P. Donoghue, "Reach and grasp by people with tetraplegia using a neurally controlled robotic arm.," *Nature*, vol. 485, pp. 372–5, may 2012.

[5] H.-S. P. Wong and S. Salahuddin, "Memory leads the way to better computing," *Nature Nanotechnology*, vol. 10, no. March, pp. 191–194, 2015.

[6] H.-S. P. Wong, S. Raoux, S. Kim, J. Liang, J. P. Reifenberg, B. Rajendran, M. Asheghi, and K. E. Goodson, "Phase Change Memory," *Proceedings of the IEEE*, vol. 98, pp. 2201–2227, dec 2010.

[7] H.-S. P. Wong, H.-Y. Lee, S. Yu, Y.-S. Chen, Y. Wu, P.-S. Chen, B. Lee, F. T. Chen, and M.-J. Tsai, "Metal‚ÄìOxide RRAM," *Proceedings of the IEEE*, vol. 100, pp. 1951–1970, jun 2012.

[8] R. Waser, R. Dittmann, G. Staikov, and K. Szot, "Redox-Based Resistive Switching Memories - Nanoionic Mechanisms, Prospects, and Challenges," *Advanced Materials*, vol. 21, pp. 2632–2663, jul 2009.

[9] R. L. Stamps, S. Breitkreutz, J. Åkerman, A. V. Chumak, Y. Otani, G. E. W. Bauer, J.-U. Thiele, M. Bowen, S. A. Majetich, M. Kläui, I. L. Prejbeanu, B. Dieny, N. M. Dempsey,

and B. Hillebrands, “The 2014 Magnetism Roadmap,” *Journal of Physics D: Applied Physics*, vol. 47, no. 33, p. 333001, 2014.

[10] J. Müller, P. Polakowski, S. Mueller, and T. Mikolajick, “Ferroelectric Hafnium Oxide Based Materials and Devices: Assessment of Current Status and Future Prospects,” *ECS Journal of Solid State Science and Technology*, vol. 4, no. 5, p. N30, 2015.

[11] J. Muller, “Ferroelectric hafnium oxide : A CMOS compatible and highly scalable approach to future ferroelectric memories,” *Electron Devices Meeting (IEDM), 2013 IEEE International*, no. 9, pp. 10.8.1 – 10.8.4, 2013.

[12] S. Park, M. Yang, and H. Ju, “A non-linear ReRAM cell with sub-1μA ultralow operating current for high density vertical resistive memory (VRRAM),” *... (IEDM), 2012 IEEE ...*, pp. 501–504, 2012.

[13] A. Chen, “A review of emerging non-volatile memory (NVM) technologies and applications,” *Solid-State Electronics*, vol. 125, pp. 25–38, 2016.

[14] M. Suri, O. Bichler, D. Querlioz, B. TraoreÃÅ, O. Cueto, L. Perniola, V. Sousa, D. Vuillaume, C. Gamrat, and B. DeSalvo, “Physical aspects of low power synapses based on phase change memory devices,” *Journal of Applied Physics*, vol. 112, no. 5, p. 054904, 2012.

[15] O. Bichler, M. Suri, D. Querlioz, D. Vuillaume, B. DeSalvo, and C. Gamrat, “Visual Pattern Extraction Using Energy-Efficient ‚Äú2-PCM Synapse‚Äù Neuromorphic Architecture,” *IEEE Transactions on Electron Devices*, vol. 59, pp. 2206–2214, aug 2012.

[16] Y. Wu, S. Yu, H.-S. P. Wong, Y.-S. Chen, H.-Y. Lee, S.-M. Wang, P.-Y. Gu, F. Chen, and M.-J. Tsai, “AlOx-Based Resistive Switching Device with Gradual Resistance Modulation for Neuromorphic Device Application,” *2012 4th IEEE International Memory Workshop*, vol. 1, pp. 1–4, may 2012.

[17] M. Suri, O. Bichler, D. Querlioz, G. Palma, E. Vianello, D. Vuillaume, C. Gamrat, and B. DeSalvo, “CBRAM devices as binary synapses for low-power stochastic neuromorphic systems: Auditory (Cochlea) and visual (Retina) cognitive processing applications,” *2012 International Electron Devices Meeting*, pp. 10.3.1–10.3.4, dec 2012.

[18] A. F. Vincent, J. Larroque, W. S. Zhao, N. B. Romdhane, O. Bichler, C. Gamrat, J. Klein, and D. Querlioz, “Spin-Transfer Torque Magnetic Memory as a Stochastic Memristive Synapse,” vol. 1, pp. 1074–1077, 2014.

[19] A. Sengupta and K. Roy, “Short-Term Plasticity and Long-Term Potentiation in Magnetic Tunnel Junctions: Towards Volatile Synapses,” *Physical Review Applied*, vol. 5, no. 2, 2016.

[20] L. T. Clark, R. Grondin, and S. K. Dey, "Integrated Circuit Neural Networks Using Ferroelectric Analog Memory," *IPCCC*, pp. 736–742, 1992.

[21] Y. Kaneko and Y. Nishitani, "Neural network based on a three-terminal ferroelectric memristor to enable on-chip pattern recognition," *. . .), 2013 Symposium on*, vol. 99, pp. 2012–2013, 2013.

[22] S. Boyn, J. Grollier, G. Lecerf, B. Xu, N. Locatelli, S. Fusil, S. Girod, C. Carrétéro, K. Garcia, S. Xavier, J. Tomas, L. Bellaiche, M. Bibes, A. Barthélémy, S. Saïghi, and V. Garcia, "Learning through ferroelectric domain dynamics in solid-state synapses," *Nature Communications*, vol. 8, p. 14736, 2017.

[23] D. Cattaert and A. E. Manira, "Shunting versus inactivation: analysis of presynaptic inhibitory mechanisms in primary afferents of the crayfish," *The Journal of neuroscience*, vol. 19, no. 14, pp. 6079–6089, 1999.

[24] D. a. Henze, Z. Borhegyi, J. Csicsvari, a. Mamiya, K. D. Harris, and G. Buzsáki, "Intracellular features predicted by extracellular recordings in the hippocampus in vivo.," *Journal of neurophysiology*, vol. 84, no. 1, pp. 390–400, 2000.

[25] M. V. Tsodyks and H. Markram, "The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability," *Proc. Natl. Acad. Sci. USA*, vol. 94, pp. 719–723, 1997.

[26] J. Wessberg, C. R. Stambaugh, J. D. Kralik, P. D. Beck, M. Laubach, J. K. Chapin, J. Kim, S. J. Biggs, M. a. Srinivasan, and M. a. Nicolelis, "Real-time prediction of hand trajectory by ensembles of cortical neurons in primates.," *Nature*, vol. 408, pp. 361–5, nov 2000.

[27] P. J. Ifft, S. Shokur, Z. Li, M. a. Lebedev, and M. a. L. Nicolelis, "A brain-machine interface enables bimanual arm movements in monkeys.," *Science translational medicine*, vol. 5, p. 210ra154, nov 2013.

[28] M. E. Spira and A. Hai, "Multi-electrode array technologies for neuroscience and cardiology.," *Nature nanotechnology*, vol. 8, pp. 83–94, feb 2013.

[29] M. Yin, D. a. Borton, J. Komar, N. Agha, Y. Lu, H. Li, J. Laurens, Y. Lang, Q. Li, C. Bull, L. Larson, D. Rosler, E. Bezard, G. Courtine, and A. V. Nurmikko, "Wireless neurosensor for full-spectrum electrophysiology recordings during free behavior.," *Neuron*, vol. 84, pp. 1170–82, dec 2014.

[30] M. Abeles and M. G. Jr, "Multispike train analysis," *Proceedings of the IEEE*, vol. 65, no. 5, 1977.

[31] J. von Neumann, "First Draft of a Report on the EDAVAC," *American Mathematical Society*, vol. 15, no. 1, pp. 1–10, 1945.

[32] F. Walter, F. Röhrbein, and A. Knoll, "Neuromorphic implementations of neurobiological learning algorithms for spiking neural networks," *Neural Networks*, vol. 72, pp. 152–167, 2015.

[33] a. B. Schwartz, "Direct cortical representation of drawing.," *Science (New York, N.Y.)*, vol. 265, pp. 540–2, jul 1994.

[34] G. Buzsáki and A. Draguhn, "Neuronal oscillations in cortical networks.," *Science (New York, N.Y.)*, vol. 304, pp. 1926–9, jun 2004.

[35] Y. Sakurai, K. Song, S. Tachibana, and S. Takahashi, "Volitional enhancement of firing synchrony and oscillation by neuronal operant conditioning: interaction with neurorehabilitation and brain-machine interface," *Frontiers in Systems Neuroscience*, vol. 8, no. February, p. 11, 2014.

[36] B. Walmsley, F. Edwards, and D. Tracey, "The probabilistic nature of synaptic transmission at a mammalian excitatory central synapse," *J. Neurosci.*, vol. 7, no. 4, pp. 1037–1046, 1987.

[37] A. Neishabouri and A. A. Faisal, "Axonal Noise as a Source of Synaptic Variability," *PLoS Computational Biology*, vol. 10, no. 5, 2014.

[38] L. L. R. L. Squire, "Mechanisms of memory," *Science*, vol. 232, no. 1983, 1986.

[39] D. Purves and J. W. Lichtman, "Elimination of synapses in the developing nervous system.," *Science (New York, N.Y.)*, vol. 210, pp. 153–7, oct 1980.

[40] C. W. Cotman and M. Nieto-Sampedro, "The cell biology of synaptic plasticity.," *Science (New York, N.Y.)*, vol. 225, pp. 1287–1294, sep 1984.

[41] W. G. Regehr, "Short-Term Presynaptic Plasticity," pp. 1–19, 2012.

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

[43] H. Markram, J. Lubke, M. Frotscher, B. Sakmann, J. Lu, M. Frotscher, and B. Sakmann, "Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs," *Science*, vol. 275, no. 5297, pp. 213–215, 1997.

[44] G. Le Masson, S. Renaud-Le Masson, D. Debay, and T. Bal, "Feedback inhibition controls spike transfer in hybrid thalamic circuits.," *Nature*, vol. 417, pp. 854–8, jun 2002.

[45] P. K. Stanton, "LTD, LTP, and the sliding threshold for long-term synaptic plasticity," *Hippocampus*, vol. 6, no. 1, pp. 35–42, 1996.

[46] S. Kellis, L. Sorensen, F. Darvas, C. Sayres, K. O. Neill, R. B. Brown, P. House, J. Ojemann, and B. Greger, "Multi-scale analysis of neural activity in humans: Implications for micro-scale electrocorticography," *Clinical Neurophysiology*, vol. 127, no. 1, pp. 591–601, 2016.

[47] G. Buzsáki, C. a. Anastassiou, and C. Koch, "The origin of extracellular fields and currents–EEG, ECoG, LFP and spikes.," *Nature reviews. Neuroscience*, vol. 13, pp. 407–20, jun 2012.

[48] G. Buzsáki, "Large-scale recording of neuronal ensembles.," *Nature neuroscience*, vol. 7, pp. 446–51, may 2004.

[49] A. G. Zippo, P. Romanelli, N. R. Torres Martinez, G. C. Caramenti, A. L. Benabid, and G. E. M. Biella, "A novel wireless recording and stimulating multichannel epicortical grid for supplementing or enhancing the sensory-motor functions in monkey (Macaca fascicularis)," *Frontiers in Systems Neuroscience*, vol. 9, no. May, pp. 1–12, 2015.

[50] A. a. Prinz, L. F. Abbott, and E. Marder, "The dynamic clamp comes of age.," *Trends in neurosciences*, vol. 27, pp. 218–24, apr 2004.

[51] A. Cutrone, J. Del Valle, D. Santos, J. Badia, C. Filippeschi, S. Micera, X. Navarro, and S. Bossi, "A three-dimensional self-opening intraneural peripheral interface (SELINE).," *Journal of neural engineering*, vol. 12, no. 1, p. 016016, 2015.

[52] D. Hill, S. Mehta, and D. Kleinfeld, "Quality metrics to accompany spike sorting of extracellular signals," *The Journal of Neuroscience*, vol. 31, no. 24, pp. 8699–8705, 2011.

[53] R. Q. Quiroga, "Spike sorting.," *Current biology : CB*, vol. 22, pp. R45–6, jan 2012.

[54] R. Bestel, A. W. Daus, and C. Thielemann, "A novel automated spike sorting algorithm with adaptable feature extraction.," *Journal of neuroscience methods*, vol. 211, pp. 168–78, oct 2012.

[55] C. Pedreira, J. Martinez, M. J. Ison, and R. Quian Quiroga, "How many neurons can we see with current spike sorting algorithms?," *Journal of neuroscience methods*, vol. 211, pp. 58–65, oct 2012.

[56] M. Lewicki, "A review of methods for spike sorting: the detection and classification of neural action potentials," *Network: Computation in Neural Systems*, vol. 9, no. January, 1998.

[57] J. M. A. Tanskanen, F. E. Kapucu, and J. A. K. Hyttinen, "On the Threshold Based Neuronal Spike Detection , and an Objective Criterion for Setting the Threshold," *Neural Engineering*, pp. 22–24, 2015.

[58] T. Datta-Chaudhuri, "An active micro-electrode array with spike detection and asynchronous readout," ... *Circuits and Systems* ..., pp. 3–6, 2014.

[59] H. Takahashi, M. Suezawa, and K. Sumino, "Charge-state-dependent activation energy for diffusion of iron in silicon," *Physical Review B*, vol. 46, no. 3, pp. 1882–1885, 1992.

[60] A. M. Mamlouk, H. Sharp, K. M. L. Menne, U. G. Hofmann, and T. Martinetz, "Unsupervised spike sorting with ICA and its evaluation using GENESIS simulations," *Neurocomputing*, vol. 65-66, no. SPEC. ISS., pp. 275–282, 2005.

[61] D. A. Adamos, E. K. Kosmidis, and G. Theophilidis, "Performance evaluation of PCA-based spike sorting algorithms," *Computer Methods and Programs in Biomedicine*, vol. 91, no. 3, pp. 232–244, 2008.

[62] M. Delescluse and C. Pouzat, "Efficient spike-sorting of multi-state neurons using inter-spike intervals information," *Journal of Neuroscience Methods*, vol. 150, no. 1, pp. 16–29, 2006.

[63] A. Bar-Hillel, A. Spiro, and E. Stark, "Spike sorting: Bayesian clustering of non-stationary data," *Journal of Neuroscience Methods*, vol. 157, no. 2, pp. 303–316, 2006.

[64] F. Wood and M. J. Black, "A nonparametric Bayesian alternative to spike sorting," *Journal of neuroscience* ..., vol. 173, no. 1, pp. 13–23, 2008.

[65] S. Gibson, J. Judy, and D. Marković, "Comparison of spike-sorting algorithms for future hardware implementation," *Engineering in Medicine and* ..., pp. 5015–5020, 2008.

[66] A. Oliynyk, C. Bonifazzi, F. Montani, and L. Fadiga, "Automatic online spike sorting with singular value decomposition and fuzzy C-mean clustering.," *BMC neuroscience*, vol. 13, no. 1, p. 96, 2012.

[67] T. Levi, J.-F. Beche, S. Bonnet, and R. Escola, "METHODS AND DEVICES FOR PROCESSING PULSE SIGNALS, AND IN PARTICULAR NEURAL ACTION POTENTIAL SIGNALS," 2012.

[68] A. Kamboh and A. J. Mason, "On-chip feature extraction for spike sorting in high density implantable neural recording systems," *Biomedical Circuits and Systems* ..., pp. 13–16, 2010.

[69] B. Yu, T. Mak, X. Li, F. Xia, A. Yakovlev, Y. Sun, and C.-S. Poon, "Real-Time FPGA-Based Multichannel Spike Sorting Using Hebbian Eigenfilters," *Ieee Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 1, no. 4, pp. 502–515, 2011.

[70] W.-J. Hwang, W.-H. Lee, S.-J. Lin, and S.-Y. Lai, "Efficient architecture for spike sorting in reconfigurable hardware.," *Sensors (Basel, Switzerland)*, vol. 13, pp. 14860–87, jan 2013.

[71] J. Dragas, D. Jackel, A. Hierlemann, and F. Franke, "Complexity Optimisation and High-Throughput Low-Latency Hardware Implementation of a Multi-Electrode Spike-Sorting Algorithm," vol. 23, no. 2, pp. 149–158, 2014.

[72] K. Kim and S. Kim, "Neural spike sorting under nearly 0-dB signal-to-noise ratio using nonlinear energy operator and artificial neural-network classifier," *Biomedical Engineering, IEEE Transactions . . .* , vol. 47, no. 10, pp. 1406–1411, 2000.

[73] T. I. Aksenova, O. K. Chibirova, O. a. Dryga, I. V. Tetko, A.-L. Benabid, and A. E. Villa, "An unsupervised automatic method for sorting neuronal spike waveforms in awake and freely moving animals," *Methods*, vol. 30, pp. 178–187, jun 2003.

[74] O. K. Chibirova, T. I. Aksenova, A.-L. Benabid, S. Chabardes, S. Larouche, J. Rouat, and A. E. P. Villa, "Unsupervised Spike Sorting of extracellular electrophysiological recording in subthalamic nucleus of Parkinsonian patients.," *Bio Systems*, vol. 79, no. 1-3, pp. 159–71, 2005.

[75] F. Öhberg and H. Johansson, "A neural network approach to real-time spike discrimination during simultaneous recording from several multi-unit nerve filaments," *Journal of neuroscience . . .* , vol. 64, 1996.

[76] Y. Yang, A. J. Mason, and S. Member, "On-Chip Spike Clustering & Classification using Self Organizing Map for Neural Recording Implants," pp. 145–148, 2011.

[77] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[78] M. Aghagolzadeh and K. Oweiss, "Compressed and distributed sensing of neuronal activity for real time spike train decoding," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 17, no. 2, pp. 116–127, 2009.

[79] R. Quiroga, Z. Nadasdy, and Y. Ben-Shaul, "Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering," *Neural computation*, vol. 1687, pp. 1661–1687, 2004.

[80] D. Rinberg, W. Bialek, H. Davidowitz, and N. Tishby, "Spike sorting in the frequency domain with overlap detection," *arXiv preprint physics/ ...*, p. 30, jun 2003.

[81] B. Zhang, Z. Jiang, Q. Wang, J.-s. Seo, and M. Seok, "A neuromorphic neural spike clustering processor for deep-brain sensing and stimulation systems," 2015.

[82] H. G. Rey, C. Pedreira, and R. Q. Quiroga, "Past, present and future of spike sorting techniques," *Brain Research Bulletin*, pp. 1–12, 2015.

[83] J. K. Chapin, K. a. Moxon, R. S. Markowitz, and M. a. Nicolelis, "Real-time control of a robot arm using simultaneously recorded neurons in the motor cortex.," *Nature neuroscience*, vol. 2, pp. 664–70, jul 1999.

[84] A. K. Bansal, W. Truccolo, C. E. Vargas-Irwin, and J. P. Donoghue, "Decoding 3D reach and grasp from hybrid signals in motor and premotor cortices: spikes, multiunit activity, and local field potentials," 2012.

[85] M. Spüler, A. Walter, A. Ramos Murguialday, G. Naros, N. Birbaumer, A. Gharabaghi, W. Rosenstiel, and M. Bogdan, "Decoding of motor intentions from epidural ECoG recordings in severely paralyzed chronic stroke patients.," *Journal of neural engineering*, vol. 11, no. 6, p. 066008, 2014.

[86] B. Wodlinger, J. E. Downey, E. C. Tyler-Kabara, A. B. Schwartz, M. L. Boninger, and J. L. Collinger, "Ten-dimensional anthropomorphic arm control in a human brain,àímachine interface: difficulties, solutions, and limitations," *Journal of Neural Engineering*, vol. 12, no. 1, p. 016011, 2015.

[87] S. Todorova, P. Sadtler, A. Batista, S. Chase, and V. Ventura, "To sort or not to sort: the impact of spike-sorting on neural decoding performance.," *Journal of neural engineering*, vol. 11, no. 5, p. 056005, 2014.

[88] K. D. Harris, D. a. Henze, J. Csicsvari, H. Hirase, and G. Buzsáki, "Accuracy of tetrode spike separation as determined by simultaneous intracellular and extracellular measurements.," *Journal of neurophysiology*, vol. 84, no. 1, pp. 401–414, 2000.

[89] R. Quian Quiroga and S. Panzeri, "Extracting information from neuronal populations: information theory and decoding approaches.," *Nature reviews. Neuroscience*, vol. 10, pp. 173–85, mar 2009.

[90] J. L. Collinger, B. Wodlinger, J. E. Downey, W. Wang, E. C. Tyler-Kabara, D. J. Weber, A. J. C. McMorland, M. Velliste, M. L. Boninger, and A. B. Schwartz, "High-performance neuroprosthetic control by an individual with tetraplegia.," *Lancet*, vol. 381, pp. 557–64, feb 2013.

[91] S. F. Cogan, "Neural stimulation and recording electrodes," *Annual review of biomedical engineering*, vol. 10, pp. 275–309, 2008.

[92] V. Gilja, C. Pandarinath, C. H. Blabe, P. Nuyujukian, J. D. Simeral, A. a. Sarma, B. L. Sorice, J. a. Perge, B. Jarosiewicz, L. R. Hochberg, K. V. Shenoy, and J. M. Henderson, "Clinical translation of a high-performance neural prosthesis," *Nature Medicine*, vol. 21, no. 10, pp. 1142–1145, 2015.

[93] A. Chen, "Emerging Nonvolatile memory (NVM) technologies," *European Solid-State Circuits Conference*, pp. 109–113, 2015.

[94] L. Chua, "Memristor-The missing circuit element," *IEEE Transactions on Circuit Theory*, vol. 18, no. 5, pp. 507–519, 1971.

[95] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, "The missing memristor found.," *Nature*, vol. 453, no. 7191, pp. 80–3, 2008.

[96] W. W. Koelmans, A. Sebastian, V. P. Jonnalagadda, D. Krebs, L. Dellmann, and E. Eleftheriou, "Projected phase-change memory devices.," *Nature communications*, vol. 6, no. May, pp. 1–7, 2015.

[97] Y. Lin, Y. Chen, F. Lee, M. BrightSky, H. Lung, and C. Lam, "A Simple New Write Scheme for Low Latency Operation of Phase Change Memory," vol. 294, no. 2011, pp. 2011–2012, 2012.

[98] H. Lung, Y. Ho, Y. Zhu, W. Chien, S. Kim, W. Kim, H. Cheng, a. Ray, M. Brightsky, R. Bruce, C. Yeh, and C. Lam, "A novel low power phase change memory using inter-granular switching," *2016 IEEE Symposium on VLSI Technology*, vol. 692, no. 2014, pp. 1–2, 2016.

[99] I. S. Kim, S. L. Cho, D. H. Im, E. H. Cho, D. H. Kim, G. H. Oh, D. H. Ahn, S. O. Park, S. W. Nam, J. T. Moon, and C. H. Chung, "High performance PRAM cell scalable to sub-20nm technology with below 4F2 cell size, extendable to DRAM applications," *Digest of Technical Papers - Symposium on VLSI Technology*, pp. 203–204, 2010.

[100] J. Gibbons and W. Beadle, "Switching properties of thin NiO films," *Solid-State Electronics*, vol. 7, no. 2, pp. 785–797, 1964.

[101] L. Goux, Y.-Y. Chen, L. Pantisano, X.-P. Wang, G. Groeseneken, M. Jurczak, and D. J. Wouters, "On the Gradual Unipolar and Bipolar Resistive Switching of TiN\HfO[sub 2]\Pt Memory Systems," *Electrochemical and Solid-State Letters*, vol. 13, no. 6, p. G54, 2010.

[102] G. Bersuker, D. C. Gilmer, D. Veksler, P. Kirsch, L. Vandelli, a. Padovani, L. Larcher, K. McKenna, a. Shluger, V. Iglesias, M. Porti, and M. NafriÃÅa, "Metal oxide resistive memory switching mechanism based on conductive filament properties," *Journal of Applied Physics*, vol. 110, no. 12, p. 124518, 2011.

[103] U. Russo, D. Ielmini, C. Cagli, and A. L. Lacaita, "Filament conduction and reset mechanism in NiO-based resistive-switching memory (RRAM) devices," *IEEE Transactions on Electron Devices*, vol. 56, no. 2, pp. 186–192, 2009.

[104] L. Goux, A. Fantini, G. Kar, Y. Y. Chen, N. Jossart, R. Degraeve, S. Clima, B. Govoreanu, G. Lorenzo, G. Pourtois, D. J. Wouters, J. a. Kittl, L. Altimime, and M. Jurczak, "Ultralow sub-500nA operating current high-performance TiN\Al 2O 3\HfO 2\Hf\TiN bipolar RRAM achieved through understanding-based stack-engineering," *Digest of Technical Papers - Symposium on VLSI Technology*, pp. 159–160, 2012.

[105] Y. S. Chen, H. Y. Lee, P. S. Chen, W. S. Chen, K. H. Tsai, P. Y. Gu, T. Y. Wu, C. H. Tsai, S. Z. Rahaman, Y. D. Lin, F. Chen, M. J. Tsai, and T. K. Ku, "Novel defects-trapping TaOX/HfOX RRAM with reliable self-compliance, high nonlinearity, and ultra-low current," *IEEE Electron Device Letters*, vol. 35, no. 2, pp. 202–204, 2014.

[106] W. Kim, J. Kim, and J. Moon, "Effect of Inserting Al2O3 Layer and Device Structure in HfO2-Based ReRAM for Low Power Operation," ... *(IMW), 2012 4th* ..., pp. 3–6, 2012.

[107] Y. Wu, B. Lee, and H. P. Wong, "Al2O3-Based RRAM Using Atomic Layer Deposition (ALD) With 1uA RESET Current," *IEEE Electron Device Letters*, vol. 31, no. 12, pp. 1449–1451, 2010.

[108] Y. Wu, B. Lee, and H. Wong, "Ultra-low power Al2O3-based RRAM with 1μA reset current," *VLSI Technology Systems and* ..., pp. 136–137, 2010.

[109] C. Y. Chen, L. Goux, A. Fantini, R. Degraeve, A. Redolfi, G. Groeseneken, and M. Jurczak, "Engineering of a TiN\Al2O3\(Hf,Al)O2\Ta2O5\Hf RRAM cell for Fast Operation at Low Current," pp. 262–265, 2015.

[110] B. Govoreanu and G. Kar, "10x10nm2 Hf/HfOx crossbar resistive RAM with excellent performance, reliability and low-energy operation," in *IEDM*, pp. 729–732, 2011.

[111] Y.-B. Kim, S. R. Lee, D. Lee, C. B. Lee, M. Chang, J. H. Hur, M.-J. Lee, G.-S. Park, C. J. Kim, U.-I. Chung, I.-K. Yoo, and K. Kim, "Bi-layered RRAM with unlimited endurance and extremely uniform switching," *Symposium on VLSI Technology - Digest of Technical Papers*, pp. 52–53, 2011.

[112] L. Zhao, H. Chen, S. Wu, and Z. Jiang, "Improved multi-level control of RRAM using pulse-train programming," ... *Technical Program-* ..., pp. 8–9, 2014.

[113] M. Barlas, B. Traoré, L. Grenouillet, S. Bernasconi, P. Blaise, M. Alayan, B. Sklenard, E. Jalaguier, P. Rodriguez, F. Mazen, E. Vilain, M. Guillermet, S. Jeannot, E. Vianello, and L. Perniola, "Impact of Si / Al implantation on the forming voltage and pre-forming conduction modes in HfO 2 based OxRAM cells," pp. 168–171, 2016.

[114] S. Yu, X. Guan, and H. S. P. Wong, "On the stochastic nature of resistive switching in metal oxide RRAM: Physical modeling, Monte Carlo simulation, and experimental characterization," in *Technical Digest - International Electron Devices Meeting, IEDM*, pp. 413–416, 2011.

[115] L. Goux, N. Raghavan, a. Fantini, R. Nigon, S. Strangio, R. Degraeve, G. Kar, Y. Y. Chen, F. De Stefano, V. V. Afanas'ev, and M. Jurczak, "On the bipolar resistive-switching characteristics of Al2O3- and HfO2-based memory cells operated in the soft-breakdown regime," *Journal of Applied Physics*, vol. 116, p. 134502, oct 2014.

[116] D. Garbin, E. Vianello, Q. Rafhay, M. Azzaz, P. Candelier, B. DeSalvo, G. Ghibaudo, and L. Perniola, "Resistive memory variability: A simplified trap-assisted tunneling model," *Solid-State Electronics*, vol. 115, pp. 126–132, 2016.

[117] A. Fantini, G. Gorine, R. Degraeve, L. Goux, C. Y. Chen, A. Redolfi, S. Clima, A. Cabrini, G. Torelli, and M. Jurczak, "Intrinsic program instability in HfO2 RRAM and consequences on program algorithms," *Technical Digest - International Electron Devices Meeting, IEDM*, vol. 2016-Febru, pp. 7.5.1–7.5.4, 2016.

[118] S. Yu and H. Wong, "Compact modeling of conducting-bridge random-access memory (CBRAM)," *Electron Devices, IEEE Transactions on*, vol. 58, no. 5, pp. 1352–1360, 2011.

[119] R. Waser, "Resistive non-volatile memory devices (Invited Paper)," *Microelectronic Engineering*, vol. 86, no. 7-9, pp. 1925–1928, 2009.

[120] U. Celano, L. Goux, R. Degraeve, A. Fantini, O. Richard, H. Bender, M. Jurczak, and W. Vandervorst, "Imaging the Three-Dimensional Conductive Channel in Filamentary-Based Oxide Resistive Switching Memory," *Nano Letters*, p. acs.nanolett.5b03078, 2015.

[121] U. Celano, L. Goux, A. Belmonte, K. Opsomer, A. Franquet, A. Schulze, C. Detavernier, O. Richard, H. Bender, M. Jurczak, and W. Vandervorst, "Three-dimensional observation of the conductive filament in nanoscaled resistive memory devices.," *Nano letters*, vol. 14, pp. 2401–6, may 2014.

[122] U. Celano, L. Goux, A. Belmonte, G. Giammaria, K. Opsomer, C. Detavernier, O. Richard, F. Irrera, M. Jurczak, and W. Vandervorst, "Progressive vs . Abrupt reset behavior in Conductive Bridging devices : a C-AFM tomography study .," pp. 351–354, 2014.

[123] R. Waser and M. Aono, "Nanoionics-based resistive switching memories.," *Nature materials*, vol. 6, pp. 833–40, nov 2007.

[124] M. Barci, J. Guy, G. Molas, E. Vianello, a. Toffoli, J. Cluzel, a. Roule, M. Bernard, C. Sabbione, L. Perniola, and B. De Salvo, "Impact of SET and RESET conditions on CBRAM high temperature data retention," *IEEE International Reliability Physics Symposium Proceedings*, pp. 3–6, 2014.

[125] A. Belmonte, U. Celano, A. Redolfi, A. Fantini, R. Muller, W. Vandervorst, M. Houssa, M. Jurczak, and L. Goux, "Analysis of the excellent memory disturb characteristics of a hourglass-shaped filament in Al<inf>2</inf>O<inf>3</inf>/Cu-based CBRAM devices," *IEEE Transactions on Electron Devices*, vol. 62, no. 6, pp. 2007–2013, 2015.

[126] Y. Huai, "Spin-Transfer Torque MRAM (STT-MRAM): Challenges and Prospects," *AAPPS bulletin*, vol. 18, no. 6, pp. 33–40, 2008.

[127] a. V. Khvalkovskiy, D. Apalkov, S. Watts, R. Chepulskii, R. S. Beach, A. Ong, X. Tang, A. Driskill-Smith, W. H. Butler, P. B. Visscher, D. Lottis, E. Chen, V. Nikitin, and M. Krounbi, "Basic principles of STT-MRAM cell operation in memory arrays," *Journal of Physics D: Applied Physics*, vol. 46, no. 7, p. 074001, 2013.

[128] N. Nagel, T. Mikolajick, I. Kasko, W. Hartner, M. Moert, C.-u. Pinnow, C. Dehm, and C. Mazure, "An Overview of FeRAM Technology for High Density Applications Nicolas Nagel, Thomas Mikolajick, Igor Kasko, Walter Hartner, Manfred Moert, Cay-Uwe Pinnow, Christine Dehm and Carlos Mazure," *Materials Research*, vol. 655, pp. 1–10, 2001.

[129] T. Mikolajick and C.-U. Pinnow, "The Future of Nonvolatile Memories," *Non-Volatile Memory Technology Symposium*, pp. 4–6, 2002.

[130] S. H. Jo, T. Kumar, S. Narayanan, W. D. Lu, H. Nazarian, and S. Clara, "3D-stackable Crossbar Resistive Memory based on Field Assisted Superlinear Threshold (FAST) Selector," *Electron Devices Meeting (IEDM), 2014 IEEE International*, no. 408, pp. 3–5, 2014.

[131] E. T. E. Peter Clarke, Analog Editor, "Crossbar ReRAM in Production at SMIC."

[132] E. T. E. Peter Clarke, Analog Editor, "Patent Search Supports View 3D XPoint Based on Phase-Change."

[133] C.-w. Hsu, C.-c. Wan, I.-t. Wang, M.-c. Chen, C.-l. Lo, Y.-j. Lee, W.-y. Jang, C.-h. Lin, and T.-h. Hou, "3D Vertical TaOx/TiO2 RRAM with over 103 Self-Rectifying Ratio and Sub-uA Operating Current," vol. 2, pp. 264–267, 2013.

[134] L. O. Chua and S. M. Kang, "Memristive Devices and Systems," *Proceedings of the IEEE*, vol. 64, no. 2, pp. 209–223, 1976.

[135] G. S. Snider, "Spike-timing-dependent learning in memristive nanodevices," in *2008 IEEE/ACM International Symposium on Nanoscale Architectures, NANOARCH 2008*, pp. 85–92, 2008.

[136] F. Galluppi, X. Lagorce, E. Stromatias, M. Pfeiffer, L. A. Plana, S. B. Furber, and R. B. Benosman, "A framework for plasticity implementation on the SpiNNaker neural architecture," *Frontiers in Neuroscience*, vol. 9, no. JAN, pp. 1–20, 2015.

[137] B. Vogginger, R. Schüffny, A. Lansner, L. Cederström, J. Partzsch, and S. Höppner, "Reducing the computational footprint for real-time BCPNN learning," *Frontiers in Neuroscience*, vol. 9, no. JAN, pp. 1–16, 2015.

[138] D. Querlioz, P. Dollfus, O. Bichler, and C. Gamrat, "Learning with memristive devices: How should we model their behavior?," *Proceedings of the 2011 IEEE/ACM International Symposium on Nanoscale Architectures, NANOARCH 2011*, pp. 150–156, 2011.

[139] G. Palma, M. Suri, D. Querlioz, E. Vianello, and B. De Salvo, "Stochastic neuron design using conductive bridge RAM," *2013 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*, pp. 95–100, jul 2013.

[140] P. Stoliar, J. Tranchant, B. Corraze, E. Janod, M.-P. Besland, F. Tesler, M. Rozenberg, and L. Cario, "A Leaky-Integrate-and-Fire Neuron Analog Realized with a Mott Insulator," *Advanced Functional Materials*, pp. 1–7, 2017.

[141] G. W. Burr, R. M. Shelby, A. Sebastian, S. Kim, S. Kim, S. Sidler, K. Virwani, M. Ishii, P. Narayanan, A. Fumarola, L. L. Sanches, I. Boybat, M. Le Gallo, K. Moon, J. Woo, H. Hwang, and Y. Leblebici, "Neuromorphic computing using non-volatile memory," *Advances in Physics: X*, vol. 2, no. 1, pp. 89–124, 2017.

[142] P. Y. Chen and S. Yu, "Partition SRAM and RRAM based synaptic arrays for neuro-inspired computing," *Proceedings - IEEE International Symposium on Circuits and Systems*, vol. 2016-July, pp. 2310–2313, 2016.

[143] S. Saïghi, C. G. Mayr, T. Serrano-Gotarredona, H. Schmidt, G. Lecerf, J. Tomas, J. Grollier, S. Boyn, A. F. Vincent, D. Querlioz, S. La Barbera, F. Alibart, D. Vuillaume, O. Bichler, C. Gamrat, and B. Linares-Barranco, "Plasticity in memristive devices for spiking neural networks.," *Frontiers in neuroscience*, vol. 9, p. 51, jan 2015.

[144] G. Indiveri, B. Linares-Barranco, R. Legenstein, G. Deligeorgis, and T. Prodromakis, "Integration of nanoscale memristor synapses in neuromorphic computing architectures.," *Nanotechnology*, vol. 24, no. 38, p. 384010, 2013.

[145] A. Redaelli, A. Pirovano, and F. Pellizzer, "Electronic switching effect and phase-change transition in chalcogenide materials," *IEEE Electron Device . . .* , vol. 25, no. 10, pp. 684–686, 2004.

[146] M. Suri, V. Sousa, L. Perniola, D. Vuillaume, and B. DeSalvo, "Phase change memory for synaptic plasticity application in neuromorphic systems," *The 2011 International Joint Conference on Neural Networks*, pp. 619–624, jul 2011.

[147] M. Suri, O. Bichler, D. Querlioz, O. Cueto, L. Perniola, V. Sousa, D. Vuillaume, C. Gamrat, and B. DeSalvo, "Phase change memory as synapse for ultra-dense neuromorphic systems: Application to complex visual pattern extraction," *2011 International Electron Devices Meeting*, pp. 4.4.1–4.4.4, 2011.

[148] O. Bichler, D. Roclin, C. Gamrat, and D. Querlioz, "Design exploration methodology for memristor-based spiking neuromorphic architectures with the Xnet event-driven simulator," *Proceedings of the 2013 IEEE/ACM International Symposium on Nanoscale Architectures, NANOARCH 2013*, no. 318597, pp. 7–12, 2013.

[149] G. W. Burr, R. M. Shelby, J. W. Jang, R. S. Shenoy, P. Narayanan, K. Virwani, E. U. Giacometti, B. Kurdi, and H. Hwang, "Experimental demonstration and tolerancing of a large-scale neural network (165,000 synapses), using phase-change memory as the synaptic weight element," *IEDM*, vol. 95120, pp. 0–2, 2014.

[150] G. W. Burr, R. M. Shelby, S. Sidler, C. Di Nolfo, J. Jang, I. Boybat, R. S. Shenoy, P. Narayanan, K. Virwani, E. U. Giacometti, B. N. Kurdi, and H. Hwang, "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*, vol. 62, no. 11, pp. 3498–3507, 2015.

[151] G. W. Burr, P. Narayanan, R. M. Shelby, S. Sidler, I. Boybat, C. Di Nolfo, and Y. Leblebici, "Large-scale neural networks implemented with non-volatile memory as the synaptic weight element: Comparative performance analysis (accuracy, speed, and power)," *Technical Digest - International Electron Devices Meeting, IEDM*, vol. 2016-Febru, no. 408, pp. 4.4.1–4.4.4, 2016.

[152] M. Suri, O. Bichler, Q. Hubert, L. Perniola, V. Sousa, C. Jahan, D. Vuillaume, C. Gamrat, and B. DeSalvo, "Interface Engineering of PCM for Improved Synaptic Performance in Neuromorphic Systems," *2012 4th IEEE International Memory Workshop*, pp. 1–4, may 2012.

[153] M. Suri, O. Bichler, Q. Hubert, L. Perniola, V. Sousa, C. Jahan, D. Vuillaume, C. Gamrat, and B. DeSalvo, "Addition of HfO2 interface layer for improved synaptic performance

of phase change memory (PCM) devices," *Solid-State Electronics*, vol. 79, pp. 227–232, jan 2013.

[154] M. Suri, D. Garbin, O. Bichler, D. Querlioz, D. Vuillaume, C. Gamrat, and B. DeSalvo, "Impact of PCM resistance-drift in neuromorphic systems and drift-mitigation strategy," *2013 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*, pp. 140–145, jul 2013.

[155] D. Kuzum, R. G. D. Jeyasingh, S. Yu, and H.-S. P. Wong, "Low-Energy Robust Neuromorphic Computation Using Synaptic Devices," *IEEE Transactions on Electron Devices*, vol. 59, pp. 3489–3494, dec 2012.

[156] Y. Li, Y. Zhong, L. Xu, J. Zhang, X. Xu, H. Sun, and X. Miao, "Ultrafast synaptic events in a chalcogenide memristor.," *Scientific reports*, vol. 3, p. 1619, jan 2013.

[157] D. Garbin, M. Suri, O. Bichler, and D. Querlioz, "Probabilistic neuromorphic system using binary phase-change memory (pcm) synapses: Detailed power consumption analysis," ... *(NANO), 2013 IEEE* ..., pp. 91–94, 2013.

[158] S. Wozniak, T. Tuma, A. Pantazi, and E. Eleftheriou, "Learning spatio-temporal patterns in the presence of input noise using phase-change memristors," *Proceedings - IEEE International Symposium on Circuits and Systems*, vol. 2016-July, pp. 365–368, 2016.

[159] K. Seo, I. Kim, S. Jung, M. Jo, S. Park, J. Park, J. Shin, K. P. Biju, J. Kong, K. Lee, B. Lee, and H. Hwang, "Analog memory and spike-timing-dependent plasticity characteristics of a nanoscale titanium oxide bilayer resistive switching device.," *Nanotechnology*, vol. 22, p. 254023, jun 2011.

[160] S. Yu, Y. Wu, and R. Jeyasingh, "An electronic synapse device based on metal oxide resistive switching memory for neuromorphic computation," *Electron Devices, IEEE* ..., vol. 58, no. 8, pp. 2729–2737, 2011.

[161] S. Yu, B. Gao, Z. Fang, H. Yu, J. Kang, and H.-S. P. Wong, "A low energy oxide-based electronic synaptic device for neuromorphic visual systems with tolerance to device variation.," *Advanced materials (Deerfield Beach, Fla.)*, vol. 25, pp. 1774–9, mar 2013.

[162] S. Ambrogio, S. Balatti, F. Nardi, S. Facchinetti, and D. Ielmini, "Spike-timing dependent plasticity in a transistor-selected resistive switching memory," *Nanotechnology*, vol. 24, no. 38, p. 384012, 2013.

[163] S. Park, J. Noh, M.-L. Choo, A. M. Sheri, M. Chang, Y.-B. Kim, C. J. Kim, M. Jeon, B.-G. Lee, B. H. Lee, and H. Hwang, "Nanoscale RRAM-based synaptic electronics: toward a neuromorphic computing device.," *Nanotechnology*, vol. 24, p. 384009, sep 2013.

[164] S. Park, a. Sheri, J. Kim, J. Noh, J. Jang, M. Jeon, B. H. R. Lee, B. H. R. Lee, B. H. R. Lee, and H. Hwang, "Neuromorphic speech systems using advanced ReRAM-based synapse," *2013 IEEE International Electron Devices Meeting*, pp. 25.6.1–25.6.4, dec 2013.

[165] O. Kavehei, "Highly Scalable Neuromorphic Hardware with 1-bit Stochastic nano-Synapses," *arXiv preprint arXiv:1309.6419*, pp. 1–9, 2013.

[166] M. Prezioso, F. Merrikh-Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev, and D. B. Strukov, "Training and operation of an integrated neuromorphic network based on metal-oxide memristors," *Nature*, vol. 521, pp. 61–64, may 2015.

[167] B. Yan, A. M. Mahmoud, J. J. Yang, Q. Wu, Y. Chen, and H. H. Li, "A neuromorphic ASIC design using one-selector-one-memristor crossbar," *Proceedings - IEEE International Symposium on Circuits and Systems*, vol. 2016-July, pp. 1390–1393, 2016.

[168] J. Bill and R. Legenstein, "A compound memristive synapse model for statistical learning through STDP in spiking neural networks," *Frontiers in Neuroscience*, vol. 8, no. December, pp. 1–18, 2014.

[169] M. Prezioso, Y. Zhong, D. Gavrilov, F. Merrikh-Bayat, B. Hoskins, G. Adam, K. Likharev, and D. Strukov, "Spiking neuromorphic networks with metal-oxide memristors," *Proceedings - IEEE International Symposium on Circuits and Systems*, vol. 2016-July, pp. 177–180, 2016.

[170] M. Prezioso, F. Merrikh Bayat, B. Hoskins, K. Likharev, and D. Strukov, "Self-Adaptive Spike-Time-Dependent Plasticity of Metal-Oxide Memristors," *Scientific Reports*, vol. 6, no. February, p. 21331, 2016.

[171] D. Garbin, E. Vianello, O. Bichler, Q. Rafhay, C. Gamrat, G. Ghibaudo, B. Desalvo, and L. Perniola, "HfO2-Based OxRAM Devices as Synapses for Convolutional Neural Networks," *IEEE Transactions on Electron Devices*, pp. 1–8, 2015.

[172] E. Covi, S. Brivio, M. Fanciulli, and S. Spiga, "Synaptic potentiation and depression in Al:HfO2-based memristor," *Microelectronic Engineering*, vol. 147, pp. 41–44, nov 2015.

[173] E. Covi, S. Brivio, A. Serb, T. Prodromakis, M. Fanciulli, and S. Spiga, "HfO2-based memristors for neuromorphic applications," *Proceedings - IEEE International Symposium on Circuits and Systems*, vol. 2016-July, pp. 393–396, 2016.

[174] T. Chang, S.-H. Jo, and W. Lu, "Short-term memory to long-term memory transition in a nanoscale memristor.," *ACS nano*, vol. 5, pp. 7669–76, sep 2011.

[175] R. Yang, K. Terabe, G. Liu, T. Tsuruoka, T. Hasegawa, J. K. Gimzewski, and M. Aono, "On-demand nanodevice with electrical and neuromorphic multifunction realized by local ion migration.," *ACS nano*, vol. 6, pp. 9515–21, nov 2012.

[176] S. Kim, C. Du, P. Sheridan, W. Ma, S. Choi, and W. D. Lu, "Experimental demonstration of a second-order memristor and its ability to biorealistically implement synaptic plasticity," *Nano Letters*, vol. 15, no. 3, pp. 2203–2211, 2015.

[177] R. Berdan, E. Vasilaki, A. Khiat, G. Indiveri, A. Serb, and T. Prodromakis, "Emulating short-term synaptic dynamics with memristive devices," *Scientific Reports*, vol. 6, no. November 2015, p. 18639, 2016.

[178] V. Milo, G. Pedretti, R. Carboni, A. Calderoni, N. Ramaswamy, S. Ambrogio, and D. Ielmini, "Demonstration of hybrid CMOS/RRAM neural networks with spike time/rate-dependent plasticity," *IEDM*, 2016.

[179] Z. Wang, S. Ambrogio, S. Balatti, and D. Ielmini, "A 2-transistor/1-resistor artificial synapse capable of communication and stochastic learning in neuromorphic systems," *Frontiers in Neuroscience*, vol. 9, no. JAN, pp. 1–11, 2015.

[180] D. Ielmini, S. Ambrogio, V. Milo, S. Balatti, and Z. Q. Wang, "Neuromorphic computing with hybrid memristive/CMOS synapses for real-time learning," *Proceedings - IEEE International Symposium on Circuits and Systems*, vol. 2016-July, pp. 1386–1389, 2016.

[181] H. Mostafa, C. Mayr, and G. Indiveri, "Beyond spike-timing dependent plasticity in memristor crossbar arrays," *Proceedings - IEEE International Symposium on Circuits and Systems*, vol. 2016-July, pp. 926–929, 2016.

[182] W. Zhao, J. Portal, W. Kang, M. Moreau, Y. Zhang, H. Aziza, J.-O. Klein, Z. Wang, D. Querlioz, D. Deleruyelle, M. Bocquet, D. Ravelosona, C. Muller, and C. Chappert, "Design and analysis of crossbar architecture based on complementary resistive switching nonvolatile memory cells," *Journal of Parallel and Distributed Computing*, vol. 74, pp. 2484–2496, jun 2014.

[183] I. T. Wang, Y. C. Lin, Y. F. Wang, C. W. Hsu, and T. H. Hou, "3D synaptic architecture with ultralow sub-10 fJ energy per spike for neuromorphic computation," in *Technical Digest - International Electron Devices Meeting, IEDM*, vol. 2015-Febru, pp. 28.5.1–28.5.4, 2015.

[184] G. Piccolboni, G. Molas, J. M. Portal, R. Coquand, M. Bocquet, D. Garbin, E. Vianello, C. Carabasse, V. Delaye, C. Pellissier, T. Magis, C. Cagli, M. Gely, O. Cueto, D. Deleruyelle, G. Ghibaudo, B. De Salvo, and L. Perniola, "Investigation of the potentialities of Vertical Resistive RAM (VRRAM) for neuromorphic applications," *Technical Digest - International Electron Devices Meeting, IEDM*, vol. 2016-Febru, pp. 17.2.1–17.2.4, 2016.

[185] D. Mahalanabis, M. Sivaraj, W. Chen, S. Shah, H. J. Barnaby, M. N. Kozicki, J. B. Christen, and S. Vrudhula, "Demonstration of spike timing dependent plasticity in CBRAM devices with silicon neurons," *Proceedings - IEEE International Symposium on Circuits and Systems*, vol. 2016-July, pp. 2314–2317, 2016.

[186] M. Suri, D. Querlioz, and O. Bichler, "Bio-inspired stochastic computing using binary CBRAM synapses," *Electron Devices, ...*, vol. 60, no. 7, pp. 2402–2409, 2013.

[187] S. La Barbera, D. Vuillaume, and F. Alibart, "Filamentary switching: Synaptic plasticity through device volatility," *ACS Nano*, vol. 9, no. 1, pp. 941–949, 2015.

[188] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu, "Nanoscale memristor device as synapse in neuromorphic systems.," *Nano letters*, vol. 10, pp. 1297–301, apr 2010.

[189] S. Yu and H.-S. Philip Wong, "Modeling the switching dynamics of programmable-metallization-cell (PMC) memory and its application as synapse device for a neuromorphic computation system," *2010 International Electron Devices Meeting*, pp. 22.1.1–22.1.4, dec 2010.

[190] S.-J. Choi, G.-B. Kim, K. Lee, K.-H. Kim, W.-Y. Yang, S. Cho, H.-J. Bae, D.-S. Seo, S.-I. Kim, and K.-J. Lee, "Synaptic behaviors of a single metal,Äìoxide,Äìmetal resistive device," *Applied Physics A*, vol. 102, pp. 1019–1025, jan 2011.

[191] H. Choi, H. Jung, J. Lee, J. Yoon, J. Park, D.-j. Seong, W. Lee, M. Hasan, G.-Y. Jung, and H. Hwang, "An electrically modifiable synapse array of resistive switching memory.," *Nanotechnology*, vol. 20, p. 345201, aug 2009.

[192] W. Lu, K. Kim, T. Chang, and S. Gaba, "Two-terminal resistive switches (memristors) for memory and logic applications," ... *Automation Conference (ASP* ..., pp. 217–223, 2011.

[193] C. Zhang, Y.-T. Tai, J. Shang, G. Liu, K.-L. Wang, C. Hsu, X. Yi, X. Yang, W. Xue, H. Tan, S. Guo, L. Pan, and R.-W. Li, "Synaptic plasticity and learning behaviours in flexible artificial synapse based on polymer/viologen system," *J. Mater. Chem. C*, vol. 4, no. 15, pp. 3217–3223, 2016.

[194] A. F. Vincent, S. Member, N. Locatelli, J.-o. Klein, W. S. Zhao, S. Member, S. Galdin-retailleau, and D. Querlioz, "Analytical Macrospin Modeling of the Stochastic Switching Time of Spin-Transfer Torque Devices," vol. 62, no. 1, pp. 164–170, 2015.

[195] A. F. Vincent, J. Larroque, N. Locatelli, N. B. Romdhane, O. Bichler, C. Gamrat, W. S. Zhao, J. O. Klein, S. Galdin-Retailleau, and D. Querlioz, "Spin-transfer torque magnetic mem-

ory as a stochastic memristive synapse for neuromorphic systems," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 9, no. 2, pp. 166–174, 2015.

[196] A. Thomas, S. Niehörster, S. Fabretti, N. Shepheard, O. Kuschel, K. Küpper, J. Wollschläger, P. Krzysteczko, and E. Chicca, "Tunnel junction based memristors as artificial synapses," *Frontiers in Neuroscience*, vol. 9, no. JUN, pp. 1–9, 2015.

[197] H. J. Kelley, "Gradient Theory of Optimal Flight Paths," *ARS Journal*, vol. 30, no. 10, pp. 947–954, 1960.

[198] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[199] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[200] T. Masquelier, R. Guyonneau, and S. J. Thorpe, "Spike timing dependent plasticity finds the start of repeating patterns in continuous spike trains.," *PloS one*, vol. 3, p. e1377, jan 2008.

[201] T. Masquelier, R. Guyonneau, and S. J. Thorpe, "Competitive STDP-based spike pattern learning.," *Neural computation*, vol. 21, pp. 1259–76, may 2009.

[202] K. Dhoble and N. Nuntalid, "Online spatio-temporal pattern recognition with evolving spiking neural networks utilising address event representation, rank order, and temporal spike learning," ... *Networks (IJCNN), The* ..., pp. 10–15, 2012.

[203] D. Roclin, O. Bichler, C. Gamrat, S. J. Thorpe, and J. O. Klein, "Design study of efficient digital order-based STDP neuron implementations for extracting temporal features," in *Proceedings of the International Joint Conference on Neural Networks*, 2013.

[204] T. Masquelier and S. J. Thorpe, "Unsupervised learning of visual features through spike timing dependent plasticity.," *PLoS computational biology*, vol. 3, p. e31, feb 2007.

[205] S. J. Thorpe, A. Brilhault, and J. A. Perez-Carrasco, "Suggestions for a biologically inspired spiking retina using order-based coding," in *ISCAS 2010 - 2010 IEEE International Symposium on Circuits and Systems: Nano-Bio Circuit Fabrics and Systems*, pp. 265–268, 2010.

[206] O. Bichler, D. Querlioz, S. J. Thorpe, J.-P. Bourgoin, and C. Gamrat, "Extraction of temporally correlated features from dynamic vision sensors with spike-timing-dependent plasticity.," *Neural networks : the official journal of the International Neural Network Society*, vol. 32, pp. 339–48, aug 2012.

[207] X. Lagorce, S. H. Ieng, X. Clady, M. Pfeiffer, and R. B. Benosman, "Spatiotemporal features for asynchronous event-based data," *Frontiers in Neuroscience*, vol. 9, no. FEB, pp. 1–13, 2015.

[208] V. Chan, S. C. Liu, and A. van Schaik, "AER EAR: A matched silicon cochlea pair with address event representation interface," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 54, no. 1, pp. 48–59, 2007.

[209] E. Covi, S. Brivio, A. Serb, T. Prodromakis, M. Fanciulli, and S. Spiga, "Analog memristive synapse in spiking networks implementing unsupervised learning," *Frontiers in Neuroscience*, vol. 10, no. OCT, pp. 1–13, 2016.

[210] A. Serb, J. Bill, A. Khiat, R. Berdan, R. Legenstein, and T. Prodromakis, "Unsupervised learning in probabilistic neural networks with multi-state metal-oxide memristive synapses," *Nature Communications*, vol. 7, p. 12611, 2016.

[211] H. Mostafa, A. Khiat, A. Serb, C. G. Mayr, G. Indiveri, and T. Prodromakis, "Implementation of a spike-based perceptron learning rule using TiO2,àíx memristors," *Frontiers in Neuroscience*, vol. 9, no. October, pp. 1–11, 2015.

[212] D. Querlioz, O. Bichler, and C. Gamrat, "Simulation of a memristor-based spiking neural network immune to device variations," *Proceedings of the International Joint Conference on Neural Networks*, pp. 1775–1781, 2011.

[213] D. Querlioz, O. Bichler, P. Dollfus, and C. Gamrat, "Immunity to device variations in a spiking neural network with memristive nanodevices," *IEEE Transactions on Nanotechnology*, vol. 12, no. 3, pp. 288–295, 2013.

[214] D. Garbin and O. Bichler, "Variability-tolerant convolutional neural network for pattern recognition applications based on oxram synapses," . . . *(IEDM), 2014 IEEE* . . . , pp. 8–10, 2014.

[215] E. Vianello, D. Garbin, N. Jovanovic, O. Bichler, O. Thomas, B. De Salvo, and L. Perniola, "Oxide based Resistive Memories for Low Power Embedded Applications and Neuromorphic Systems," in *ECS*, vol. 69, pp. 3–10, 2015.

[216] Y. Wang, L. Xia, T. Tang, B. Li, S. Yao, M. Cheng, and H. Yang, "Low Power Convolutional Neural Networks on a Chip," *IEEE International Symposium on Computer Architecture*, no. 1, pp. 129–132, 2016.

[217] G. Piccolboni, G. Molas, D. Garbin, T. Werner, E. Vianello, B. D. Salvo, G. Ghibaudo, and L. Perniola, "Investigation of variability in Vertical Resistive RAM (VRRAM): Physical Model," in *SSDM*, vol. 1, pp. 3–4, 2016.

[218] R. Q. Quiroga, L. Reddy, G. Kreiman, C. Koch, and I. Fried, "Invariant visual representation by single neurons in the human brain.," *Nature*, vol. 435, pp. 1102–7, jun 2005.

[219] J. Secco, M. Farina, D. Demarchi, F. Corinto, and M. Gilli, "Memristor cellular automata for image pattern recognition and clinical applications," *Proceedings - IEEE International Symposium on Circuits and Systems*, vol. 2016-July, pp. 1378–1381, 2016.

[220] F. Corradi and G. Indiveri, "A Neuromorphic Event-Based Neural Recording System for Smart Brain-Machine-Interfaces," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 9, no. 5, pp. 699–709, 2015.

[221] I. Gupta, A. Serb, A. Khiat, and T. Prodromakis, "Practical operation considerations for memristive integrating sensors," *Proceedings - IEEE International Symposium on Circuits and Systems*, vol. 2016-July, no. i, pp. 2322–2325, 2016.

[222] R. Ananthanarayanan, S. K. Esser, H. D. Simon, and D. S. Modha, "The Cat is Out of the Bag : Cortical Simulations with 10ˆ9 Neurons , 10ˆ13 Synapses," *Matrix*, vol. 2, no. c, pp. 1–12, 2009.

[223] P. a. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. K. Esser, R. Appuswamy, B. Taba, A. Amir, M. D. Flickner, W. P. Risk, R. Manohar, and D. S. Modha, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, pp. 668–673, aug 2014.

[224] D. Kuzum, S. Yu, and H.-S. P. Wong, "Synaptic electronics: materials, devices and applications.," *Nanotechnology*, vol. 24, p. 382001, sep 2013.

[225] E. Vianello, O. Thomas, G. Molas, D. Garbin, G. Palma, and L. Perniola, "Resistive memories for ultra-low-power embedded computing design," *IEDM*, pp. 144–147, 2014.

[226] D. Ielmini, F. Nardi, and S. Balatti, "Evidence for voltage-driven set/reset processes in bipolar switching RRAM," *IEEE Transactions on Electron Devices*, vol. 59, no. 8, pp. 2049–2056, 2012.

[227] A. Fantini, L. Goux, R. Degraeve, D. J. Wouters, N. Raghavan, G. Kar, A. Belmonte, Y. Y. Chen, B. Govoreanu, and M. Jurczak, "Intrinsic switching variability in HfO2 RRAM," *2013 5th IEEE International Memory Workshop, IMW 2013*, pp. 30–33, 2013.

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

[229] E. M. Izhikevich, "Simple model of spiking neurons.," *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, vol. 14, pp. 1569–72, jan 2003.

[230] W. Gerstner and W. M. Kistler, *Spiking Neuron Models*.
Cambridge University Press, 2002.

[231] a. Joubert, B. Belhadj, O. Temam, and R. Heliot, "Hardware spiking neurons design: Analog or digital?," *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–5, 2012.

[232] D. Cattaert, J.-P. Delbecque, D. H. Edwards, and F. a. Issa, "Social interactions determine postural network sensitivity to 5-HT.," *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 30, pp. 5603–16, apr 2010.

[233] J. Gasthaus, Y. W. Teh, F. Wood, and G. Dilan, "Dependent Dirichlet Process Spike Sorting," in *Neural Information Processing Systems (NIPS)*, pp. 1–8, 2008.

[234] J. A. Gasthaus and F. Wood, "Spike sorting using time-varying Dirichlet process mixture models," 2008.

[235] M. Mahvash and A. C. Parker, "Synaptic variability in a cortical neuromorphic circuit," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 3, pp. 397–409, 2013.

[236] C. H. Bennett, S. La Barbera, A. F. Vincent, F. Alibart, and D. Querlioz, "Exploiting the Short-term to Long-term Plasticity Transition in Memristive Nanodevice Learning Architectures," *Arxiv*, pp. 947–954, 2016.

[237] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128 x 128 120 dB 15 us latency asynchronous temporal contrast vision sensor," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, 2008.

[238] M. Noack, J. Partzsch, C. G. Mayr, S. Hänzsche, S. Scholze, S. Höppner, G. Ellguth, and R. Schüffny, "Switched-capacitor realization of presynaptic short-term-plasticity and stop-learning synapses in 28 nm CMOS," *Frontiers in Neuroscience*, vol. 9, no. FEB, pp. 1–14, 2015.

[239] N. Qiao, H. Mostafa, F. Corradi, M. Osswald, F. Stefanini, D. Sumislawska, and G. Indiveri, "A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128K synapses," *Frontiers in Neuroscience*, vol. 9, no. APR, pp. 1–17, 2015.

[240] R. Escolá, C. Pouzat, A. Chaffiol, B. Yvert, I. E. Magnin, and R. Guillemaud, "SIMONE: a realistic neural network simulator to reproduce MEA-based recordings.," *IEEE transactions on neural systems and rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society*, vol. 16, pp. 149–60, apr 2008.