



# Technologies émergentes de mémoire résistive pour les systèmes et application neuromorphique

Manan Suri

## ► To cite this version:

Manan Suri. Technologies émergentes de mémoire résistive pour les systèmes et application neuromorphique. Autre. Université de Grenoble, 2013. Français. NNT : 2013GRENT023 . tel-00935190

**HAL Id: tel-00935190**

**<https://theses.hal.science/tel-00935190>**

Submitted on 23 Jan 2014

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

## THÈSE

Pour obtenir le grade de

## DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

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

Arrêté ministériel : 7 août 2006

Présentée par

**Manan SURİ**

Thèse dirigée par **Dr. Barbara DESALVO**  
et codirigée par **Dr. Dominique VUILLAUME**

préparée au sein **CEA-LETI**  
et de **EEATS, Grenoble**

# Technologies Émergentes de Mémoire Résistive pour les Systèmes et Applications Neu- romorphiques

Thèse soutenue publiquement le **18 September 2013**,  
devant le jury composé de :

**M Daniele IELMINI**

Prof., Politecnico di Milano, Rapporteur

**M Giacomo INDIVERI**

Prof., Swiss Federal Institute of Technology in Zurich (ETH), Rapporteur

**M Philippe CANDELIER**

Dr., ST Microelectronics, Crolles, Examineur

**Mme Rose-Marie SAUVAGE**

Dr., DGA-France, Examineur

**M Gérard GHIBAUDO**

Dr., Université de Grenoble (IMEP-LAHC), Président

**Mme Barbara DESALVO**

Dr., CEA-LETI, Grenoble, Directeur de thèse

**M Dominique VUILLAUME**

Dr., Institute for Electronics, Microelectronics and Nanotechnology (IEMN), Villeneuve d'Ascq, Co-Directeur de thèse

**M Christian GAMRAT**

CEA-LIST, Gif-sur-Yvette, Invité



## Abstract

# EMERGING RESISTIVE MEMORY TECHNOLOGY FOR NEUROMORPHIC SYSTEMS AND APPLICATIONS

Research in the field of neuromorphic- and cognitive- computing has generated a lot of interest in recent years. With potential application in fields such as large-scale data driven computing, robotics, intelligent autonomous systems to name a few, bio-inspired computing paradigms are being investigated as the next generation (post-Moore, non-Von Neumann) ultra-low power computing solutions. In this work we discuss the role that different emerging non-volatile resistive memory technologies (RRAM), specifically (i) Phase Change Memory (PCM), (ii) Conductive-Bridge Memory (CBRAM) and Metal-Oxide based Memory (OXRAM) can play in dedicated neuromorphic hardware. We focus on the emulation of synaptic plasticity effects such as long-term potentiation (LTP), long term depression (LTD) and spike-timing dependent plasticity (STDP) with RRAM synapses. We developed novel low-power architectures, programming methodologies, and simplified STDP-like learning rules, optimized specifically for some RRAM technologies. We show the implementation of large-scale energy efficient neuromorphic systems with two different approaches (i) deterministic multi-level synapses and (ii) stochastic-binary synapses. Prototype applications such as complex visual- and auditory- pattern extraction are also shown using feed-forward spiking neural networks (SNN). We also introduce a novel methodology to design low-area efficient stochastic neurons that exploit intrinsic physical effects of CBRAM devices.

---

## RÉSUMÉ

# TECHNOLOGIES ÉMERGENTES DE MÉMOIRE RÉSISTIVE POUR LES SYSTÈMES ET APPLICATIONS NEUROMOR- PHIQUES

La recherche dans le domaine de l'informatique neuro-inspirée suscite beaucoup d'intérêt depuis quelques années. Avec des applications potentielles dans des domaines tels que le traitement de données à grande échelle, la robotique ou encore les systèmes autonomes intelligents pour ne citer qu'eux, des paradigmes de calcul bio-inspirés sont étudiés pour la prochaine génération solutions informatiques (post-Moore, non-Von Neumann) ultra-basse consommation. Dans ce travail, nous discutons les rôles que les différentes technologies de mémoire résistive non-volatiles émergentes (RRAM), notamment (i) Phase Change Memory (PCM), (ii) Conductive-Bridge Memory (CBRAM) et de la mémoire basée sur une structure Metal-Oxide (OXRAM) peuvent jouer dans des dispositifs neuromorphiques dédiés. Nous nous concentrons sur l'émulation des effets de plasticité synaptique comme la potentialisation à long terme (Long Term Potentiation, LTP), la dépression à long terme (Long Term Depression, LTD) et la théorie STDP (Spike-Timing Dependent Plasticity) avec des synapses RRAM. Nous avons développé à la fois de nouvelles architectures de faiblement énergivore, des méthodologies de programmation ainsi que des règles d'apprentissages simplifiées inspirées de la théorie STDP spécifiquement optimisées pour certaines technologies RRAM. Nous montrons l'implémentation de systèmes neuromorphiques à grande échelle et efficace énergétiquement selon deux approches différentes: (i) des synapses multi-niveaux déterministes et (ii) des synapses stochastiques binaires. Des prototypes d'applications telles que l'extraction de schéma visuel et auditif complexe sont également montrés en utilisant des réseaux de neurones impulsionnels (Feed-forward Spiking Neural Network, SNN). Nous introduisons également une nouvelle méthodologie pour concevoir des neurones stochastiques très compacts qui exploitent les caractéristiques physiques intrinsèques des appareils CBRAM.



## Acknowledgements

### OFFICIAL

In the order of decreasing bureaucratic significance, I would like to thank the following entities and people for their support, without which the current manuscript that you are reading wouldn't have existed. First I thank CEA-Grenoble, DGA-France and University of Grenoble-INPG for supporting my PhD scholarship and providing me the laboratory resources to perform this research. I then thank my PhD directors Dr. Barbara DeSalvo and Dr. Dominique Vuillaume, who have been excellent supportive guides. I would like to thank our collaborators, specially Dr. Olivier Bichler, Dr. Damien Querlioz and Dr. Christian Gamrat, with whom we have had extremely fruitful collaboration and strong team work over the last three years. I would like to thank the entire LTMA team which has been a wonderful family (especially Luca, Veronique, Elisa, Gabriel, Carine, John-Francois and Ludovic Poupinet (ex-LTMA)). I thank the thesis jury members for finding time to review this manuscript. Finally I thank my parents for fabricating me. I also thank my sister, our pet dog, and the rest of my family for being there when I needed them.

### UNOFFICIAL

Writing an acknowledgement is the hardest part of the thesis. It's like picking a drop of water and asking it to name and acknowledge every atom that it is composed of. If one looks at the universe, it is obvious that as individual humans we are minuscule entities. The sheer scale and multitude of existence entails that, when we act, or don't act, we are constantly under

the influence of numerous factors. We don't live in an isolated environment or a Faraday cage. While we can perceive and realize some of the things that influence us and our actions, there are numerous others that we barely perceive or realize. Thus, even if I write down every single name that has influenced this thesis or the last three years of my life (including all 2nd and 3rd order effects), it will still be incomplete. That being said, I will start with thanking my PhD colleagues/friends - Quentin, Gabriele, Giorgio, Veeresh, Cuiqin, Boubacar, Daniele, Marinela, Thérèse, Thanasis, Yann and Santosh. From my advisor Barbara I learnt important qualities such as - (i) Being self-critical and looking at things objectively. (ii) Pinpointing what actually matters; Barbara has a hawk's eye, from a great height she can see a small fish in the vast sea. In other words, when you work on strongly interdisciplinary topics it's easy to get overwhelmed by the amount of information and lose track of your target. But if you understand what actually matters the most, at different levels of detail, it keeps you on track. PhD is like a roller coaster ride, and one needs to have some seat-belts to keep his/her emotions contained. For me the two seat-belts were music and spirituality. I am thankful to my band members- Koceila, Tapas and Antoine for the music. If you want to draw a simple conclusion from this long and boring acknowledgement- The real fuel and driving force behind this thesis were the Indian and Mexican restaurants in Grenoble. Regular Spicy food and excess of Nutella are the pillars of good research.

---

---

कर्मण्यकर्म यः पश्येदकर्मणि च कर्म यः ।  
स बुद्धिमान्मनुष्येषु स युक्तः कुत्सन्कर्मकृत् ॥ १८ ॥

*Shrimad Bhagwad Gita [Chapter 4: Verse 18]*

*For French friends...*

**“ Celui qui voit l'inaction dans l'action, et l'action dans l'inaction  
est une personne sage, est un yogi et a tout accompli ”**

*For Italian friends...*

**“ L'uomo che vede l'inazione nell'azione, e l'azione nell'inazione  
si distingue per la sua saggezza, esiste trascendentalmente,  
un perfetto esecutore di tutte le azioni ”**

*For all others...*

**“ He who sees inaction in action, and action in inaction  
Is spiritually wise, transcendently situated a perfect performer of all  
actions ”**

Dedicated to all my teachers, who I came across at different times, and in different forms, during the last 26 years of my life. Some young, some old, some friendly, some hostile, some by intent, some by chance, some whom I know, some whom I don't, some who know me, some who don't, some living and some even not.

---

## MANUSCRIPT OUTLINE

This dissertation was written and submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy (Nanoelectronics and Nanotechnology) in The University of Grenoble, 2013. The topic addressed in the manuscript focuses on the use of emerging resistive memory technology for neuromorphic systems and applications. Chapter.1, begins with the motivation behind pursuing R&D in the field of neuromorphic systems. It then focuses on some basic concepts from neurobiology. A review of state-of-the art hardware implementation of biological synapses and their limitations are discussed. The concept of emerging non-volatile resistive memory technology is introduced. Towards the end of the chapter, we briefly summarize the scope and the overall strategy adopted for the research conducted during this PhD thesis.

In Chapter.2, we discuss how Phase Change Memory (PCM) technology can be used to emulate biological synapses in large-scale neuromorphic systems with low-power dissipation and easy to implement programming methodology.

In Chapter.3, we discuss how filamentary-switching type of memory devices can be used to emulate biological synapses in large-scale neuromorphic systems. The first part of the chapter focuses on Ag/GeS<sub>2</sub> based Conductive-bridge (CBRAM) technology, while the second part focusses on HfO<sub>x</sub> based resistive metal-oxide (OXRAM) technology.

In Chapter.4, where we describe how RRAM devices can be used to design innovative neuron structures. We present an original methodology to design hybrid neuron circuits (CMOS + non volatile resistive memory) with stochastic firing behaviour. Finally the manuscript ends with a general conclusion and overall perspective on the topic.

Chapter.5, provides an overall conclusion and perspective of the research conducted for this thesis. A brief comparison of the three synaptic technologies is provided, followed by a description of the on-going activities and the ones that need further investigation. Finally the chapter ends by highlighting some issues requiring more attention to enable further progress in the field of neuromorphic or cognitive hardware.

---

# Contents

<b>1</b>	<b>Background</b>	<b>1</b>
1.1	Neuromorphic Systems . . . . .	1
1.1.1	Historical Perspective . . . . .	3
1.1.2	Advantages . . . . .	5
1.1.3	Applications . . . . .	7
1.2	Neurobiology Basics . . . . .	8
1.2.1	Neuron, Synapse and Spike . . . . .	8
1.2.2	Synaptic Plasticity and STDP . . . . .	12
1.2.3	Retina: The Natural Visual Processing System . . . . .	14
1.2.4	Cochlea: The Natural Auditory Processing System . . . . .	17
1.3	Simplified Electrical Modeling . . . . .	20
1.4	Nanoscale Hardware Emulation of Synapses . . . . .	22
1.4.1	VLSI-technology . . . . .	22
1.4.1.1	Floating-gate Synapses . . . . .	22
1.4.1.2	Dynamic Random Access Memory (DRAM) or Capacitive Synapses . . . . .	24
1.4.1.3	Static Random Access Memory (SRAM) Synapses . . . . .	25
1.4.1.4	Limitations of VLSI type synapses . . . . .	27
1.4.2	Exotic Device Synapses . . . . .	28
1.4.3	Resistive Memory Technology (RRAM) . . . . .	30
1.4.3.1	Memistor Synapse (The father of Memristor or RRAM) . . . . .	33
1.5	Scope and approach of this work . . . . .	35
1.6	Conclusion . . . . .	37



## CONTENTS

---

<b>2</b>	<b>Phase Change Memory Synapses</b>	<b>39</b>
2.1	PCM Working Principle . . . . .	39
2.2	State-of-Art PCM Synapses . . . . .	40
2.3	Device and Electrical Characterization . . . . .	43
2.3.1	LTP Experiments . . . . .	46
2.3.2	LTD Experiments . . . . .	47
2.3.3	Mixed Tests . . . . .	48
2.4	Physical Simulation . . . . .	51
2.5	Modeling . . . . .	56
2.5.1	Behavioral model for system level simulations . . . . .	56
2.5.2	Circuit-compatible model . . . . .	56
2.6	PCM Interface Engineering . . . . .	60
2.7	The "2-PCM Synapse" . . . . .	63
2.7.1	Simplified STDP-rule . . . . .	65
2.7.2	Programming Scheme . . . . .	66
2.7.2.1	Read . . . . .	66
2.7.2.2	Write . . . . .	66
2.7.2.3	Refresh . . . . .	68
2.8	Complex Visual Pattern Extraction Simulations . . . . .	72
2.8.1	Network and the Stimuli . . . . .	72
2.8.2	Neuron and Synapses . . . . .	73
2.8.3	Learning Performance . . . . .	75
2.8.4	Energy/Power Consumption Analysis . . . . .	79
2.9	Resistance-Drift and Mitigation Strategy . . . . .	81
2.9.1	"Binary PCM Synapse" . . . . .	82
2.9.1.1	Programming Scheme . . . . .	82
2.9.1.2	Analysis . . . . .	84
2.10	Conclusion . . . . .	91
<b>3</b>	<b>Filamentary-Switching Type Synapses</b>	<b>93</b>
3.1	CBRAM Technology . . . . .	93
3.1.1	CBRAM state-of-art Synapses . . . . .	94
3.1.2	Device and Electrical Characterization . . . . .	96

3.1.3	Limitations on LTD emulation . . . . .	97
3.1.4	Deterministic and Probabilistic Switching . . . . .	98
3.1.5	Stochastic STDP and Programming Methodology . . . . .	102
3.1.6	Auditory and Visual Processing Simulations . . . . .	105
3.2	OXRAM Technology . . . . .	111
3.2.1	State-of-art OXRAM Synapses . . . . .	111
3.2.2	Device and Electrical Characterization . . . . .	113
3.2.3	LTD Experiments: $R_{OFF}$ Modulation . . . . .	113
3.2.4	LTP Experiments: $R_{ON}$ modulation . . . . .	118
3.2.5	Binary operation . . . . .	119
3.2.6	Learning Simulations . . . . .	120
3.3	Conclusion . . . . .	122
<b>4</b>	<b>Using RRAM for Neuron Design</b>	<b>123</b>
4.1	Introduction . . . . .	123
4.2	CBRAM Stochastic Effects . . . . .	124
4.3	Stochastic Neuron Design . . . . .	126
4.3.1	Integrate and Fire Neuron . . . . .	126
4.3.2	Stochastic-Integrate and Fire principle and circuit . . . . .	126
4.4	Results and Discussion . . . . .	132
4.4.1	Set- and Reset- Operation . . . . .	132
4.4.2	Parameter Constraints . . . . .	133
4.4.3	Energy Consumption . . . . .	134
4.5	Conclusion . . . . .	135
<b>5</b>	<b>Conclusions and Perspective</b>	<b>137</b>
5.1	Conclusion . . . . .	137
5.2	Which one is better . . . . .	139
5.2.1	Intermediate Resistance States . . . . .	139
5.2.2	Energy/Power . . . . .	139
5.2.3	Endurance . . . . .	140
5.2.4	Speed . . . . .	140
5.2.5	Resistance Window . . . . .	141
5.3	On-Going and Next Steps . . . . .	141

## CONTENTS

---

5.4 The road ahead...	143
<b>A List of Patents and Publications</b>	<b>145</b>
A.1 Patents	145
A.2 Book Chapter	145
A.3 Conference and Journal Papers	145
<b>B Résumé en Français</b>	<b>149</b>
B.1 Chapitre I: Découverte	149
B.2 Chapitre II: Synapses avec des Mémoires à Changement de Phase	149
B.3 Chapitre III: Synapses avec des Mémoires à ‘Filamentary-switching’	155
B.4 Chapitre IV: Utiliser des RRAM pour la conception de neurones	160
B.5 Chapitre V: Conclusions et Perspectives	161
<b>List of Figures</b>	<b>165</b>
<b>List of Tables</b>	<b>181</b>
<b>Bibliography</b>	<b>183</b>

“No decision is right or wrong by itself...  
what you do after taking the decision defines it”

# 1

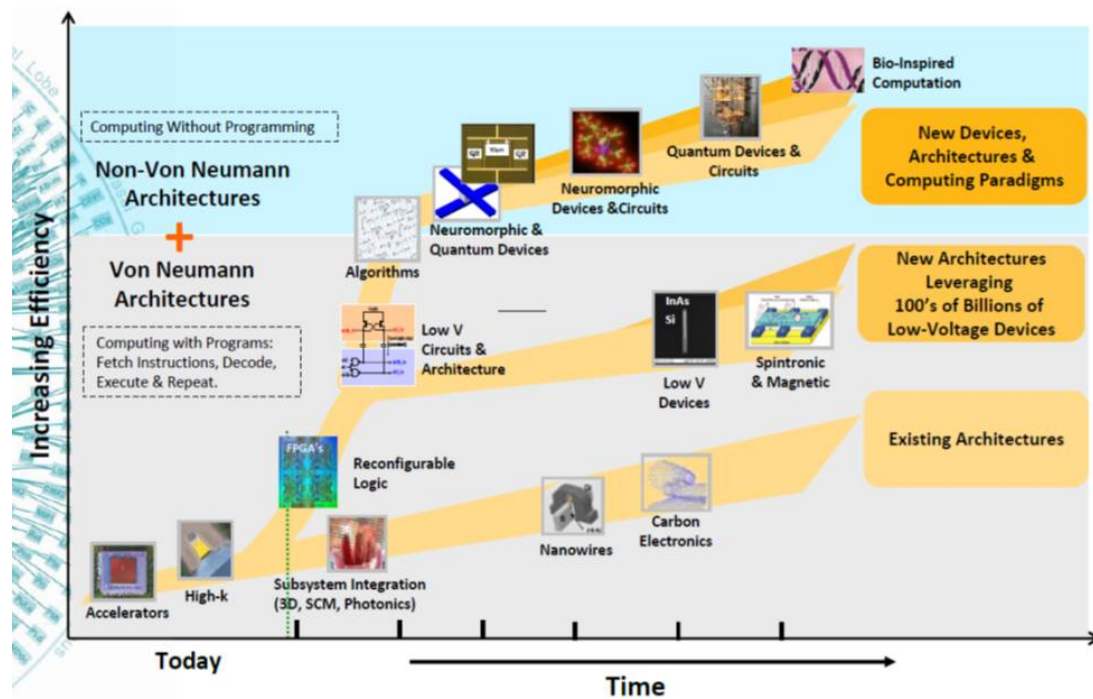
## Background

This chapter begins with a motivation behind pursuing R&D in the field of neuromorphic systems. We then focus on some basic concepts from neurobiology. A review of state-of-the art hardware implementation of biological synapses and their limitations are discussed. The concept of emerging non-volatile resistive memory technology is introduced. Towards the end of the chapter, we briefly summarize the scope and the overall strategy adopted for the research conducted during this PhD thesis.

### 1.1 Neuromorphic Systems

Neuromorphic hardware refers to an emerging field of hardware design that takes its inspiration from biological neural architectures and computations occurring inside the mammalian nervous system or the cerebral cortex. It is a strongly interdisciplinary field comprising principles and knowledge from neurobiology, computational neuroscience, computer science, machine learning, VLSI circuit design, and more recently nanotechnology. Unlike conventional Von-Neumann computing hardware (i.e Processors, DSPs, GPUs FPGAs), neuromorphic computing is different, as memory (storage) and processing are not completely isolated tasks in the later. Memory is intelligent and participates in processing of information. Neuromorphic computing may also referred to as Cognitive computing. Neuromorphic and bio-inspired computing paradigms have been proposed as the third generation of computing or the future successors of moore type von-neumann machines (Fig.1.1).

## 1. BACKGROUND



**Figure 1.1:** Proposed future computing roadmap with emerging beyond-moore technologies (adapted from IBM research colloquia-2012, Madrid, M. Ritter et. al.).

### 1.1.1 Historical Perspective

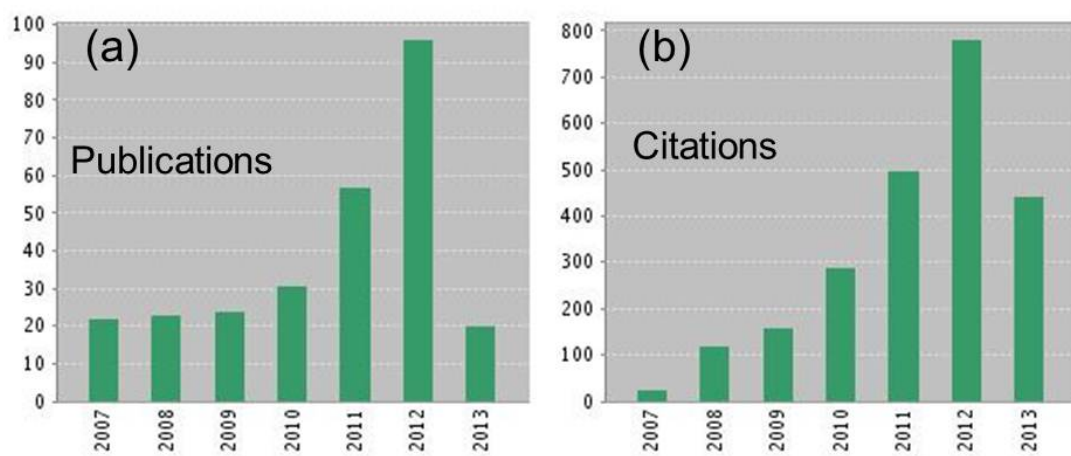
Historically the roots of neuromorphic hardware or neuro-inspired computing can be traced back to the works of physiologists McCulloch and Pitts, who came up with an interesting neuron model in 1943 [1]. They proposed a neuron model with two weighted inputs and one output. It was governed by a simple binary activation function. In 1958, Rosenblatt formulated the next milestone in the form of the Perceptron [2], or the first neuromorphic engine, which still holds as a very central concept in the field of artificial neural networks. The field was relatively stagnant through the 70s as key issues surfaced regarding the limitations of computational machines that processed neural networks [3]. Firstly, single-layer neural networks were incapable of processing the exclusive-or circuit. More importantly the computers of the time were not efficient enough to handle the long run time required by large neural networks. The advent of greater processing power in computers, and advances with the backpropagation algorithm [4], brought back some interest in the field. The 80s saw the rise of parallel distributed processing systems to efficiently simulate neural processes, mainly under the concept of connectionism [5]. The pioneering work of Carver Mead brought VLSI design to the forefront for neuro-inspired designs [6], when he designed the first silicon retina and neural learning chips in silicon.

Several interesting demonstrations of neurocomputers surfaced in the period from 80s to early 90s. For instance, IBM demonstrated a neuro-inspired vector classifier engine, known as ZISC (zero instruction set computing) processor [7], developed by Guy Paillet, who later formed a neuromorphic chip company called CogniMem Technologies Inc. Intel demonstrated the ETANN (Electrically Trainable Artificial Neural Network) chip with 10240 floating-gate synapses in 1989 [8]. L-Neuro by Philips, ANNA by AT&T, SYNAPSE 1 by Siemens [9], and MIND-1024 of CEA [10], were some other demonstrations of neurocomputers in that period.

However, advances in neuroscience in the 90s, particularly the interest in LTP/LTD and learning rules like STDP brought another turning point in the field [11]. The weaknesses of the perceptron model could now be overcome by using time critical spike based neural coding. This followed by the advances in the field of emerging non-volatile resistive memory (RRAM) technologies (also commonly and vaguely defined as memristors), sparked enormous renewed interest in the field of neuromorphic hardware in the 2000s.

## 1. BACKGROUND

---



**Figure 1.2:** Data obtained from the Web of Knowledge using the search expressions: Topic=(neuromorphic and memristor) OR Topic=(neuromorphic and RRAM) OR Topic=(neuromorphic and PCM) OR Topic=(neuromorphic and Phase change) OR Topic=(neuromorphic and resistive switching) OR Topic=(neuromorphic and magnetic) OR Topic=(phase change memory and synapse) OR Topic=(conductive bridge memory and synapse) OR Topic=(PCM and synapse) OR Topic=(CBRAM and synapse) OR Topic=(RRAM and synapse) OR Topic=(OxRAM and synapse) OR Topic=(OxRAM and neuromorphic) for the time period Jan 2007- April 2013 (a) Publications, (b) Citations.

The inherently similar properties of two-terminal nanoscale RRAM devices and biological synapses field them as the ultimate 'synaptic' candidate for building ultra-dense large scale neuromorphic systems.

The steep interest can be gauged by the fact that several large international projects such as the Blue-Brain Project (IBM/EPFL), Spinnaker (Manchester/ARM), Brain-ScaleS (Heidelberg), Neurogrid (Stanford) and SYNAPSE (DARPA) have been floated over the last 10 years. With the most recent ones being the 'Human Connectome Project' (US), the BRAIN Initiative (US) and the massive European flagship- 'Human Brain Project' (HBP) with an estimated budget exceeding 1.19 billion euros and a time span of 10 years. Fig.1.2, shows the upward rising trend and renewed interest in recent years.

### 1.1.2 Advantages

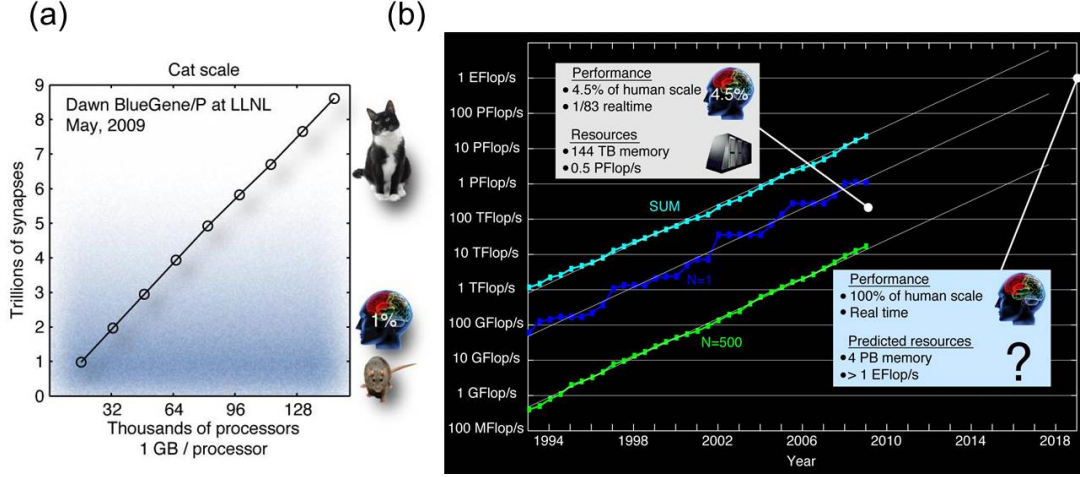
Apart from an improving historical trend, there are more concrete reasons that justify R&D for the development of special-purpose dedicated neuromorphic hardware. Neuromorphic computing offers several advantages compared to conventional von-neumann computing paradigms, such as-

- Low power/energy dissipation
- High scalability
- High fault-tolerance and robustness to variability
- Efficient handling of complex non-linear computations
- Programming free unsupervised learning
- High adaptability and re-configurability

While emulation of neural networks in software and Von-Neumann type hardware has been around for a while, they fail to realize the true potential of bio-inspired computing in terms of low power dissipation, scalability, reconfigurability and low instruction execution redundancy [13]. The human brain is a prime example of extreme biological efficiency on all these accounts- it consumes only about 20 W of power and occupies just about 2 L volume. Fig.1.3a shows the enormous number of CPUs required



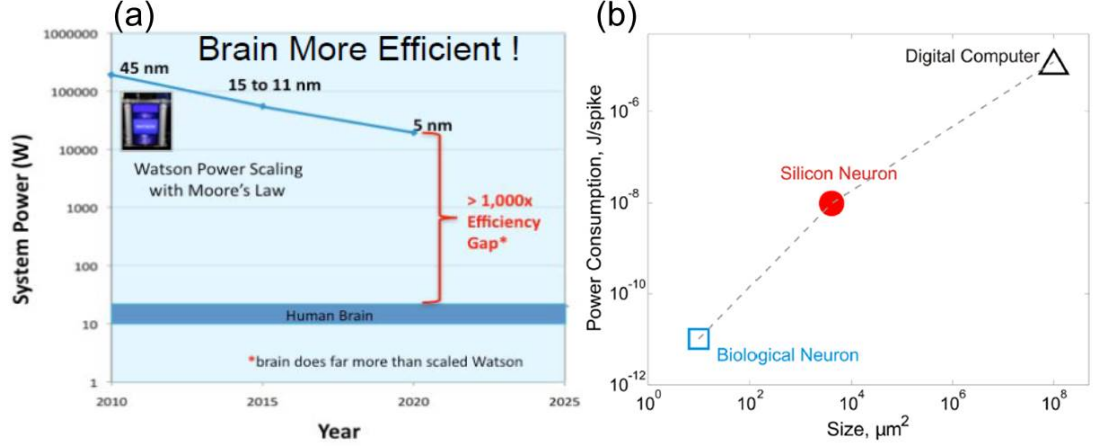
## 1. BACKGROUND



**Figure 1.3:** (a) Number of synapses Vs number of processing cores required for cat scale brain simulations using IBM Blue Gene supercomputer, (b) Growth of Top 500 supercomputers overlaid with recent IBM results and projection for realtime human-scale cortical simulation. Green line (lower) shows the 500th fastest supercomputer, dark blue line (middle) shows the fastest supercomputer, and light blue line (upper) shows the summed power of the top 500 machines [12].

to simulate just 1% of the human brain. Fig.1.3b, shows the extrapolated trend for the best supercomputers to perform full human brain scale simulations is real time. The trend predicts that for a full brain scale real-time simulation, 4 PB of memory and more than 1 EFlops/s of processing would be required [12]. Just for a 4.5 % human brain scale simulation the IBM Blue-gene supercomputer requires 144 TB memory, 0.5 PFlops/s processing, and about 1 Megawatt power.

Fig.1.4a, shows a power/energy comparison between the existing digital von-neumann systems and the brain. Even with strong Moore scaling till 2020's there will remain a huge power efficiency gap of more than 1000x. Fig.1.4b, outlines the power consumption difference between different neurons. A biological neuron consumes approximately  $3.84 \times 10^8$  ATP molecules in generating a spike. Assuming 30-45 kJ released per mole of ATP, the energy cost of a neuronal spike is in the order of  $10^{-11}$  J. The density of neurons under cortical surface in various mammalian species is roughly 100,000/mm<sup>2</sup>, which translates to a span of about 10  $\mu\text{m}^2$  per neuron. Silicon neurons have power consumption in the order of  $10^{-8}$  J/spike on a biological timescale. For example, an Integrate-and-Fire neuron circuit consumes 3-15 nJ at 100 Hz and a compact neuron



**Figure 1.4:** (a) System power scaling for IBM Watson supercomputer w.r.t human brain, (adapted from IBM research colloquia-2012, Madrid, Ritter et. al.). (b) Biological and silicon neurons have much better power and space efficiencies than digital computers [14].

model consumes 8.5-9.0 pJ at 1 MHz, which translates to 85-90 nJ at 100 Hz. For silicon neurons, the on-chip neuron area is estimated to be about 4,000  $\mu\text{m}^2$ . The power efficiency of digital computers is estimated to be  $10^{-3}$  to  $10^{-7}$  J/spike. Most current multi-core digital microprocessor chips have dimensions from 263 to 692 mm<sup>2</sup>. A single core has an average size from 50 to 90 mm<sup>2</sup> [14].

To emulate massively parallel asynchronous neural networks, the Von-Neumann architecture requires very high bandwidths (GHz) to transmit spikes to-and-fro between the memory and the processor. This leads to high power dissipation. The true potential of bio-inspired learning rules can be realized only if they are implemented on optimized special purpose hardware which can provide direct one-to-one mapping with the learning algorithms running on it [15].

### 1.1.3 Applications

Bio-inspired computing paradigms and neuromorphic hardware has a far reaching potential application base. Software based artificial neural networks are already being used efficiently in fields such as pattern- classification, extraction, recognition, machine-learning, machine-vision, robotics, optimization, prediction, natural language processing (NLP) and data-mining [16], [17]. Big-data analytics, data-center applications, and intelligent autonomous systems are new emerging fields where neuromorphic hard-

## 1. BACKGROUND

---

ware can play a significant role. Heterogeneous multi-core architectures with efficient neural-network accelerators have also been proposed in the recent years [18]. Neuromorphic concepts are also being explored for defense and security applications such as autonomous navigation [19], use in drones, crypt-analysis and cryptography. Neuromorphic hardware can also be used for health-care applications such as future generation prosthetics [20], brain-machine interfaces (BMI), and even serve as a reconfigurable simulation platform for neuroscientists.







### 1.2 Neurobiology Basics

Neurobiology is an extremely vast and complicated field of science. This section introduces some basic concepts and building blocks of a neuromorphic system such as neurons, synapses, spikes (action-potentials) and synaptic plasticity. We briefly describe how real stimuli or sensory information is converted to action potentials or spike trains for using the simplified examples of the mammalian retina for visual, and the cochlea for auditory processing respectively.

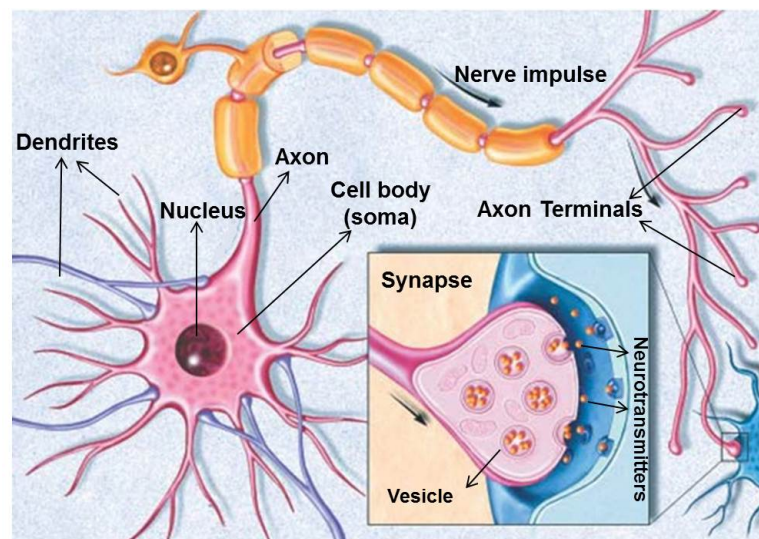
#### 1.2.1 Neuron, Synapse and Spike

A neuron is an electrically excitable cell, the basic building block of the nervous system. Neurons process, and transmit information between each other through detailed electrical and chemical signaling mechanisms. Neurons connect to each other with the help of synapses forming neural networks which perform different functions inside the brain such as vision, auditory perception, memory, movement, speech and communication with different body parts. It is estimated that there are about  $10^{11}$  neurons, and  $10^{15}$  synapses connecting them, to form various neural networks in the human cerebral cortex [21]. Increasing number of neurons and high synaptic connectivity leads to higher overall intelligence of the organism (Fig.1.5).

As shown in Fig.1.6, a neuron consists of three main parts- a cell body called the soma, the dendrites, and the axon. Dendrites are filaments that arise from the soma branching multiple times. The cell body (soma) is the metabolic center of the cell and it contains the cell nucleus. The nucleus stores the genes of the neuron. Dendrites act as the receivers for incoming signals from other nerve cells. The axon is the main

Animal	Neurons	Synapses	
Trichoplax	0	0	
Roundworm	300	$10^3$	
Honey bee	960,000	$10^9$	
Mouse	75,000,000	$10^{11}$	
Cat	1,000,000,000	$10^{12}$ – $10^{13}$	
Human	85,000,000,000	$10^{14}$ – $10^{15}$	

**Figure 1.5:** Species with increasing intelligence, number of neurons and synapses. Neuron and synapse numbers extracted from [22], [23], [24], [25], [26].



**Figure 1.6:** Illustration showing the basic structure of a neuron cell. Inset shows a zoom of the biological synapse. Adapted and modified from [27].

## 1. BACKGROUND

---

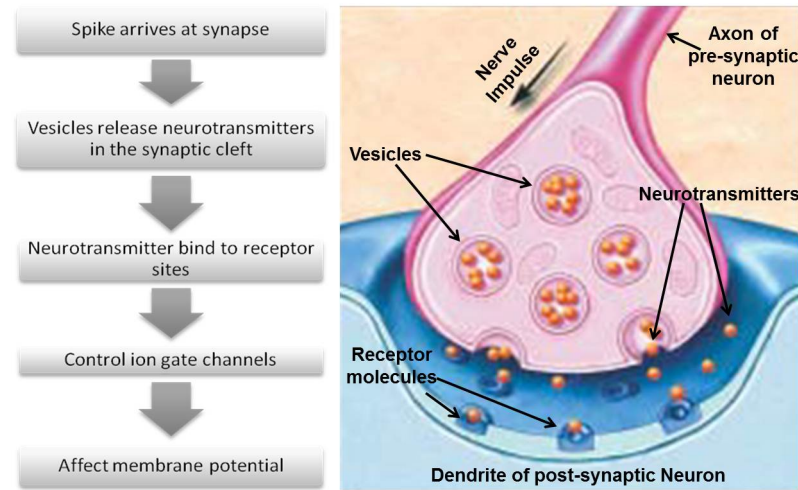
conducting unit, extruding from the soma, that is responsible for carrying electrical signals (called as action-potentials or spikes) to other neurons.

The spikes are rapid, transient, all-or-none nerve impulses, with an amplitude of 100 mV and a duration of about 1 ms (Fig.1.8). The neuron is surrounded by a plasma-membrane, made up of bilayer phospholipid molecules. The bilayer molecules make the membrane impermeable to the flow of ions. The impermeable membrane gives rise to a potential-gradient across the neuron and its extra-cellular medium due to differential ionic concentrations. In the unperturbed or equilibrium state, the neuron membrane stays polarized at a resting value of -70 mV [21]. However, the membrane is embedded with two special protein structures: namely ion-pumps, and voltage gated ion-channels, that allow the flow of ions in and out of the neuron under specific conditions. Ions such as  $\text{Na}^+$ ,  $\text{K}^+$ ,  $\text{Cl}^-$  and  $\text{Ca}^{2+}$ , play an essential role in the generation and propagation of the action-potentials.

Synapse is the region where the axon-terminal of one neuron comes close to the start of the dendrites of another neuron (see inset of Fig.1.6). Functionally, the synapse acts as a complex communication channel or the conductance medium between any two neurons, through which neuron signals are transmitted. The neuron which sends the spike to a synapse is termed as the pre-synaptic neuron, while the one that receives the spike is called the post-synaptic neuron. A single neuron in the cerebellum can have about  $10^3$  -  $10^4$  synapses, thus leading to massive parallel connectivity in the brain. Synapses and synaptic transmission can be either chemical or electrical in nature.

Electrical synapses are faster compared to chemical synapses. The signal transmission delay for electrical synapses is about 0.2 ms, compared to 2 ms for chemical synapses [28]. Electrical synapses are common in neural systems that require fast response time, such as defensive reflexes. An electrical synapse is a mechanical and electrically conductive link between two neurons that is formed at a narrow gap between the pre- and post- synaptic neurons known as a gap junction. Unlike chemical synapses, electrical synapses do not have gain. The post-synaptic signal is either of the same strength or smaller than the original signal.

Chemical synapses allow neurons to form circuits within the central nervous system that are crucial for biological computations that underlie perception and thought. The process of chemical synaptic transmission is summarized in Fig.1.7. The process begins with the action potential traveling along the axon of the pre-synaptic neuron, until it



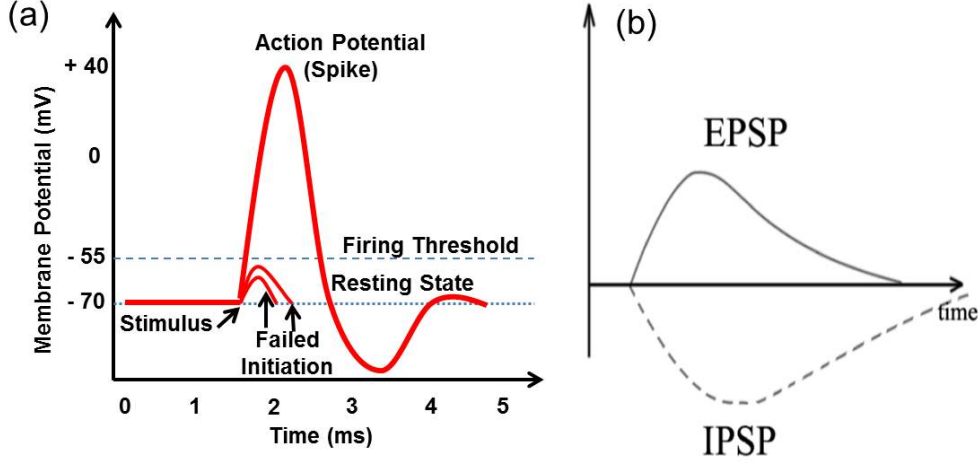
**Figure 1.7:** Illustration showing the chemical synaptic transmission process. Adapted and modified from [27].

reaches the synapse. Electrical depolarization of the membrane at the synapse leads to opening of ion-channels that are permeable to calcium ions. Calcium ions flow inside the pre-synaptic neuron, rapidly increasing the  $\text{Ca}^{2+}$  concentration. The high calcium concentration activates calcium-sensitive proteins attached to special cell structures (called vesicles) that contain chemical neurotransmitters. The neurotransmitters are then released into the synaptic cleft, the narrow space between the membranes of the pre- and post-synaptic neurons. The neurotransmitter diffuses within the cleft and binds to specific receptor molecules located in the post-synaptic neuron membrane.

Binding of neurotransmitter activates receptor sites of the post-synaptic neuron. Activation of receptors may lead to opening of ion-channels in the postsynaptic cell membrane, causing ions to enter or exit the cell, thus changing the resting membrane potential. The resulting change in the membrane voltage is defined as post-synaptic potential (PSP). If the PSP depolarizes the membrane of the post-synaptic neuron, it is called an Excitatory Post Synaptic Potential (EPSP). While, if the PSP hyperpolarizes the cell membrane, it is defined as an Inhibitory Post Synaptic Potential (IPSP). Depending on the type of PSP that a synapse generates, it can be classified as an excitatory- or inhibitory- synapse.

A neuron constantly integrates or sums all the incoming PSPs, that it receives at its dendrites, from several pre-synaptic neurons. The incoming EPSPs and IPSPs lead

## 1. BACKGROUND



**Figure 1.8:** (a) Illustration of neuron Action-Potential (spike). (b) EPSP and IPSP, adapted from [29].

to a change in the resting potential of the membrane. When the membrane potential depolarizes beyond  $-55$  mV, it leads to spiking or action potential generation inside the post-synaptic neuron. Thus, a neuron would spike only if the following criteria is satisfied by eq.1.1

$$\Sigma(\text{EPSP}) - \Sigma(\text{IPSP}) > (-55\text{mV}) \quad (1.1)$$

Fig.1.8 shows that if the resultant stimuli at the post-synaptic neuron is less than the firing threshold ( $-55$  mV), it leads to a failed initiation and no spike. In case of a failed initiation, the ion-pumps and the voltage gated ion-channels restore the membrane potential back to the resting value of  $-70$  mV. An interesting attribute of synaptic transmission is that it has been shown to be stochastic in nature due to probabilistic release of neurotransmitters [30].

### 1.2.2 Synaptic Plasticity and STDP

The strength (or weight) of a synapse is defined by the intensity of change that it can induce in the membrane potential of a post-synaptic neuron. Within a neural network synaptic strength may differ from one synapse to another, and evolve with time, depending on the nature of stimuli. The ability of a synapse to change its strength, in response to neuronal stimuli is defined as synaptic plasticity. Increase of synaptic strength is defined as synaptic-potential, while decrease is defined as synaptic-

depression. Synaptic plasticity effects, can either be shortterm (lasting for few seconds to minutes) or long-term (few hours to days). Different underlying mechanisms such as- changes in the quantity of released neurotransmitters, and changes in the response activity of the receptors, cooperate to achieve synaptic plasticity. Plasticity effects in both excitatory and inhibitory synapses have been linked to the flow of calcium ions [31].

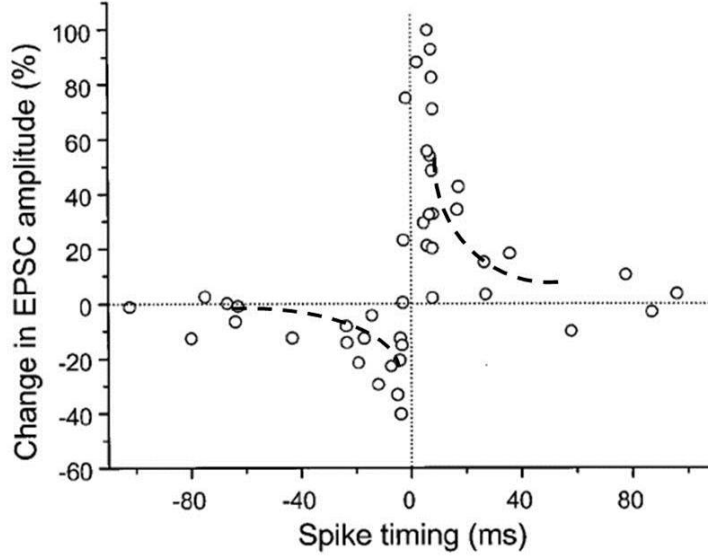
Learning and memory are believed to result from the effects of long-term synaptic plasticity such as long-term depression (LTD) and long-term potentiation (LTP)[32]. Both LTP and LTD are governed by multiple mechanisms that vary by species and the region of the brain in which they occur. In LTP the enhanced communication is predominantly carried out by improving the post-synaptic neuron's sensitivity to signals received from the pre-synaptic neuron. LTP increases the activity of the existing receptors, and the total number of receptors on the post-synaptic neuron membrane. While, LTD is thought to result mainly from a decrease in post-synaptic receptor density [33]. Conventionally, synaptic plasticity has been understood and formulated to be bi-directional, continuous, finely graded or analog levels of synaptic conductance states [34]. However recent neurobiological studies [35], indicate that bi-directional synaptic plasticity may be composed of discrete, non-graded and more digital or binary-like (all or none) synaptic conductance states [36].

Spike-timing dependent plasticity (STDP) is a biological process or learning-rule that adjusts the efficacy of synapses based on the relative timing of spiking of the pre- and post-synaptic neurons. According to STDP, if the pre-synaptic neuron spikes before the post-synaptic neuron, the synapse is potentiated. Whereas if the post-synaptic neuron spikes before the pre-synaptic neuron, the synaptic connection is depressed or weakened (LTD) [38]. Fig.1.9 shows the experimentally observed classical anti-symmetric STDP rule in cultured hippocampus neurons [37]. Note that the relative change of synaptic strength is more profound if the time difference ( $\Delta t$ ) between the spikes is smaller. As  $\Delta t$  increases, the effect of LTD and LTP becomes less profound like an exponential decay. STDP may vary depending upon the type of synapse and the region of the brain [39]. For classical anti-symmetric STDP rule, the width of the temporal windows for LTD and LTP are roughly equal in hippocampal excitatory synapses, whereas in the case of neocortical synapses the LTD timing window is considerably wider [40]. For some synapses [41], the STDP timing windows are inverted



## 1. BACKGROUND

---



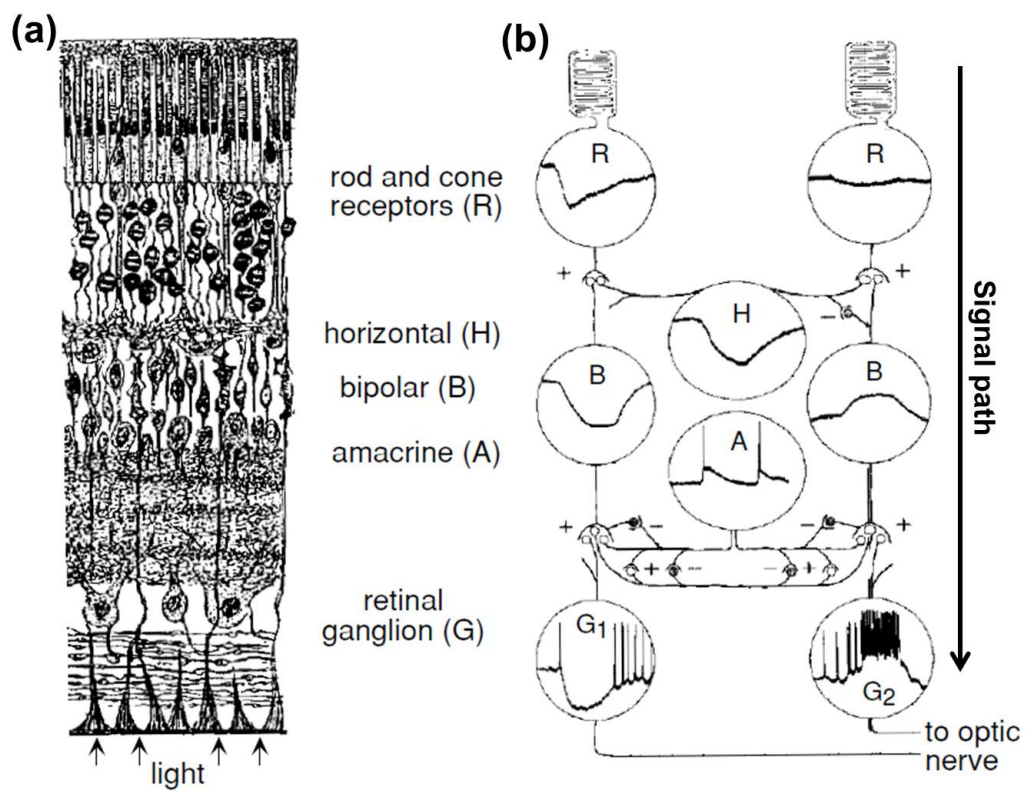
**Figure 1.9:** Experimentally observed, classical anti-symmetric STDP rule, in cultured hippocampus neurons.  $\Delta t < 0$  implies LTD while  $\Delta t > 0$  implies LTP [37]. Change in EPSP amplitude is indicative of change in synaptic strength.

compared to the form of STDP shown in Fig.1.9. Different forms of symmetric-STDP rules have also been shown in literature [42].

### 1.2.3 Retina: The Natural Visual Processing System

Fig1.10, shows an anatomical diagram of the retina and the signal pathway for visual stimuli. Light stimuli (photons) is first converted into electrical signals and then to a sequence or train of action potentials (spikes) at the output of the retina. The retina consists of three major types of neuron cells; (i) photoreceptors (rods and cones), (ii) intermediate-neurons (bipolar, horizontal and amacrine), and (iii) ganglion cells. Light is first converted into electrical signals, through complex biochemical processes, occurring inside the rods and cones. The rods are mainly responsible for night time vision, have a high sensitivity to light and high amplification. The cones are primarily responsible for color vision, more suited for day-time, are less sensitive to light, and have lower amplification [32].

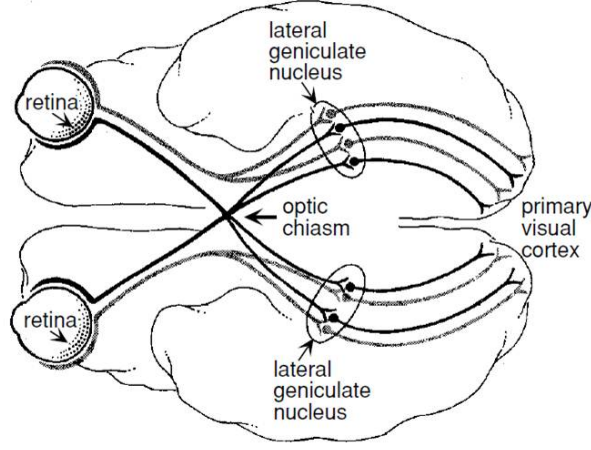
The electrical signals generated at the photoreceptor cells are passed to the intermediate-neurons (bipolar, horizontal and amacrine cells) through synaptic connections. The intermediate-neurons then pass the signals to the ganglion cells. The amacrine, bipolar



**Figure 1.10:** (a) Illustration showing different types of cells in the retina (b) Anatomical diagram of visual stimuli signal pathway starting at the photoreceptors and ending at the optic-nerve. Adapted from [32].

## 1. BACKGROUND

---



**Figure 1.11:** Pathway from the retina through the LGN of the thalamus to the primary visual cortex in the human brain [32].

and horizontal cells combine signals from several photoreceptors in such a way that the electrical responses evoked in ganglion cells depend critically on the precise spatial and temporal patterns of the light that stimulates the retina [32]. The retina compresses visual information by a factor of 100, as the number of photoreceptor cells is approximately 100 million while the number of nerve fibers comprising the optic nerve is only one million.

The ganglion cells are responsible for producing the final output of the retina and their axons converge to the optic nerve. Based on the type of ganglion cell, there can be different output spike patterns for a given stimuli. The spatial region inside which a ganglion cell is sensitive to any stimuli is defined as its receptive-field. For most ganglion cells the receptive field is divided into two parts: a circular zone at the center, called the receptive field center, and the remaining area of the field, called the surround. ON-type ganglion cells fire frequently if their receptive field center is illuminated, while OFF-type ganglion cells fire frequently only if their receptive field surround is illuminated. In Fig.1.10b, the ganglion cell G1 is OFF-type, while G2 is ON-type.

The optic nerve conducts the output spike trains from the ganglion cells to the region known as lateral geniculate nucleus (LGN) of the thalamus. The LGN acts as the relay station between the retina and the visual cortex (Fig.1.11). The visual

cortex is one of the most studied and well-understood cortical system in primates. It consists of several layers; V1, V2 etc. Information inside the visual cortex is propagated in a hierarchical manner mainly in one direction ('Feedforward') [11]. Functionally, the neurons in the layer V1 respond as highly selective spatiotemporal filters. Their receptive fields can be typically modeled by Gabor filters [43], which are sensitive to spatial and temporal orientation (or movement). The neurons of the second layer (V2) are functionally responsible for higher tasks such as encoding of complex shapes, combination of directions, edge detection, and surface segmentation [44].

Synaptic plasticity and STDP are believed to play an important role in the learning of complex intermediate features in visual data in an unsupervised manner [45]. It has been shown [46] that receptor fields similar to the ones found in V1 can emerge naturally through STDP on sufficiently large visual stimuli.

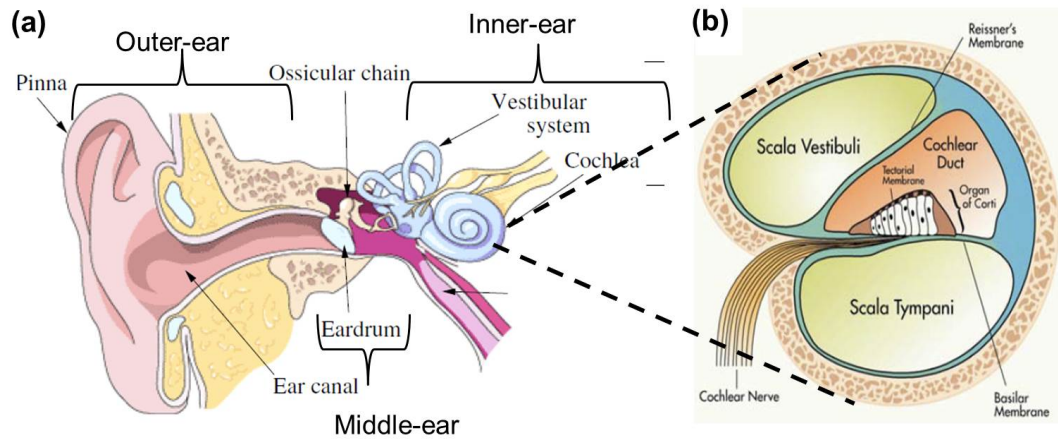
### 1.2.4 Cochlea: The Natural Auditory Processing System

The human ear can be broadly divided in three regions (outer-, middle- and inner-ear). Hearing starts with the capture of sound in the outer-ear (Fig.1.12a). Sound waves and mechanical energy flow through the middle-ear to the inner-ear (cochlea), where it is transduced in to electrical neural signals and coding. The complex auditory pathways of the brain stem mediate certain functions, such as the localization of sound sources, and forward auditory information to the cerebral cortex. Several distinct brain areas analyze sound to detect the complex patterns characteristic of speech.

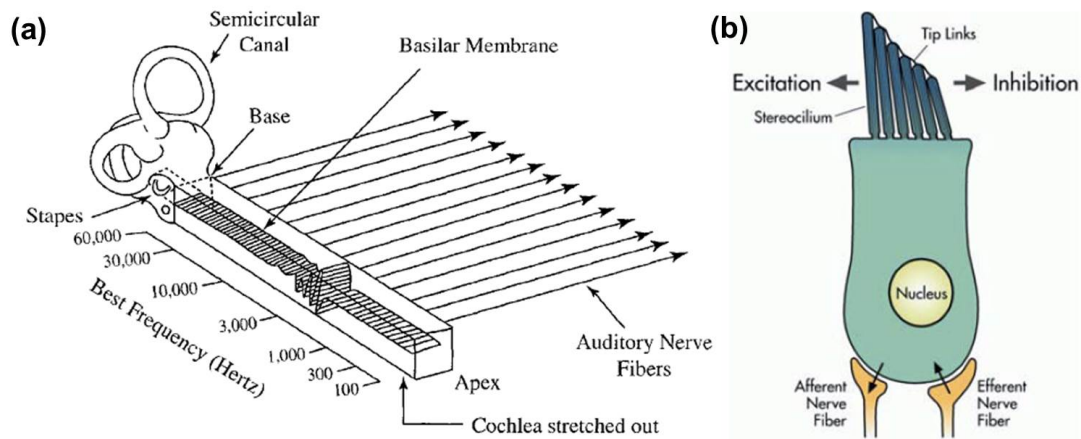
The human cochlea consists of coils with progressively diminishing diameter, stacked in a conical structure like a snail's shell. The interior of the cochlea contains three fluid-filled tubes, wound helically around a conical bony core called the modiolus. In a cross-sectional view (Fig.1.12b), the uppermost fluid-filled tube is the scala-vestibule, the middle tube is scala-media, and the lowermost tube is called the scala-tympani. A thin membrane (Reissner's membrane) separates the scala-media from the scala-vestibuli. The basilar membrane, which forms the partition between the scala-media and the scala-tympani, is a complex structure where the transduction of auditory-to-electrical signals occurs.

The basilar membrane acts as a mechanical analyzer of sound frequencies. Its mechanical properties vary continuously along the cochlea's length. As the cochlear chambers become progressively larger from the organ's apex toward its base the basilar

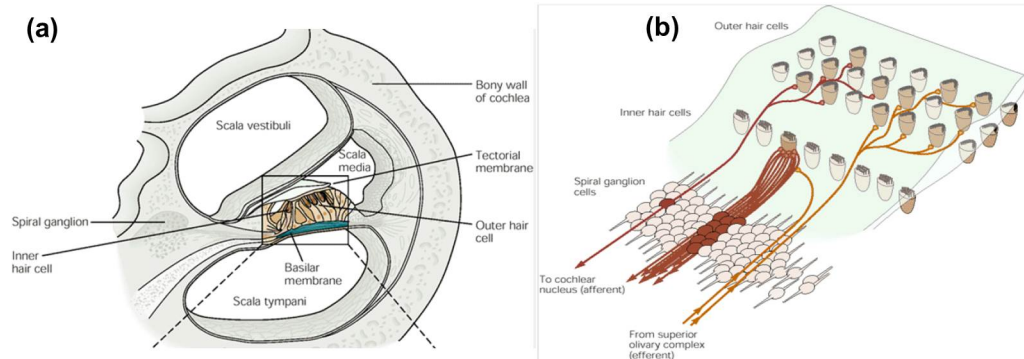
## 1. BACKGROUND



**Figure 1.12:** (a) Illustration of the human ear and (b) cross-section of the cochlea, adapted from [29].



**Figure 1.13:** (a) Illustration of uncoiled basilar membrane with different frequency sensitive regions, adapted from [47] (b) inner hair cell, movement of the stereocilium leads to generation of receptor potentials, adapted from [29]



**Figure 1.14:** (a) Illustration showing the organ of Corti in the cochlear cross-section. (b) zoomed view of the organ of Corti showing location of the inner hair cells, adapted from [21]

membrane decreases in width. The membrane is relatively thin and floppy at the apex of the cochlea but thicker and tauter towards the base. Such variation in mechanical properties accounts for the fact that the basilar membrane is tuned to a progression of frequencies along its length [21]. At the apex of the human cochlea the partition responds best to the lower frequencies of the order of 20 Hz, while at the opposite end, the membrane responds to higher frequencies around 20 kHz (Fig.1.13a). The relation between characteristic frequency and position upon the basilar membrane varies smoothly and monotonically but is not linear. Instead, the logarithm of the best frequency is roughly proportional to the distance from the cochlea's apex.

The organ of Corti is an important receptor part of the inner ear. It extends as an epithelial ridge along the length of the basilar membrane (Fig.1.14). It contains approximately 16,000 hair cells innervated by about 30,000 afferent nerve fibers, which carry information into the brain along the eighth cranial nerve. Like the basilar membrane, both the hair cells and the auditory nerve fibers are tonotopically organized: At any position along the basilar membrane they are optimally sensitive to a particular frequency, and these frequencies are logarithmically mapped in ascending order from the cochlea's apex to its base. The organ of Corti contains two types of hair cells (Fig.1.13b). The inner hair cells form a single row of approximately 3500 cells. Farther from the helical axis of the cochlear spiral lie rows of about 12,000 outer hair cells [21].

When the basilar membrane vibrates in response to a sound, the organ of Corti is also carried with it. This leads to deflection of hair bundles (Fig.1.13a). The me-

## 1. BACKGROUND

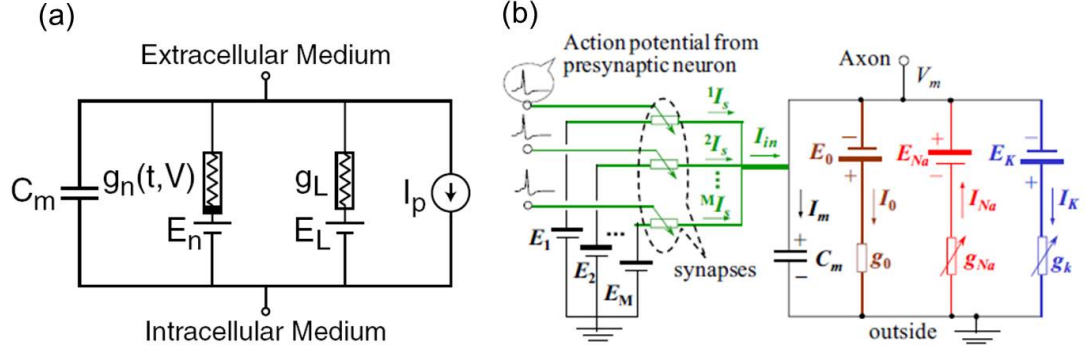
---

chanical deflection of the hair bundle is the stimulus that excites each hair cell of the cochlea. This deflection leads to generation of receptor potentials. The receptor potentials of inner hair cells can be as great as 25 mV in amplitude. An upward movement of the basilar membrane leads to depolarization of the cells, whereas a downward deflection leads to hyperpolarization. Due to the tonotopic arrangement of the basilar membrane, every hair cell is most sensitive to stimulation at a specific frequency. On average, successive inner hair cells differ in characteristic frequency by about 0.2%. Information flows from cochlear hair cells to neurons whose cell bodies lie in the cochlear ganglion. Since this ganglion follows a spiral course within the bony core (modiolus) of the cochlear spiral, it is also called the spiral ganglion (Fig.1.14). About 30,000 ganglion cells connect the hair cells of each inner ear. Each axon connects a single hair cell, but each inner hair cell directs its output to several nerve fibers, on average about 10. This arrangement has important consequences. Firstly, the neural information from which hearing arises originates almost entirely at inner hair cells, which dominate the input to cochlear ganglion cells. Secondly, the output of each inner hair cell is sampled by many nerve fibers, which independently encode information about the frequency and intensity of sound.

Each hair cell therefore forwards information of somewhat differing nature to the brain along separate axons. Finally, at any point along the cochlear spiral, or at any position within the spiral ganglion, neurons respond best to stimulation at the characteristic frequency of the contiguous hair cells. The central nervous system can get information about sound stimulus frequency in two ways. Firstly, a spatial code; the neurons are arrayed in a tonotopic map such that position is related to characteristic frequency. Secondly, a temporal code; the neurons fire at a rate reflecting the frequency of the stimulus.

### 1.3 Simplified Electrical Modeling

Numerous models of biological neurons and synapses, with varying degrees of complexity and abstraction, exist in literature. The complexity and the choice of a model depends on the application. For better understanding the working of the biological neurons or to simulate biology it is essential to have a detailed model which takes in



**Figure 1.15:** (a) Simplified circuit equivalent of the Hodgkin-Huxley (HH) neuron model. (b) Circuit model with synapses as variable programmable resistors [49].

account the dynamics at the level of individual ion-channels and underlying biophysical mechanisms. While for the purpose of bio-inspired or neuromorphic computing, which is more closely related to the scope of the work presented in this thesis, simple behavioral models are sufficient.

One of the earliest and simplest neuron models is the Integrate-and-Fire (IF) neuron model shown as early as 1907 [48]. In this model a neuron is represented by a simple capacitive differential eq.1.2-

$$I(t) = C_m \cdot \frac{dV_m}{dt} \quad (1.2)$$

Where,  $C_m$  denotes the neuron membrane capacitance. According to the IF model, the neuron constantly sums or integrates the incoming pre-synaptic currents and fires (generates action potential) when the membrane voltage reaches a certain firing threshold voltage ( $V_{th}$ ). An advanced and more relevant form of the IF model is the Leaky-Integrate-and-Fire (LIF) model, described by the eq.1.3-

$$I(t) - \frac{V_m(t)}{R_m} = C_m \cdot \frac{dV_m(t)}{dt} \quad (1.3)$$

The LIF model takes in account the leakage-effect of the neuron membrane potential by drift of some ions, assuming that the neuron membrane is not a perfect insulator.  $R_m$  denotes the membrane resistance. For the neuron to fire, the accumulated input should exceed the threshold  $I_{th} > V_{th}/R_m$ . Several CMOS-VLSI hardware implementations of functional IF and LIF neuron models have been described in literature [50].



## 1. BACKGROUND

---

A more detailed model is the Hodgkin-Huxley neuron model (HH). It describes the action potential by a coupled set of four ordinary differential equations [51]. Fig.1.15a, shows the simplified circuit equivalent of the HH model. The bilayer phospholipid membrane is represented as a capacitance ( $C_m$ ). Voltage-gated ion-channels are represented by nonlinear electrical conductances ( $g_n$ , where  $n$  is the specific ion- channel for  $\text{Na}^+$ ,  $\text{K}^+$ ), the conductance is a function of voltage and time. The electrochemical gradients driving the flow of ions are represented by batteries ( $E_n$  and  $E_L$ ). Ion-pumps are modeled by current sources ( $I_p$ ). Interestingly, neurons have also been modeled as pulse frequency signal processing devices, and synapses as variable programmable resistors (Fig.1.15b) [49].

### 1.4 Nanoscale Hardware Emulation of Synapses

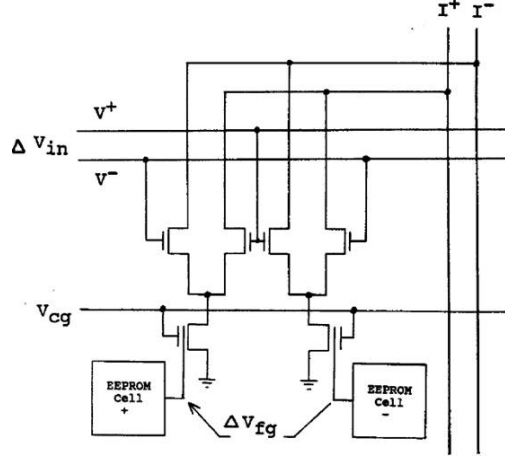
Several different hardware embodiments of artificial neural networks exist in literature. In this section we summarize some state-of-art hardware implementations of synapses based on (i) VLSI-technology and (ii) Exotic devices. We outline some limitations of these approaches and introduce the concept of emerging non-volatile Resistive Memory (RRAM) technology and its advantages. The underlying or unifying theme in most of the embodiments discussed in this section is that the synapse is broadly treated as a non-volatile, programmable resistor.

#### 1.4.1 VLSI-technology

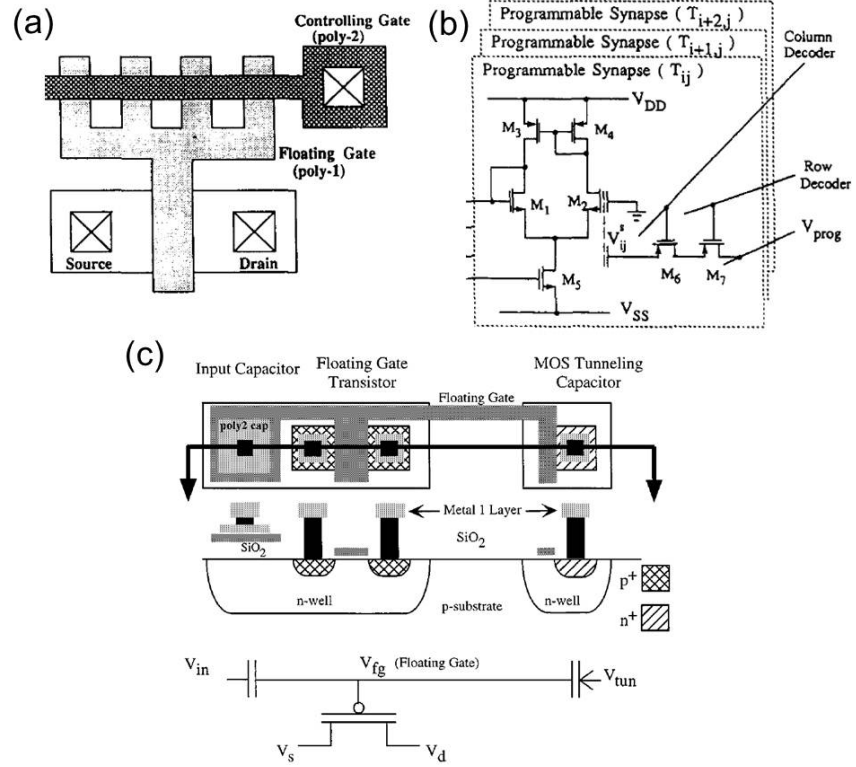
These include emulation of synaptic behavior with VLSI structures such as floating-gate transistors, DRAM and SRAM.

##### 1.4.1.1 Floating-gate Synapses

The 10240 synapses in Intels ETANN chip were realized using EEPROM cells (see Fig.1.16) [8]. For each synapse circuit in ETANN, a pair of EEPROM cells are incorporated in which a differential voltage representing the weight may be stored or adjusted. Electrons are added to or removed from the floating gates in the EEPROM cells by Fowler-Nordheim tunneling. A desired differential floating-gate voltage can be attained by monitoring the conductances of the respective EEPROM MOSFETs.

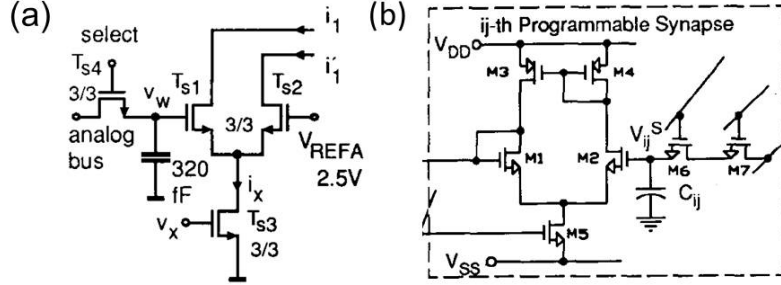


**Figure 1.16:** Intel ETANN synapse structure implemented using EEPROM floating-gate devices [8].



**Figure 1.17:** (a) Layout of poly silicon floating-gate synaptic device [52]. (b) circuit schematic of floating-gate synapse with transconductance amplifier [52]. (c) layout of floating-gate pFET synaptic device [53].

## 1. BACKGROUND



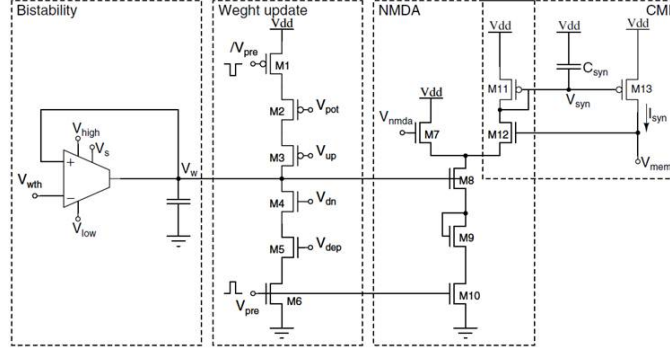
**Figure 1.18:** (a) circuit schematic of analog DRAM based synapse with three additional transistors [55]. (b) circuit schematic of DRAM based synapse with transconductance amplifier [56].

Double-poly floating gate transistors along with transconductance amplifiers (Fig.1.17a,b) were used by Lee et. al [52], for implementing VLSI synapse circuits. In this approach the synaptic weight was programmed using Fowler-Nordheim tunneling, and the neural computation is interrupted for the duration of the applied programming voltages. Correlation learning rules have also been demonstrated on synapses made of floating-gate pFET type structures, as shown in Fig.1.17c, [53]. In this approach Fowler-Nordheim tunneling is used to remove charge from the floating-gate and thus increase the synapse channel current. Conversely, pFET hot-electron injection is used to add charge to the floating-gate and decrease the synapse channel current. More recently synaptic plasticity effects like LTP/LTD and the STDP learning rule were demonstrated on floating-gate pFET structures with the help of additional pre-synaptic computational circuitry [54].

### 1.4.1.2 Dynamic Random Access Memory (DRAM) or Capacitive Synapses

Different DRAM (or capacitor based) synaptic hardware implementations utilizing both analog and digital types of storage have been proposed in literature. Jerzy et.al, demonstrated an analog multilayer perceptron network with back-propagation algorithm using nonlinear DRAM synapses [55]. Their synapse consists of a storage capacitor and 3 additional transistors (Fig.1.18a).

The DRAM based analog synaptic weight storage suffers from capacitive discharge, need for frequent refresh, noise induced from the switching transistors and errors due to clock feedthrough [55]. Lee et.al, proposed a DRAM based synapse with 7 additional



**Figure 1.19:** Capacitor based synapse with additional learning and weight update circuit blocks [58].

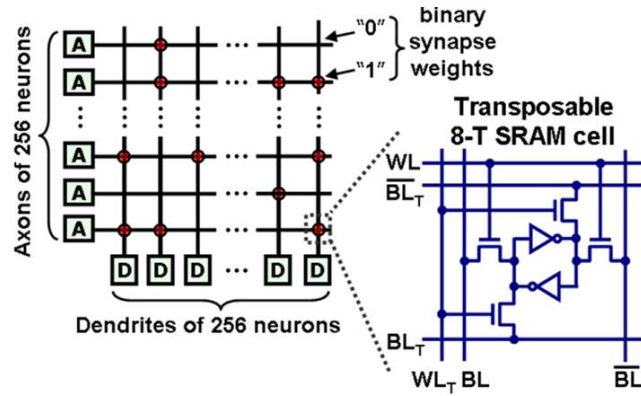
transistors or a transconductance amplifier (Fig.1.18b). They show a 8-bit synaptic weight accuracy with a 0.2 s refresh cycle. Additional decoder circuitry is required to program the synapse weight voltage. Takao et.al, demonstrated a digital chip architecture with  $10^6$  synapses [57]. They use an on-chip DRAM cell array to digitally store 8-bit synaptic weights with automatic refreshing circuits. In some capacitive implementations there are additional weight update circuits inside the synaptic block, like the one shown in Fig.1.19, [58]. The additional circuits are needed to implement the learning rules. Similar of circuits with learning functionality are also shown in [59], [50].

### 1.4.1.3 Static Random Access Memory (SRAM) Synapses

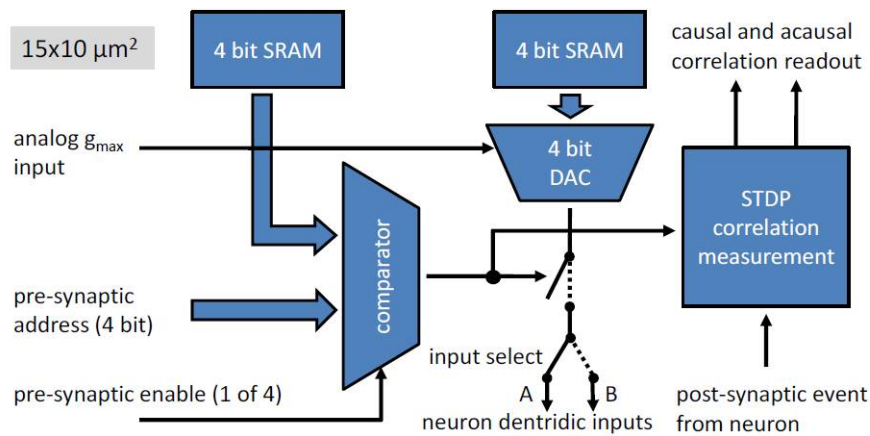
More recently, the use of standard and modified SRAM cells has also been proposed for synaptic emulation. IBM proposed a modified 8-T transposable SRAM cell (Fig.1.20) in their digital neurosynaptic core [60]. Their modified 8-T structure enables single-cycle write and read access in both row and column directions. The cell area is  $1.6 \mu\text{m}^2$  in 45 nm node. The 8-T SRAM synapses have binary weights which are probabilistically controlled. A variant structure containing 4-bit analog weight was also implemented [60].

4-bit SRAM cells are also used to store individual synaptic weights in the wafer-scale FACETS neuromorphic project [61]. Fig.1.21 shows the schematic diagram of a single synapse for the FACETS project. Two types of plasticity rules: short-term depression (STD) and STDP are implemented using the FACETS synapses.

## 1. BACKGROUND



**Figure 1.20:** IBM's 45 nm node neurosynaptic core and 8-T transposable modified SRAM cell [60].



**Figure 1.21:** Synapse schematic comprising of 4-bit SRAM cells for the wafer-scale FACETS neuromorphic project [61].

The Spinnaker approach uses specially designed hardware synaptic channels with off-chip mobile DDR SDRAM memory with a 1 GB capacity. Synaptic weights use a large, concurrently-accessed global memory for long-term storage. Since the SDRAM resides off-chip, it is easy to expand available global memory simply by using a larger memory device [62].

### 1.4.1.4 Limitations of VLSI type synapses

While synaptic emulations that use VLSI constituents discussed in the previous section (like floating-gate transistors, DRAMs, SRAMs, DDR-SDRAM) are tempting to use, considering the availability of standardized design tools and a mature fabrication process, there exist several limitations. Floating-gate devices are not ideal for mapping bio-inspired learning rules because unlike biological synapses they are 3-terminal. During synaptic learning individual synapses may undergo weight modification asynchronously, which is not very easy to do with the available addressing schemes for large Flash arrays. Floating-gate devices also require high operating voltages. In many cases, additional pre-synaptic circuitry is required to implement timing dependent learning rules, due to the difference in the charging and discharging physics of the floating gate devices. The pulse shapes used to program floating-gate devices are complicated. Endurance of even state-of-the-art floating-gate devices (Flash) is not very high. Due to the operating physics, there exists an inherent limitation on the frequency of programming synapses based on floating-gate FETs.

The DRAM or capacitor based synapses require frequent refresh cycles to retain the synaptic weight. In most of the capacitor based demonstrations, a single synapse circuit needs more than 10 additional transistors to implement learning rules, as shown in sec.1.4.1. The capacitor is also an area consuming entity for the circuit. The SRAM based synapses further suffer due to disadvantage in terms of area consumption and volatility. When the network is turned off, the synaptic weights are lost, and so they need to be stored to some offline memory during or after the learning. Reloading of the synaptic weights during learning operation will lead to additional power dissipation and silicon area overhead. These limitations lay the basis for the interest in synaptic emulation with new types of emerging non-volatile memory technology (RRAM) as described in sec.1.4.3.

## 1. BACKGROUND

---

### 1.4.2 Exotic Device Synapses

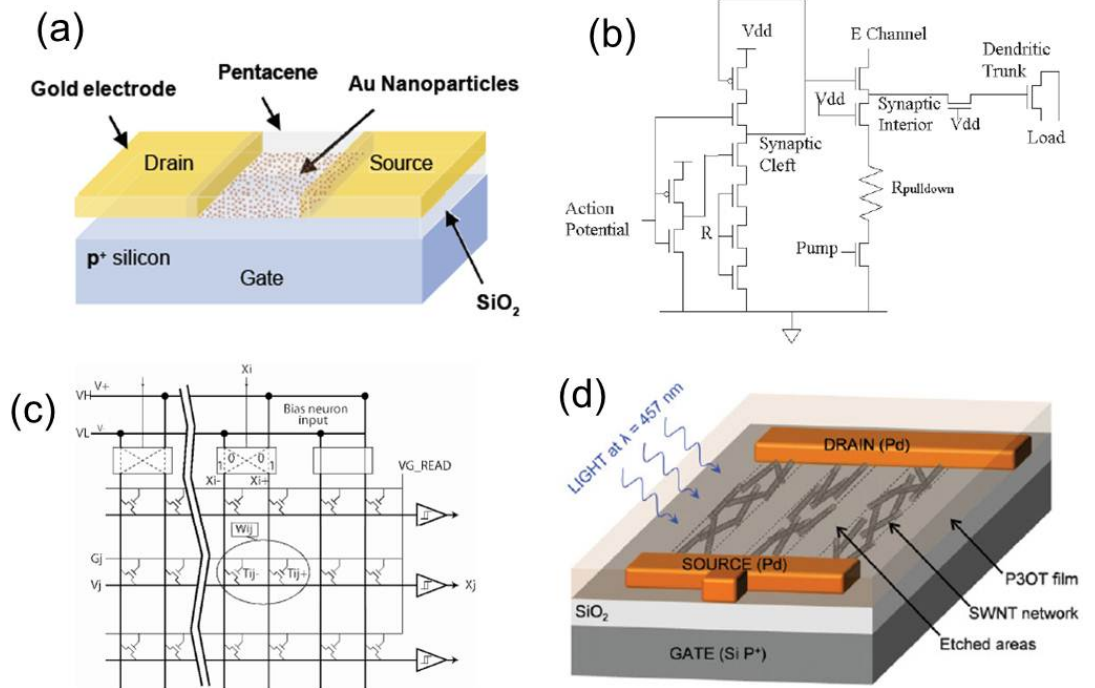
Several interesting exotic devices such as organic-transistors, single-electron transistors, optically-gated transistors, atomic-switches and even thin-film transistors have been used to implement synaptic behavior.

Alibart et. al, propose an organic nanoparticle field-effect transistor (NOMFET) that can emulate effects such as synaptic short-term plasticity effects (potentiation/depression) and STDP based learning rules [63]. The NOMFET structure (Fig.1.22a) exploits (i) the transconductance gain of the transistor and (ii) the memory effect due to charges stored in the nano-particles (NPs). The NPs are used as nanoscale capacitors to store the electrical charges and they are embedded into an organic semiconductor layer of pentacene. The transconductance of the transistor can be dynamically tuned by the amount of charge in the NPs. More recently the NOMFET based synapses were also used to demonstrate associative learning based on Pavlov's dog experiment [64].

A.K Friesz used SPICE models to propose a carbon nanotube based synapse circuit (Fig.1.22b) [65]. The output of their CNT synapse circuit produces excitatory post-synaptic potentials (EPSP). Carbon nanotube transistors with optically controlled gates (OG-CNTFETs) have also been proposed by different groups for synaptic emulation (see Fig.1.22c,d) [66], [67]. Agnus et.al [67], show that the conductivity of the OG-CNTFETat can be controlled independently using either a gate potential or illumination at a wavelength corresponding to an absorption peak of the polymer.

Recently, 2-terminal "atomic switch" structures consisting metal electrodes, a nanogap and an Ag<sub>2</sub>S electrolyte layer (see Fig.1.23a), have been shown to emulate short-term synaptic plasticity and LTP type of effects [68], [69]. Avizienis et. al, [70] fabricated massively interconnected silver nanowire networks (Fig.1.23b) functionalized with interfacial Ag/Ag<sub>2</sub>S/Ag atomic switches. Cantley et.al, used spice models to demonstrate hybrid synapse circuits comprising of nano-crystalline ambipolar silicon thin-film transistors (TFT) and memristive devices [71].

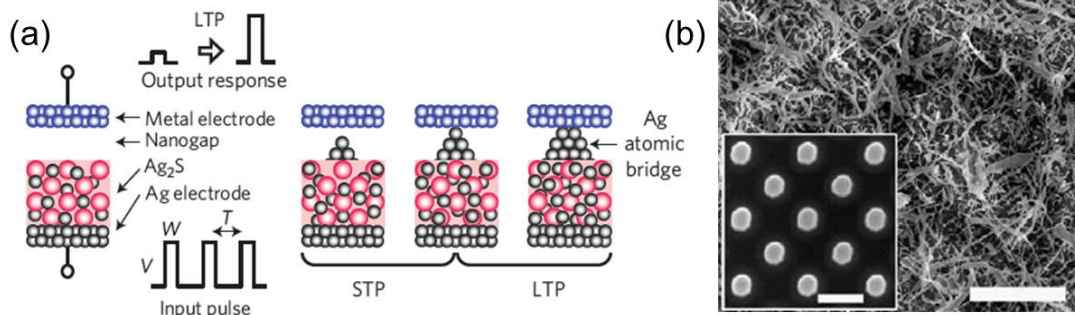
The exotic devices discussed herein suffer from limitations such as complicated fabrication process, poor CMOS compatibility, low technological maturity, and high voltage operation in some cases.



**Figure 1.22:** (a) Physical structure of the NOMFET. It is composed of a p<sup>+</sup> doped bottom-gate covered with silicon oxide (200 nm). Source and drain electrodes are made of gold and Au NPs (20 nm diameter) are deposited on the interelectrode gap (5  $\mu$ m), before the pentacene deposition [64]. (b) The carbon nanotube synapse circuit [65]. (c) Neural Network Crossbar with OG-CNTFET synapse [66]. (d) Schematic representation of a nanotube network-based OG-CNTFET [67].



## 1. BACKGROUND

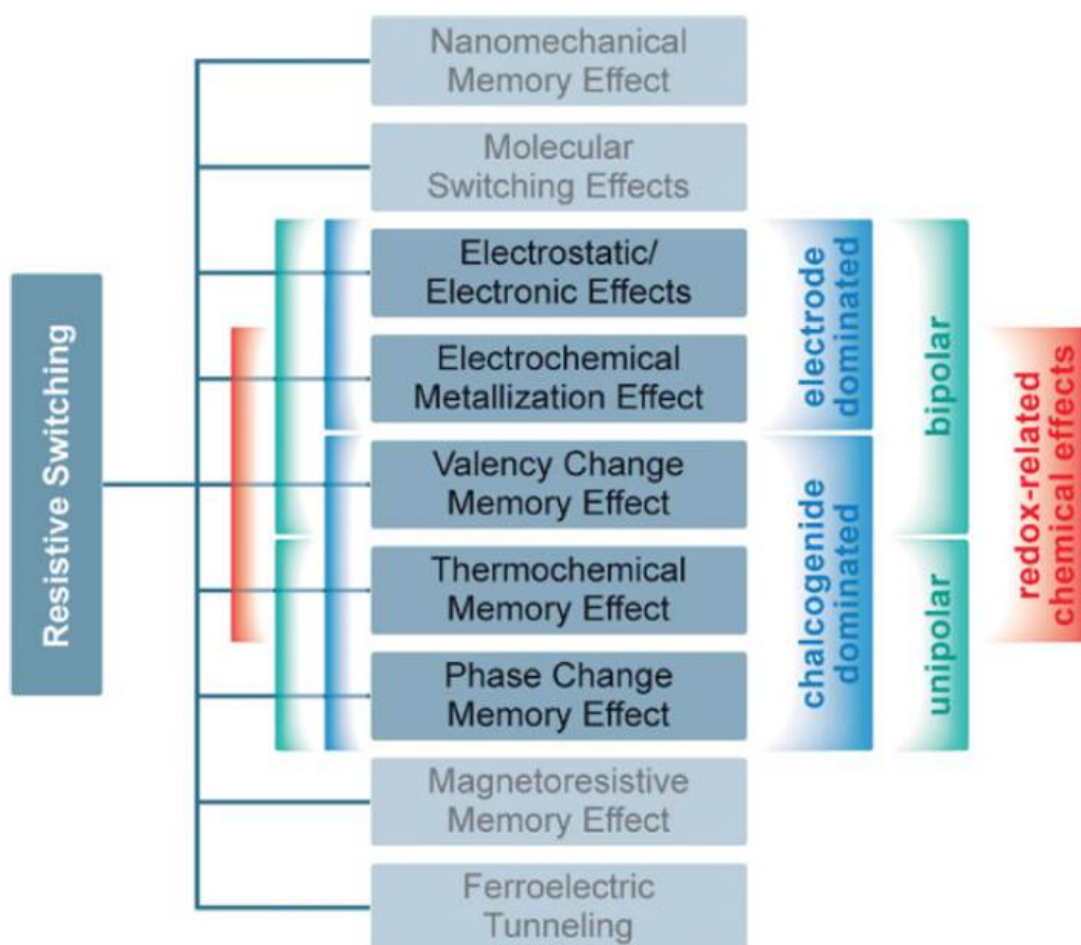


**Figure 1.23:** (a) Schematics of a  $\text{Ag}_2\text{S}$  atomic switch inorganic synapse. Application of input pulses causes the precipitation of Ag atoms from the  $\text{Ag}_2\text{S}$  electrode, resulting in the formation of a Ag atomic bridge between the  $\text{Ag}_2\text{S}$  electrode and a counter metal electrode. When the precipitated Ag atoms do not form a bridge, the inorganic synapse works as STP. After an atomic bridge is formed, it works as LTP [68]. (b) SEM image of complex Ag networks produced by reaction of aqueous  $\text{AgNO}_3$  with (inset) lithographically patterned Cu seed posts [70].

### 1.4.3 Resistive Memory Technology (RRAM)

Resistive random access memory (RRAM) is an umbrella term for emerging non-volatile memory (NVM) devices and concepts based on electrically switchable resistance states. The motivation behind the development of RRAM technologies is to overcome the limitations of existing VLSI memory concepts. A RRAM cell is generally a capacitor-like MIM structure, composed of an insulating material ‘I’ sandwiched between two metallic electrodes ‘M’ [72]. The MIM cells can be electrically reversibly switched between two or more different resistance states by applying appropriate programming voltages or currents. The programmed resistance states are non-volatile. Based on the type of material stack and the underlying physics of operation, the RRAM devices can be classified in several categories. Fig.1.24, shows different types of emerging RRAM technologies classified on the basis of the underlying resistance-switching physics. RRAM is also vaguely defined as ‘memristor’ or ‘ReRAM’.

This thesis focuses on three specific types of RRAM technologies: (i) unipolar Phase Change Memory (PCM), based on phase change effects in chalcogenide layers, (ii) bipolar Conductive Bridge Memory (CBRAM), based on electrochemical metallization effect, and (iii) bipolar Oxide based resistive memory (OXRAM), based on valency change/electrostatic memory effects.



**Figure 1.24:** Classification of the resistive switching effects which are considered for non-volatile memory applications [72].

## 1. BACKGROUND

---

**Table 1.1:** Comparison of emerging RRAM technology with Standard VLSI technologies. Adapted from ITRS-2012. (Values indicated for PCM and Redox are the best demonstrated.) Redox includes both CBRAM and OXRAM devices.

Parameter	DRAM	SRAM	NOR Flash	NAND Flash	PCM	Redox (OX/CB)
Cell Area	6F <sup>2</sup>	140F <sup>2</sup>	10F <sup>2</sup>	5F <sup>2</sup>	4F <sup>2</sup>	4F <sup>2</sup>
Feature Size (nm)	36	45	90	22	20	9
Read Time (ns)	<10	0.2	10	50	12	<50
Write-Erase Time	<10 ns	0.2	1 $\mu$ s / 10 ms	1 ms / 0.1 ms	50 ns / 120 ns	0.3 ns
Write Voltage (V)	2.5	1	12	15	3	0.6/-0.2
Read Voltage (V)	1.8	1	2	2	1.2	0.15
Write Energy (J)	5E-15	5E-16	1E-10	>2E- 16	6E-12	1E-13
Endurance	64 ms	NA	>10 years	>10 years	>10 years	>10 years

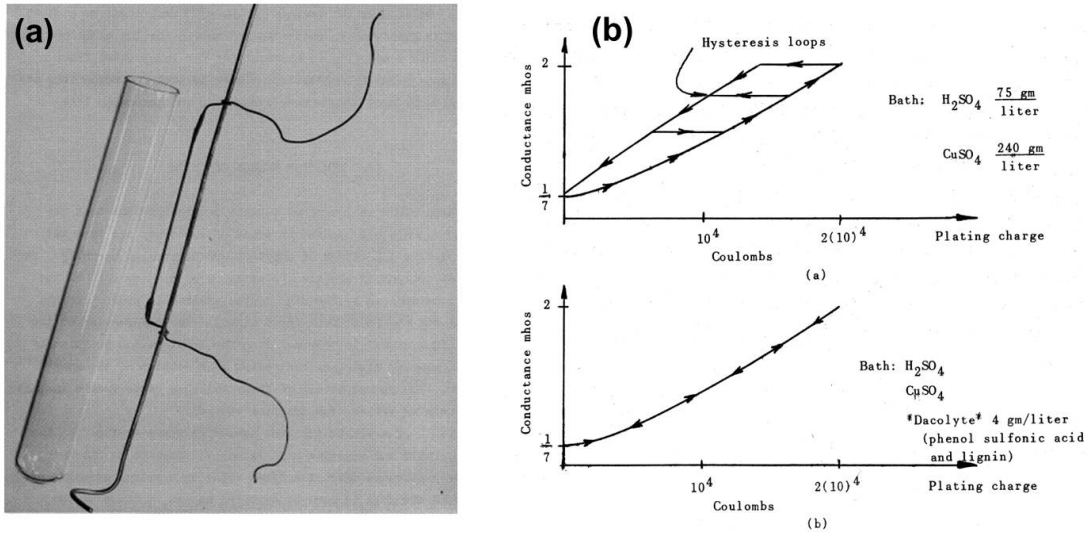
RRAM technologies offer interesting attributes both for replacing standard VLSI type memories and also for emulating synaptic functionality in large-scale silicon based neuromorphic systems (see Tab.1.1). Some promising features of RRAM are: full CMOS compatibility, cheap fabrication, high integration density, low-power operation, high endurance, high temperature retention and multi-level operation [73], [74], [75]. The two terminal RRAM devices can be integrated in 2D or 3D architectures with-selector device configuration (1 Transistor/Diode - 1 Resistor) or selector-free configuration (1 Resistor) [72]. The detailed RRAM working, and state-of-art synaptic implementations with RRAM devices is discussed in chapters 2 and 3.

### 1.4.3.1 Memistor Synapse (The father of Memristor or RRAM)

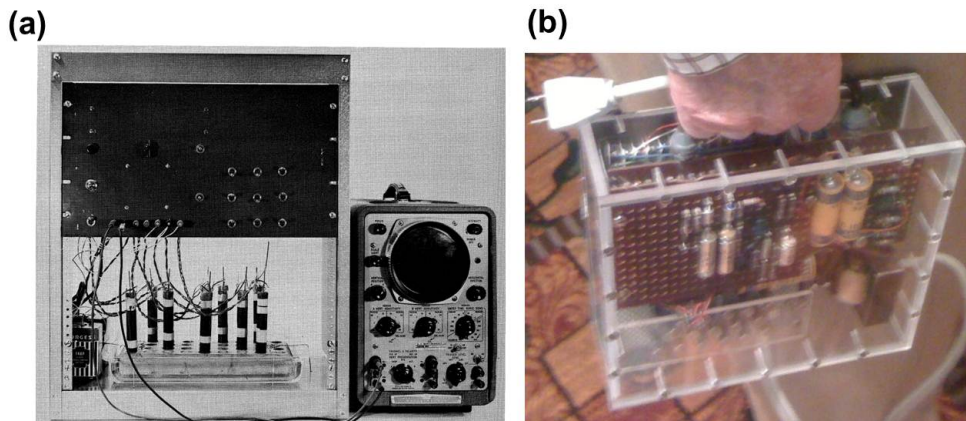
The memistor is one of the earliest electronic devices developed specially for emulation of synaptic functionality in artificial neural networks. It was first proposed and demonstrated in 1960 [76], by Bernard Widrow and Ted Hoff (who later became one of the inventors of the microprocessor), used as synapse in their pattern classification ADALINE neural architecture [77]. Memistor is a three-terminal device, not to be confused with the two-terminal memristor first theoretically postulated by Leon Chua in 1971 [78], and later experimentally claimed by HP labs [79], rather it is a predecessor to both of them.

Memistor working is based on reversible electroplating reactions. Fig.1.25a, shows the photograph of Widrows memistor made of pencil led graphite and a supporting copper rod. Resistance is controlled (or programmed) by electroplating copper from a copper sulphate-sulphuric acid solution on a resistive substrate (graphite). Change in memistor conductance with application of plating current and a hysteresis effect is shown in Fig.1.25b. Fig.1.26a shows the original ADALINE neural architecture with a 3x3 memistor array and 1 neuron, developed in 1960. Fig.1.26b shows a more recent and compact version of the same. Inspired from widrow's electroplating based memistor, a fully solid state memistor for neural networks was demonstrated in 1990 [80]. It is a 3-terminal device based on tungsten-oxide ( $\text{WO}_3$ ), Ni-electrodes and a Al-gate Fig.1.27a. A voltage controlled, reversible injection of  $\text{H}^+$  ions in electrochromic thin films of  $\text{WO}_3$  is utilized to modulate its resistance. A hygroscopic thin film of  $\text{Cr}_2\text{O}_3$  is the source of  $\text{H}^+$  ions. The resistance of the device can be modulated over four orders of programming window. The programming speed can be modulated by

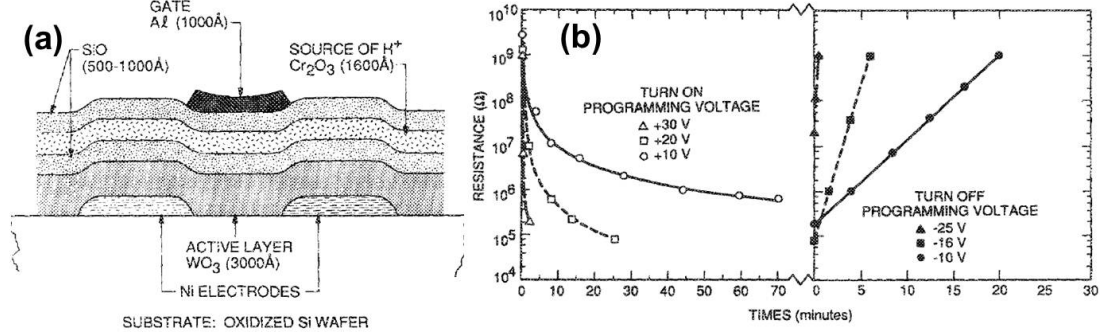
## 1. BACKGROUND



**Figure 1.25:** (a) Photo of the constituents of the copper-sulphate based memistor device. (b) Characteristic programming curves showing hysteresis loop in the memistor devices, adapted from [76].



**Figure 1.26:** (a) Photo of the ADALINE architecture with 1 neuron and 3x3 memistor synapses [76]. (b) Recent photo of the ADALINE system containing memistors taken at IJCNN-2011.



**Figure 1.27:** (a) Cross-section schematic of the tungsten oxide based 3-terminal memistor. (b) Programming characteristics of the solid state memistor device, adapted from [80].

control voltage. Fig.1.27b shows the time-dependent programming characteristics of the solid-state memistors.

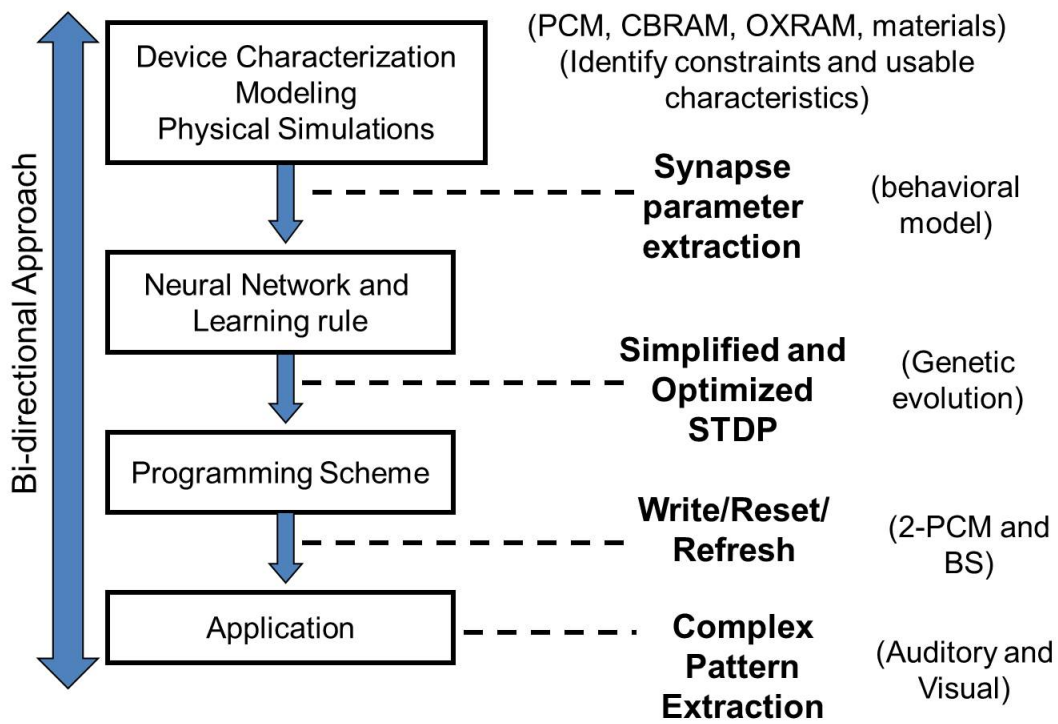
## 1.5 Scope and approach of this work

The work done in this PhD thesis focuses on the development of a complete "synapse-solution" for using specific RRAM devices inside large-scale ultra-low power neuromorphic systems.

The "synapse-solution" comprises of different ingredients starting from individual devices, circuits, programming-schemes and learning rules. For each of the three RRAM technologies investigated in this work (i.e. PCM, CBRAM and OXRAM), we begin with investigating the basic working and characteristics of the devices. We then identify what device characteristics can be directly used in a neural learning environment, and which ones are not usable. At the device level we identify some material and physical characteristics that can be tuned or engineered to optimize the individual synaptic performance. At the architectural level, we propose specific connection topologies and detailed programming methodologies (Read, Write, Refresh) for building low-power neuromorphic systems. At the level of algorithms or learning we propose the use of simplified and optimized learning rules. Finally we demonstrate relevant applications such as complex auditory and visual pattern extraction/recognition. The overall strategy for developing a perfect hardware "synapse-solution" should be of bi-directional nature (see Fig.1.28). The bottom-up approach begins with an individual synaptic device, while the top-down approach is more application centric.

## 1. BACKGROUND

---



**Figure 1.28:** Bi-directional strategy (Top-down + Bottom-up) adopted for the work presented in this PhD thesis. To develop the ideal "synapse-solution" optimization and fine-tuning was performed at different levels such as architectures, learning-rules and programming-schemes.(BS: Binary Stochastic synapses).

## **1.6 Conclusion**

In this chapter, we looked at some key motivations behind R&D in the field of neuromorphic computing. We then summarized the main biological concepts relevant for the purpose of this work. In the last section of the chapter, we discussed several different hardware implementations of synapses and the positioning of emerging RRAM technologies. The following chapters specifically focus on synaptic emulation using individual RRAM technologies (PCM, CBRAM and OXRAM).



## 1. BACKGROUND

---

“An experiment is something which everyone believes  
except the person who did it,  
A simulation is something that no one believes...  
except the person who did it.”

## 2

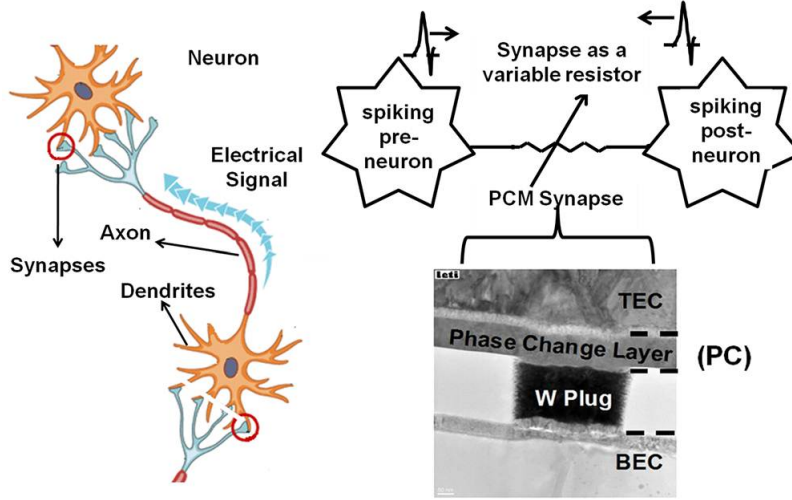
# Phase Change Memory Synapses

This chapter discusses how Phase Change Memory (PCM) technology can be used to emulate biological synapses in large-scale neuromorphic systems with low-power dissipation and easy to implement programming methodology.

## 2.1 PCM Working Principle

PCM devices consist of an active chalcogenide layer sandwiched between two metal electrodes. The working principle exploits reversible and nonvolatile phase-change phenomenon inside chalcogenide layers such as  $\text{Ge}_2\text{Sb}_2\text{Te}_5$  (GST). The high resistive amorphous phase is usually defined as the RESET state, while the low-resistive crystalline phase as the SET state. When a bias is applied across the two electrodes of the PCM, current flows through the metallic heater and the chalcogenide layer, causing joule-heating. Depending on the pulse-duration, fall-time edge, and the amplitude of the current flowing through the device, crystalline, amorphous, or partially crystalline and partially amorphous regions can be created inside the chalcogenide layer. If the chalcogenide layer is melted and quenched quickly, it does not get sufficient time to re-organize itself into a crystalline structure and thus amorphous regions are created. If the chalcogenide layer is heated, between the glass-transition and the melting temperature, for sufficiently long time it leads to crystallization [73]. Intermediate resistance states can also be obtained by tuning the programming conditions and controlling the volume of amorphous/crystalline volume fraction created inside the chalcogenide layer [73].

## 2. PHASE CHANGE MEMORY SYNAPSES

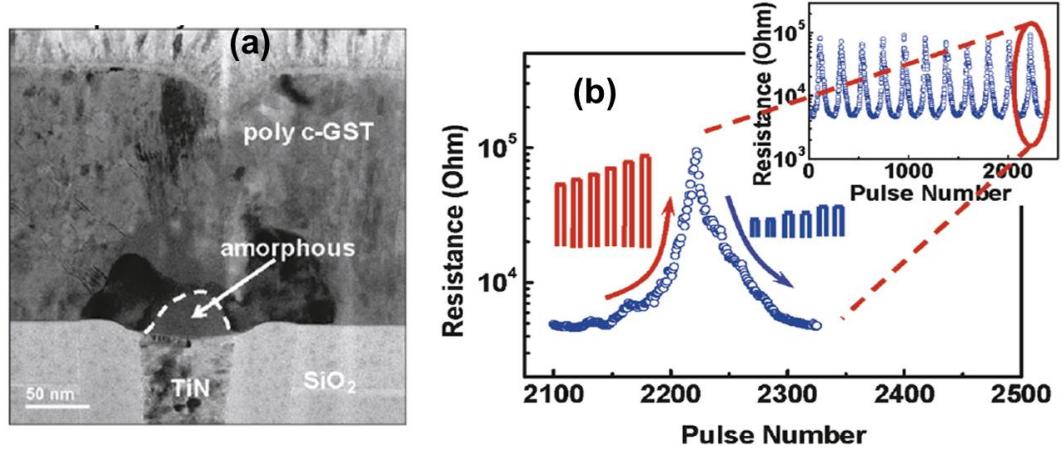


**Figure 2.1:** Illustration of biological synapse and the equivalent PCM synapse in a neural circuit connecting a spiking pre- and post- neuron [81]. TEM cross-section image of the GST PCM devices fabricated for this study is shown.

### 2.2 State-of-Art PCM Synapses

Kuzum et.al [82] have shown the emulation of progressive LTP- and LTD- like effects, by programming their PCM synapses, with different voltage pulse trains of increasing amplitude (Fig.2.2b).

Kuzum et. al [83] have shown emulation of asymmetric and symmetric forms of STDP using PCM synapses. For implementing the STDP rules they propose the interaction of a specially designed pre-neuron spike (a combination of two or three different pulse trains) and a simple post-neuron spike (Fig.2.3a) at the PCM synapse. Half of the pre-neuron spike consists of a depressing pulse train of increasing amplitude, while the other half consists of a potentiating pulse train of decreasing amplitude. The post neuron spike is a single pulse of inverted polarity. Only an overlap of the pre- and post-neuron spike leads to a change in the synaptic weight. When the post neuron spike arrives late, it overlaps and adds with the potentiating half of the pre-neuron spike. The resultant pulse leads to potentiation or conductance increase of the PCM device. Similarly if the post-neuron spike arrives before the pre-neuron spike, it will overlap and add with the depressing half of the pre-spike, thus resulting in synaptic depression resistance increase of the PCM device. A similar approach by IBM [84] outlines the use of a select transistor connected across one terminal of the PCM device (1T-1R). In



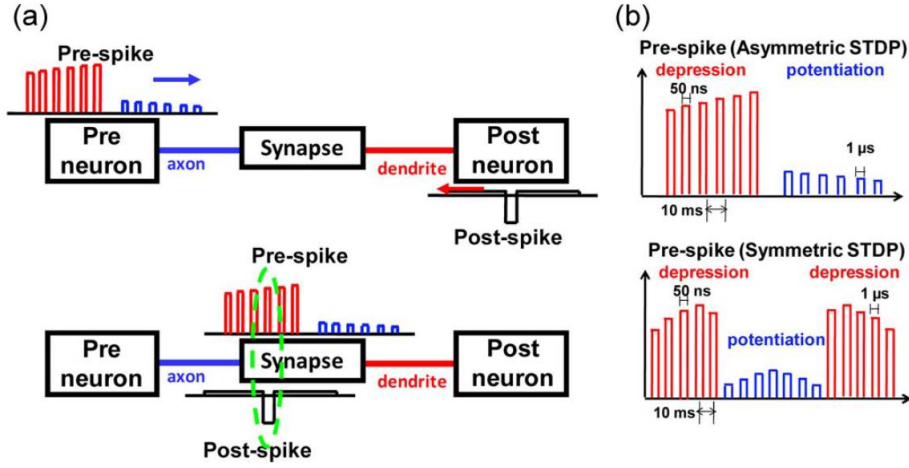
**Figure 2.2:** (a) Cross-section of the GST-PCM device showing amorphous and crystalline regions. (b) Potentiation and depression using voltage pulse trains with changing amplitude, adapted from [82].

this case the post-neuron spike is a combination of depressing and potentiating pulse trains. While the pre-neuron spike is a simple gating pulse for controlling the select transistor.

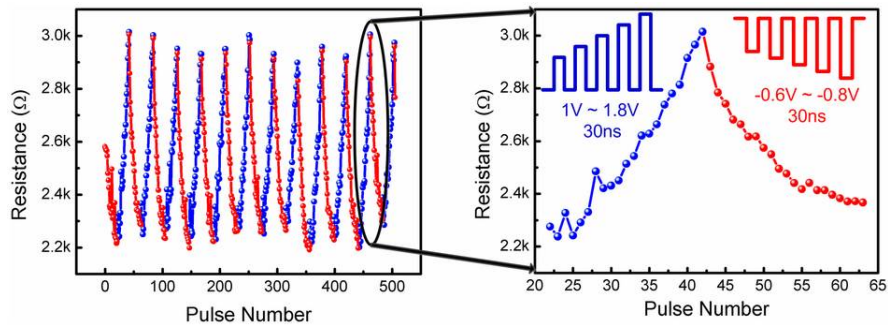
The approaches used to implement STDP, discussed here suffer from the issue of complicated pulse shapes. Implementing pulse trains that are not of identical amplitude would lead to additional circuitry in the CMOS neuron circuits. Such pulse schemes when implemented in large scale neural systems would lead to excessive power dissipation and capacitive line charging across the synaptic crossbars or arrays. Pulse-trains are not energy efficient, as they include application of several pulses even when a particular synaptic event does not occur. In the following sections of this chapter, we present our novel solution (2-PCM Synapse) for overcoming these limitations.

Recently, different forms of STDP emulation over small resistance modulation windows was demonstrated [85] in GST based PCM devices by using bipolar programming pulses (Fig.2.4). The authors claim that mechanism for resistance modulation is different from the one based on joule heating and actual phase-change in GST. They rather explain the resistance changes on basis of charge trapping and releasing in crystalline GST [85]. It is worth mentioning that PCM devices and chalcogenide based systems have also been proposed to emulate certain neuron-like characteristics. Ovshinsky, the father of modern day PCM, was also among the first ones to suggest the possibility of

## 2. PHASE CHANGE MEMORY SYNAPSES



**Figure 2.3:** (a) Spike scheme with set and reset pulse trains. (b) Spikes for different forms of STDP [83].



**Figure 2.4:** PCM conductance increases or decreases in response to negative or positive pulses, respectively. The pulse amplitudes vary with identical 30 ns widths [85].

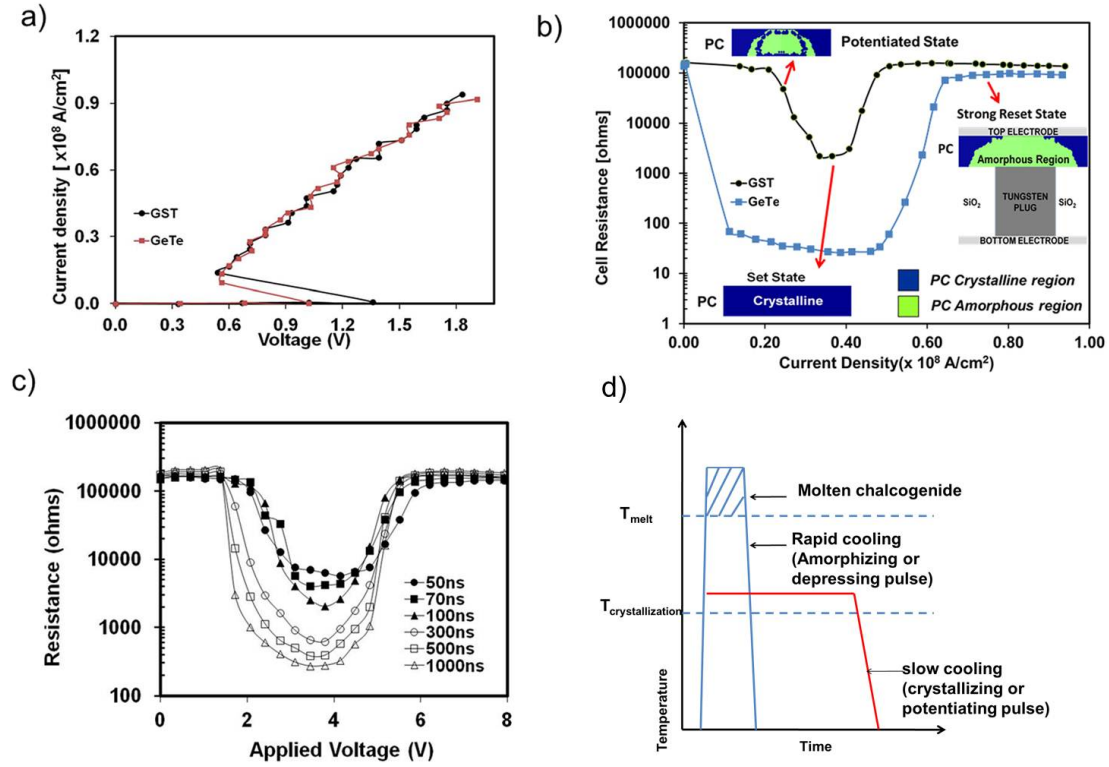
using PCM for cognitive processing applications or ovonic computers [86]. The threshold switching feature of PCM is central to the concept of ovonic cognitive processing. Ovshinsky shows that the gradual reset-to-set transition of PCM devices by application of partially crystallizing pulses can be exploited in a manner analogous to the firing action of LIF and IF neurons [86]. More recently C.D. Wright et.al demonstrated the realization of a phase change processor exploiting the non-volatility and threshold-switching property of PCM and chalcogenide layers [87]. Their processor can perform bio-inspired and simple arithmetic functions like addition, subtraction, multiplication, and division with simultaneous storage of results.

## 2.3 Device and Electrical Characterization

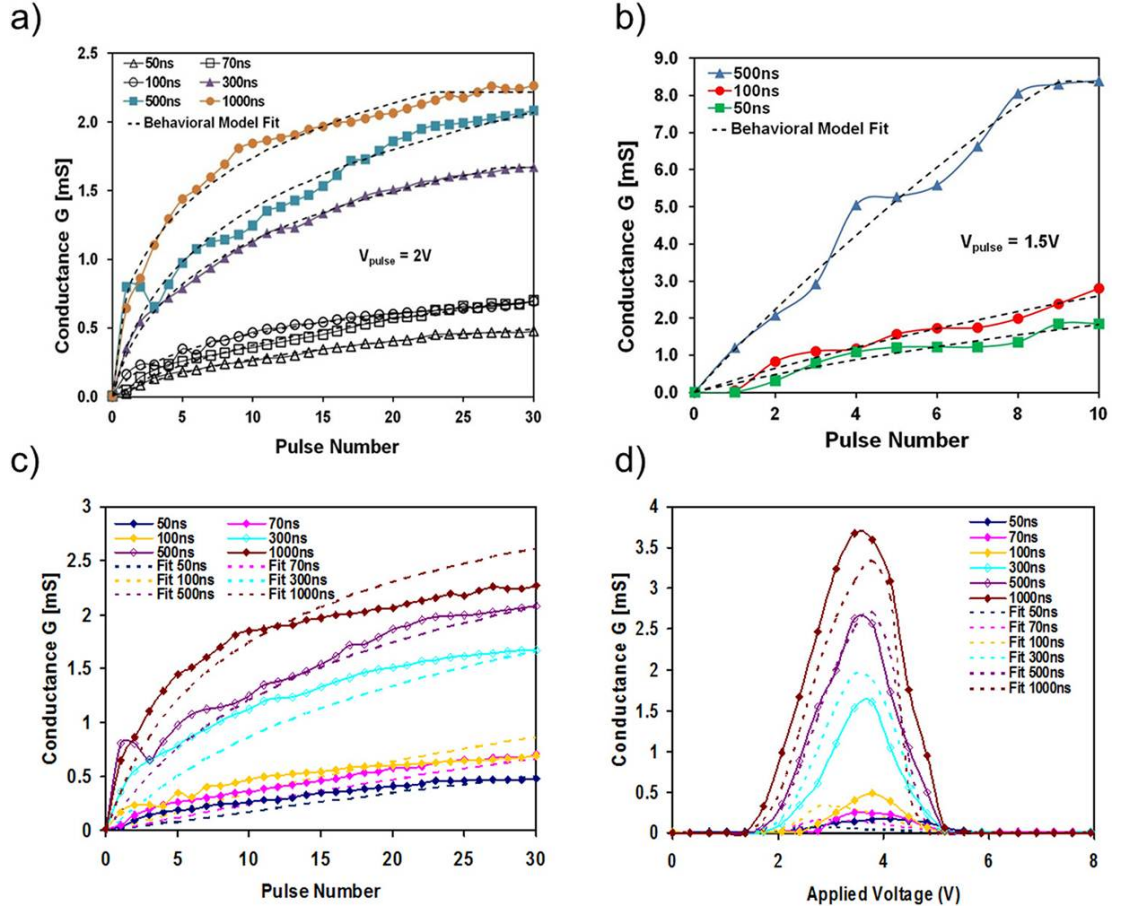
At LETI, we fabricated lance-type PCM test devices (Fig.2.1), with a 100 nm-thick phase change layer and a 300 nm-diameter tungsten plug, were fabricated and characterized. Two different chalcogenide materials were integrated: nucleation dominated  $\text{Ge}_2\text{Sb}_2\text{Te}_5$  (GST) and growth-dominated GeTe [88]. GST and GeTe were chosen to examine how materials with different crystallization parameters would impact the synaptic behavior.

For all the measurements, a current limiting resistance of  $100\ \Omega$  was connected in series with the top pad of the PCM device. The electrical test setup is described in detail in [89], where bottom pad is grounded and the signal is applied to the top pad. Throughout this chapter, an increase in the PCM conductance is referred to as synaptic potentiation (or long term potentiation, LTP [90]), while a decrease in PCM conductance is referred to as synaptic depression (or long term depression, LTD [90]). Fig.2.5a and Fig.2.5b show the measured current-voltage (I-V) and resistance-current (R-I) curves, for GST and GeTe PCM devices. In Fig.2.5a (I-V curve), the device was initially reset to the amorphous state. As the applied voltage is increased, electronic switching occurs at the threshold voltage, and the amorphous region becomes conductive (the so called volatile electronic-switching [91]). Fig.2.5b (R-I curve), illustrates well the difference between a SET state (crystalline), a potentiated state (partially crystalline), and a strong RESET state (large amorphous region). At the beginning of the test, the device was reset to a high resistive amorphous state using a strong

## 2. PHASE CHANGE MEMORY SYNAPSES



**Figure 2.5:** (a) IV characteristics for PCM devices with 100 nm thick GST and GeTe layer starting from initially amorphous phase. (b) R-I characteristics of GST and GeTe PCM devices, with inset showing the PCM phase of intermediate resistance states. (c) R-V curves for GST devices with six different pulse widths. Read pulse = 0.1 V, 1 ms. Legend indicates applied pulse widths. (d) Temperature Vs Time profile for PCM programming pulses [81].



**Figure 2.6:** (a) Experimental LTP characteristics of GST PCM devices. For each curve, first a reset pulse (7 V, 100 ns) is applied followed by 30 consecutive identical potentiating pulses (2 V). Dotted lines correspond to the behavioral model fit described in Eq.2.3 and eq.2.4. (b) Experimental LTP characteristics of GeTe PCM devices. (c) LTP simulations for GST devices using circuit-compatible model. (d) Conductance evolution as a function of the applied voltage for GST devices with six different pulse widths, using circuit-compatible model (sec.2.5.2). Legends in Figs.2.6(a–d) indicate pulse widths [81].



## 2. PHASE CHANGE MEMORY SYNAPSES

---

reset-pulse (7 V, 100 ns, rise/fall time = 10 ns). This was followed by the application of a programming pulse. After the programming pulse, a small reading voltage of 0.1 V was applied to measure the device resistance. Fig.2.5c, shows the characteristic resistance-voltage (R-V) curves for GST based PCM devices, for different programming pulse widths. For a given pulse amplitude, the resistance decreases much faster for longer pulse widths. Thus, by using the combination of right pulse width and right pulse amplitude, PCM resistance (or conductance) can be modulated, as an analog to synaptic weights. For both Fig.2.5b and Fig.2.5c the rise and fall times of the set pulse are always 10 ns. The reset-to-set transition in GeTe is more abrupt compared to GST; GeTe being a growth dominated material crystallizes much faster compared to GST which is nucleation dominated. On the other hand the set-to-reset transition appears to show more gradual resistance change for both GST and GeTe as it is possible to obtain different intermediate resistance values by controlling the volume of the amorphous region created inside the phase change layer post-melting.

### 2.3.1 LTP Experiments

LTP can be emulated if the PCM is progressively programmed along the RESET-to-SET transition (or amorphous-to-crystalline). LTD can be emulated if the PCM is progressively programmed from the SET-to-RESET transition (or crystalline-to-amorphous). In order to emulate spiking neural networks (SNN), it is desired that both LTP and LTD can be achieved by application of simple and identical pulses. Generation of complex types of spikes or pulses would require additional complexity in the neuron circuit design. Two types of pulses can be defined for our purpose: depressing or amorphizing pulse (reset) and a potentiating or partially crystallizing pulse (a weak set). Fig.2.6a and Fig.2.6b show LTP-like conductance variation of PCM devices with GST and GeTe, respectively. Initially, the devices were programmed to a high resistive state by using a strong depressing pulse (7 V, 100 ns). This was followed by the application of several identical potentiating pulses, which are simple rectangular voltage pulses with a rise and fall times of 10 ns (2 V for GST, 1.5 V for GeTe). The voltage amplitude of the potentiating pulses is chosen such that the resulting current flowing through the device is just sufficient to switch the device and cause minimum amount of crystallization with the application of each pulse. Nucleation-dominated behavior leads to a more gradual conductance change in GST, when compared to GeTe (GeTe being

growth dominated). The saturation of the conductance programming window in GeTe occurs in less than a third of the total number of pulses required for GST. From the viewpoint of storage capacity (in the form of synaptic weights) inside a neural network, a GST synapse seems superior to a GeTe synapse, as GST offers a higher number of intermediate conductance states.

### 2.3.2 LTD Experiments

Fig.2.9 shows the emulation of LTD-like behavior on PCM devices. The devices were first crystallized by applying a strong potentiating pulse (2 V, 1  $\mu$ s), followed by the application of several identical depressing pulses (7 V, 50 ns). It was not possible to obtain a gradual synaptic depression by applying identical depressing pulses. The LTD experiment seems more like an abrupt binary process, with negligible intermediate conductance states. Multi-physical simulations (described in Sec.2.4) were performed to interpret the LTD experiment. From these simulations (shown embedded in Fig.2.9), we observed that the volume of the molten region created after the application of each pulse remains almost the same if the pulses are identical. We also performed LTD simulations with consecutive pulses of increasing amplitude (not shown here), and observed a gradual decrease in the device conductance with the application of each pulse, in agreement with other papers [82]. A strong depressing pulse would melt a larger region compared to a weak depressing pulse, creating a larger amorphous cap and a lower value of final device conductance. Thus, in order to obtain gradual synaptic depression behavior, the amplitude of the consecutive pulses should increase progressively. This is also in agreement with the set-to-reset transition seen in Fig.2.5b. Innovative pulse sequences like the ones suggested in [82] could help in achieving LTD with multiple intermediate states. Nevertheless, implementing such complex pulse-schemes with varying amplitudes can lead to practical problems, such as capacitive line charging and high power dissipation, when implemented in large scale neural systems. Moreover, the generation of non-identical pulses would lead to an augmented complexity in the design of the CMOS neuron circuits. Additionally, emulation of LTD or synaptic depression on PCM devices is significantly more energy consuming compared to the emulation of synaptic potentiation, as the former requires amorphization (occurring at a higher temperature compared to crystallization). In the following sections, we introduce

## 2. PHASE CHANGE MEMORY SYNAPSES

---

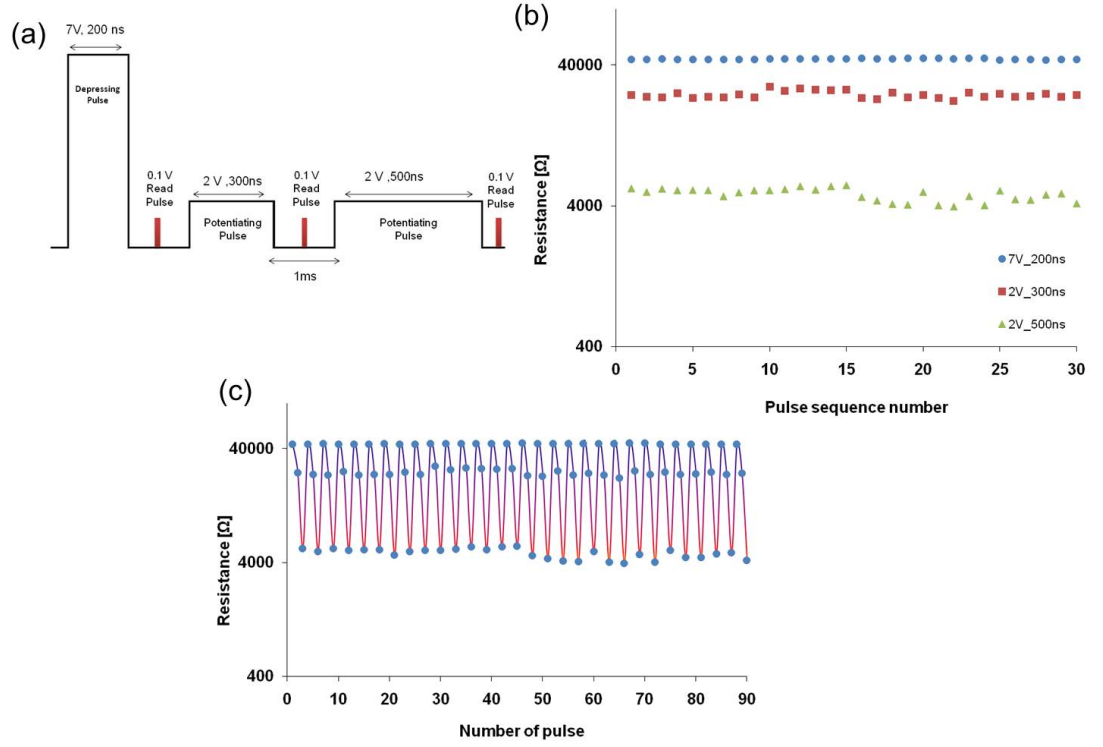
a new solution (The 2-PCM Synapse) to emulate LTD/LTP with PCM devices, thus overcoming the limitations discussed above.

### 2.3.3 Mixed Tests

Within an evolving neural network, a synapse might undergo potentiation or depression (with varying intensities) or a combination of both, depending upon factors such as spiking characteristics of the post-and pre-synaptic neurons. In order to emulate such realistic synaptic modifications, we performed different hypothetical test situations for the PCM synapse. One such test is shown in Fig.2.7. In this test, we show the PCM synapse undergoing three modifications. A single test sequence comprises of a strong potentiating event (as LTP) followed by a moderate and a strong depressing event (as LTD). The strong LTP event is performed by applying a long potentiating pulse of 2 V during 4  $\mu$ s. The moderate and strong depressing events are performed by applying 6 V, during 100 ns and 7 V during 300 ns, respectively. Thus one sequence of the test consists of 3 distinct pulses. After each pulse, a small reading voltage of 0.1 V is applied to measure the device resistance. The entire sequence was repeated 30 times. Fig.2.7b, shows the change in resistance (or synaptic weight) of the PCM with every sequence, and Fig.2.7c displays the variation in resistance with every individual pulse of the 30 sequences. The resistance values obtained were well reproducible. In another similar test (Fig.2.8), we modified the combination of LTP and LTD events. In this test a single sequence comprises two identical strong depressing events followed by a moderate potentiating event. The depressing events were performed by applying a depressing pulse of 5.5 V during 70 ns. The moderate potentiating event was emulated using a pulse of 1.5 V and 1  $\mu$ s.

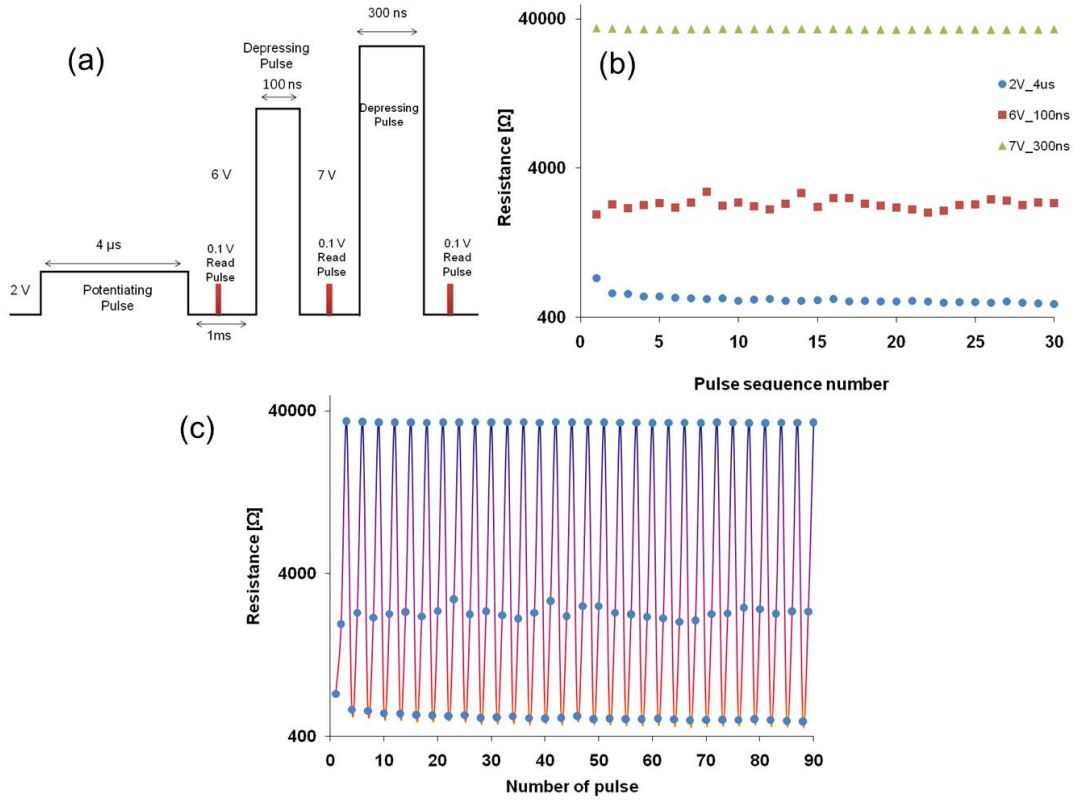
Notice that the difference in PCM resistance values after application of the two LTP pulses used in Fig.2.7 is much more than the ones obtained in Fig.2.8. This is due to the fact that in the later case the depressing pulses are identical. Also, according to the test shown in Fig.2.7, the device attains much lower values of resistance compared to Fig.2.8, on the application of the potentiating pulse. This is also in agreement with the trend shown in Fig.2.5c, as in the case of Fig.2.7, the width of the potentiating pulse is four times the width used in the case of Fig.2.8.

## 2.3 Device and Electrical Characterization

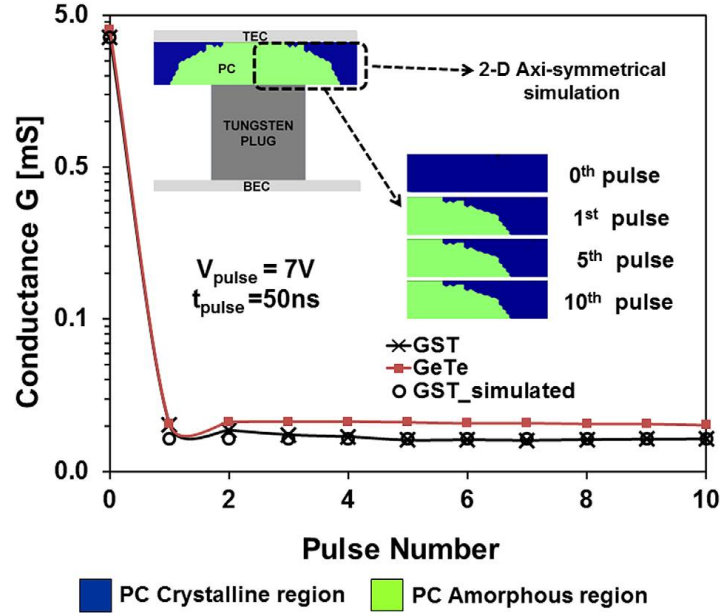


**Figure 2.7:** . (a) Illustration showing a single test sequence applied to the PCM device, consisting of one potentiating and two depressing events of varying intensity. (b) Variation of PCM resistance for all the 30 test sequences. The circle indicate resistance after application of the potentiating pulse, while the square and triangle indicate the reading values after application of 1st and 2nd depressing pulses respectively. (c) Variation of PCM resistance with every individual pulse [92].

## 2. PHASE CHANGE MEMORY SYNAPSES



**Figure 2.8:** (a) Illustration showing a single test sequence, consisting of two identical depressing pulses and one potentiating pulse. (b) Variation of PCM resistance for all the 30 test sequences. The circle indicate resistance after application of the potentiating pulse, while the square and triangle indicate the reading values after application of 1st and 2nd depressing pulses respectively. (c) Variation of PCM resistance with every individual pulse [92].



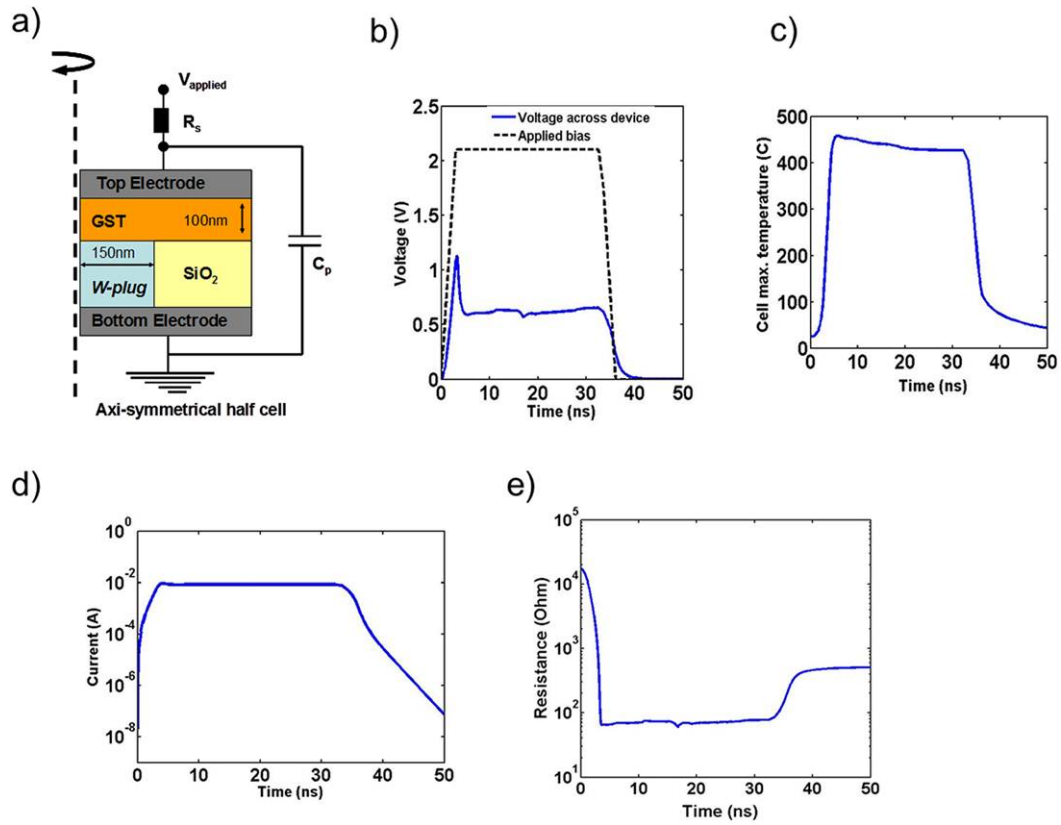
**Figure 2.9:** Experimental LTD characteristics of GST and GeTe PCM devices. Inset shows simulated phase morphology of GST layer after the application of consecutive depressing pulses [81].

## 2.4 Physical Simulation

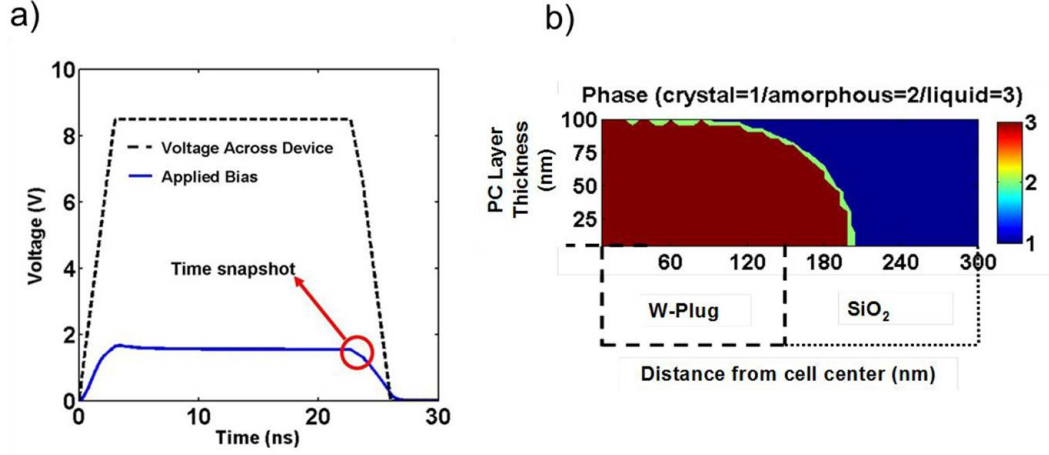
In this section, we present electro-thermal simulations to study the underlying physics of the LTP and LTD experiments described in Sec.2.3. The motivation for such a study is to optimize the synaptic characteristics by engineering the crystallization properties of the phase change materials. Crystallization properties of phase change materials can be altered to some extent by doping [93],[94], modifying the stoichiometric compositions, [95], [96], [97] and interface engineering [98]. A better control over the crystallization parameters by materials engineering is also in the interest of multi-level PCM implementation [99] and enhanced data-retention characteristics [100]. All the simulations were performed for the GST based PCM devices.

The electro-thermal simulator, developed in MATLAB and C++, was described in detail in [101]. The simulations were performed in the 2D axi-symmetrical coordinate reference system. In all the simulations, a series load resistance,  $R_s$ , and a parasitic capacitance  $C_p$  (Fig.2.10a) were considered. Fig.2.10(b-e) show the time evolution of several simulated electro-thermal parameters for a PCM device (initially amorphous),

## 2. PHASE CHANGE MEMORY SYNAPSES



**Figure 2.10:** (a) 2D Axi-symmetrical half cell description used for physical simulations. (b) Simulated time evolution of applied voltage pulse and drop across the device for a potentiating pulse. (c) Simulated maximum temperature in GST layer with the applied pulse. (d) Simulated current passing through the device during the applied pulse. (e) Simulated resistance of the device with the applied pulse [81].



**Figure 2.11:** (a) Simulated depressing (reset) pulse indicating the instance of time snapshot. (b) Time snapshot of the simulated phase morphology of the GST phase change layer [81].

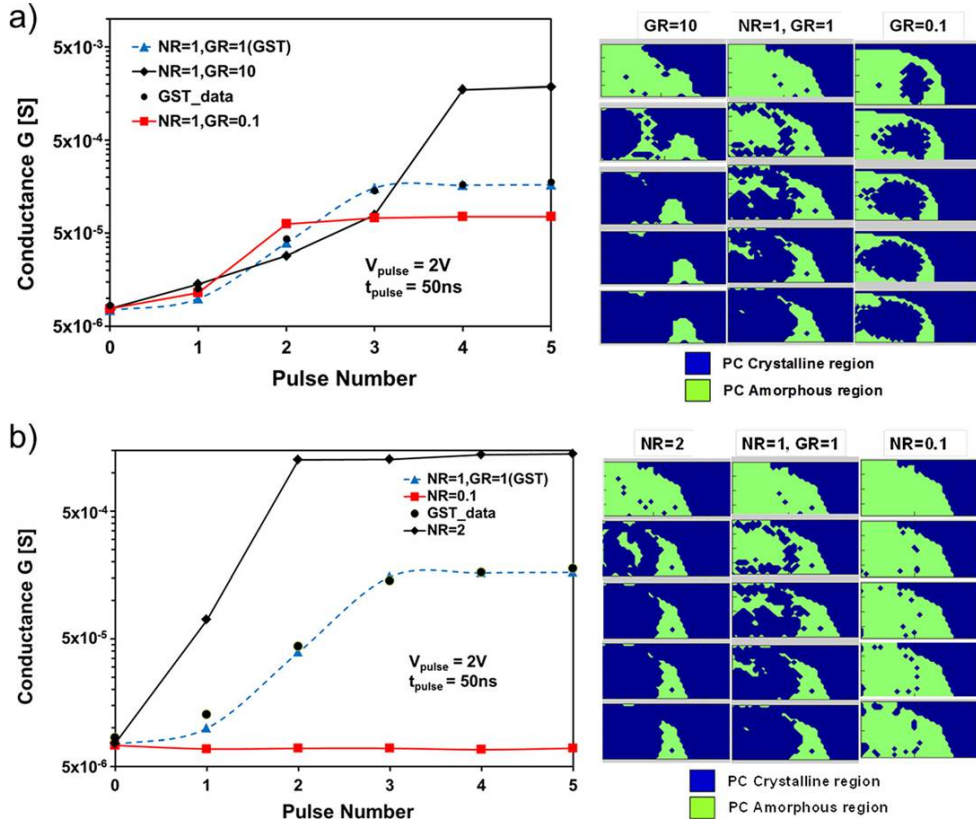
during the application of a potentiating pulse (2.1 V, 30 ns). The evolution of the voltage drop across the device, the current, the resistance, and the maximum cell temperature is shown. Fig.2.11 shows a time snapshot of the GST phases during the application of a depressing pulse (8 V, 25 ns), starting from an initially crystalline state. The selected time is the beginning of the quenching process (i.e., the falling edge of the reset pulse, Fig.2.11a). Before this instant, a mushroom shaped melted mass of GST (brown color) in the region right above the W-plug is seen. As the quenching progresses, the GST amorphizes, moving inwards from the melted crystalline interface towards the center of the mushroom.

Formation of thin amorphous GST (green color) can be seen at melted-crystalline interface in Fig.2.11b. The first few points in the LTP curves (for GST) shown in (Fig.2.6a) are crucial in determining the total number of intermediate conductance states that can be obtained for a given programming-window. The maximum change in conductance was indeed observed to occur during the application of the first 5 pulses for GST. In order to better understand the variation of conductance during the application of first few pulses in the LTP experiment, we performed the simulations shown in Fig.2.12.

The nucleation rate ( $I$ ) and growth velocity ( $V$ ) in the phase change layer were modeled using eq.2.1 and eq.2.2, respectively, adapted from [101]



## 2. PHASE CHANGE MEMORY SYNAPSES



**Figure 2.12:** (a) Simulated LTP curves while fixing the nucleation rate (NR) and varying the growth rate GR compared to GST (taken as reference:  $GR = 1$ ,  $NR = 1$ ). Corresponding simulations of GST layer morphology are shown (0th pulse: reset; 1st-5th: potentiating). (b) Simulated LTP curves while fixing the growth rate ( $GR = 1$ ) and varying the nucleation rate (NR) compared to GST (taken as reference material:  $NR = 1$ ,  $GR = 1$ ). Corresponding simulation of GST layer morphology are also shown [81].

$$I = N_a \cdot O_n \cdot \gamma \cdot Z \cdot \exp\left(-\frac{\Delta G^*}{kT}\right) \quad (2.1)$$

$$V = \gamma \cdot d \cdot \left(1 - \exp\left(\frac{-\Delta G_v}{RT} \cdot \frac{M}{\rho}\right)\right) \quad (2.2)$$

where  $N_a$  is the number of nucleation sites,  $\gamma$  atomic vibration frequency,  $\Delta G$  the free energy,  $Z$  Zeldovitch parameter,  $O_n$  number of atoms at critical nucleus surface,  $M$  the molar mass,  $d$  inter-atomic distance,  $q$  volumic mass, and  $\Delta G_v$  the difference in Gibbs free energy between the amorphous and the crystalline phases. The calculation of each parameter is detailed in [101].

For the simulations shown in Fig.2.12, the fitting of the GST-LTP experimental data (corresponding to the 50 ns pulse width, Fig.2.6a) was defined as the reference nucleation rate ( $NR = 1$ ) and the reference growth velocity ( $GR = 1$ ). LTP simulations (Fig.2.12a) with artificially increased ( $GR = 10$ ) and artificially reduced ( $GR = 0.1$ ) growth velocities with respect to GST ( $GR = 1$ ) were performed, keeping the nucleation rate constant. Similarly, LTP simulations with artificially increased ( $NR = 2$ ) and artificially reduced ( $NR = 0.1$ ) nucleation rates were also performed (Fig.2.12b). The artificial boost or decrease in  $NR$  and  $GR$  was performed by directly multiplying eq.2.1 and eq.2.2 with a constant value. Three major observations were made. First, the maximum value of conductance was reached in fewer pulses if either the growth or the nucleation rate were enhanced. Second, the shape of the bulk amorphous region created after application of the initial reset pulse had a strong dependence upon the values of the growth and the nucleation rates. It is not straightforward to decouple the effect of the nucleation and of the growth parameters as the shape of the amorphous region or morphology changes after the application of each potentiating pulse [102]. A high growth rate ( $GR = 10$ ) leads to a strong crystal growth from the amorphous-crystalline interface during the falling edge of the reset pulse, thus distorting the mushroom-like shape of the amorphous region. A low growth rate ( $GR = 0.1$ ) leads to a more uniform mushroom shape of the amorphous region. Finally, after the application of the first potentiating pulse, conductance was more sensitive to changes in the nucleation rate compared to growth. Fig.2.12 also shows the strong impact of nucleation rate and growth velocity on the morphological evolution of the crystalline phase inside phase change layer.

### 2.5 Modeling

#### 2.5.1 Behavioral model for system level simulations

In order to model the millions of PCM synapses in large scale neural networks, and thus to evaluate the potential of PCM synaptic technology, a computationally efficient model of its behavior is particularly desirable. For this purpose, we introduced the following phenomenological equation (eq.2.3) to model the LTP characteristics of the GST and GeTe devices during a LTP pulse:

$$\frac{dG}{dt} = \alpha \cdot \exp\left(-\beta \cdot \frac{G - G_{\min}}{G_{\max} - G_{\min}}\right) \quad (2.3)$$

where  $G$  is the device conductance,  $\alpha$  and  $\beta$  are fitting parameters.  $G_{\min}$  and  $G_{\max}$  are the minimum and maximum values of device conductance, respectively. This equation was originally introduced to model memristive devices [103]. To model the conductance change  $\Delta G$  after the application of a short LTP pulse of duration  $\Delta t$ , eq.2.3 may be integrated as in eq.2.4

$$\Delta G = \alpha \cdot \Delta t \cdot \exp\left(-\beta \cdot \frac{G - G_{\min}}{G_{\max} - G_{\min}}\right) \quad (2.4)$$

These equations gave a very satisfactory fit of our measurements for both GST and GeTe devices, as shown in Fig.2.6a and Fig.2.6b. This is valuable because the shape of the potentiation curve should have a serious impact on the system performance as suggested by the works in computational neuroscience [104]. For GST and GeTe, we used unique sets of parameters ( $\alpha$ ,  $\beta$ ,  $G_{\min}$ , and  $G_{\max}$ ) for different pulse widths. Tab.2.1 lists the values of the parameters used for the fitting two pulse widths for GST and GeTe shown in Fig.2.6a and Fig.2.6b, respectively.

#### 2.5.2 Circuit-compatible model

To design hybrid neural circuits consisting of CMOS neurons and PCM synapses, a circuit-compatible model for the PCM is required. We thus developed a circuit compatible PCM model, specifically tailored to capture the progressive character of the LTP experiments shown in the previous sections. This simple circuit-compatible model, inspired by [105], was developed using the VHDL-AMS language and includes both LTP and LTD. The simulations were performed with the Cadence AMS simulator. Fig.2.6c

**Table 2.1:** Fitting parameters of the behavioral model for 300 ns GST LTP curve and 100 ns GeTe LTP curve shown in Fig.2.6a and Fig.2.6b respectively.

Parameters	GST (300 ns)	GeTe (100 ns)
Gmin ( $\mu$ S)	8.50	8.33
Gmax (mS)	2.3	2.9
$\alpha$ (S/s)	1100	3300
$\beta$	-3.8	-0.55

shows the simulated LTP curves for six different pulse widths for GST. Tab.2.1 lists all the constants and fitting parameters used in the circuit-compatible model. The model consists of three parts: electrical, thermal, and phase-change. For the electrical part, an Ohmic relationship between current and voltage is assumed (eq.2.5):

$$V = i \cdot R_{gst} \quad (2.5)$$

We preferred not to include the threshold switching effect in this model, as its primary purpose is to emulate LTP behavior during learning in large-scale neural networks simulations, where simulation efficiency is essential.  $R_{gst}$  is the low field resistance of the device, which consists of the sum of the GST layer resistance and the bottom and top electrodes resistance  $R_s$ . The resistance of the phase change layer (eq.2.6) is a function of the amorphous volume fraction  $C_a$ :

$$R_{gst} = R_s + R_{0c}^{1-C_a} \cdot R_{0a}^{C_a} \quad (2.6)$$

$R_{0c}$  and  $R_{0a}$  correspond to the resistances of the fully crystallized and fully amorphized states, respectively. We used a logarithmic interpolation, which is intermediate between the series and parallel cases [105], as this led to the best fitting for our GST devices. In order to evaluate the impact of resistance drift [106] on the stability of the learning, we included a behavioral modeling of this phenomenon, which can be optionally enabled or disabled in the simulations. To do so,  $R_{0a}$  is replaced with  $R_a$  as in eq.2.7:

$$R_a = R_{0a} \cdot \left( \frac{t}{t_0} \right)^{C_{a.dr}} \quad (2.7)$$

## 2. PHASE CHANGE MEMORY SYNAPSES

---

**Table 2.2:** Parameters used for the GST compact model simulations shown in Fig.2.6c.

Parameter	value	Description
Electrical Model		
$R_s$	100 $\Omega$	Serial resistance (top and bottom electrodes)
$R_{0a}$	159 k $\Omega$	Resistance of the fully amorphized state
$R_{0c}$	135 k $\Omega$	Resistance of the fully crystallized state
Thermal Model		
$R_{ta0}$	15.9 x 10 <sup>3</sup> W/K	Thermal resistance of the fully crystallized state
$R_{tc0}$	5.07 x 10 <sup>3</sup> W/K	Thermal resistance of the fully amorphized state
$T_0$	300 K	Ambient temperature
Phase-Change Model		
$E_a$	0.335	Fitting parameter for crystallization rate at high temperature
$E_b$	5.77 x 10 <sup>3</sup>	Fitting parameter for crystallization rate at low temperature
$T_g$	380 K	Lowest temperature at which crystallization can occur
$T_m$	698 K	Melting temperature of the phase change material
$\tau_a$	8.17 x 10 <sup>-13</sup> s <sup>-1</sup>	Amorphization rate (fitting)
$\tau_c$	1.10 x 10 <sup>-6</sup> s <sup>-1</sup>	Crystallization rate (fitting)
dr	0.03	Drift coefficient

where  $t_0$  is the time, at which the latest phase change occurred (i.e., when  $T_b$  last crossed  $T_g$ ).

In the thermal part of the model, the electrical power  $P_t$  and the temperature of the phase-change material  $T_b$  are connected by the eq.2.8, with  $T_0$  the ambient temperature:

$$T_b = T_0 + P_t \cdot R_{tgst} \quad (2.8)$$

The thermal resistance of the phase change layer  $R_{tgst}$  is described by the following eq.2.9:

$$R_{tgst} = (1 - C_a) \cdot R_{tco} + C_a \cdot R_{tao} \quad (2.9)$$

where  $R_{tco}$  and  $R_{tao}$  are the thermal resistances for the completely crystallized and completely amorphized states. Due to the threshold switching effect, the electrical power during phase change is essentially independent on the amorphous ratio. The electrical power is therefore calculated using the fully crystallized phase-change resistance (eq.2.10), for which no threshold switching occurs, instead of the low field resistance

$$P_t = \frac{V^2}{R_s + R_{0c}} \quad (2.10)$$

The phase-change part of the model uses behavioral equations. Amorphization occurs when  $T_b$  is higher than  $T_m$ , the melting temperature. The amorphization rate is assumed to increase linearly with the temperature, and is zero when  $T_b$  is equal to  $T_m$ , thus ensuring continuity with the crystallization rate at this temperature. This leads to the eq.2.11:

$$\frac{dC_a}{dt} = \frac{1}{\tau_a} \cdot \frac{T_b - T_m}{T_m} \quad (2.11)$$

when  $T_b > T_m$

The equation modeling the crystallization rate (eq.2.12) does not attempt to model nucleation-driven and growth-driven rate separately. The expression we used is, however, reminiscent to growth rate modeling. It includes a term  $C_a^2$ , because crystallization rate typically depends on the amorphous crystal interface surface for growth-driven process (and volume for nucleation-driven process).

$$\frac{dC_a}{dt} = \frac{C_a^2}{\tau_c} \cdot (1 - \exp(E_a \cdot \frac{T_b - T_m}{T_b})) \cdot \exp(-\frac{E_b}{T_b}) \quad (2.12)$$

## 2. PHASE CHANGE MEMORY SYNAPSES

---

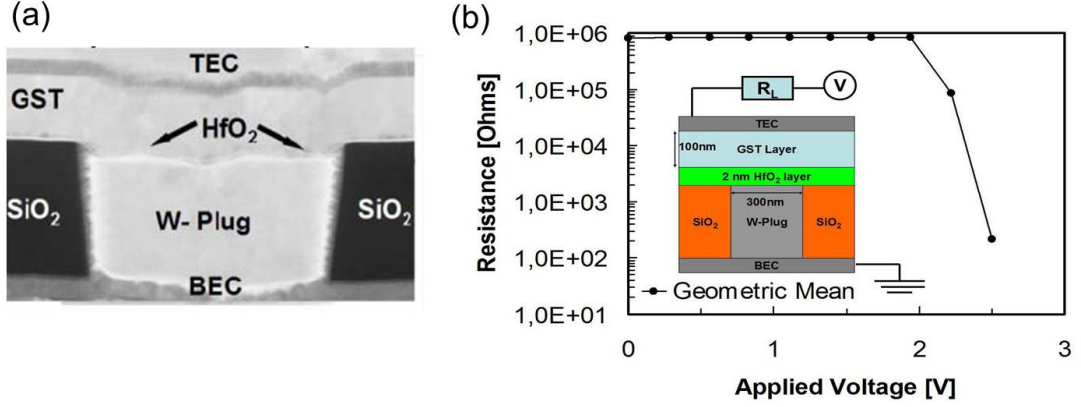
when  $T_b < T_m$  and  $T_b > T_g$

This model, with parameters listed in Tab.2.1, gives good fitting results for the LTP curves (Fig.2.6c), although the fit is not as excellent as the behavioral model (Fig.2.6a). Simulation of the conductance evolution as a function of the applied voltage with the same parameters is shown in Fig.2.6d. Although the fitting is less accurate for shorter pulses, all the curves in Fig.2.6 were fitted with a single set of parameters. The model captures the correct behavior of the PCM for a relatively wide range of measurements with a small number of semi-physical parameters. It is therefore adapted for fast exploration and easier circuit design where PCM devices are employed to emulate millions of synapses.

### 2.6 PCM Interface Engineering

It is important to note that as a neural network undergoes learning, the synapses (or PCM devices) may undergo continuous programming (i.e. frequent LTP/LTD events) and get saturated. To overcome the issue of synaptic weight saturation, we defined a refresh-sequence, described in detail in Sec.2.7. We show in Sec.2.8, that in order to enable continuous synaptic learning and to reduce the system power dissipation with PCM synapses, it is desirable to increase the number of intermediate resistance states in the LTP-curves of the PCM devices (Fig.2.6a,b). In other words, if the PCM device takes much longer to attain its minimum resistance value or if it crystallizes slowly, it leads to improved power/energy efficiency of the neural network. In this section we show that by addition of a thin  $\text{HfO}_2$  interfacial layer to standard GST-PCM devices: (i) the number of intermediate resistance states (or synaptic weights) in the LTP-curves can be increased, and (ii) the set/reset current for individual programming events can be significantly decreased, thus improving the performance both at the level of the full neuromorphic system and also the individual PCM-synapses.

Lance-type PCM devices with a 300 nm diameter tungsten (W) heater-plug and 100 nm-thick GST layer were fabricated. A  $\text{HfO}_2$  layer of 2 nm thickness, was deposited between the heater plug and the GST layer by atomic layer deposition (ALD) (Fig.2.13a). Due to the insulating  $\text{HfO}_2$  layer, the PCM device is initially in a high resistive state and cannot be programmed. To enable programming a forming step to

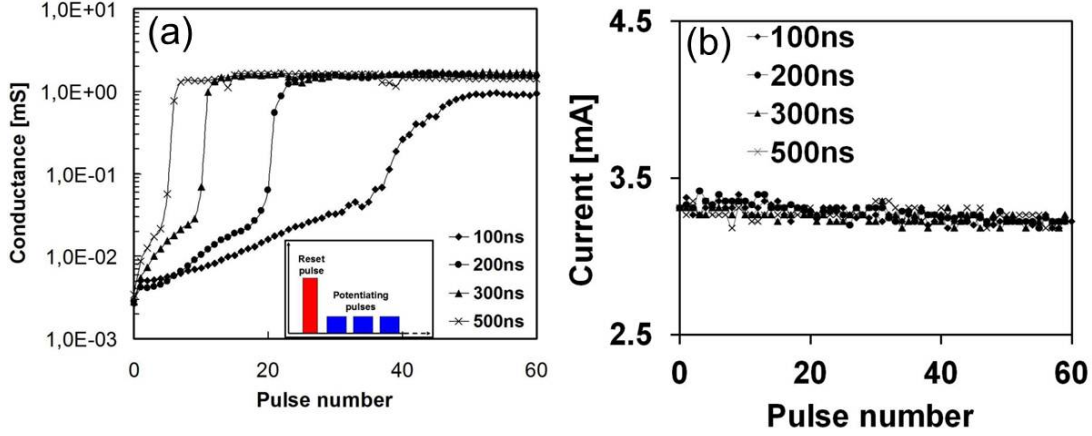


**Figure 2.13:** (a) TEM of the GST PCM device with  $\text{HfO}_2$  interface layer. (b) Resistance versus voltage applied curve, for the interface layer (2nm thick  $\text{HfO}_2$ ) PCM device, during the initial forming step[98].

breakdown the  $\text{HfO}_2$  layer is required [107]. In this step, shown in Fig.2.13b, a staircase sequence of increasing voltage pulses is applied to the device (0-3 V). Initially the device resistance is very high (about 1 M $\Omega$ ), the breakdown occurs when the device resistance jumps to a lower value (few k $\Omega$ ). The LTP-curve for the interface PCM devices is shown in Fig.2.14a. Initially the device is amorphized to a low conductive state by applying a reset pulse (4 V, 100 ns). This is followed by the application of several potentiating pulses (partially crystallizing) of 2.1 V and the same pulse width. The conductance curves shown in Fig.2.14a are obtained by applying trains of pulses with different pulse widths. The conductance gradually increases with the application of each pulse, showing a strong influence of the pulse-width. The conductance window of the devices with the interface layer saturates with the application of more than double the number of potentiating pulses compared to GST devices without interface layer (Fig.2.6a). Several material related physical or chemical factors may explain the possible cause for more number of intermediate points in the LTP curves for the interface devices. For instance, interface layers can impact the crystallization kinetics affecting activation energies associated with growth and nucleation sites [108], [109]. Diffusion from the interfacial oxide inside the phase change layer can also lead to a change in the properties of the reset-to-set transition. Pinpointing the exact cause for the increase in the number of LTP points would require a more detailed material analysis of the devices. Fig.2.14b shows that for each applied pulse, the measured currents in the



## 2. PHASE CHANGE MEMORY SYNAPSES

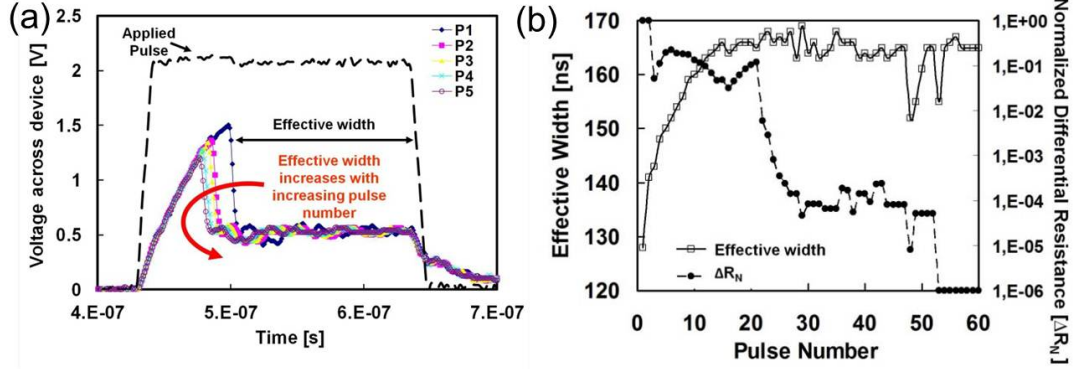


**Figure 2.14:** (a) LTP curves for the GST devices with 2 nm thick HfO<sub>2</sub> layer. Inset shows the test procedure (Reset pulse: 4V/100ns and potentiating pulse: 2.1 V. (b) Current values for each pulse of the LTP test shown in (a). [98].

LTP experiment remains constant. The programming current values were decreased by more than 50% compared to standard GST devices. The set/reset current decreases as the effective contact area between the W-plug and the GST layer scales and better thermal insulation is achieved [107]. Although the actual W-plug diameter is 300 nm, the forming step leads to formation of small nanoscale conductive filament(s) inside the insulating HfO<sub>2</sub> layer. Thus the effective contact area for the flow of current to the GST layer is limited to the dimensions of the nanoscale filaments and not the entire W-plug [107].

Fig.2.15a, shows the experimentally acquired waveforms for the first five potentiating pulses applied in the LTP experiment for the 200 ns pulse width. Effective pulse-width can be defined as the difference between the actual pulse-width applied, and the time spent before the occurrence of the electronic switching-event in the amorphous region. Notice that with the application of each pulse, the effective pulse-width keeps increasing. Effective pulse width is important because crystallization takes place only in the duration after the electronic switch has occurred. Before the electronic switch, almost negligible current passes through the cell and the temperature of the phase change layer stays below the crystallization temperature.

Fig.2.15b, shows the plot of effective pulse width and the normalized differential resistance ( $\Delta R_N$ ) for each point in the LTP curve for the 200 ns pulse width.  $\Delta R$  is



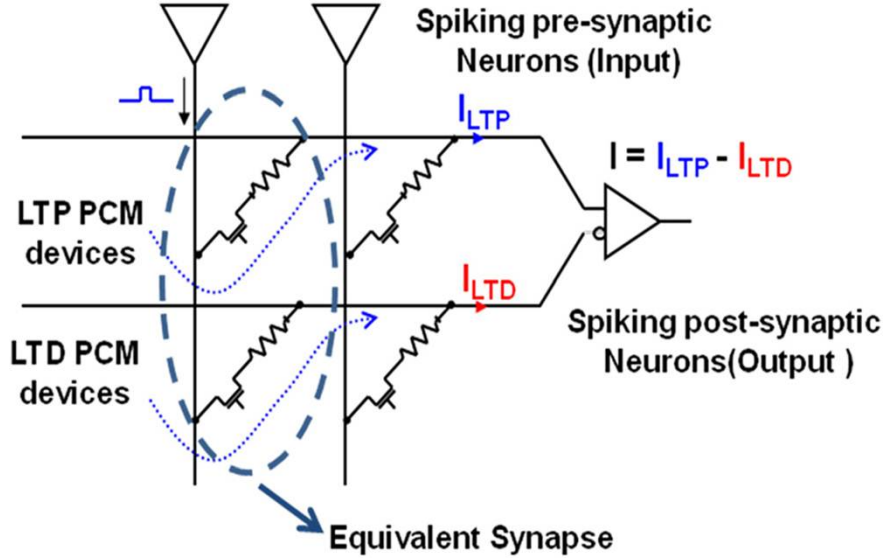
**Figure 2.15:** (a) Experimentally acquired waveforms for the applied voltage pulse and the actual voltage drop across the PCM devices. The waveforms were acquired for the first 5 pulses of the 200 ns LTP curve shown in Fig.2.14a. (b) Graph showing a plot of effective pulse width and the normalized differential resistance (for 200 ns LTP curve shown in Fig.2.14). The resistance values were normalized with respect to maximum value of  $\Delta R$ , [98].

defined as  $R_i - R_{i+1}$ , for the  $i$ th pulse in the LTP curve. The normalization is done with respect to the maximum value of  $\Delta R$ . As the LTP experiment progresses the value of  $\Delta R_N$  decreases, this trend was observed for all the applied pulse widths. The decrease in resistance between two pulses depends upon the effective pulse-width, the pulse amplitude, and the instantaneous morphology of the crystalline and amorphous regions inside the phase change layer. With the application of successive potentiating pulses, the amorphous region inside the phase change layer decreases, new conductive percolation paths are formed and the existing ones are strengthened. Thus there is a steady decrease in the value of  $\Delta R$ .

## 2.7 The "2-PCM Synapse"

To emulate synaptic behavior (i.e. gradual synaptic potentiation, depression and STDP-like rules), using PCM devices as synapses, we developed a novel low-power methodology called the "2-PCM Synapse" (Fig2.16) [110]. In this approach, we use two PCM devices to implement a single synapse and connect them in a complementary configuration to the post-synaptic output neuron. Both PCM devices are initialized to a high resistive amorphous state. When the synapse needs to be potentiated, the so-called LTP device undergoes a partial crystallization, increasing the equivalent weight of the

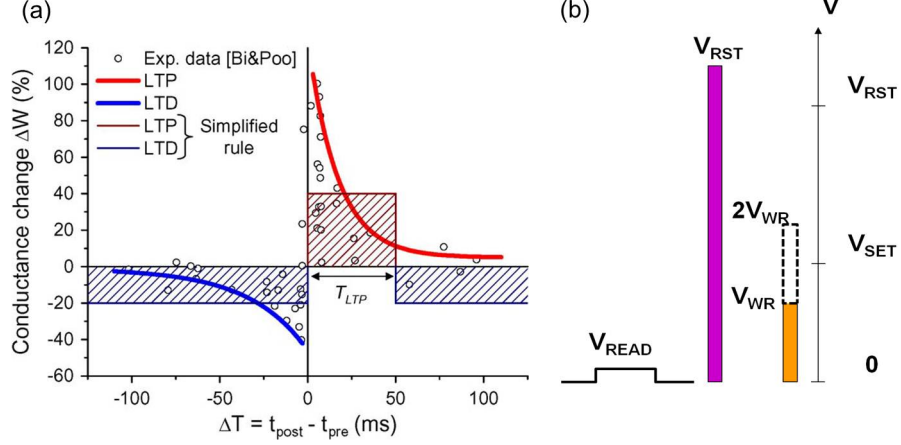
## 2. PHASE CHANGE MEMORY SYNAPSES



**Figure 2.16:** The "2-PCM Synapse" concept schematic. The contribution of the current flowing through the LTP device is positive, while that of the LTD device is negative, towards the integration in the output neuron [110].

synapse. Similarly, when the synapse must be depressed, the LTD device is crystallized. As the LTD device has a negative contribution to the neurons integration, the equivalent weight of the synapse is reduced.

Important benefits of the "2-PCM Synapse" approach are the following: Firstly, exploiting mostly the crystallization phenomenon of chalcogenide devices, it allows defining a programming methodology which uses identical neuron spikes (or pulses) to obtain both gradual LTP and LTD. In sec.2.3.2, we showed that gradual LTD-like effect is not achievable with application of identical programming pulses for PCM devices. To this purpose, the programming schemes described in [82] utilize variable pulses with changing amplitudes. In fact, generation of such non-identical pulses lead to the increased complexity of pre/post CMOS neuron circuits, added parasitic effects, such as capacitive line charging, and excessive power dissipation in the neural network. These limitations can be overcome by the crystallization dominated "2-PCM Synapse" methodology. Secondly, the "2-PCM Synapse" is a low-power approach, because majority of the synaptic events are achieved through crystallization, which is a less energy consuming process for PCM compared to amorphization. The current required for amorphization is typically 510 times higher than for crystallization, even for state-of-



**Figure 2.17:** (a) Biological STDP (from [37]) and simplified STDP used in the proposed PCM implementation. In the simplified rule, a synapse receiving a postsynaptic spike with no presynaptic spike in the LTP window undergoes an LTD regardless of the existence of a presynaptic spike [111]. (b) Write, Reset and Read pulses for the programming scheme proposed in sec.2.7.2, [110].

the art PCM devices (see Fig.2.28). Another inherent advantage of our approach is the decrease of the impact of PCM resistance-drift on the stored synaptic information discussed in detail sec.2.9.

### 2.7.1 Simplified STDP-rule

Reproducing the complex  $\delta T$  dependent biological STDP learning rule (Fig.1.9a) with PCM is not straightforward. The programming pulses need to be tuned both in duration and in amplitude depending on the exact timing difference between the pre- and post- synaptic spike events [82]. The benefit of this biomimetic approach (where biology is matched as closely as possible) is that it does not require making assumptions on how the synapses will be exploited for learning in the final system. Indeed, if the biological low-level synaptic update rule is replicated with reasonable accuracy in the electronic system, there is a significant chance that any higher level learning or computation occurring in biological neural networks will be reproducible in artificial hardware as well. There are, however, benefits in designing specific STDP rules targeted towards specific applications. The exact shape of the STDP learning rule may not be required to capture the correct computational behavior. To go further, the measured STDP curve in vitro might not reflect the actual behavior of the neurons in

## 2. PHASE CHANGE MEMORY SYNAPSES

---

vivo [112]. Finally, there is not just one STDP rule: a broad family of synaptic update characteristics in function of the prepost synaptic time difference were recorded [113]. Here, we utilize a novel simplified and optimized STDP learning rule that is easy and efficient to implement with PCM devices. In this STDP rule, all the synapses of a neuron are equally depressed upon receiving a postsynaptic spike, except for the synapses that were activated with a presynaptic spike a short time before, which are strongly potentiated [114]. Contrary to a biological synapse, the magnitude of the potentiation or depression is independent on the relative timing between the presynaptic spike and the postsynaptic spike, as shown in Fig.2.17a.

### 2.7.2 Programming Scheme

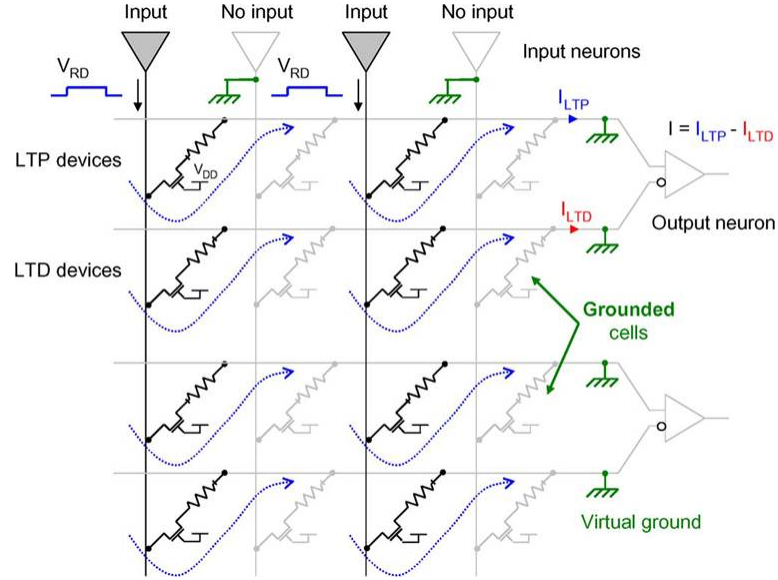
In this section we present the detailed programming scheme for implementing the simplified STDP rule on the "2-PCM Synapse" architecture. We discuss the cases of PCM devices with selector (1T-1R) and without selector (1R). The pulse shapes for the Read, Write and Reset pulses is indicated in Fig.2.17b.

#### 2.7.2.1 Read

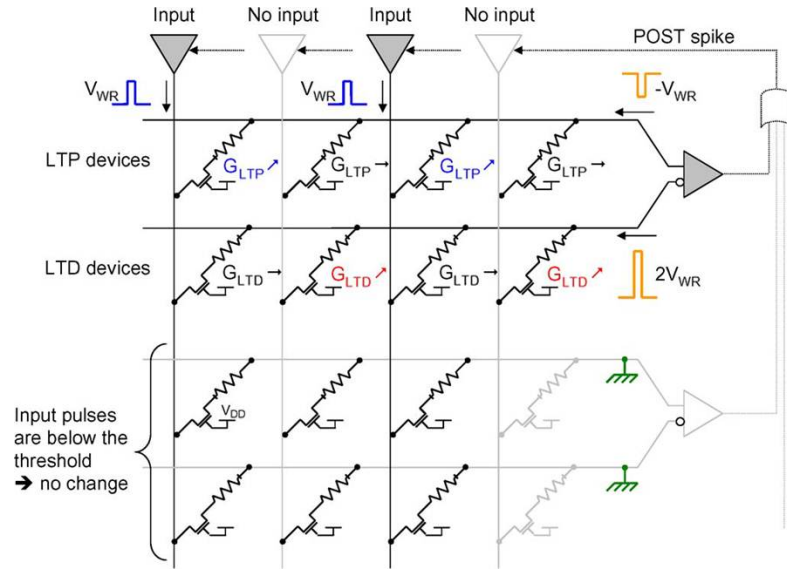
The read operation described here (Fig.2.18), is the normal operation of the network between two output neuron activations. When an input neuron receives an incoming event, it generates a small voltage pulse that is propagated to all its output neurons through its synapses. The resulting current flowing to each output neuron is the difference between the current from LTP and the LTD devices. The read pulse amplitude and duration can be minimal, as long as it allows reasonably accurate reading of the low-field resistance of the PCM. The output neurons are of type LIF [50]. When the integrated current reaches the neuron's threshold, the network enters a write mode operation to update the synaptic weights through the simplified STDP rule. Each time an input neuron is activated, it enters or re-enters a LTP internal state for the duration of the STDP LTP window ( $T_{LTP}$ ), as shown in Fig.2.17a.

#### 2.7.2.2 Write

When an output neuron fires, it transmits a post-spike signal to every input neuron, signaling write operations. In write operations, input neurons generate a LTP pulse of



**Figure 2.18:** Read Operations. Current from both LTP and LTD PCM devices is integrated in the output neuron, with a positive and negative contribution, respectively [111].



**Figure 2.19:** Write operations based on the simplified-STDP rule. For a specific PCM,  $G \nearrow$  denotes an increase in conductance (thus, partial crystallization of the device), while  $G \rightarrow$  denotes no change in conductance [111].

## 2. PHASE CHANGE MEMORY SYNAPSES

---

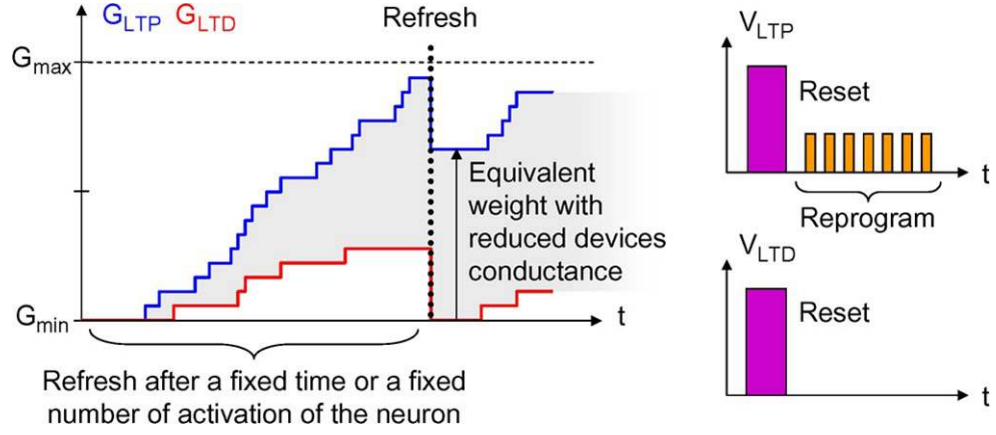
amplitude  $V_{WR}$  such that  $V_{WR} < V_{SET} < 2.V_{WR}$ , only if they are in the LTP state. The output firing neuron generates a negative feedback pulse  $-V_{WR}$  for the LTP devices and a positive feedback pulse  $2.V_{WR}$  for the LTD devices. When a LTP pulse interacts with the feedback pulses, the effective voltage across the LTP device is  $2.V_{WR} > V_{SET}$  and the voltage across the LTD device is  $V_{WR} < V_{SET}$ . The conductance of the LTP device is then increased. If there is no LTP pulse for a given input, it means that pre-post spike timing difference is not within the LTP window and thus the conductance of the LTD device must be increased according to our simplified STDP rule. This is indeed this case, as the voltage across the LTP device is  $-V_{WR} > -V_{SET}$  and the voltage across the LTD device is  $2.V_{WR} > V_{SET}$  (Fig.2.19).

Selector devices are not required for the write operations, as the input LTP pulse amplitude is below the SET threshold of the PCM devices, so the synaptic weights of the other output neurons is not affected. The LTP pulses may however significantly alter the integration value of other output neurons. This is a non-issue in the proposed architecture as lateral inhibition is implemented: when a neuron fires, integration of the others is disabled for an inhibit refractory period  $T_{inhibit}$  [114].

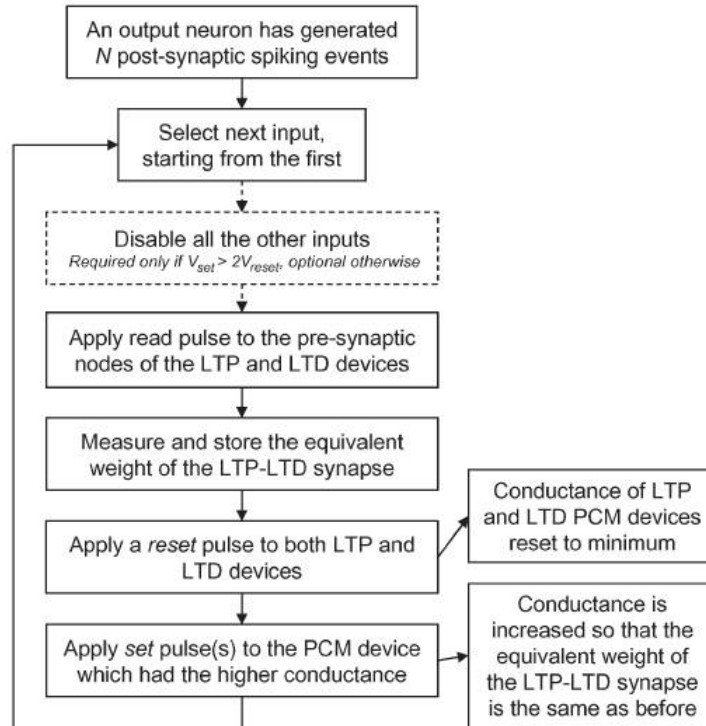
### 2.7.2.3 Refresh

Because the conductance of the PCM devices gradually increases during the learning, a refresh mechanism is introduced to reduce the conductance of LTP and LTD devices while keeping the weight of the equivalent synapse unchanged. The principle of the refresh operation is shown in figure 2.20. When one of the two devices reaches its maximum conductance, they are both reset and the one that had the higher conductance undergoes a series of SET pulses until the equivalent weight is reached again. Because one of the devices stays at minimum conductance, this mechanism enables continued evolution of the weights.

Knowing the average number of conductance steps  $N$  achievable for a given PCM technology, with a given SET pulse duration and amplitude, a refresh operation is necessary after  $N$  potentiations or  $N$  depressions of the synapse (whichever comes first, assuming that one of the devices is initially at minimum conductance). Therefore, output neurons can initiate a refresh operation on their synapses after a fixed number of activations, which would be  $N$  in the worst case. Although such a simple mechanism would certainly involve a substantial amount of unnecessary resets, as few synapses



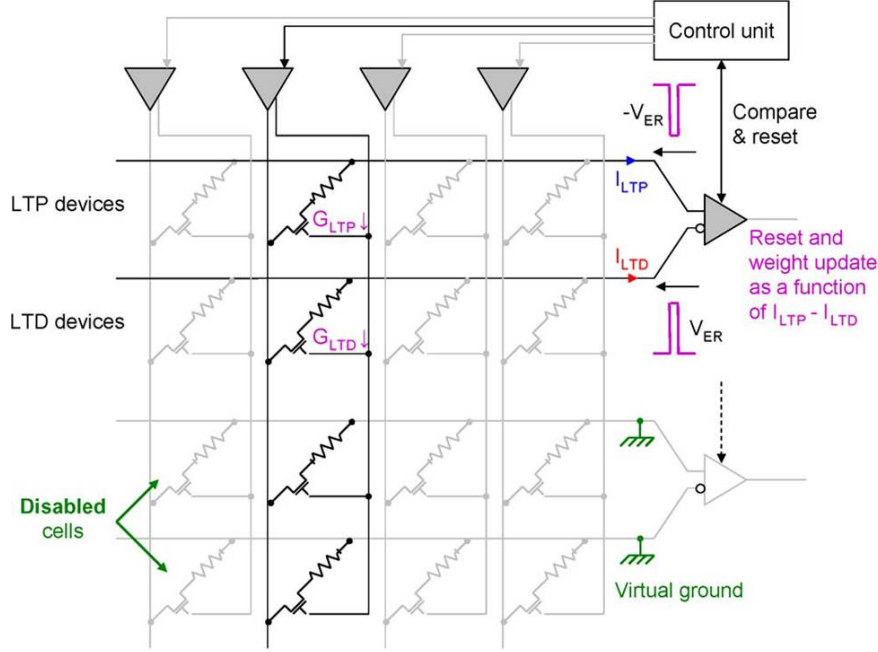
**Figure 2.20:** Refresh principle: The two devices forming a synapse are reset, and the one that had the higher conductance is reprogrammed such that the equivalent weight of the synapse stays unchanged [111].



**Figure 2.21:** Refresh-operation flowchart [111].



## 2. PHASE CHANGE MEMORY SYNAPSES



**Figure 2.22:** Refresh-operation: RESET pulses generation to re-initialize the LTP and LTD devices conductance to the minimum when  $V_{\text{RESET}} > 2.V_{\text{SET}}$ . [111].

would undergo  $N$  potentiations or  $N$  depressions in a row, it does not require permanent monitoring of the state of the LTP and LTD devices.  $N$  can be high (value approaching 100 is shown in [84]), thus reducing the time/energy overhead cost to a minimum. Simulations show that even  $N = 10$  incurs only a marginal cost for the system on a real-life learning experiment with almost 2,000,000 synapses (see sec.2.8).

Refresh operations are described in the diagram in Fig.2.21. The synapses are read, reset and reprogrammed sequentially. The other neurons are disabled during the process. To strongly amorphize the PCM, a RESET pulse of amplitude  $V_{\text{RESET}}$  has to be applied across the device, as shown in Fig.2.22. If  $V_{\text{RESET}} < 2.V_{\text{SET}}$ , a voltage of  $V_{\text{RESET}}$  across the PCM can be obtained with the interaction of two pulses of amplitude  $V_{\text{ER}}$  such that  $V_{\text{ER}} < V_{\text{SET}} < 2.V_{\text{ER}}$ , as shown in Fig.2.23. In this case, the voltage across the other synapses in the crossbar is always below the SET threshold and their conductance is not affected.

Therefore, if the condition  $V_{\text{RESET}} < 2.V_{\text{SET}}$  is true, no selector device is required for refresh operations. It is noteworthy that this condition is usually verified for scaled down PCM devices [115]. As neither the read nor the write operations actually require



## 2. PHASE CHANGE MEMORY SYNAPSES

---

one, selector devices could be eliminated completely. This would theoretically allow the highest possible PCM integration density in a crossbar and free the underlying CMOS layer for neuron integration.

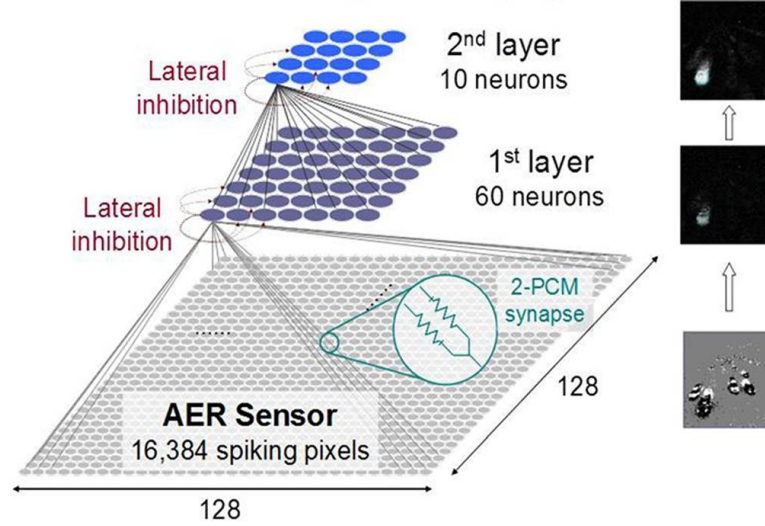
If  $V_{\text{RESET}} > 2.V_{\text{SET}}$ , the  $V_{\text{RESET}}$  voltage cannot be obtained with two pulses of amplitude below  $V_{\text{SET}}$ . Selectors are thus required to disable the other PCM devices and prevent their conductance to be altered, when the 2-PCM synapse being refreshed is reset, as shown in Fig.2.22 (disabled PCM devices are grayed).

### 2.8 Complex Visual Pattern Extraction Simulations

In this section, we present the results of a large scale learning simulation of our STDP implementation with the "2-PCM Synapse" architecture. We used a special purpose C++ event-based simulator called XNET, that was developed to simulate large scale spiking neural networks (SNN) based on use of resistive memory (RRAM) devices as synapses [114], [116]. In the simulations, real-time, asynchronous complex video data recorded from an artificial silicon-retina sensor [117] is presented to our SNN with PCM synapses. The network undergoes learning with STDP and is able to extract complex repetitive patterns from the visual stimuli in a fully unsupervised manner. In the present case, the learned patterns are car trajectories, which can be used to detect, track or count cars.

#### 2.8.1 Network and the Stimuli

Fig.2.24 shows the topological view of the simulated two-layer feed-forward SNN. It is a fully connected network, with 60 neurons in the first layer and 10 neurons in the second. The bottommost layer represents incoming stimuli from a  $128 \times 128$  pixels Address Event Representation (AER) dynamic vision sensor [117]. A pixel generates an event each time the relative change of its illumination intensity reaches a positive or a negative threshold. Therefore, depending on the sign of the intensity change, events can be of either type ON or type OFF, corresponding to a increase or a decrease in pixel illumination, respectively. There are two synapses per pixel, one for each event type. The total number of synapses in this system is thus  $2 \times 128 \times 128 \times 60 + 60 \times 10 = 1,966,680$  and thus 3,933,360 PCM devices (2 PCM/synapse).



**Figure 2.24:** 2-Layer Spiking Neural Network (SNN) topology used in simulation. The network is fully connected and each pixel of the  $128 \times 128$  pixel AER dynamic vision sensor (DVS-retina) is connected to every neuron of the 1st layer through two synapses, receiving positive and negative change in illumination events respectively. Lateral inhibition is also implemented for both layers [110].

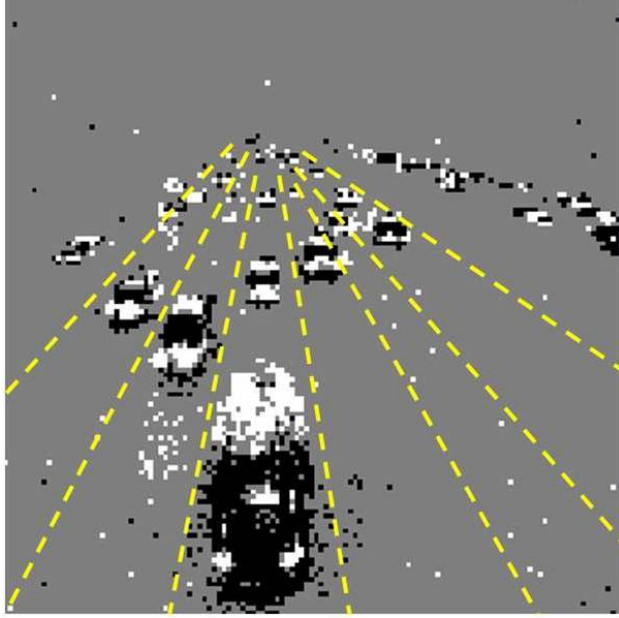
The visual stimuli used was recorded from the TMPDIFF128 DVS sensor [117]. It represents cars passing under a bridge over the 210 freeway in Pasadena. The sequence is 78.5 s in duration, containing a total of 5.2 M events, with an average event rate of 66.1 k events per second. Fig.2.25 shows a rendering of the sequence, where the traffic lanes have been marked.

### 2.8.2 Neuron and Synapses

The neurons used in simulations are described by a standard LIF neuron model. Tab.2.3, provides all the neuron parameters used for the simulation with GST and GeTe PCM devices.  $I_{thres}$ , denotes the neuron threshold firing current, expressed in Siemens (S) to make it independent of the read pulse's voltage and duration.  $T_{LTP}$ , is the LTP time window for the STDP rule (Fig.2.17a).  $T_{refrac}$ , is the neuron refractory period (time during which it cannot fire).  $T_{inhibit}$ , is the inhibitory time, when a neuron spikes, it disables all the other neurons during a period  $T_{inhibit}$ , during which no incoming spike is integrated (Lateral-inhibition).  $\tau_{leak}$  is the neuron leaky time-constant.  $N$ , is the number of activations of an output neuron required to initiate the

## 2. PHASE CHANGE MEMORY SYNAPSES

---



**Figure 2.25:** AER video data snapshot. Cars passing on a freeway recorded with the DVS-sensor described in [117]. The original video has no lane demarcations, yellow demarking lines were drawn later for lane-clarity [110].

refresh operations. LTP/LTD is the relative strength or equivalent weight change) of LTP compared to LTD. The LTP/LTD ratio of 2 used in our simulations ensures that repetitively potentiated synapses converge to their maximum equivalent weight quickly enough for the neuron to become selective to a traffic lane. This can be implemented by adding a current gain of 2 on the LTP input of the neurons.

All the parameters are obtained through genetic evolution optimization algorithm, as described in [114]. The optimized parameter values may change depending on: (i) characteristics of the synaptic devices (for example- Tab.2.1), (ii) the specific type of learning application (car detection in this case), and (iii) the stimuli dynamics, in this case it corresponds to the average spiking activity generated by the cars at the bottom of the retina (where activity is maximal due to the perspective). More detailed information on the meaning and the optimization of the parameters for the learning can be found in [114].

To compare the performance of the different PCM synapses, we repeated the simulations with LTP characteristics of all the 3 types of PCM devices (i.e. GST, GeTe and GST+HfO<sub>2</sub>). The behavioral model described in sec.2.5 is used for fitting the experi-

## 2.8 Complex Visual Pattern Extraction Simulations

**Table 2.3:** Neuron parameters for the learning. A different set of parameters is used depending on the PCM materials. See [114] for a detailed explanation of the parameters.

Parameter	GST		GeTe	
	1st Layer	2nd Layer	1st Layer	2nd Layer
$I_{thres}$	2.49 S	0.00437 S	2.50 S	0.00431 S
$T_{LTP}$	7.59 ms	7.12 ms	11.5 ms	12.9 ms
$T_{refrac}$	554 ms	410 ms	524 ms	393 ms
$T_{inhibit}$	15.7 ms	56.5 ms	11.8 ms	70.9 ms
$\tau_{leak}$	100 ms	821 ms	115 ms	714 ms
$N$	30		10	
$LTP/LTD$	2.0		2.0	

mental LTP-curves of the PCM synapses. Tab.2.1, shown earlier provides the values of the parameters used for describing the GST and GeTe synapses in XNET simulator. Several different factors can lead to device-to-device and cycle-to-cycle variability in PCM devices [118].

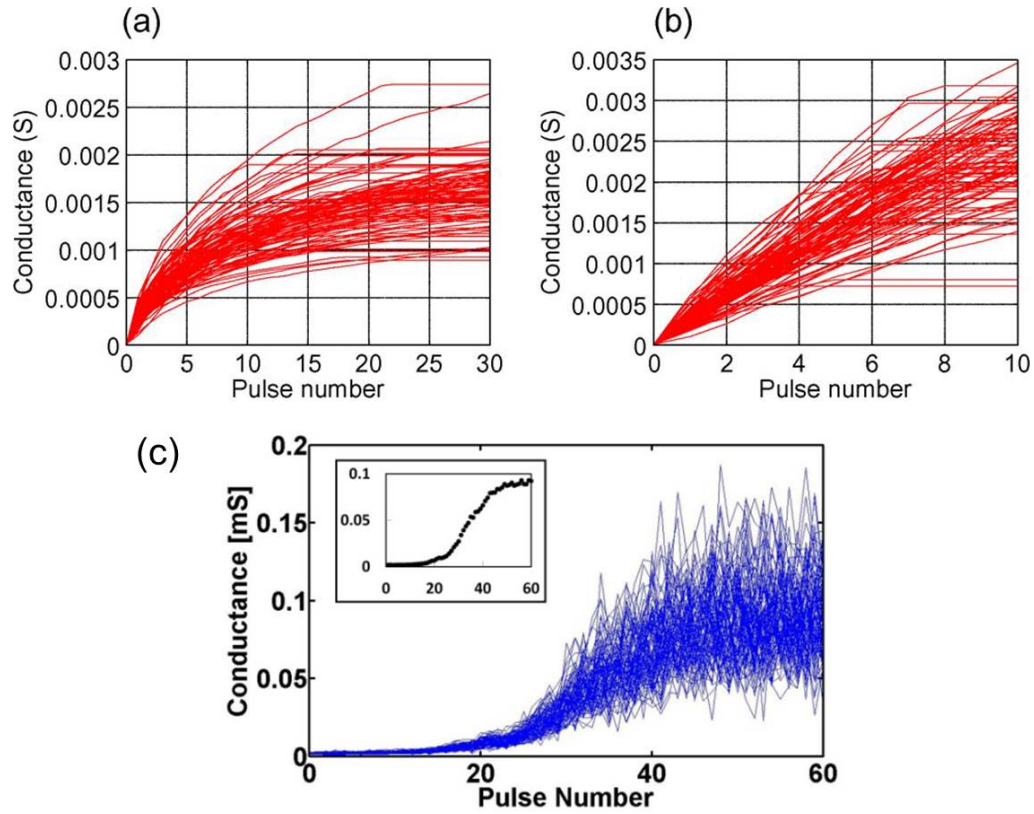
To evaluate the robustness to synaptic variability of our neuromorphic system, the simulated neural network included a pessimistic 20% dispersion (meaning that the standard deviation of every parameter is 20% of their mean value) for all the parameters of the behavioral model, described in sec.2.5, (i.e  $G_{min}$ ,  $G_{max}$ ,  $\alpha$  and  $\beta$ ). Fig.2.26 shows how the synapse LTP-curves look inside the simulator after adding the variability (100 different sets of parameters obtained by adding 20% dispersion from the values extracted from the fitting of Fig.2.6 and Fig.2.14). In our simulations, the parameters of a PCM device are changed each time it is refreshed. The 20% dispersion can therefore be seen as a representative of both device-to-device, and also cycle-to-cycle variability.

### 2.8.3 Learning Performance

Figure 2.27 shows the learning results for the AER dataset. The neurons are fully selective to single lane trajectories after only 8 presentations of the sequence, corresponding to approximatively 10 minutes of real-time traffic. STDP learning and lateral

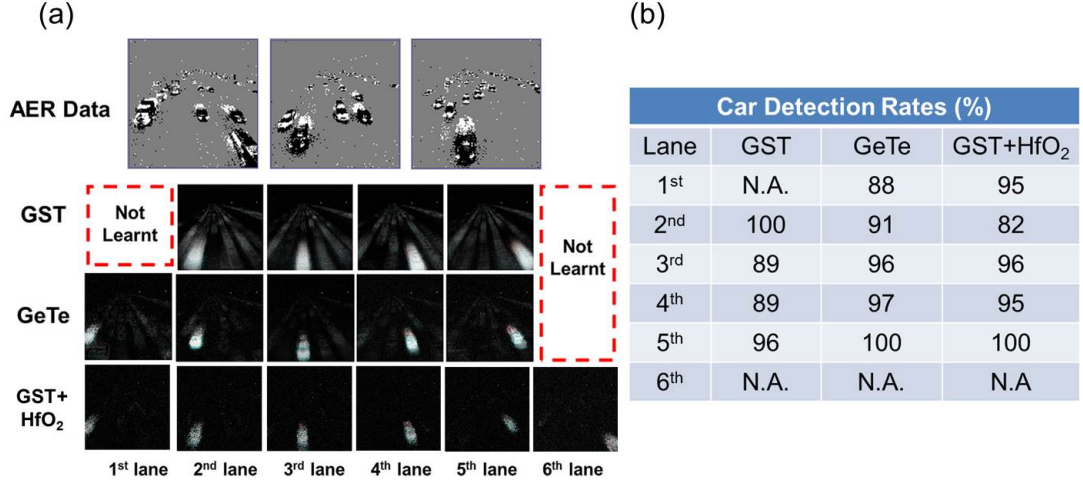
## 2. PHASE CHANGE MEMORY SYNAPSES

---



**Figure 2.26:** How synapses actually look inside the neural network: Simulated variability for (a) GST- (b) GeTe- and (c)Interface- PCM devices. The plots show the increase in conductance as a function of the number of SET pulses (LTP-curves), for 100 different sets of parameters, obtained by applying 20% dispersion (standard deviation of the mean value) from values extracted from fitting [98], [111]. Inset of (c) shows the experimentally obtained LTP-curve for a 200 ns potentiating pulse.

## 2.8 Complex Visual Pattern Extraction Simulations



**Figure 2.27:** Learning Results for GST, GeTe and Interface synapses: (a) Final output neuron sensitivity patterns for the 6 lanes and (b) Lane-specific car detection rates, [110], [98]. Lane-6 is not learnt by any neuron.

inhibition can be disabled altogether for continuous car detection afterward. Output neurons in the second layer are able to detect cars in 4/6 for systems based on GST-PCM synapses, and 5/6 lanes for systems based on GeTe-PCM and Interface layer-PCM synapses, respectively. The sixth lane is never learned, because it is at the very right of the retina and cars activate less pixels over their trajectory than those on other lanes. Over the learned lanes, the average detection rate is above 92%, with no false positive (i.e. neurons fire only once per car and they never fire for cars passing on a different lane than the one they learned). Learning statistics are given in table 2.4: the synaptic weight update frequency (or post-synaptic frequency) is of the order of 0.1 hertz and the average pre-synaptic frequency is around 2 Hz. The average frequencies are similar for the two layers.

The frequency of potentiating pulses (SET) per device was about 55 times higher than the frequency of RESET pulses for Interface-layer PCM based system, 25 times higher for GST-PCM based system, and about 10 times higher for the GeTe-PCM based system. This is consistent with the fact that refresh operations were initiated after 60 activations for a given output neuron with Interface devices, 30 for GST devices, and only 10 activations for GeTe. As mentioned earlier, this result suggests that the efficiency of the system can be further increased by engineering the PCM device with



## 2. PHASE CHANGE MEMORY SYNAPSES

---

**Table 2.4:** Learning statistics, over the whole learning duration ( $8 \times 85 = 680$  s). The SET pulses number includes both the write pulses for the learning and the additional pulses to reprogram the equivalent synaptic weight during refresh operations (Fig.2.20).

	/device	/device (max)	/device/s	Overall
	GST (2 V / 300 ns LTP pulses)			
Read pulses	1,265	160,488	1.9	$4.97 \times 10^9$
SET pulses	106	430	0.16	$4.16 \times 10^8$
RESET pulses	4.2	7	0.0062	$1.65 \times 10^7$
	GeTe (1.5 V / 100 ns LTP pulses)			
Read pulses	1,265	160,488	1.86	$4.97 \times 10^9$
SET pulses	190	740	0.28	$7.48 \times 10^8$
RESET pulses	20	37	0.030	$7.99 \times 10^7$
	GST+HfO <sub>2</sub> (interface) (2.1 V / 200 ns LTP pulses)			
Read pulses	1,265	160,488	1.9	$4.97 \times 10^9$
SET pulses	144	430	0.21	$5.64 \times 10^8$
RESET pulses	2.6	7	0.0038	$1.6 \times 10^7$

## 2.8 Complex Visual Pattern Extraction Simulations

**Table 2.5:** Energy statistics and synaptic power for the test case described in table 2.4, by using voltage and current values extracted from literature.

PCM Technology	$E_{\text{RESET}}$ (pJ)	$E_{\text{SET}}$ (pJ)	Power ( $\mu\text{W}$ )
GST-PCM	1552	121	112
Jiale Liang, VLSIT 2011 [119]	1.2	0.045	0.056
Xiong, Science 2011 [120]	0.1	0.03	0.02
Pirovano, ESSDERC 2007 [115]	24	4.9	3.6
D.H. Im, IEDM 2008 [121]	5.6	0.9	0.68

the optimum conductance window and phase change material stack, to maximize the number of conductance levels reachable with a series of identical SET pulses.

### 2.8.4 Energy/Power Consumption Analysis

Using the learning statistics from tab.2.4, we made a rough estimate of the power consumed for the programming of the PCM synapses:

$$E_{\text{total}} = E_{\text{SET}} \cdot N_{\text{total SET pulses}} + E_{\text{RESET}} \cdot N_{\text{total RESET pulses}} \quad (2.13)$$

$$\text{with } E_{\text{SET}} \approx V_{\text{SET}} \cdot I_{\text{SET}} \cdot t_{\text{SET}} \quad (2.14)$$

$$\text{with } E_{\text{RESET}} \approx V_{\text{RESET}} \cdot I_{\text{RESET}} \cdot t_{\text{RESET}} \quad (2.15)$$

With the SET and RESET voltages and currents measured on our GST devices and  $t_{\text{SET}} = 30 \text{ ns}$ ,  $t_{\text{RESET}} = 50 \text{ ns}$ ,  $E_{\text{SET}} \approx 121 \text{ pJ}$  and  $E_{\text{RESET}} \approx 1,552 \text{ pJ}$ . Using these values, the estimated synaptic programming power consumption for GST based learning is  $112 \mu\text{W}$ . The total synaptic programming power consumption for the interface layer (GST +  $\text{HfO}_2$ ) systems was almost halved to  $60 \mu\text{W}$ . The power consumption decreases with interface devices as they require less frequent refresh (RESET), and the current for individual SET/RESET programming is decreased due to the interface layer as explained in sec.2.6. We did not include the read energy in the calculation as it proved

## 2. PHASE CHANGE MEMORY SYNAPSES

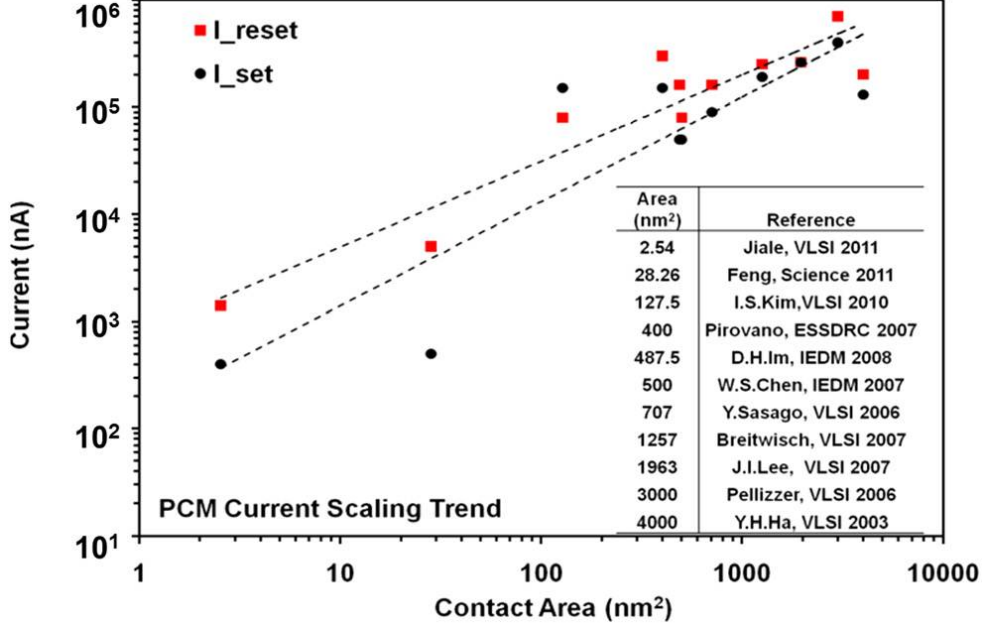


Figure 2.28: PCM programming current scaling trend [110].

to be negligible. Indeed, in the worst case, the total read energy would be as follows:

$$E_{\text{total read}} = E_{\text{read}_{\text{max}}} \cdot N_{\text{total read pulses}} \quad (2.16)$$

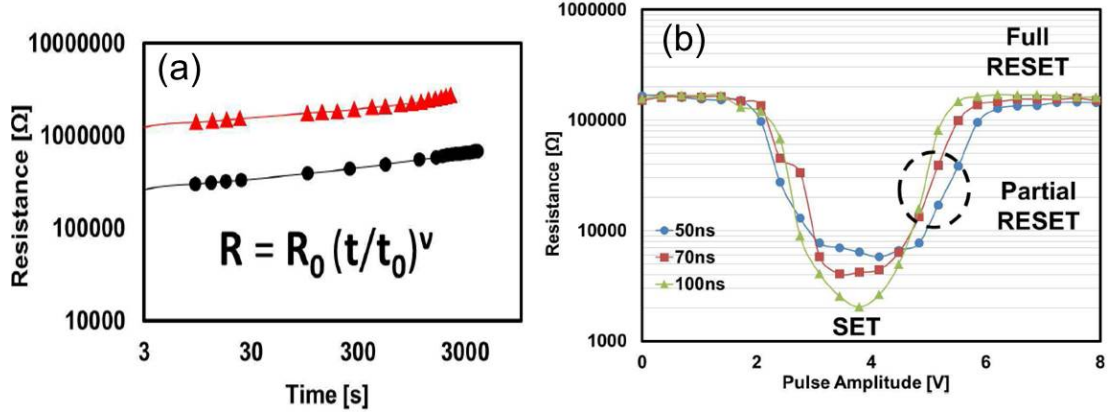
$$\text{with } E_{\text{read}_{\text{max}}} \approx V_{\text{read}}^2 \cdot G_{\text{max}} \cdot t_{\text{read}} \quad (2.17)$$

With  $V_{\text{read}} = 0.1$  V and  $t_{\text{read}} = 10$  ns, we estimated  $E_{\text{read}_{\text{max}}} \approx 0.17$  pJ and  $E_{\text{total read}} \approx 0.8$   $\mu$ W.

This calculation does not include dissipation in the CMOS circuitry for the neurons and also neglects the extra energy required for capacitive line charging in the crossbar, which can be significant in modern technology.

Fig.2.28 shows that on average, the current required for RESET and SET scales almost linearly with the PCM area. Tab.2.5 shows estimations of the synaptic power consumption with several published devices. With extremely scaled PCM technologies, a power consumption as low as 100 nW seems achievable for the  $\sim 2$  million synapses with continuous STDP learning. If learning would only occur for limited amounts of time, the energy consumption could be orders of magnitude lower.

With an average SET / RESET frequency per device of the order of 1 Hz, continuous learning for over 3 years would require an endurance of  $10^8$  cycles, which is easily



**Figure 2.29:** (a) Resistance drift with time for different initial programming conditions. Measurement was carried out on GST PCM devices. Equation governing the drift dynamics is also shown. (b) Experimental Resistance-Voltage curves for different programming pulse widths on GST PCM devices [122].

achievable with PCM [73]. Performance degradation would not be drastic even if the synaptic devices fail, thanks to the high level of fault tolerance of this kind of neural networks [116]. The strong synaptic variability taken in account in all the simulations, as shown in Fig.2.26, validates the robustness of our neural network.

## 2.9 Resistance-Drift and Mitigation Strategy

In PCM devices, the amorphous or high-resistance states are not entirely stable. Melt-quenched amorphous regions created inside the chalcogenide layer undergo structural relaxations and the resistance of PCM device tends to increase with time (known as resistance-drift). The resistance-drift follows an empirical exponential rule which depends upon the initial programmed resistance and a parameter known as the drift-coefficient ( $\nu$ ) [123]. The crystalline or low-resistance states of PCM are shown to be free from resistance-drift [124]. Fig.2.29a shows the resistance-drift measured in our GST-PCM devices and the equation governing the drift dynamics. The devices were programmed in two different initial (reset) states and the resistance was read at different time intervals. Inside a neural network such resistance-drift may cause undesired change of trained synaptic weights.

In this section, we discuss how the "2-PCM Synapse" architecture is inherently tolerant to PCM resistance-drift. We also introduce an alternative "Binary-PCM

## 2. PHASE CHANGE MEMORY SYNAPSES

---

Synapse” architecture and programming methodology with a stochastic STDP learning rule, which can strongly mitigate the effects of PCM resistance drift.

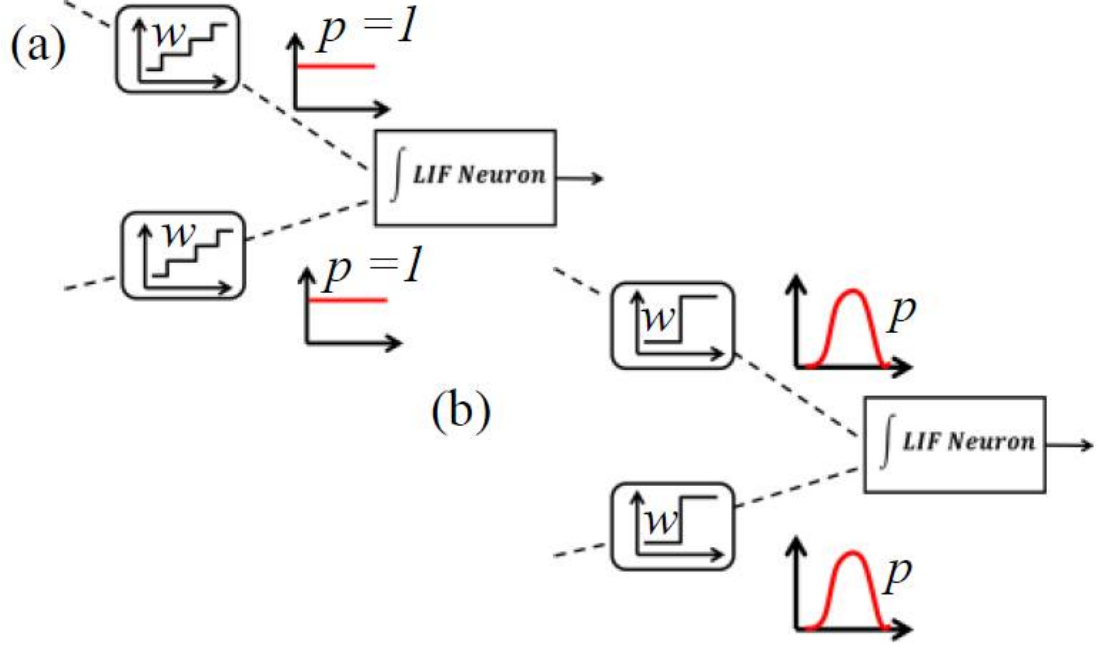
### 2.9.1 ”Binary PCM Synapse”

In this alternative approach, we propose to use PCM devices as binary synapses but with a stochastic-STDP rule. In this architecture, there will be 1-PCM device per synapse. Two resistance states (or weights) can be defined for the PCM synapse. The high-resistance state should be chosen such that it is a partial-reset (Fig.2.29b) state. The partial-reset state should lie in the negligible or low-driftable region. For GST based devices a resistance value  $< 50 \text{ k}\Omega$  will lie in low or negligible drift regime [124], [123]. At the system level, a functional equivalence [125] exists between deterministic multi-level and stochastic-binary synapses (Fig.2.30). In the case of supervised NN, several works have exploited this concept [126],[127],[128]. In this work, we use a similar approach for a fully unsupervised SNN. Our approach is also motivated by some works from biology [129], which suggest that STDP learning might be a partially stochastic process in nature.

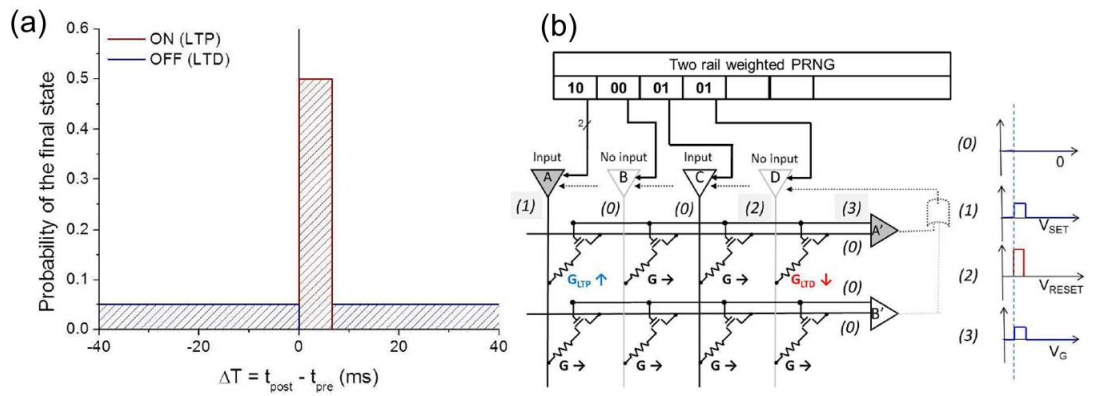
#### 2.9.1.1 Programming Scheme

Fig.2.31a shows an example of our stochastic-STDP rule. The y-axis represents a probability to switch from the set-to-reset or reset-to-set states for the PCM synapses. Fig. 2.31b shows the architecture and the programming scheme required to implement the stochastic learning with binary PCM synapses.

The stochasticity is controlled by an extrinsic PRNG (pseudo-random number generator) circuit [131]. The PRNG circuit controls the probability of LTP and LTD with a 2-bit signal. Initially, the input neurons (A-D) generate small read pulses when they encounter any stimuli event. The read current is integrated in the output neurons A’ and B’. When the output neuron reaches its firing threshold it generates a feedback signal(3) and a post-spike signal. In the example shown in Fig.2.31b, the output neuron A’ fires and B’ doesn’t fire. The signal(3) activates the gates of all the select transistors on the synaptic line connected to A’. If LTP is to be implemented the Input neuron will send a signal(1), as shown for Input neuron A in this example. In the case of LTD the input neuron will send a signal(2), as shown for the input neuron D. The probabilities of LTP/LTD can be tuned according to the learning rule (Fig. 2.31a).



**Figure 2.30:** Illustration depicting functional equivalence of deterministic multi-level and stochastic binary synapses.  $p$  indicates probability of change in conductance or switching [130].



**Figure 2.31:** (a) Simplified stochastic STDP learning rule. On corresponds to set and Off to reset of the PCM synapse. (b) Schematic of the Binary-PCM Synapse architecture and the proposed programming-scheme [122].

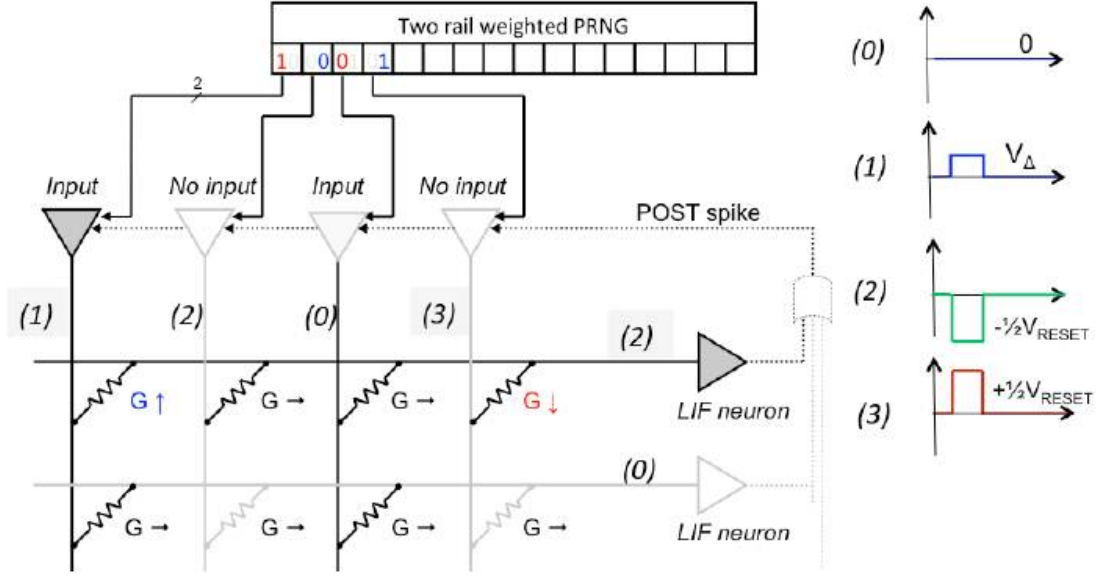
## 2. PHASE CHANGE MEMORY SYNAPSES

---

The programming scheme described herein can also be adapted for a selector-free configuration by changing the signals(1)-(4). Selector-free programming for "Binary-PCM Synapse" is shown in Fig.2.32. Whenever a post-synaptic neuron fires, a feedback pulse  $1/2 V_{RESET}$  is fed back to all the synapses connected to it. If  $1/2 V_{RESET} < V_{SET}$ , the signal will not affect the resistive state of the connected synapses by its own because its amplitude will be under the set threshold. At the same time, a write-mode signal is provided to all pre-synaptic neurons so that they will fire according to the probabilities given by the STDP rule and implemented by means of the PRNG block. If a pre-synaptic neuron was active in the LTP window, there is a  $p_{LTP}$  probability for a  $V_{\Delta}$  and signal to be fired. It will interact with the feedback signal so that the actual voltage drop across the corresponding synapse is  $V_{SET} = V_{\Delta} - (1/2 V_{RESET})$  and the synapse is switched to the ON state. The amplitude of the  $V_{\Delta}$  pulse on its own is not large enough to program the other connected synapses. If a pre-synaptic neuron's last activity is outside the LTP time window, its output will be a  $+1/2 V_{RESET}$  pulse with a  $p_{LTD}$  probability or a  $-1/2 V_{RESET}$  pulse with a  $(1-p_{LTD})$  probability. The positive pulse will interact with the feedback resulting in a pulse of amplitude  $V_{RESET} = +1/2 V_{RESET} - (-1/2 V_{RESET})$ , while the negative pulse will result in a voltage drop across the device that is negligible, thus keeping the resistance of the cell unaltered.

### 2.9.1.2 Analysis

In order to study the impact of PCM resistance-drift in our network we first classify its operation in two different modes: (a) Learning-mode and (b) Read-mode. In learning mode the synaptic programming is enabled and the network is trained using various datasets or stimuli. It is only during this mode that the synaptic weights can be changed. For the "2-PCM Synapse" architecture, in the learning-mode, the PCM are typically not experiencing drift. Only, whenever a refresh-sequence is applied to a synapse, it pushes the respective PCM devices into the driftable region. In the refresh-sequence (sec.2.7.2.3), both devices are reset and one of them is reprogrammed to a lower resistance state. For the device which stays in the fully reset state, drift is irrelevant: the more reset it is, the better. The other device, which is reprogrammed to a high resistance but intermediate state, may experience drift until it encounters sufficient crystallizing events (LTP or LTD) that push it in the non-driftable region.



**Figure 2.32:** Schematic of the Binary-PCM Synapse architecture and the proposed programming-scheme for selector-free configuration.

Such drift can delay learning, and its occurrence can be attributed to the refresh-frequency. More precisely, the drift in learning will be a consequence of the on-going competition between the refresh-frequency and the set-frequency.

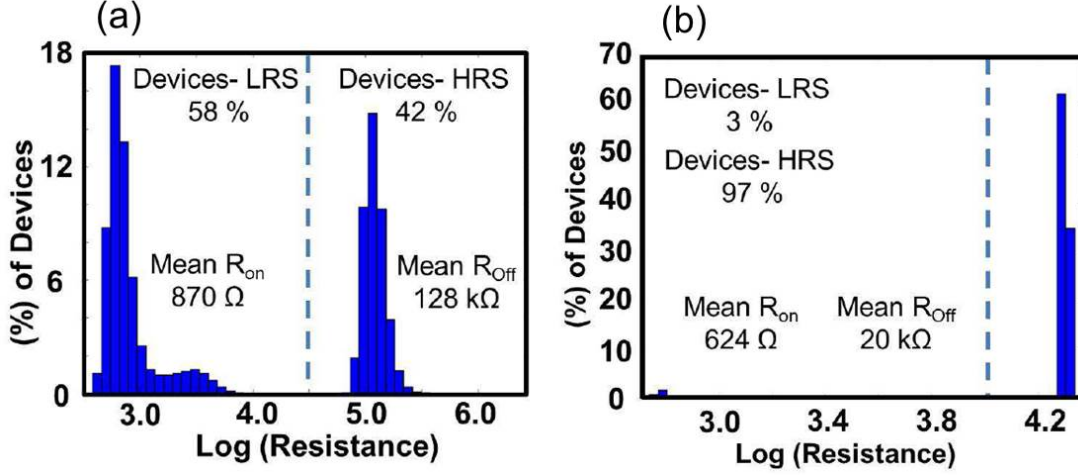
Set-frequency depends on (1) nature of the learning rule and (2) nature of the stimuli used for training, while the refresh-frequency mainly depends on the type of chalcogenide material used, as shown in sec.2.8.3.

Drift is more dramatic in the case of the read-mode compared to learning mode, since the system has no means to compensate for it through learning. In the read-mode, synaptic programming is disabled and a pre-trained neural network is used to identify patterns in new datasets or stimuli without changing the synaptic weights. Thus impact of resistance-drift in the read-mode is proportional to the final weight distribution of the synapses at the end of the training, and the time interval after which the network is operated in read-mode post training.

Synaptic weight distribution at the end of learning mode gives the number of synapses that are left in the high-resistance or driftable state. An inherent advantage of the "2-PCM Synapse" approach, compared to the methodology used in [82], is that we implement both potentiation and depression by crystallization. Thus the



## 2. PHASE CHANGE MEMORY SYNAPSES

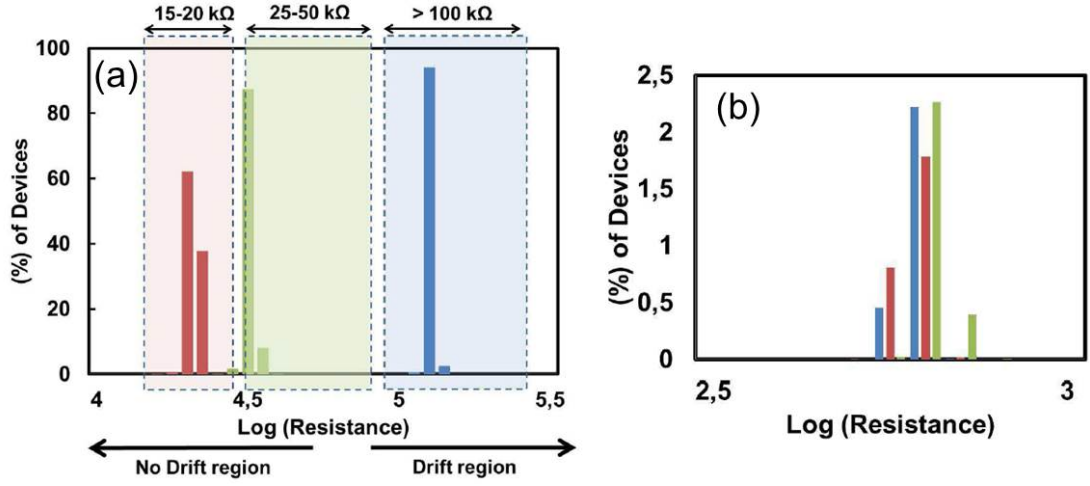


**Figure 2.33:** (a) Distribution of synaptic resistance states for the "2-PCM Synapse" architecture at the end of the visual learning simulation. (b) Distribution of synaptic resistance states for the "Binary-PCM Synapse" architecture with 20 k $\Omega$  mean  $R_{off}$  [122].

majority of PCM devices at the end of the learning are programmed in low resistance or non-driftable states. This is irrespective of the fact that we use a pre-dominantly depression (LTD) based learning rule. As crystalline or low resistance states are more stable and immune to the resistance drift [123], [124], [132] the "2-PCM Synapse" diminishes the loss of synaptic information in the read-mode.

Fig.2.33a, shows the final synaptic resistance (weights) distribution at the end of the cars learning simulation. About 60% of the devices are in the non-driftable region. The strong reduction in percentage of devices in driftable region is more evident from Fig.2.35. We can see that the number of set events is about ten times greater than number of reset events for the '2-PCM Synapse' as LTD is also implemented by set or crystallization.

In the case of "Binary-PCM Synapse" architecture the impact of drift in learning-mode can be fully mitigated if the reset state of the PCM devices is tuned carefully to a partial-reset state (negligible drift region). Fig.2.33b, shows the final synaptic resistance distribution at the end of learning when the simulation was performed for the "Binary-PCM Synapse" architecture with stochastic learning. In this simulation the mean reset state was defined as 20 k $\Omega$ , which lies in the non-driftable region. At the end of the learning, about 97% of the synapses are in the reset state. This is due to the



**Figure 2.34:** (a) Distribution of synapses in off-state for the "Binary-PCM Synapse" and (b) Distribution of synapses in on-state, for the PCM synapses with mean  $R_{off}$  values of 20 kΩ, 30 kΩ and 123 kΩ [122].

strongly LTD dominant nature of our learning rule (Fig.2.31a). Even though majority of the synapses are in reset state, they will not drift as they lie in the non-driftable region.

We performed the cars-learning simulations for the "Binary-PCM Synapse" architecture with 3 different PCM reset resistance states, keeping the set state constant: (1) negligible drift region (mean  $R_{off}$  = 20 kΩ), (2) Low drift region (mean  $R_{off}$  = 30 kΩ), and (3) high drift region (mean  $R_{off}$  > 100 kΩ). The final synaptic resistance distributions for the reset devices and the set devices in the 3 cases are shown in Fig.2.34a, and Fig.2.34b, respectively.

Fig.2.35, presents a comprehensive comparison of the learning performance, statistics and energy/power dissipation for the visual-pattern extraction simulations between the "2-PCM Synapse" and the "Binary-PCM Synapse" architectures. In fig.2.35, as we move from the "2-PCM Synapse" architecture to the "Binary-PCM Synapse", the number of read-events becomes half as the number of PCM devices is also halved. However the read frequency/device/s stays constant. The read-frequency stays constant as it depends on the nature of stimuli used for training the network. In the case of "Binary-PCM Synapse" the set and reset events are a direct representative of the number of LTP and LTD events. However in the case of "2-PCM Synapse" the reset events represent the number of refresh-sequences while the set events denote both LTP

## 2. PHASE CHANGE MEMORY SYNAPSES

Quantity	"2-PCM Synapse"	"Binary-PCM Synapse"	
	Roff = 128 K $\Omega$	Roff = 20 K $\Omega$	Roff = 30 K $\Omega$
Total Read	$4.97 \times 10^9$	$2.48 \times 10^9$	$2.48 \times 10^9$
Total Set	$4.16 \times 10^8$	$5.13 \times 10^5$	$4.84 \times 10^5$
Total Reset	$1.65 \times 10^7$	$1.90 \times 10^7$	$1.78 \times 10^7$
Frequencies of events ( event / device / sec)			
F read/d/s	1.9	1.9	1.9
F set/d/s	0.16	0.00038	0.00036
F reset/d/s	0.0062	0.014	0.013
Energy/ Power Consumption			
Set Energy	38.1 mJ	0.01 mJ	0.16 mJ
Reset Energy	38 mJ	21.8 mJ	25.7 mJ
Total Energy	76.1 mJ	22.0 mJ	25.9 mJ
Total Power	112 $\mu$ W	32.4 $\mu$ W	38.1 $\mu$ W
Read Energy	402 $\mu$ J	51 $\mu$ J	48 $\mu$ J
Read Power	600 nW	75 nW	70 nW

Car Detection Rate (%)		
Lane	Roff = 20k $\Omega$	Roff = 30k $\Omega$
1	NA	NA
2	97	98
3	93	92
4	94	93
5	99	96
6	NA	NA

**Figure 2.35:** (Left) Comparison of learning statistics for the "2-PCM Synapse" and "Binary-PCM Synapse" architectures. (Right) Car detection rates for the "Binary-PCM Synapse" architecture [122]. For both statistics two cases of "Binary-PCM Synapse" are shown (with mean Roff = 20 k $\Omega$  and 30 k $\Omega$ ).

## 2.9 Resistance-Drift and Mitigation Strategy

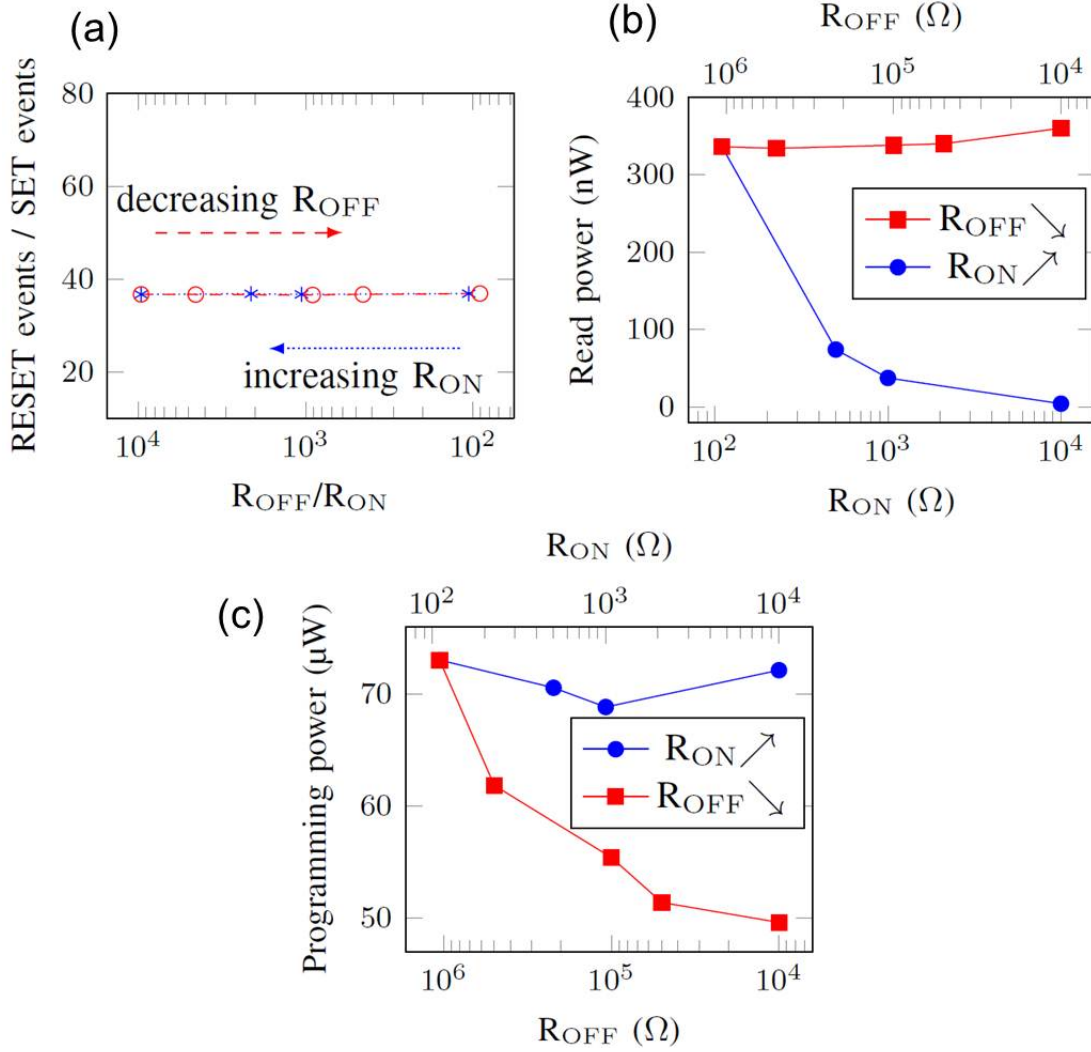
and LTD. The energy consumption decreases in the case of ‘Binary-PCM Synapse’ as the current required to program partial-reset states (20 k $\Omega$  and 30 k $\Omega$ ) is much less compared to the current required to program a strong reset state (128 k $\Omega$ ). However the energy consumption doesn’t decrease drastically as the number of reset-events increases in the ‘Binary-PCM Synapse’ architecture. In both cases (R<sub>off</sub>:20 k $\Omega$ , 30 k $\Omega$ ), the power consumption by the PCM devices during learning remains low (<80  $\mu$ W).

From Fig.2.5b,c we can see that it is possible to tune different levels of PCM resistance windows by changing the programming conditions. Thus we performed several parametric simulations to study the impact of PCM resistance window on the system power dissipation and learning performance. In the first case we keep the R<sub>ON</sub> constant at 110  $\Omega$ , and change the R<sub>OFF</sub>. In the second case we fix R<sub>OFF</sub> to 1.06 M $\Omega$  and change the R<sub>ON</sub>. For all simulations, the the average detection rate was 94%. Obtaining a high detection rate with an unsupervised system and binary synapses is a strong accomplishment from a machine learning point of view. Binary synapses appear especially fit to process this kind of highly dynamic video data.

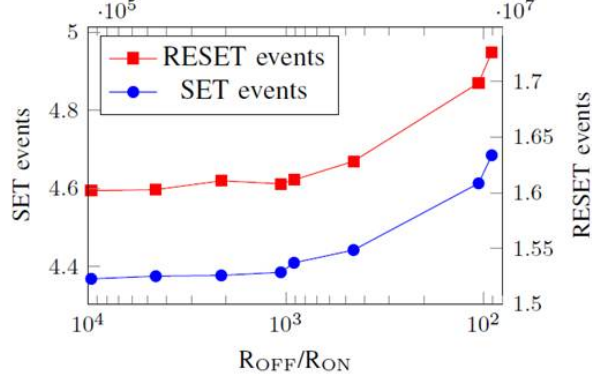
Fig.2.36a shows that the ratio between the total number of RESET and SET events remains constant when the resistance window changes. Indeed, this means that the programming activity is dominated by the input stimuli and the STDP learning rule. Analysis of the learning-mode power: Fig.2.36b shows that when R<sub>OFF</sub> is decreased while keeping R<sub>ON</sub> fixed it is possible to reduce the synaptic programming power by about 32%. This is explained by the fact that smaller current values are required to obtain smaller R<sub>OFF</sub> values. Weakening the SET state (increasing R<sub>ON</sub>) does not translate into a reduction of the programming power. This is explained by two reasons: 1) the STDP rule is strongly dominated by LTD, i.e. RESET operations, rather than SET operations; 2) when the resistance window is decreased, the number of RESET events increases (Fig.2.37). So, the effect of weakening the SET conditions gets compensated by the increased number of RESET events.

R<sub>ON</sub> plays more important role in determining the read power dissipation, as it determines the current flowing into the synapses at each read pulse. As shown in Fig.2.36c, when the resistance window is reduced by increasing the R<sub>ON</sub> value (blue curve), it is possible to reduce the power consumption for read operations by 99%. We can observe that the trends in power consumption for learning-mode and read-mode are opposite, thus two strategies can be adopted for optimizing power consumption

## 2. PHASE CHANGE MEMORY SYNAPSES



**Figure 2.36:** (a) Ratio between the number of RESET and SET events as a function of the resistance window  $R_{OFF}/R_{ON}$ . (b) Programming power as a function of decreasing  $R_{OFF}$  - red line (keeping  $R_{ON} = 110 \Omega$  constant) and increasing  $R_{ON}$  - blue line, (keeping  $R_{OFF} = 1.06 \text{ M}\Omega$  constant). (c) Read power as a function of decreasing  $R_{OFF}$  - red line (keeping  $R_{ON} = 110 \Omega$  constant), and increasing  $R_{ON}$  - blue line (keeping  $R_{OFF} = 1.06 \text{ M}\Omega$  constant) [133].



**Figure 2.37:** SET and RESET events as functions of resistance window [133].

based on the usage of the system. If the network is mostly used in learning-mode, in order to minimize the system power consumption smaller values of  $R_{OFF}$  are recommended because the reduction of the RESET current has much stronger impact on the programming power in the case of PCM synapses. On the contrary, if the network is mostly used in read-mode, larger values for  $R_{ON}$  and  $R_{OFF}$  are recommended to reduce read-mode power consumption.

## 2.10 Conclusion

In this chapter we demonstrated that PCM devices could be used to emulate LTP-like and LTD-like synaptic plasticity effects. We showed that while gradual LTP can be obtained with the application of identical potentiating (crystallizing) pulses, the nature of LTD is abrupt when identical depressing (amorphizing) pulses are used. The reason for the abrupt LTD behavior was explained through experiments and multi-physical simulations. We studied the role of crystallization kinetics (growth and nucleation rates) in LTP emulation, using PCM devices fabricated with two different chalcogenide materials: nucleation dominated GST, and growth dominated GeTe. A versatile (i) behavioral model and a (ii) circuit compatible model, useful for large scale neural network simulations with PCM devices were developed.

To overcome the limitations of abrupt LTD, we developed a novel low-power architecture (“2-PCM Synapse”) and a detailed programming methodology (Read-, Write- and Refresh- protocol) for architectures (i) with selector devices (1T-1R) and (ii) without selector devices (1R). We simulated a 2-layer spiking neural network (SNN), spe-

## 2. PHASE CHANGE MEMORY SYNAPSES

---

cially designed for complex visual pattern extraction application, consisting about 4 million PCM devices and a simplified STDP learning rule. Our SNN was able to extract the orientation and shapes of moving cars on a freeway with a very high average detection rate ( $> 90\%$ ) and extremely low synaptic programming power consumption of  $112 \mu\text{W}$ . We demonstrated that by engineering the interface of GST-PCM devices (adding a 2 nm  $\text{HfO}_2$  layer), energy efficiency of our neuromorphic system can be enhanced both at the level of individual synapses and the overall system. With the interface layer, power consumption was decreased to as low as  $60 \mu\text{W}$ , while individual synaptic programming power is decreased by  $> 50\%$ . We then investigated in detail the impact of PCM resistance-drift on our neuromorphic system. We show that the “2-PCM Synapse” architecture has high tolerance towards loss of synaptic information due to resistance drift. To further mitigate the impact of resistance drift, we introduce an alternative architecture and programming methodology (called “Binary-PCM Synapse”) with a stochastic STDP learning rule. System level simulations confirmed that using the “Binary-PCM Synapse” approach doesn’t affect the learning performance. Synaptic power consumption as a function of  $R_{OFF}$  and  $R_{ON}$  values was investigated. The results show that the learning-mode power consumption can be minimized by decreasing the  $R_{OFF}$  value, while read-mode power consumption can be optimized by increasing both the  $R_{ON}$  and  $R_{OFF}$  values. The synaptic power consumption can be strongly reduced to a few 100 nWs if state-of-the-art PCM devices are used.

To go further, the novel network topology introduced in [134] can be exploited, with spatially localized neurons, providing similar learning performance with only a tenth of the synapses. Thus requiring only about 130000 synapses bringing the prospect of a practical hardware realization even closer. The chapter establishes the potential of PCM technology for future intelligent ‘ubiquitous’ embedded neuromorphic systems.

“If two wrongs don’t make a right, try three”

-Laurence Peter

## 3

# Filamentary-Switching Type Synapses

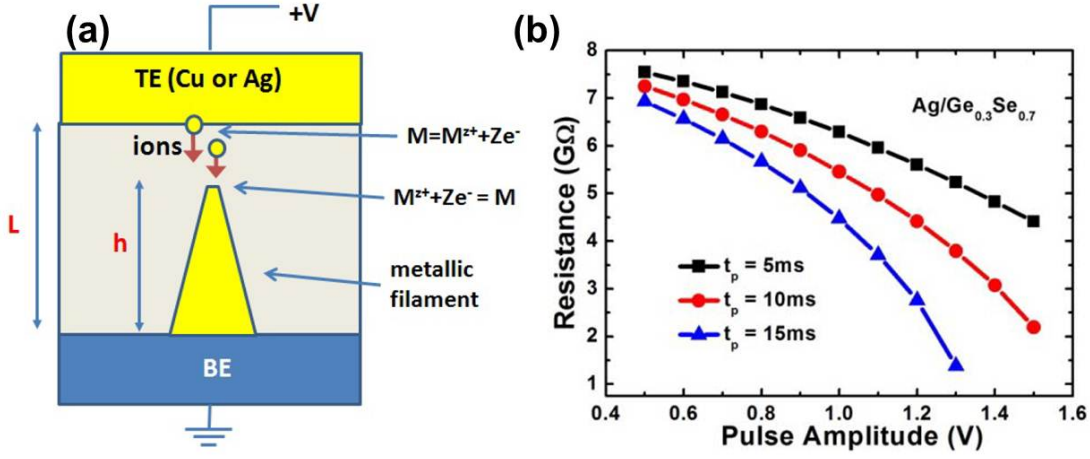
This chapter discusses how filamentary-switching type of memory devices can be used to emulate biological synapses in large-scale neuromorphic systems. The first part of the chapter focuses on Ag/GeS<sub>2</sub> based Conductive-bridge (CBRAM) technology, while the second part focusses on HfO<sub>x</sub> based resistive metal-oxide (OXRAM) technology.

### 3.1 CBRAM Technology

CBRAM also known as programmable metalization cell (PMC), consists of a solid electrolyte layer sandwiched between two metal electrodes. The working principle is understood to be based on reversible electrochemical redox reactions [135]. Usually the top electrode (anode) contains an electrochemically active layer that acts as a donor of metal ions. When voltage bias of a specific polarity is applied across the device, metal ions from the anode diffuse and drift in the electrolyte, and get reduced at the inert electrode (Fig.3.1a). A small metallic filament or dendritic nanowire of metal ions is formed between the two electrodes leading to a conductive ON state. When reverse polarity is applied the metallic filament dissolves and the device switches to OFF state. Intermediate resistance levels can be obtained by tuning the dimensions of the metallic filament[136].



### 3. FILAMENTARY-SWITCHING TYPE SYNAPSES

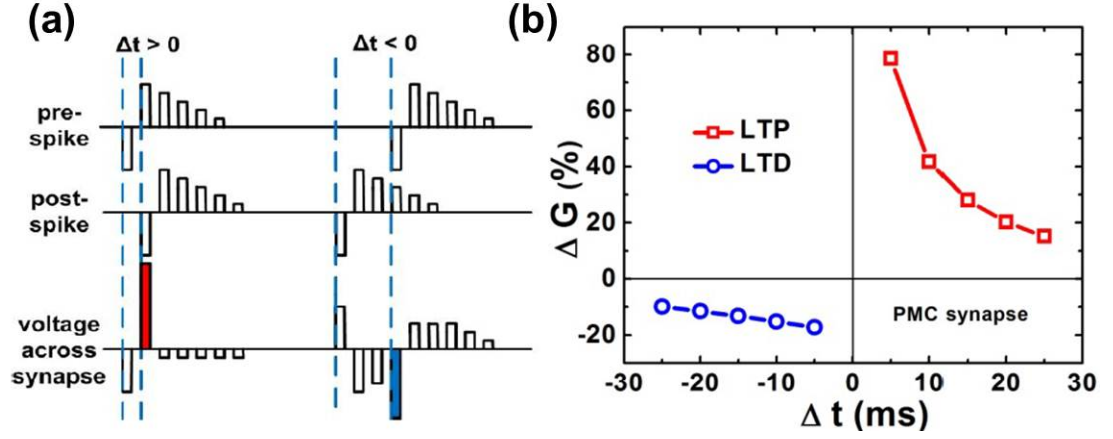


**Figure 3.1:** (a) Schematic of the CBRAM device showing the formation of the conductive-filament (b) Simulated modulation of R<sub>off</sub>, adapted from [137].

#### 3.1.1 CBRAM state-of-art Synapses

Yu et. al [137] simulated STDP emulation on CBRAM cells containing Ag top-electrode and Ge<sub>0.3</sub>Se<sub>0.7</sub> electrolyte. Fig.3.1b, shows the simulated off-state resistance modulation in Ag/Ge<sub>0.3</sub>Se<sub>0.7</sub> CBRAM devices by applying pulses with varying pulse widths and amplitudes. Fig.3.2, shows the proposed programming scheme and simulated STDP emulation for the same devices.

Fig.3.3, shows another 2-terminal CBRAM-like device used for synaptic emulation. It consists of a layered structure including a cosputtered Ag and Si active layer with a properly designed Ag/Si mixture ratio gradient that leads to the formation of a Ag-rich (high conductivity) region and a Ag-poor (low conductivity) region [138]. Unlike CBRAM, the cosputtered devices don't undergo electrochemical redox reactions and there is no formation or dissolution of a conductive filament. However due to cosputtering of Ag and Si, nanoscale Ag particles are incorporated into the Si medium during device fabrication and a uniform conduction front between the Ag-rich and Ag-poor regions is formed. Under applied bias, the continuous motion of the conduction front in the device replaces discrete, localized conducting filament formation and results in analog switching behavior with multiple intermediate resistance states [138]. Fig.3.3a, shows emulation of LTP- and LTD- like effects, when the device is programmed by a series of 100 identical potentiating (3.2 V, 300 μs) pulses followed by a series of 100 identical depressing pulses (-2.8 V, 300 μs).

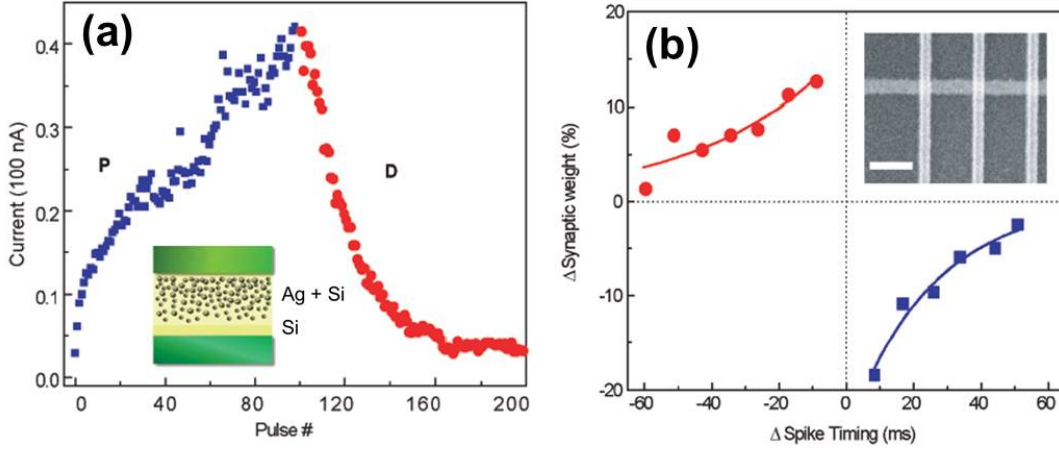


**Figure 3.2:** (a) Shapes of pre- and post- neuron spikes used to emulate STDP on Ag/Ge<sub>0.3</sub>Se<sub>0.7</sub> CBRAM devices (b) Simulated STDP-like curve for the Ag/Ge<sub>0.3</sub>Se<sub>0.7</sub> CBRAM devices, adapted from [137].

To emulate STDP with the Ag+Si/Ag devices, Jo et. al [138] implemented a special CMOS neuron circuit that converts the relative timing information of the neuron spikes into pulse width information seen by the synaptic devices. Their neuron circuit consists of two CMOS based IF-neurons (pre- and post) connected by a Ag+Si/Ag device synapse. The neuron circuit employs a time division multiplexing (TDM) approach with globally synchronized time frames to convert the spike timing information into a pulse width. Fig.3.3b, shows the measured change of the synaptic weight after each neuron spiking event obtained in the hybrid CMOS-neuron/RRAM-synapse circuit.

Almost all of these recent demonstrations of RRAM based synaptic emulation treat the synapse as a deterministic multi-valued programmable non-volatile resistor. Although such treatment is desirable, it is challenging in terms of actual implementation. Programming schemes for multi-level operation in RRAM devices are more complicated compared to binary operation. Gradual multi-level resistance modulation of RRAM synapses may require generation of successive non-identical neuron spikes (pulses with changing amplitude or width or a combination of both), thus increasing the complexity of the peripheral CMOS neuron circuits which drive the synapses. Pulse trains with increasing amplitude lead to higher power dissipation and parasitic effects on large crossbars. Another issue is that aggressive scaling leads to increased intrinsic device variability. Unavoidable variability complicates the definition and reproducibility of

### 3. FILAMENTARY-SWITCHING TYPE SYNAPSES



**Figure 3.3:** (a) Incremental increase and decrease of device conductance on application of potentiating and depressing pulses (b) Demonstration of STDP in the Ag+Si/Ag devices, inset shows a SEM image of the fabricated synapse array, adapted from [138].

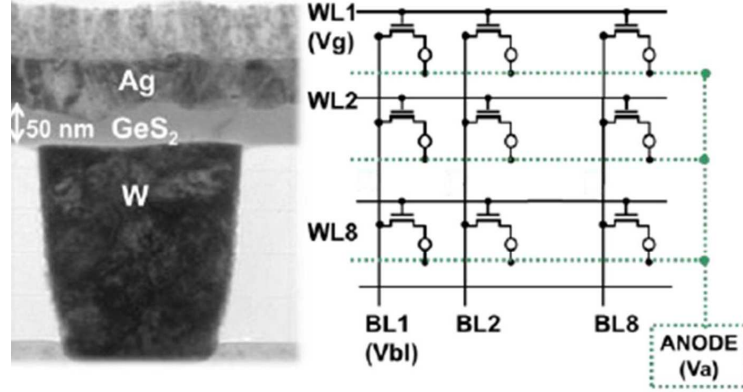
intermediate resistance states in the synaptic devices. In the following sections of this chapter, we present an alternative approach to overcome the issues with multi-level synapses. We show a neuromorphic system which uses CBRAM devices as binary synapses with a stochastic-STDP learning rule.

#### 3.1.2 Device and Electrical Characterization

1T-1R CBRAM devices (both isolated and in 8x8 matrix), integrated in standard CMOS platform [75], were fabricated and tested<sup>1</sup> (Fig. 3.4). A Tungsten (W) plug was used as bottom electrode. The solid electrolyte consisted of a 30 nm thick GeS<sub>2</sub> layer deposited by RF-PVD and a 3nm thick layer of Ag deposited by a DC PVD process. The 3 nm thick Ag layer is dissolved into the GeS<sub>2</sub> using the photo-diffusion process [139]. Then a 2<sup>nd</sup> layer of Ag about 75 nm thick was deposited to act as top electrode.

CBRAM operating principle relies on the reversible transition from high (reset) to

<sup>1</sup>In the following sections, we use the terms strong- and weak- programming conditions. However these have a relative definition with respect to the technology and materials used for fabricating the CBRAM devices. For the devices presented here, a weak-condition refers to a short pulse width ( $<10 \mu\text{s}$ ), usually  $1 \mu\text{s}$  or  $500 \text{ ns}$ , with a voltage  $<2.5 \text{ V}$  applied at the anode or the bit-line. A strong condition corresponds to a pulse width  $>10 \mu\text{s}$ .



**Figure 3.4:** (Left) TEM of the CBRAM resistor element. (Right) Circuit schematic of the 8 X 8 1T-1R CBRAM matrix. (note: the devices used in this study had a GeS<sub>2</sub> layer thickness of 30 nm. The 50 nm TEM is for illustrative purpose only [130].)

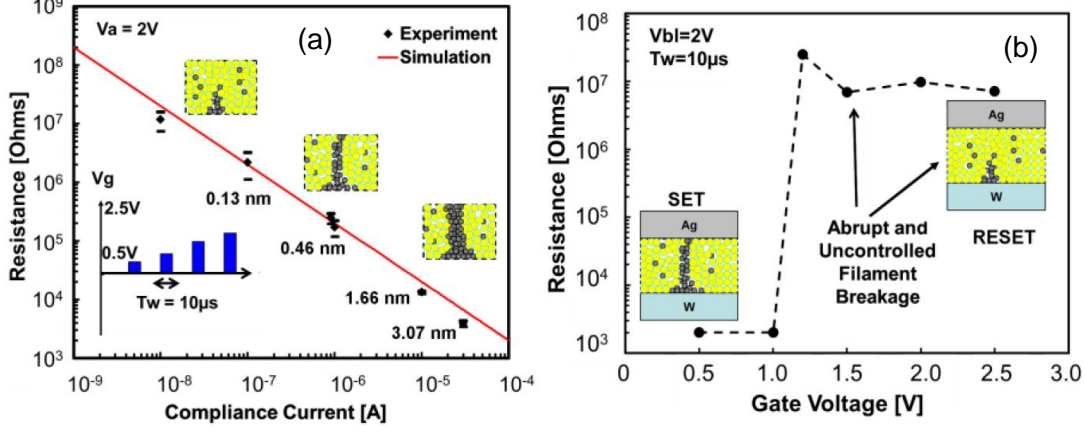
low (set) resistive states owing to the formation and dissolution of a conductive filament in the electrolyte layer. In particular, applying a positive voltage at the Ag electrode results in the drift of Ag<sup>+</sup> ions in the GeS<sub>2</sub> and discharge at the inert counter electrode (W), leading to the growth of Ag dendrites that eventually shunt the top and the bottom electrodes. Upon reversal of voltage polarity, an electrochemical dissolution of the conductive bridge occurs, resetting the system to the OFF (reset) state (Fig. 3.5). No forming step is required for this device stack. Simple fabrication, CMOS compatibility, high scalability, low power dissipation, and low operating-voltages [140] make CBRAM devices a good choice for the design of synapses in dense neuromorphic systems.

### 3.1.3 Limitations on LTD emulation

We demonstrate LTP-like behavior (i.e. gradual ON-state resistance decrease) in our GeS<sub>2</sub> based samples by applying a positive bias at the anode and gradually increasing the select transistor gate voltage (Vg) (Fig. 3.5a). This phenomenon of gradual resistance decrease can be explained with our model [136], assuming a gradual increase in the radius of the conductive filament formed during the set process. Larger gate voltages supply more metal ions leading to the formation of a larger conductive filament during the set process [141].

Nevertheless, this approach implies that each neuron must generate pulses with increasing amplitude while keeping a history of the previous state of the synaptic device,

### 3. FILAMENTARY-SWITCHING TYPE SYNAPSES

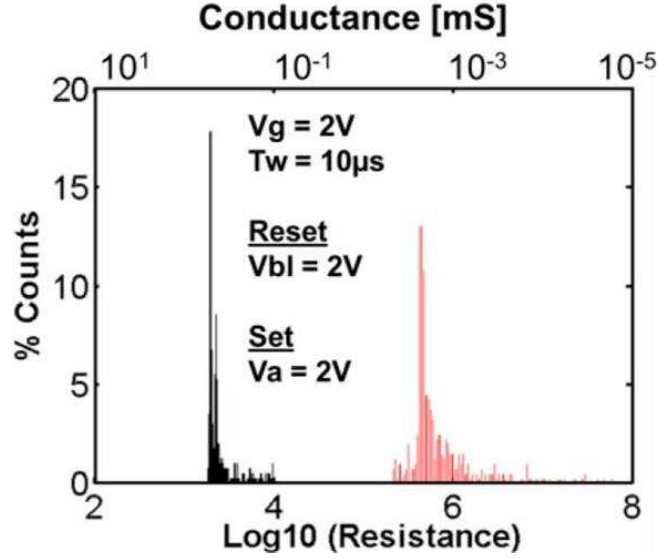


**Figure 3.5:** (a) On-state resistance modulation using current compliance. Fitting using model [136] is also shown (extracted filament radius are indicated). (b) Resistance dependence on gate voltage during the set-to-reset transition [130].

thus leading to additional overhead in the neuron circuitry. Moreover, we found it difficult to reproducibly emulate a gradual LTD-like effect using CBRAM. Fig. 3.5b shows the abrupt nature of the set-to-reset transition in our devices. Precisely controlling the dissolution of the conductive filament was not possible during the pulsed reset process. Note that for emulating a spiking neural network (SNN) it is essential that both LTP and LTD be implemented by pulse-mode programming of the synaptic devices. Pulse based synaptic programming is an analogue for the neuron spikes or action-potentials.

#### 3.1.4 Deterministic and Probabilistic Switching

Fig. 3.6 shows the On/Off resistance distributions of an isolated 1T-1R CBRAM (during repeated cycles with strong set/reset conditions). The OFF state presents a larger dispersion compared to the ON state. This can be interpreted in terms of non-uniform breaking of the filament during the reset process, due to the unavoidable defects [142],[143] close to the filament which act as preferential sites for dissolution. By fitting the Roff-spread data with our physical model [136], the distribution of the left-over filament-height was computed. Using the computed distribution of the left-over filament height and the equations in [136] we estimated the spread on the voltage ( $V_{set}$ ) and time ( $T_{set}$ ) needed for a successful consecutive set operation (Fig. 3.7). Moreover, when weak-set programming conditions are used immediately after a reset,



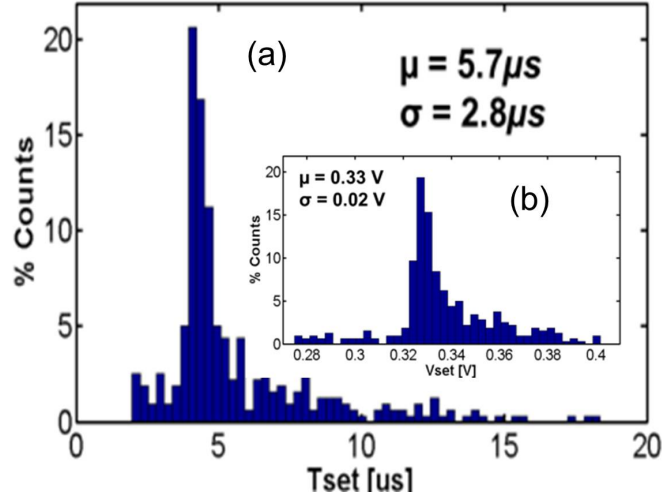
**Figure 3.6:** On/Off resistance distribution of an isolated 1T-1R device during 400 cycles when strong programming is used [130].

a probabilistic switching of the device may appear as seen in fig. 3.8. In fig. 3.8 the set operation fails in several cycles as the set-programming conditions are not strong enough to switch the device in those cycles.

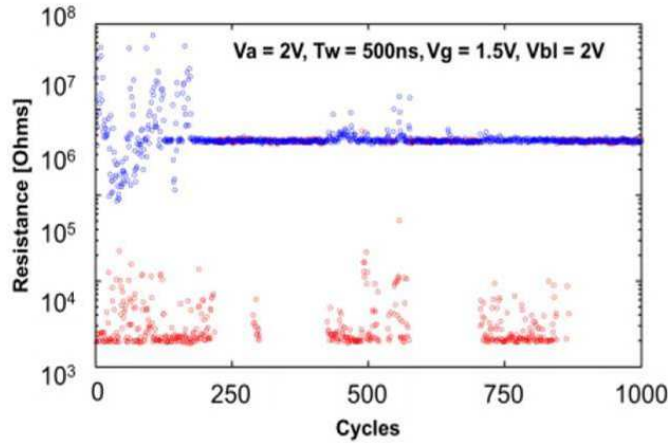
In a large-scale system, such stochastic switching behavior at weak conditions will get compounded with the inclusion of 'device-to-device' variations. To take into account the device-to-device variability, we performed similar analysis on the matrix devices. Fig. 3.9 shows the On/Off resistance distributions for all devices cycled 20 times with strong conditions. As expected, the spread on Roff values is larger compared to the Roff spread for a single device shown in fig. 3.6.

To quantify the trend of probabilistic switching (both set/reset) we designed two simple experiments: a cycling procedure with a strong-set condition and progressively weakening-reset condition was used to determine reset probability (fig. 3.10a) while a strong-reset condition and progressively weakening set condition was used to determine the set-probability (fig. 3.10b). As shown in fig. 3.10, the overall switching probability (criterion for successful switch:  $R_{off}/R_{on} > 10$ ), for 64 device matrix, increases with stronger programming conditions. It is thus conceivable to tune the CBRAM device switching probability by using the right combination of programming conditions.

### 3. FILAMENTARY-SWITCHING TYPE SYNAPSES

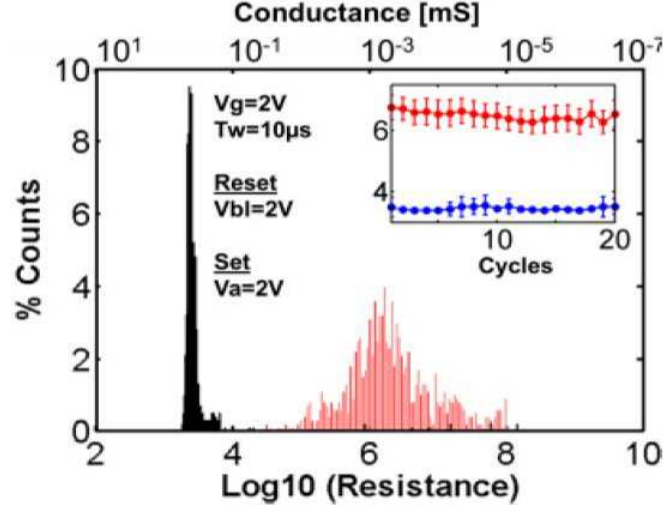


**Figure 3.7:** Computed distributions (generated using Roff data from fig. 3.6 and model [136], of: (a)  $T_{set}$  and (b)  $V_{set}$  (Inset) values for consecutive successful set operation (mean and sigma are indicated). For computing (a) the applied voltage is 1 V and for (b) a ramp rate of 1 V/s is used in the quasi-static mode [130].

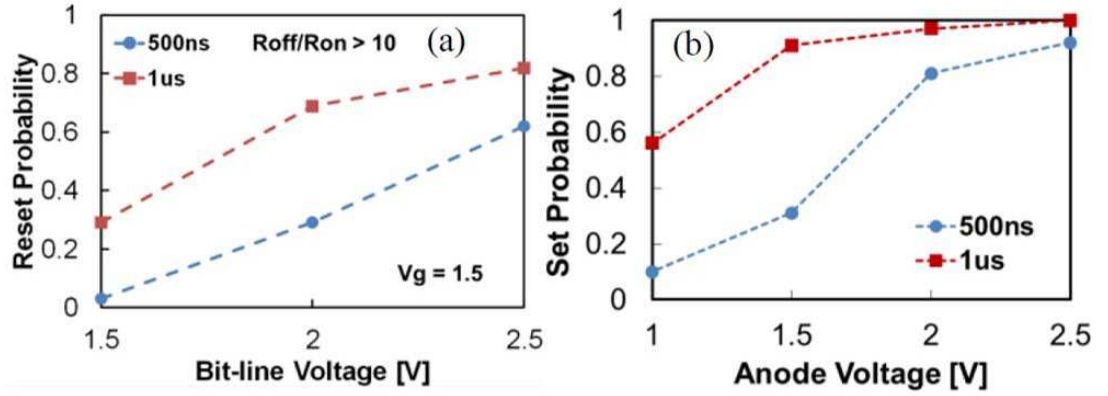


**Figure 3.8:** Stochastic switching of 1T-1R device during 1000 cycles using weak-conditions (switch-probability=0.49) [130].





**Figure 3.9:** On/Off resistance distributions of the 64 devices of the 8x8 matrix cycled 20 times. Inset shows  $R_{on}$  and  $R_{off}$  values in log scale with dispersion for each cycle [130].



**Figure 3.10:** Overall switching probability for the 64 devices of the matrix (switching being considered successful if  $R_{off}/R_{on} > 10$ ) using (a) weak-reset conditions and (b) weak-set conditions.  $V_g$  of 1.5V was used in both experiments [130].



### 3. FILAMENTARY-SWITCHING TYPE SYNAPSES

---

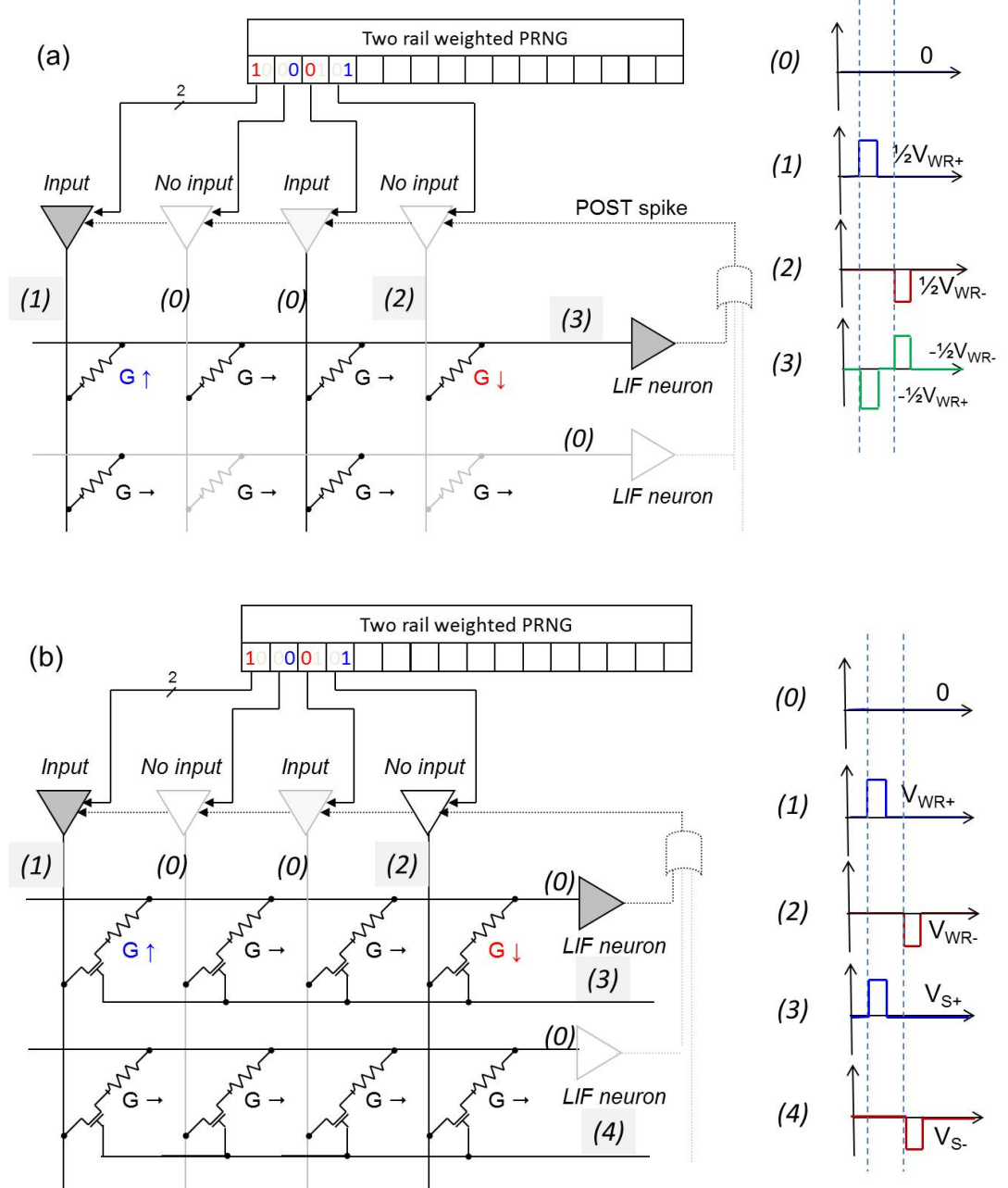
#### 3.1.5 Stochastic STDP and Programming Methodology

Fig. 3.11 shows the core circuit of our architecture. It is similar to the one that we proposed for deterministic synapses in [110],[116] but is adapted for bipolar-devices and stochastic learning rule. The core consists of three main blocks- (i) Input/Output CMOS-neuron circuits (ii) CBRAM synapse-crossbar connecting the neurons. This may be implemented without (1R) or with (1T-1R) selector devices (Fig. 3.11(a) and (b), respectively), and (iii) Pseudo-random number generator (PRNG) circuit. The PRNG block is only used for implementing optional extrinsic stochasticity as explained later. All neurons are modeled as leaky-integrate and fire (LIF) type.

Our stochastic-STDP rule (Fig. 3.12) is a simplified version of the deterministic biological STDP rule [37]. The optimization of the LTP window and neuron parameters is performed using genetic-evolution algorithm [134]. The STDP rule functions as follows: when an output neuron fires, if the input neuron was active recently (within the LTP time window) the corresponding CBRAM synapse connecting the two neurons, has a given *probability* to switch into the ON-state (probabilistic LTP). If not, the CBRAM has a given probability to switch to the OFF-state (probabilistic LTD).

Synaptic programming can be implemented using specific voltage pulses. The case without selector device is straightforward (Fig. 3.11(a)). After an output neuron spikes, it generates a specific voltage waveform (signal (3)). Additionally, the input neurons apply signal (1) if they were active recently (within the LTP time window), else they apply signal (2). The conjunction of the input and output waveforms implements STDP. In the case with selector devices (Fig. 3.11(b)), the gates are connected to the output neurons as shown. When an output neurons spikes (fires), it applies a specific voltage waveform to the gates of the selector devices (signal (3)), while non-spiking output neurons will apply signal (4) on the corresponding gates. The input neurons apply pulses similar to the case without selector devices (i.e. signals (1) and (2)). The above described signaling mechanism leads to change in synaptic conductance but does not account for probabilistic or stochastic switching. Probabilistic switching can be implemented in two ways (see Fig.3.13):

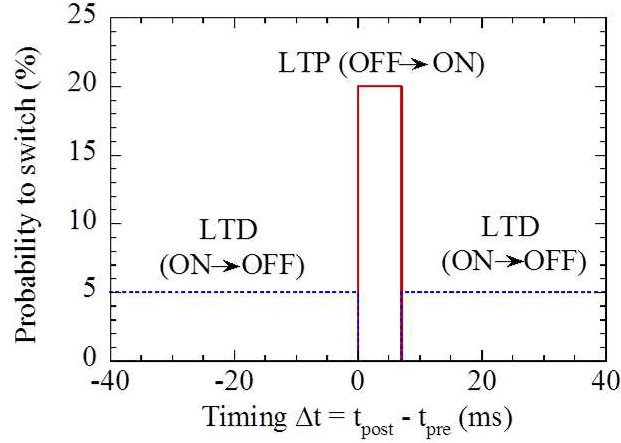
- Extrinsically, by multiplying the signal of the input spiking neuron with the PRNG output, whose signal probability can be tuned by combining with logical AND and OR operations several independent PRNGs, that can be implemented



**Figure 3.11:** (a) Circuit schematic with CBRAM synapses without selector devices, LIF neurons, in the external probability case. (b) Circuit schematic with CBRAM synapses with selector devices, LIF neurons, in the external probability case. In both cases, the presented voltages waveforms implement the simplified STDP learning rule for the CBRAMs [130].

### 3. FILAMENTARY-SWITCHING TYPE SYNAPSES

---

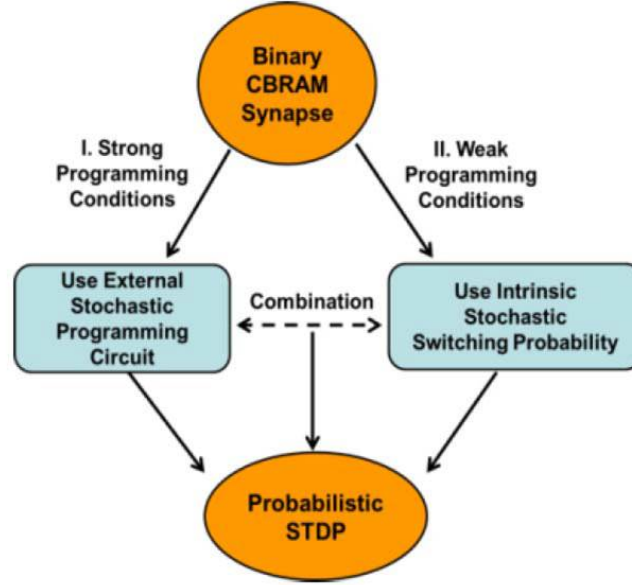


**Figure 3.12:** Probabilistic STDP learning rule (used for audio application). X-axis shows the time difference of post-and pre-neuron spike [130].

for example with linear feedback shift registers (LFSR) [131] (see Fig.3.14). This approach is illustrated in Fig. 3.11. The PRNG output allows or blocks the input neuron signals according to the defined probability levels.

- Intrinsically, by using weak programming conditions (Figures 3.8 and 3.10). In this case, the input neuron applies a weak programming signal, which leads to probabilistic switching in the CBRAM devices.

Exploiting the intrinsic CBRAM switching probability avoids the presence of the PRNG circuits, thus saving important silicon footprint. It also reduces the programming power, as the programming pulses are weaker compared to the ones used for deterministic switching. However it might be difficult to precisely control the switching probability of individual synapses using weak-conditions in a large-scale system. When weak programming conditions are used, both 'device-to-device' and 'cycle-to-cycle' variations contribute to probabilistic switching. Decoupling the effect of the two types of variations is not straightforward in filamentary type of devices (due to the spread on left-over filament height post-reset). In order to precisely control the switching probability a better understanding and modeling of the device phenomena at weak programming conditions is required. If precise values of switching probability



**Figure 3.13:** Schematic showing two different approaches for implementing stochasticity with binary synapses [130].

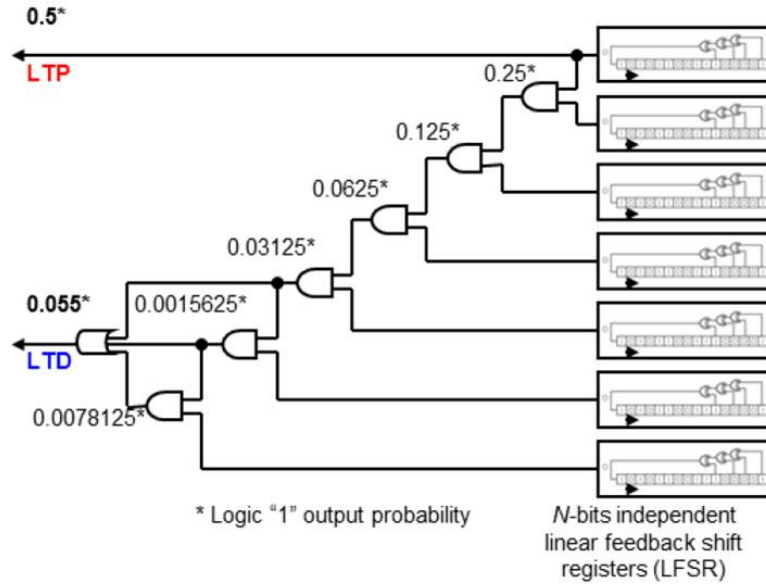
are desired then extrinsic PRNG circuits should be used. For instance a 2-bit PRNG control signal as shown in Fig. 3.11 can be used to separately tune the LTP and LTD probability.

The core with and without selector devices are equivalent from a functional point of view. Selector-free configuration is the most compact ( $4F^2$ ) and highest CBRAM integration density can be obtained with it. Although adding selector element consumes more area ( $>4F^2$ ), it helps to reduce the sneak-path leakage and unwanted device disturbs during the STDP operation which are difficult to control with just 1R devices. Since we did not fabricate a full test chip to measure the leakage and disturb effects in the 1R case, the simulations described in Section IV are based on synaptic programming methodology with-selector devices (1T-1R).

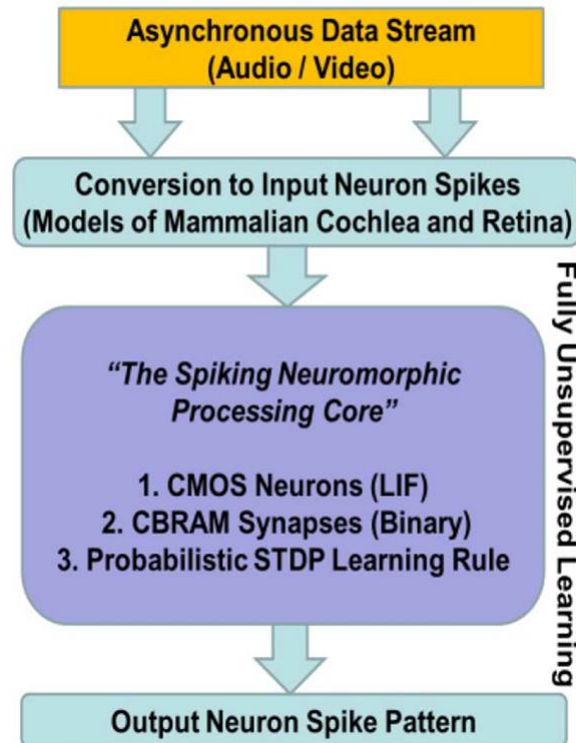
### 3.1.6 Auditory and Visual Processing Simulations

We performed full system-level simulations with our special purpose event-based Xnet simulator tool. The neuron circuits are modeled with behavioral equations as in [116],[134]. The synapses are modeled by fitting data of Fig. 3.6 and Fig. 3.9 with a log-normal distribution, in order to take into account the experimental spread in the

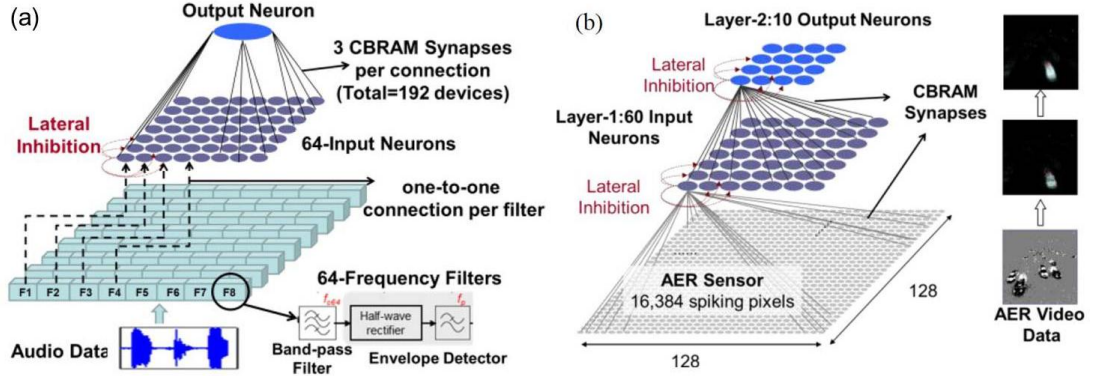
### 3. FILAMENTARY-SWITCHING TYPE SYNAPSES



**Figure 3.14:** Tunable Pseudo-random-number generator (PRNG) circuit [131], the output being tuned according to STDP in Fig.3.12.



**Figure 3.15:** Concept and data-flow of the unsupervised learning simulations [130].

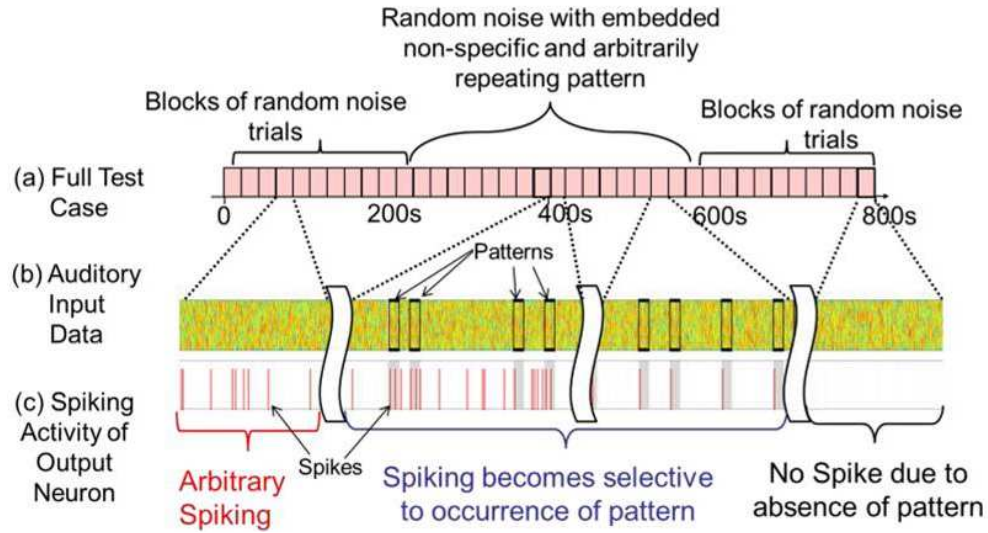


**Figure 3.16:** (a) Single-layer SNN simulated for auditory processing.(b) 2-layer SNN for visual processing.(Right) AER video data snapshot with neuron sensitivity maps [130].

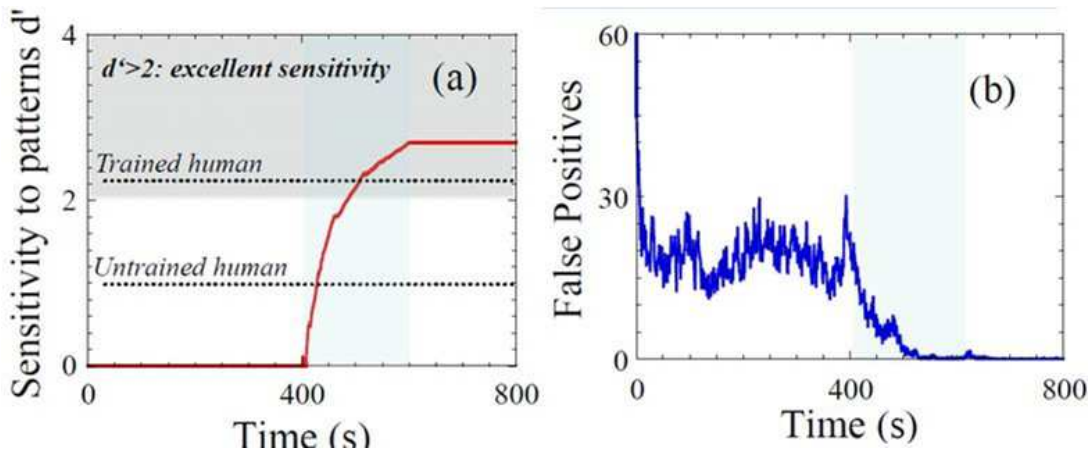
conductance parameters. Effect of both 'device-to-device' and 'cycle to cycle' variations are captured in the synapse model. Fig.3.15 summarizes the concept and data-flow path in our simulations. Two different SNN were used to process auditory and visual data. Fig. 3.16a shows the network designed to learn, extract, and recognize hidden patterns in auditory data. Temporally encoded auditory data is filtered and processed using a 64-channel silicon cochlea emulator (similar to [144], simulated within Xnet). The processed data is then presented to a single layer feed-forward SNN with 192-CBRAM synapses (i.e. every channel of the cochlea is connected to the output neuron by 3 CBRAM synapses).

Initially (from 0 to 400s), gaussian audio noise is used as input to the system, and the firing pattern of the output neuron is completely random (as seen in Fig. 3.17). Then (from 400 to 600s), an arbitrarily created pattern is embedded in the input noise data and repeated at random intervals. Within this time frame, the output neuron starts to spike predominantly when the pattern occurs, before becoming entirely selective to it at the end of the sequence. This is well seen on the sensitivity  $d'$  (a standard measurement in signal detection theory) presented in Fig. 3.18a, which grows from 0 to 2.7. By comparison, a trained human on the same problem achieves a sensitivity of approximately 2 [145]. During the same period, the number of false positives also decreases to nearly 0 (Fig. 3.18b). At the end of the test case (from 600 to 800s), pure noise (without embedded patterns) is again presented to the system. As expected, the output neuron does not activate at all, i.e. no false positive is seen (Fig. 3.17,3.18).

### 3. FILAMENTARY-SWITCHING TYPE SYNAPSES

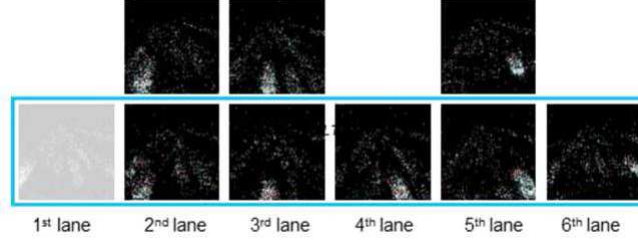


**Figure 3.17:** (a) Full auditory-data test case with noise and embedded repeated patterns. (b) Auditory input data and (c) spiking activity for selected time intervals of the full test case of the output neuron (shown in Fig.16b) [130].



**Figure 3.18:** (a) Pattern Sensitivity ( $d'$ ) for the test case shown in fig. 3.17. The system reaches a very high sensitivity ( $d' > 2$ ). (b) Number of false detections by the output neuron during the auditory learning [130].





**Figure 3.19:** Final sensitivity map of 9 output neurons from the 1st layer of the neural network shown in Fig.17b. Average detection rate for 5 lanes was 95% [130].

The total synaptic learning power ( $P_{learning}$ ) consumption (i.e. the power required to read, write and erase the CBRAMs) was extremely low ( $0.55 \mu\text{W}$  in the extrinsic probability case,  $0.15 \mu\text{W}$  in the intrinsic probability case). The estimation of synaptic learning power is described in detail in Tab.1[130], Eq.3.1-3.3, were used:

$$E_{set/reset} = V_{set/reset} \times I_{set/reset} \times t_{pulse} \quad (3.1)$$

$$E_{total} = (E_{set} \times totalsetevents) + (E_{reset} \times totalresetevents) \quad (3.2)$$

$$P_{learning} = \frac{E_{total}}{T_{learning}} \quad (3.3)$$

In the extrinsic probability case, about 90% of the energy was used to program the CBRAM devices, and about 10% to read them (while in the case of intrinsic probability it was about 81% and 19% respectively). The sound pattern extraction example can act as a prototype for implementing more complex applications such as speech recognition and sound-source localization. Fig. 3.16b shows the network simulated to process temporally encoded video data, recorded directly from an artificial silicon retina [117]. A video of cars passing on a freeway recorded in address-event-representation (AER) format by the authors of [117] is presented to a 2-layered SNN. In each layer, every input is connected to every output by a single CBRAM synapse.

The CBRAM based system learns to recognize the driving lanes, extract car-shapes (Fig. 3.19) and orientations, with more than 95% average detection rate. The total synaptic-power dissipation was  $74.2 \mu\text{W}$ , in the extrinsic probability case and  $21 \mu\text{W}$  in the intrinsic probability case. This detection rate is similar to the one that we simulated on the same video test case with a deterministic system based on multi-level PCM synapses [110],[81],[111]. The example SNN on visual pattern extraction, shown



### 3. FILAMENTARY-SWITCHING TYPE SYNAPSES

---

**Table 3.1:** Network statistics for auditory and visual learning simulations with stochastic binary CBRAM synapses [130].

Auditory Test Case	
Total Set events	102646
Total Reset events	41810
Total Read events	$2.10 \times 10^7$
Total CBRAM synapses	192
Visual Test Case	
Total Set events	449725
Total Reset events	26837412
Total Read events	$2.49 \times 10^9$
Total CBRAM synapses	about 2 million
$(I_{SET} = \mu\text{A})$	$(I_{RESET} = 90 \mu\text{A})$

here, can be used as a prototype to realize more complex functions such as image classification [45],[116], position detection and target-tracking. Tab.3.1, and tab.3.2 summarize all the network and energy statistics from the two unsupervised learning simulations.

We tested the two test applications with both extrinsic and intrinsic probability programming methodologies. Sensitivity and detection rates were nearly identical in both cases, which suggests a relative equivalence of the two approaches. Total synaptic power consumption was lower when the intrinsic probability methodology was used. This suggests that the power saved by using weak programming pulses is greater than the power dissipated due to the extra programming pulses required to implement the intrinsic probability. Additionally, we performed simulations without any intrinsic or extrinsic conductance spreads (ideal or non-variable synapses). These gave sensitivity values and detection rates similar to the ones when the spread was considered, suggesting that the experimentally measured variability in our devices had no significant impact on the overall system learning performance. This is consistent with variability-tolerance of STDP-based networks [116].

**Table 3.2:** Energy/Power statistics for auditory and visual learning simulations with stochastic binary CBRAM synapses [130].

Auditory Test Case	
Total duration	800 s
Total Energy dissipated	436 $\mu$ J
Synaptic prog. power	0.55 $\mu$ W
Visual Test Case	
Total duration	680 s
Total Energy dissipated	50.4 mJ
Synaptic prog. power	74.2 $\mu$ W

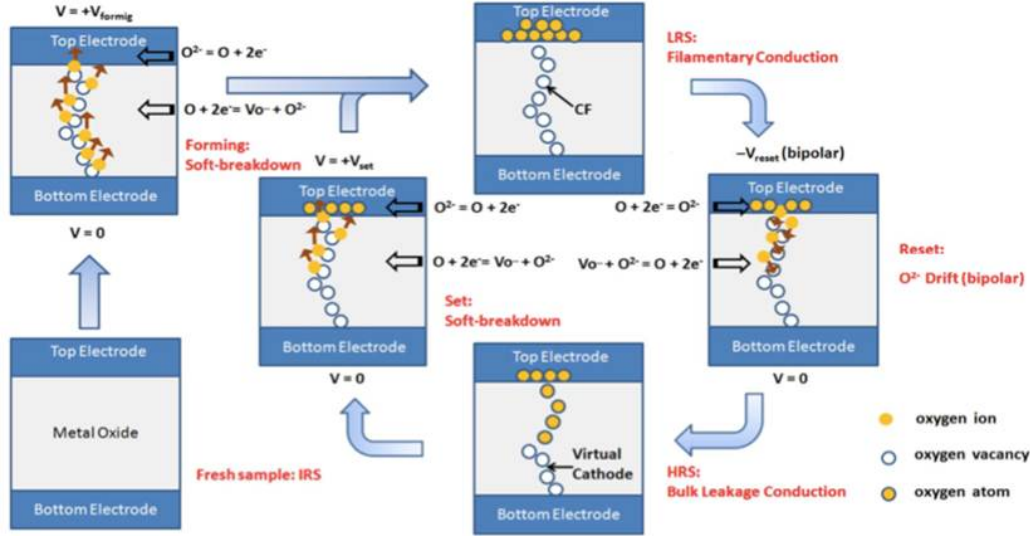
## 3.2 OXRAM Technology

Metal-oxide based resistive memory (OXRAM) has created considerable amount of interest in the non-volatile memory community in recent years [74]. Based on the choice of material stack, OXRAM devices can either function as bipolar or unipolar. The underlying physics of OXRAM devices is still debated and not completely understood, however the resistive-switching phenomenon is widely understood to be a consequence of the dynamics of oxygen defects and vacancies. In the case of bipolar devices (more widely studied), when a forming voltage is applied to a fresh device, a soft-breakdown occurs. Oxygen ions drift to the top electrode due to high electric field, where they accumulate leading to an oxygen reservoir. A conductive filament (CF) composed of oxygen vacancies is thus formed switching the device in the ON state (see Fig.3.20). In order to switch the device to OFF state, a voltage of reversed polarity is applied and oxygen ions drift back into the oxide layer, where they recombine with oxygen vacancies and disrupt the CF [74].

### 3.2.1 State-of-art OXRAM Synapses

Several material stacks, (based on Ti,  $\text{AlO}_x$ , TiN,  $\text{HfO}_x$ ,  $\text{TiO}_x$ , Pt, W, Al) and different programming strategies (such as increasing reset/set voltages, increasing compliance current, identical reset/set pulses) have been employed to demonstrate the emulation of LTP- and LTD- like effects in OXRAM devices. Tab3.3 and Fig.3.21, summarize some of these recent approaches and references.

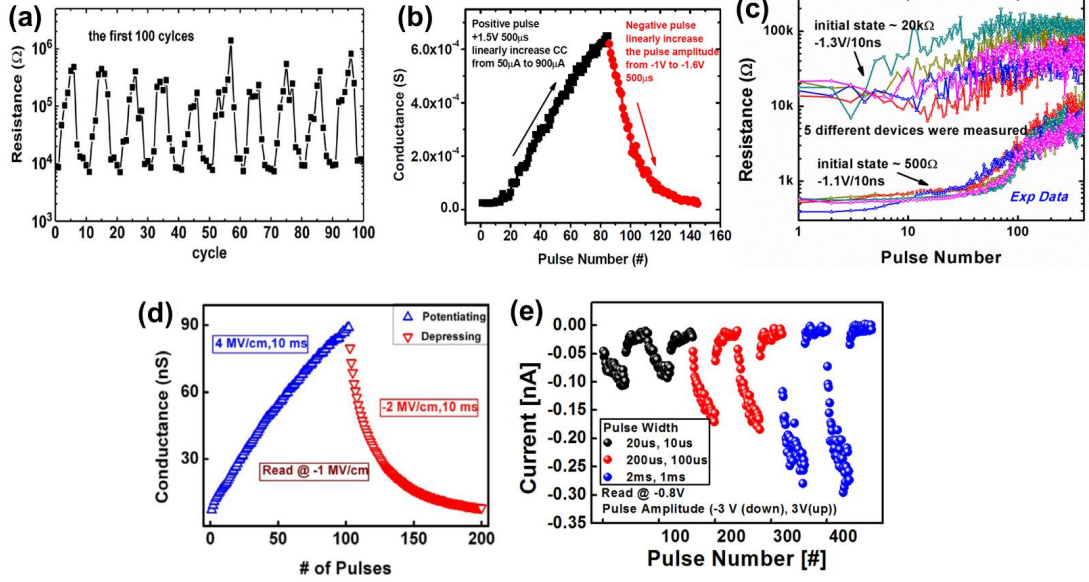
### 3. FILAMENTARY-SWITCHING TYPE SYNAPSES



**Figure 3.20:** Illustration depicting OXRAM working principle and underlying physics [74].

**Table 3.3:** Recent OXRAM based synaptic emulations

Ref	Material Stack	LTP Strategy	LTD Strategy
[146]	TiN/Ti/AlO <sub>x</sub> /TiN	Increasing set pulse amplitude	Increasing reset pulse amplitude
[147]	Ti/AlO <sub>x</sub> /TiN	Increasing Compliance Current	Increasing reset pulse amplitude
[148]	TiN/TiO <sub>x</sub> /HfO <sub>x</sub> /TiO <sub>x</sub> /HfO <sub>x</sub> /Pt	Not implemented	Identical reset pulses
[149]	Pt/Al/TiO <sub>2-x</sub> /TiO <sub>y</sub> /W	Identical set pulses	Identical reset pulses
[150]	W/Al/PCMO/Pt	Identical set pulses	Identical reset pulses



**Figure 3.21:** (a) LTD obtained by increasing RESET pulse, LTP by increasing SET pulse. [146] (b) LTD obtained by increasing the RESET pulse amplitude, LTP by increasing compliance current [147]. (c) LTD only, obtained by identical RESET pulses [148]. (d) and (e): LTD and LTP obtained by identical RESET and SET voltage [149], [150].

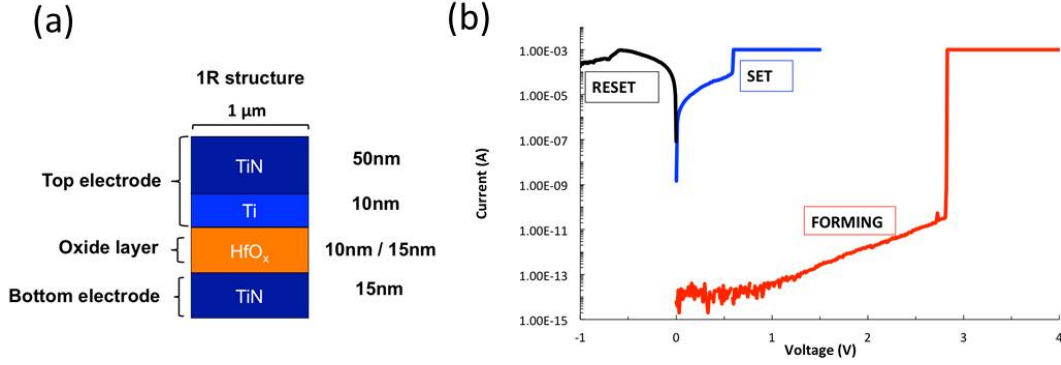
### 3.2.2 Device and Electrical Characterization

We fabricated and studied OXRAM devices composed of a TiN/Ti/HfO<sub>x</sub>/TiN stack (Fig.3.22a) with two different HfO<sub>x</sub> layer thickness (10 and 15 nm). The tested devices were composed of 1R structure, i.e. with direct access to top and bottom electrodes, without any selector device. Fig.3.22b shows the typical current-voltage (I-V) characteristic for our devices obtained in quasi-static mode (dc-bias mode) indicating the forming, RESET, and SET operations. In the forming process a voltage staircase from 0 to 4 V with a step of 0.02 V (with a current compliance of 1 mA) is applied to the devices.

### 3.2.3 LTD Experiments: R<sub>OFF</sub> Modulation

We observed that gradual R<sub>OFF</sub> modulation in our HfO<sub>x</sub> devices is reproducibly possible only with the application of varying reset conditions. Fig.3.23 shows results of R<sub>OFF</sub> modulation in quasi-static mode. Fig.3.23a, shows in detail the implemented test case. Initially formed devices are programmed to a low-resistance SET state, by apply-

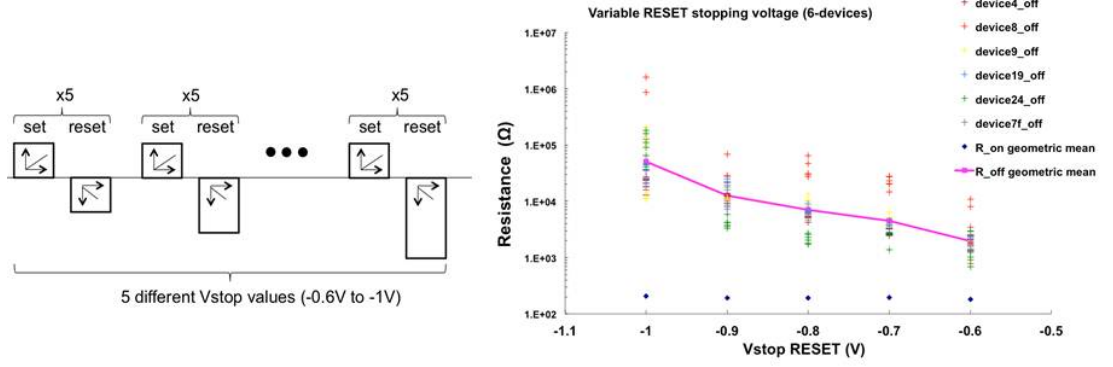
### 3. FILAMENTARY-SWITCHING TYPE SYNAPSES



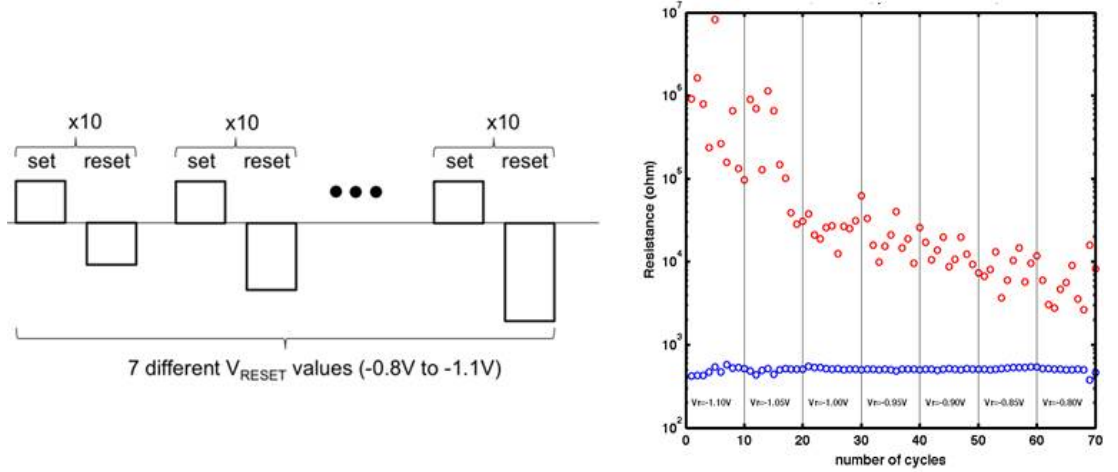
**Figure 3.22:** (a) Schematic view of the tested OXRAM memory device. (b) Typical I-V characteristic for tested OXRAM devices (10 nm HfO<sub>x</sub>).

ing a dc-SET voltage ramp. This is followed by the application of several subsequent cycles of SET/RESET operations, with a fixed SET condition, but changing RESET condition. The device resistance is read, and recorded, by applying a small reading voltage (0.1 V) between each consecutive operation.  $R_{OFF}$  modulation is achieved by changing the stopping voltage  $V_{stop}$  of the dc voltage ramp applied during the RESET operation. 5 different values of  $V_{stop}$  (in the range -0.6 V to -1 V) give rise to 5 different RESET or  $R_{OFF}$  states (Fig.3.23b). Current compliance is not required during the RESET operation, however for SET operations a compliance current ( $I_C = 1$  mA) is imposed to protect the devices from excessive flow of current in the low-resistance state and device failure. For all SET operations the  $V_{stop} = 2.5$  V. The SET operations, between two consecutive RESET operations, act as a 'refresh' operation, as they restore the devices to same initial state.

Similar tests were also performed in pulsed-mode operation. Pulsed-mode operations are more relevant for spiking neural networks. Fig.3.24a shows the test case schematic. During SET operations, a programming voltage pulse ( $V_{SET}$ ) is applied on the series combination of the OXRAM device and a 1 kΩ load resistor used as a current limiter. During RESET operations a negative voltage pulse  $V_{RESET}$  (in the range -0.8 V to -1.1 V) was applied directly on the top electrode. For each  $V_{RESET}$  value, devices were cycled 10 times with subsequent SET and RESET operations.  $V_{SET}$  in all cases was + 1.8 V. The pulse widths for both  $V_{RESET}$  and  $V_{SET} = 50$  μs. Fig.3.24b shows clear evidence of  $R_{OFF}$  modulation with  $V_{RESET}$ .

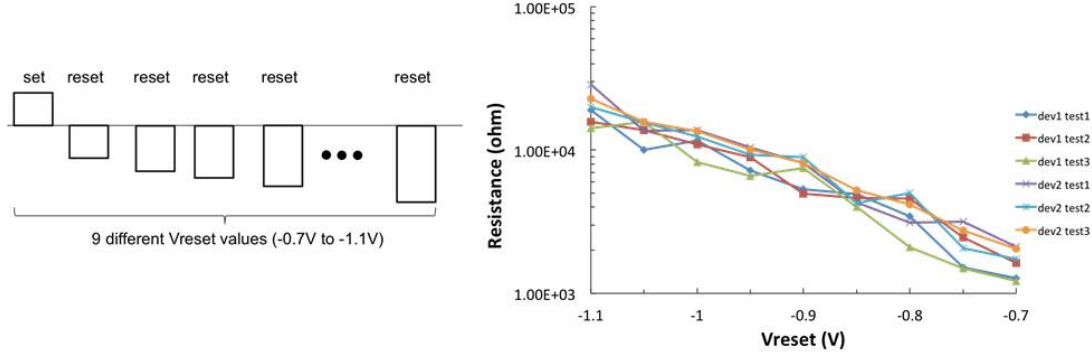


**Figure 3.23:** (a) Schematic description of test case for  $R_{OFF}$  modulation in quasi-static mode with SET 'refresh' operation. (b) Experimental data confirming  $R_{OFF}$  modulation in quasi-static mode (10 nm  $\text{HfO}_x$ ).



**Figure 3.24:** (left) Schematic description of test for  $R_{OFF}$  modulation in pulsed-mode with SET 'refresh' operation. (right) Experimental data confirming  $R_{OFF}$  modulation in pulsed-mode (10 nm  $\text{HfO}_x$ ).

### 3. FILAMENTARY-SWITCHING TYPE SYNAPSES

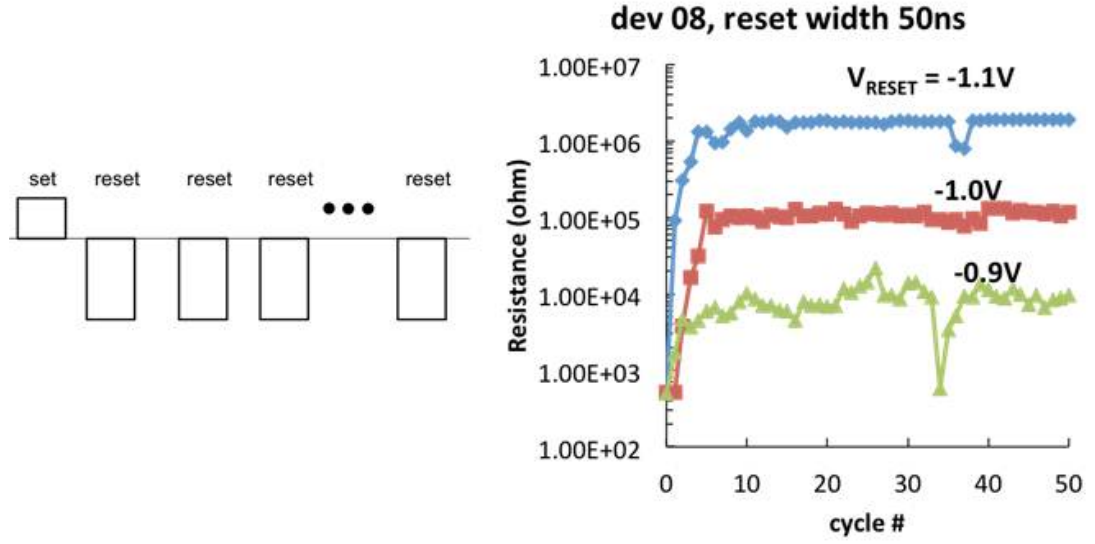


**Figure 3.25:** (left) Schematic description of test for  $R_{OFF}$  modulation in pulse mode without SET 'refresh' operation. (right) Experimental data confirming  $R_{OFF}$  modulation in pulse mode (10 nm  $\text{HfO}_x$ ).

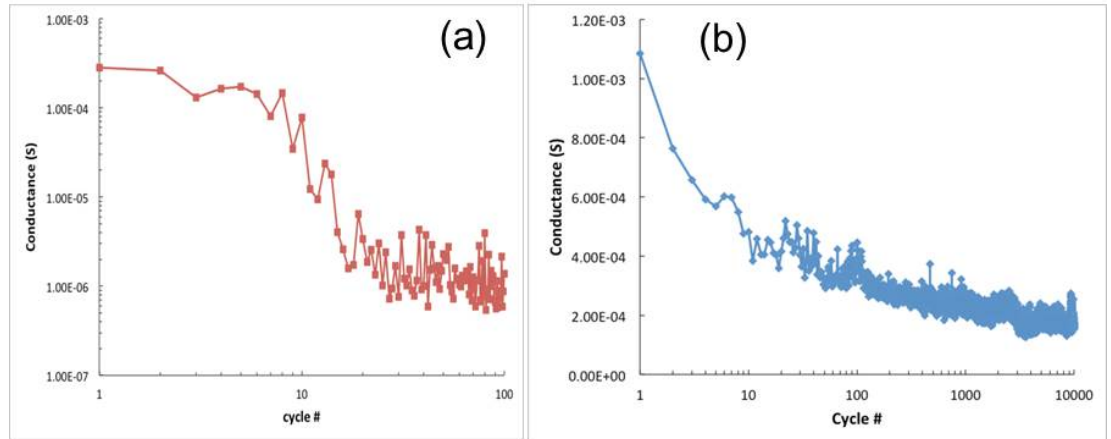
Fig.3.25, shows the results for a pulsed-mode  $R_{OFF}$  modulation test performed without the refreshing SET operation between two consecutive RESET operations. After an initial SET pulse of amplitude + 1.8 V and width 50  $\mu\text{s}$ , 9 consecutive RESET pulses of increasing amplitude and width 50  $\mu\text{s}$  were applied. The  $R_{OFF}$  modulation trend was consistent with tests shown in Fig.3.24 and Fig.3.23.

Tests with identical reset pulses didn't show very promising results in terms of gradual  $R_{OFF}$  modulation. The behavior was similar to the one observed in the case of CBRAM devices. Fig.3.26a, describes one such test case. After an initial SET operation (1.8 V, 50  $\mu\text{s}$ ), a sequence of 50 reset pulses of 50 ns width and given pulse amplitude were applied. Fig.3.26b shows the typical response of the tested devices: after very few intermediate states (3-5), the  $R_{OFF}$  value saturates to a ceiling value, that depends on the amplitude of the RESET voltage.

However, very few devices showed a gradual increase of  $R_{OFF}$  with identical RESET pulses. Fig.3.27a shows a gradual change in conductance of around 3 orders of magnitude in response to 50 subsequent reset pulses (-1 V, 50 ns) for a 10 nm thick  $\text{HfO}_x$  layer device. Fig.3.27b shows similar results for a 15 nm thick  $\text{HfO}_x$  layer device. An interesting observation is that, compared to the 10 nm thick  $\text{HfO}_x$ , the change in resistance is less than one decade, but the response is much slower (requiring 10000 reset pulses). The underlying physics for such behavior is not understood and still a matter of investigation, moreover the results shown in Fig.3.27 were not very reproducible. If in future, such effects can be obtained reproducibly, a "2-OXRAM Synapse" type



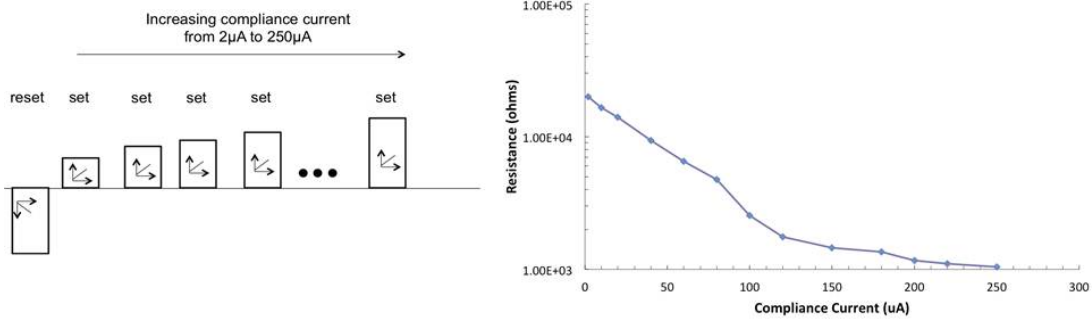
**Figure 3.26:** (left) Schematic description of test for ROFF modulation in pulse mode with identical reset-pulses. (right) Experimental data for the described test with identical reset pulses (10 nm  $\text{HfO}_x$ ).



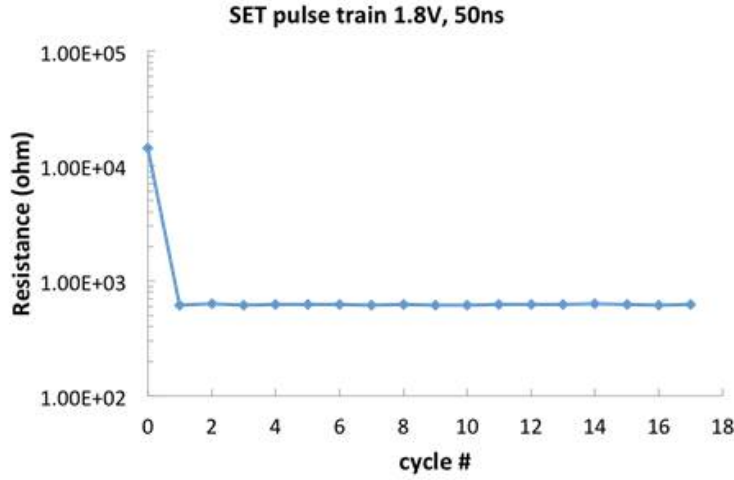
**Figure 3.27:** Cumulative device response in conductance to identical RESET pulse train for (a) 10 nm thick and (b) 15 nm thick  $\text{HfO}_x$  layer devices.



### 3. FILAMENTARY-SWITCHING TYPE SYNAPSES



**Figure 3.28:** Schematic description of test for  $R_{ON}$  modulation in quasi-static mode. (b) Experimental data showing the trend in  $R_{ON}$ .



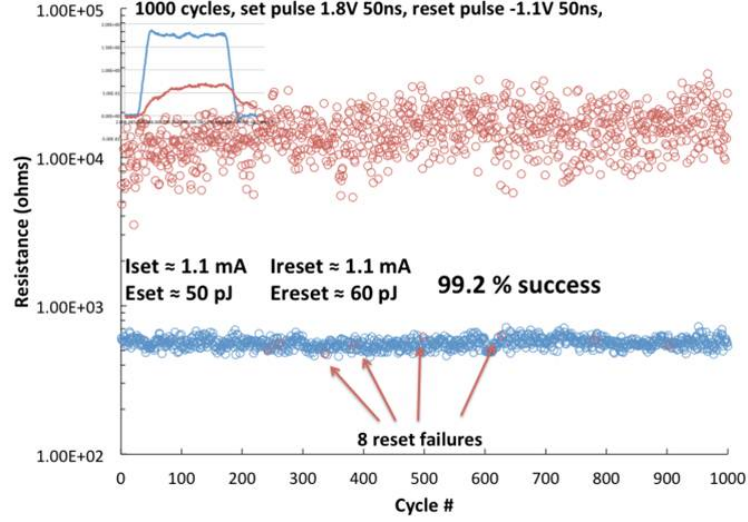
**Figure 3.29:** Device response to a pulse train of identical SET pulses.

of implementation (similar, but inverse of the "2-PCM Synapse" approach), primarily based on LTD can be envisioned.

#### 3.2.4 LTP Experiments: $R_{ON}$ modulation

LTP type of behavior or  $R_{ON}$  modulation was investigated in quasi-static mode. As shown in the schematic of Fig.3.28a, after an initial RESET operation, the device was subjected to a sequence of SET operations while increasing the imposed compliance current ( $I_C$ ) in the circuit.  $R_{ON}$  modulation of about 1 decade is possible in quasi-static mode (Fig.3.28b).

We also performed tests to check the response of device towards identical SET



**Figure 3.30:** Binary switching operation for our OXRAM devices with 'strong' programming conditions.

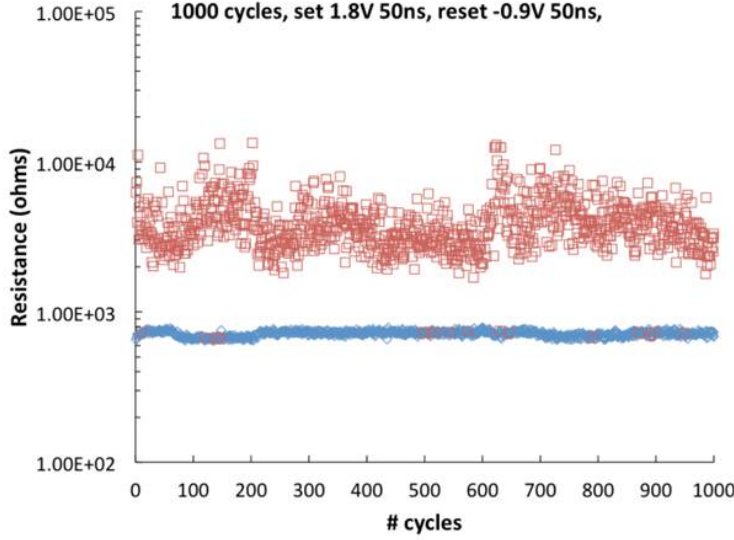
pulses. However, as shown in Fig.3.29, after the first SET operation no effect of the subsequent SET pulses is observed. After the first switch to low-resistance state, in all subsequent SET operations most of the applied voltage drops across the 1 k $\Omega$  series current limiting resistor, thus leaving the state of the OXRAM device almost unaltered. The failure to gradually modulate  $R_{ON}$  with identical pulses, in OXRAM, is similar to the one observed in case of CBRAM devices (Sec.3.1.2).

### 3.2.5 Binary operation

In the previous section, we showed that our OXRAM devices with  $\text{HfO}_x$  layer were not good enough for gradual LTP/LTD type response with application of identical programming pulses. We thus prefer to use OXRAM devices as binary synapses with stochastic learning rules, as also proposed in the case of CBRAM devices. Fig.3.30, shows a OXRAM binary cycling test with strong programming conditions. A pulse width of 50 ns was used for both SET and RESET operations.

In order to investigate if it was possible to reproduce CBRAM like intrinsic-stochasticity with OXRAM devices, we used weak RESET conditions (i.e  $V_{RESET}$  was decreased). From Fig.3.31 we see that rather than inducing RESET failures, the only thing that

### 3. FILAMENTARY-SWITCHING TYPE SYNAPSES



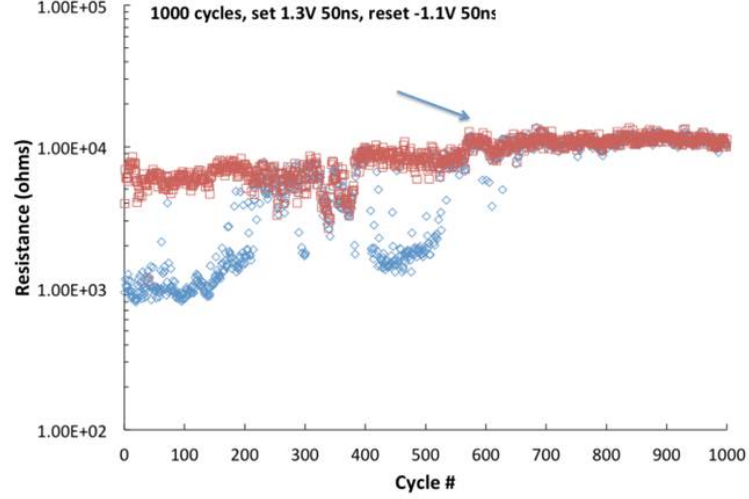
**Figure 3.31:** Binary switching operation for our OXRAM devices with 'weak' RESET programming conditions.

changes compared to Fig.3.30 is the level of  $R_{OFF}$ , thus decreasing the resistance programming window.

We then studied the effect of using weak SET conditions. Fig.3.32 shows that decreasing the applied voltage actually results in switching failures during the SET operation. However, these failures are mostly concentrated in one region of the cycling curve and not truly stochastically distributed. We thus propose to use strong programming conditions alongside the use of extrinsic PRNG circuits to externally implement stochastic-STDP, rather than relying on the use of weak programming conditions.

#### 3.2.6 Learning Simulations

We performed the cars visual pattern extraction simulations (similar to the one described in sec.3.1.6) using OXRAM binary switching data shown in Fig.3.30 and stochastic STDP rule (Fig.3.12). The learning and power statistics of the simulations are summarized in Tab.3.4 and Tab.3.5. The cars average detection rate was similar to the one for CBRAM ( $> 90\%$ ). However the total energy dissipation and power consumption was much less compared to both PCM and CBRAM devices (Total Energy = 1.6 mJ, Total Power = 2.4  $\mu$ W).



**Figure 3.32:** Binary switching operation for our OXRAM devices with 'weak' SET programming conditions.

**Table 3.4:** Learning statistics, over the whole learning duration for binary OXRAM synapses ( $8 \times 85 = 680$  s).

	/device	/device (max)	/device/s	Overall
HfO <sub>x</sub> OXRAM devices				
Read pulses	1,265	160,488	1.9	$4.97 \times 10^9$
SET pulses	0.34	32	0.0005	$6.67 \times 10^5$
RESET pulses	11	39	0.016	$2.12 \times 10^7$

**Table 3.5:** Energy/Power statistics for visual learning simulations with stochastic binary OXRAM synapses (Total learning duration = 680 s ).

Read Energy	0.04 mJ
SET Energy	0.26 mJ
RESET Energy	1.3 mJ
Total Energy dissipated	1.6 mJ
Synaptic prog. power	2.4 $\mu$ W
( $I_{SET} = 1.1$ mA)	( $I_{RESET} = 1.1$ mA)

#### 3.3 Conclusion

We proposed for the very first time a bio-inspired system with binary CBRAM synapses and stochastic STDP learning rule able to process asynchronous analog data streams for recognition and extraction of repetitive patterns in a fully unsupervised way. The demonstrated applications exhibit very high performance (auditory pattern sensitivity  $>2.5$ , video detection rate  $>95\%$ ) and ultra-low synaptic power dissipation (audio  $0.55\mu\text{W}$ , video  $74.2\mu\text{W}$ ) in the learning mode. We show different programming strategies for 1R and 1T-1R based CBRAM configurations. Intrinsic and extrinsic programming methodology for CBRAM synapses is also discussed.

We briefly investigated the possibility of implementing synaptic behavior with 1R (selector-free) OXRAM devices. Modulation of  $R_{OFF}$  both in quasi-static and pulse-mode was confirmed. Modulation of  $R_{ON}$  was demonstrated only in quasi-static mode due to the difficulty in controlling the compliance current in pulse-mode with 1R structures. Results suggest that the studied technology could be successfully used in pulsed neural networks by exploiting the gradual increase of the  $R_{OFF}$  value, however at the cost of varying or non-identical programming pulses. We also investigated binary switching characteristics of the devices. Difficulty in controlling the distribution of switching failures with weak programming conditions make it impractical to exploit the intrinsic switching probability. However, our OXRAM technology can be successfully used in probabilistic networks by using strong-programming conditions and introducing extrinsic stochasticity.

“Silence is the language of nature...  
all else is poor translation.”  
-Rumi

## 4

# Using RRAM for Neuron Design

This chapter discusses how RRAM devices can be used to design innovative neuron structures. We present an original methodology to design hybrid neuron circuits (CMOS + non volatile resistive memory) with stochastic firing behaviour. In order to implement stochastic firing, we exploit unavoidable intrinsic variability occurring in emerging non-volatile resistive memory technologies. In particular, we use the variability on the ‘time-to-set’ ( $t_{\text{set}}$ ) and ‘off-state resistance’ ( $R_{\text{OFF}}$ ) of Ag/GeS<sub>2</sub> based Conductive Bridge (CBRAM) memory devices. We propose a circuit and a novel self-programming technique for using CBRAM devices inside standard Integrate and Fire neurons. Our proposed solution is extremely compact with an additional area overhead of 1R-3T. The additional energy consumption to implement stochasticity in Integrate and Fire neurons is dominated by the CBRAM set-process.

## 4.1 Introduction

Neuromorphic computing is usually accomplished with deterministic devices and circuits. However, literature in the fields of neural networks [151],[152] and of biology [153] suggests that in many situations, actually providing a certain degree of stochastic, noisy or probabilistic behavior in their building blocks may enhance the capability and stability of neuroinspired systems. Some kind of neural networks even fundamentally rely on stochastic neurons, like Boltzmann machines [154],[155]. Finally stochastic neurons may perform signal processing in extremely noisy environments using a phenomenon known as ‘stochastic resonance’ [156],[157]. In Chapter.2 and Chapter.3, we showed

## 4. USING RRAM FOR NEURON DESIGN

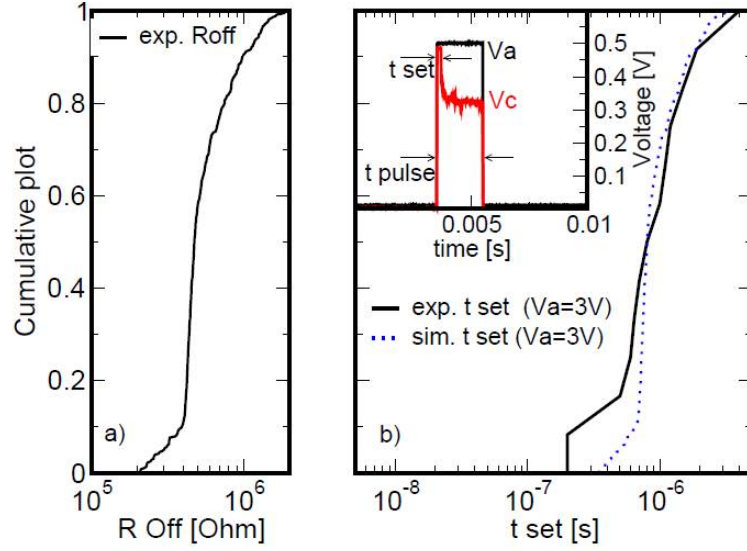
---

how unsupervised learning can be achieved with the help of stochastic learning rules and binary RRAM synapses.

Providing extrinsic stochastic behavior to neurons using pseudo-random number generator circuits (PRNG) leads to significant silicon area/power overheads. This explains interest in developing silicon neurons with an intrinsic stochastic behavior, but which may be controlled. In previous works, different techniques to implement controlled stochasticity in hardware neural networks have proposed. It is possible to exploit the thermal noise in the CMOS but this may lead to silicon overheads and unwanted correlations [151]. Other techniques exploit CMOS circuits with using noise but have significant area overhead [158], or the noise of photons with photodetectors [159] or even special kinds of ‘noisy transistors’ [160]. Finally it was proposed to use fundamentally probabilistic nanodevices like single electron transistors [161], but which might suffer from poor CMOS compatibility and room temperature operation. In this chapter, we describe an original circuit and methodology to design a neuron with stochastic firing behavior exploiting certain physical effects of emerging non-volatile resistive memory technology devices such as Conductive Bridge memory (CBRAM). There are significant advantages of our approach such as extremely low area overhead and full CMOS compatibility.

### 4.2 CBRAM Stochastic Effects

The basics and working of our CBRAM devices has already been discussed in sec.3.1.2. Here we focus on the spread in  $R_{OFF}$  values for CBRAM devices. By cycling many times our devices a statistical distribution of the high resistive state ( $R_{OFF}$ ) was obtained. Dispersion in  $R_{OFF}$  may be interpreted in terms of stochastic breaking of the filament during the reset process, due to the unavoidable defects close to the filament which act as preferential sites for dissolution. In sec.3.1.4, we showed, with the help of modeling, that a distribution in  $R_{OFF}$  leads to a spread in others physical quantities like the left-over filament height ( $h$ ) and the  $t_{set}$ . Fig. 4.1 inset shows an example of the oscilloscope trace for the evolution of voltage drop across the cell ( $V_c$ ) during a set pulse. Initially, the cell is in the high resistive state ( $R_{OFF} \simeq 10^6 \Omega$ ) and most of the applied voltage drops on the cell. Then at time  $t_{SETan}$  abrupt decrease of  $V_c$  is observed, revealing a sudden drop of the cell resistance corresponding to the switching from high to low



**Figure 4.1:** (a)  $R_{\text{Off}}$  distribution obtained in  $\text{GeS}_2$  based 1R CBRAM devices. (b) Experimental (line) and simulated (dotted)  $t_{\text{SET}}$  distribution obtained cycling the CBRAM cell with a pulse amplitude  $V_a=3\text{ V}$ . (b in the inset) Example of a typical oscilloscope trace tracking the voltage on the CBRAM ( $V_c$ ) and the applied pulse ( $V_a$ ). Between every set operation a reset operation was performed (not shown) [163].

resistive state. Starting from some of the measured values of  $R_{\text{Off}}$  (Fig. 4.1(a)) we collected the spread in  $t_{\text{SET}}$  when the applied pulses were  $V_a=3\text{ V}$  and  $t_{\text{pulse}} = 5\text{ }\mu\text{s}$  (Fig. 4.1(b)). The dotted line in Fig. 4.1(b), shows the simulated values of  $t_{\text{set}}$ . To obtain the simulated curve of  $t_{\text{set}}$ , first the distribution of  $h$  was calculated using [162]:

$$R_{\text{Off}} = \frac{\rho_{\text{on}} h + \rho_{\text{off}}(L - h)}{\pi r^2} \quad (4.1)$$

where  $\rho_{\text{on}}$  is the resistivity of the Ag-rich nanofilament,  $\rho_{\text{off}}$  is the resistivity of the  $\text{GeS}_2$ ,  $L$  is the chalcogenide thickness and  $r$  is the conductive filament radius, then  $t_{\text{SET}}$  using:

$$t_{\text{set}} = \frac{L - h}{v_h \exp\left(\frac{-E_A}{k_B T}\right) \sinh\left(\alpha q \frac{V_c - \Delta}{k_B T}\right)} \quad (4.2)$$

where  $q$  is the elementary charge,  $v_h$  is a fitting parameter for the vertical evolution velocity,  $E_A$  is the activation energy,  $k_B$  is the Boltzmann constant,  $T$  is the temperature (300 K),  $\alpha$  and  $\Delta$  are fitting parameters to take into account vertical electric field



## 4. USING RRAM FOR NEURON DESIGN

---

dependency and the overpotential that controls the kinetic of the cathodic reaction respectively (Table 4.1). In the following section we show how the spread in  $t_{\text{SET}}$  can be used to make the firing of an Integrate and Fire neuron non-deterministic.

### 4.3 Stochastic Neuron Design

#### 4.3.1 Integrate and Fire Neuron

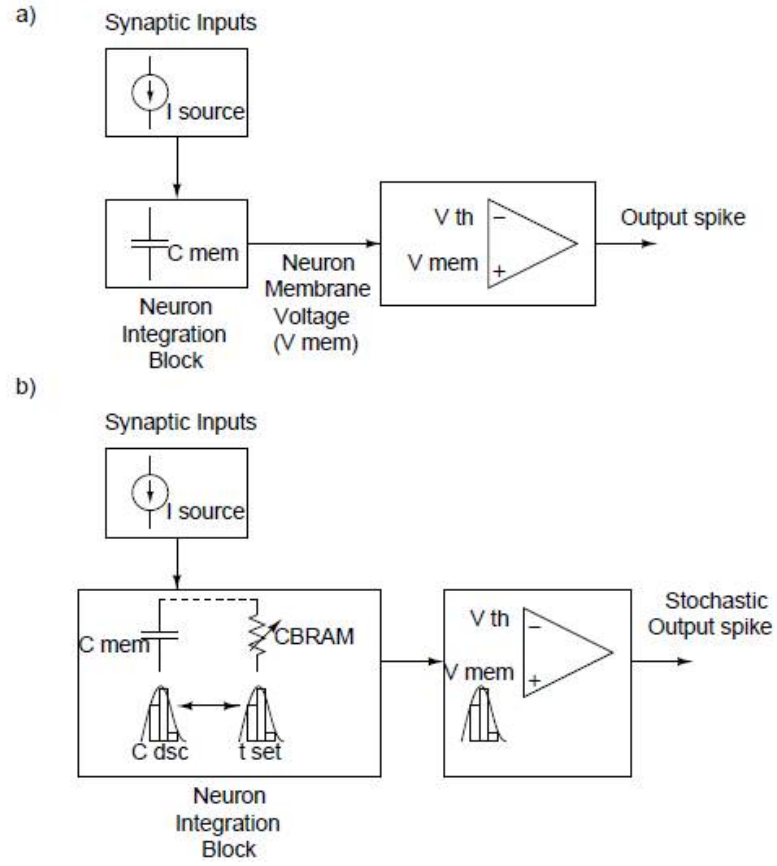
The complexity of a neuron circuit depends on the overall functionality of the neural network and of the chosen biological models. For our purpose of concept validation, we chose one of the simplest, the Integrate and Fire neuron model. Fig. 4.2(a) shows the concept of a simple Integrate and Fire neuron model. It constantly sums (integrates) the incoming synaptic-inputs or currents (excitatory and inhibitory) inside the neuron integration block using a capacitor. More advanced designs also work with this principle [50]. This integration leads to an increase in the membrane potential of the neuron  $V_{\text{mem}}$ . When the membrane potential reaches a certain threshold value  $V_{\text{th}}$ , the neuron generates an output spike (electrical signal). After the neuron has fired the membrane potential goes back to a resting value (initial state), through discharging of the capacitor  $C_{\text{mem}}$ . Usually, the output firing activity of a Integrate and Fire neuron is deterministic because the neuron fires every time the membrane potential reaches a defined threshold value.

#### 4.3.2 Stochastic-Integrate and Fire principle and circuit

To introduce non-deterministic or stochastic behavior in Integrate and Fire neuron, we propose to connect a CBRAM device to the capacitor  $C_{\text{mem}}$ , such that  $C_{\text{mem}}$  could

**Table 4.1:** Parameters used in the simulations

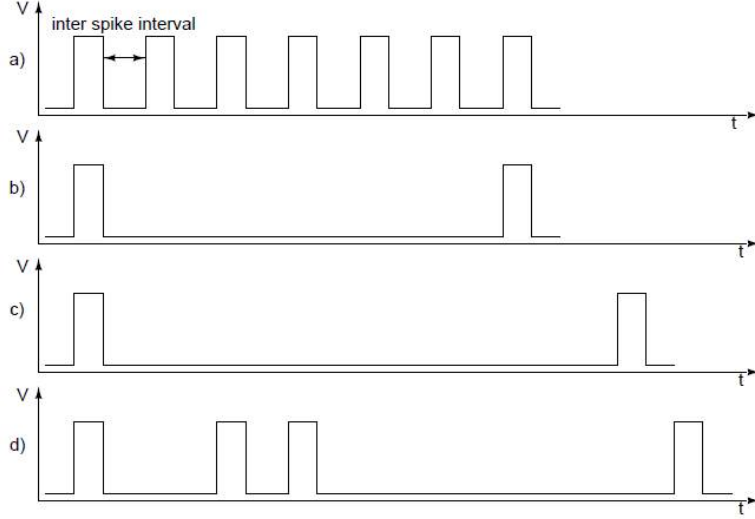
Parameter	Value	Parameter	Value
$v_h$	2 m/s	$E_A$	0.35 eV
$\rho_{\text{on}}$	$2.3 \times 10^{-6} \Omega\text{m}$	$\rho_{\text{off}}$	$10^{-3} \Omega\text{m}$
$\alpha$	0.08	$\Delta$	0.15 V
$r$	2.2 nm	L	30 nm



**Figure 4.2:** (a) Schematic image shown the basic concept of a Integrate and Fire neuron [50]. (b) Schematic showing the basic concept of our proposed Stochastic Integrate-Fire neuron (S-IF) [163].

#### 4. USING RRAM FOR NEURON DESIGN

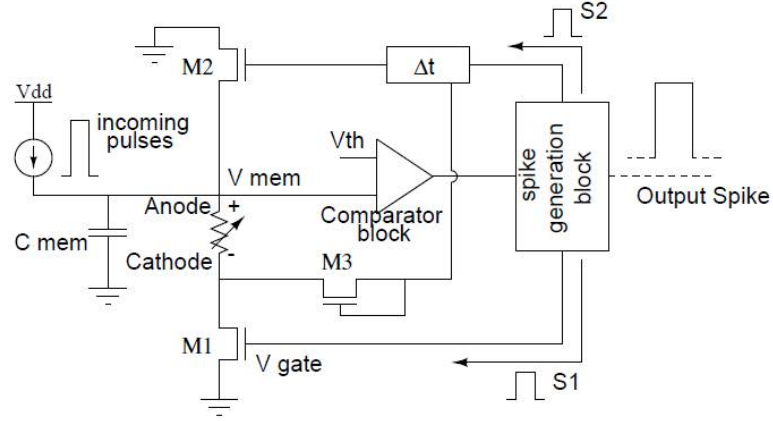
---



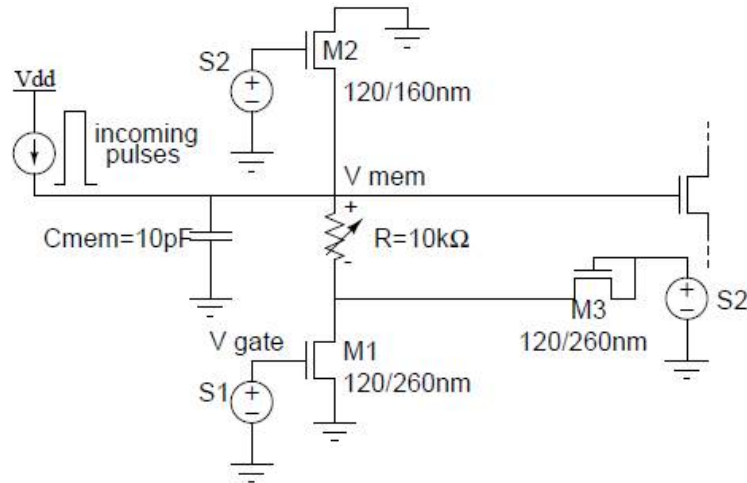
**Figure 4.3:** (a)-(d) Schematic of output neuron firing patterns for different example test cases [163].

only discharge through the CBRAM device by switching it to the low-resistive state (Fig. 4.2(b)). The anode of the CBRAM and the  $V_{\text{mem}}$  net of the capacitor should be connected. The duration for which current can flow through the low-resistive CBRAM device can be controlled using a transistor. In such a configuration, the spread on the  $t_{\text{SET}}$  of the CBRAM would translate to a spread on the discharge-time ( $t_{\text{dsc}}$ ) of the capacitor. For consecutive neuron spikes, this would lead different initial state of  $C_{\text{mem}}$ , thus making the firing of the neuron stochastic. Fig. 4.3 illustrates conceptually the impact of four different values of  $t_{\text{SET}}$  (keeping constant pre-synaptic weights), on the inter-spike interval. In case (a),  $t_{\text{SET}}$  is very long thus the capacitor has a very weak discharge. As a consequence just few additional incoming pre-neuron spikes are required to charge back the  $V_{\text{mem}}$  to the level of  $V_{\text{th}}$ , thus leading to an output pattern with the shortest inter-spike interval. In case (b),  $t_{\text{SET}}$  was the shortest, and hence the capacitor discharged the most.

Thus for this case, more incoming pre-neuron spikes are needed to recharge  $V_{\text{mem}}$ . Case (c) represents a deterministic Integrate and Fire situation with full  $V_{\text{mem}}$  discharge. Finally, case (d) depicts a situation with different  $t_{\text{SET}}$  durations for consecutive output spikes. It is a possible representation of neuron inter-spike intervals for a random sequence of  $t_{\text{SET}}$  values that can be obtained by cycling the CBRAM device



**Figure 4.4:** Proposed circuit-equivalent of the S-IF neuron [163].



**Figure 4.5:** Circuit used to demonstrate the concept of a S-IF effect when the CBRAM is in the set state [163].

#### 4. USING RRAM FOR NEURON DESIGN

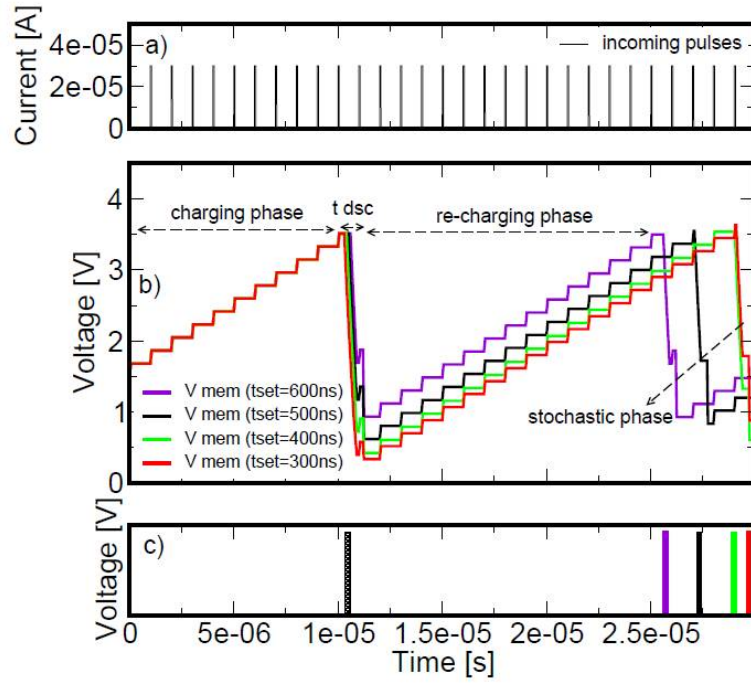
---

multiple times.

The circuit equivalent of the Stochastic-Integrate and Fire neuron concept shown in Fig. 4.2(b) is presented in Fig. 4.4. It consists of a current-source to simulate input currents coming from synapses and pre-neurons, a capacitor  $C_{\text{mem}}$  to integrate the current and build up the neuron membrane-voltage  $V_{\text{mem}}$ , a nMOS transistor M1 to perform set operation, two nMOS transistors M2 and M3 to perform the reset operation, a comparator block, a spike-generation block, a delay-element  $\Delta t$  and a CBRAM device. The delay element is used to perform the reset operation of the CBRAM device at the end of each neuron spike.

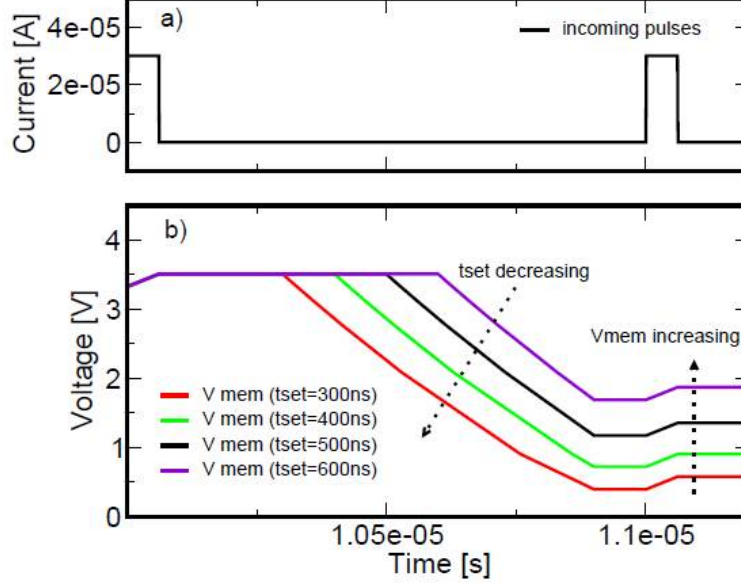
In Fig. 4.4, initially the CBRAM is in high-resistive state. As incoming pre-synaptic current is accumulated in  $C_{\text{mem}}$ ,  $V_{\text{mem}}$  would constantly build up at the anode of the CBRAM. During this time M1, M2 and M3 are off. When the neuron spikes, the spike-generation block will generate an output-spike and two additional pulsed-signals (S1, S2) going to M1 and  $\Delta t$  respectively. S1 acts as a gating signal to turn on M1.  $V_{\text{mem}}$  build-up and switching on of M1 will enable set-operation of the CBRAM since a positive voltage drop is established between the anode and the cathode. However during the set-operation, M2 and M3 are not turned on, as  $\Delta t$  delays the signal S2.

At the end of the set-operation, the signal S2 will turn on M2 and M3 thus building up the voltage at the cathode to switch the CBRAM to the off-state (reset). Thus, before the next consecutive neuron spikes the CBRAM device is automatically reset and reprogrammed to a different initial  $R_{\text{off}}$  state. Note that the flow of current through the CBRAM, during the set-operation, leads to a discharge of the capacitor  $C_{\text{mem}}$  thus decreasing the membrane voltage  $V_{\text{mem}}$ . The amount of decrease in  $V_{\text{mem}}$  can be estimated by calculating the total duration ( $t_{\text{dsc}}$ ) for which current flows through the switched CBRAM.  $t_{\text{dsc}}$  is the difference of the pulse-width of the signal S1 and the  $t_{\text{SET}}$  (inset of Fig. 4.1). Depending on the value of  $t_{\text{SET}}$  every time the neuron spikes, different amount of  $C_{\text{mem}}$  discharge will occur. Thus, in between any two firing cycles, the neuron may require different amount of incoming current to charge  $V_{\text{mem}}$  to the level of  $V_{\text{th}}$ .



**Figure 4.6:** Full evolution of  $V_{mem}$  simulating the circuit shown in Fig. 4.5. (a) Pre-neuron incoming pulses are used to build up  $V_{mem}$ . (b) Initially  $V_{mem}$  builds up as consequence of incoming currents (charging phase). Set operation lead to different discharge of  $C_{mem}$  ( $t_{dsc}$ ). During the recharging phase a different number of incoming pulses will raise  $V_{mem}$  till  $V_{th}$ . (c) Expected different inter-spike intervals depending on the  $t_{set}$  [163].

## 4. USING RRAM FOR NEURON DESIGN

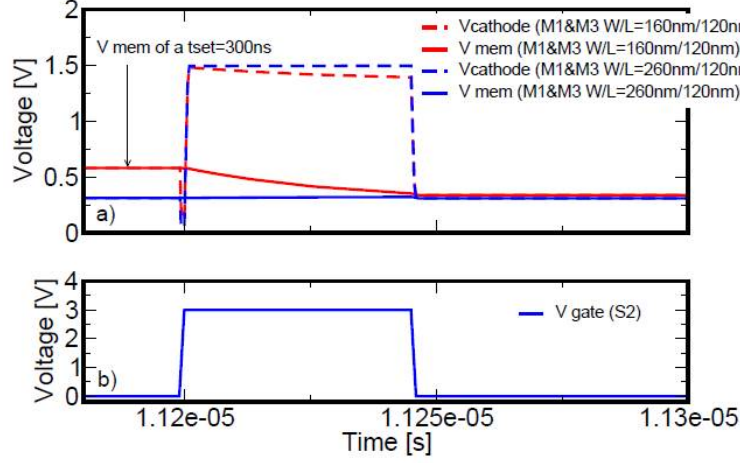


**Figure 4.7:** (a) Pre-neuron incoming pulses are used to build up  $V_{\text{mem}}$ . (b) Zoom on  $V_{\text{mem}}$  during the discharging phase for different  $t_{\text{SET}}$  in the range 300 ns - 600 ns. Lower  $t_{\text{SET}}$  leads to lower residual membrane voltage  $V_{\text{mem}}$  [163].

## 4.4 Results and Discussion

### 4.4.1 Set- and Reset- Operation

We performed SPICE transient simulation, with Eldo simulator, to validate the proposed concept using a simplified circuit shown in Fig. 4.5. Fig. 4.6(a) shows a simulated train of incoming pulses (excitatory currents) and the corresponding evolution of the  $V_{\text{mem}}$  (Fig. 4.6(b)) between two consecutive neuron spike-cycles. When  $V_{\text{mem}}$  reaches a threshold voltage  $V_{\text{th}}$  ( $V_{\text{th}} \simeq 3.5$  V in our simulation), the CBRAM device undergoes set-operation, and  $C_{\text{mem}}$  begins to discharge. Fig. 4.6(b) shows the discharging and recharging of  $C_{\text{mem}}$  for four different simulated values of  $t_{\text{SET}}$  (in the range 300 ns - 600 ns). Fig. 4.6(c), shows the expected output of the neuron. Note that different number of incoming pulses are required to reach the neuron firing threshold again, since the initial  $V_{\text{mem}}$  value is dominated by the stochasticity in  $t_{\text{SET}}$ . Five additional incoming pulses are needed to reach the threshold for the shortest value of  $t_{\text{SET}}$  (300 ns). Fig. 4.7 shows the zoomed version of  $C_{\text{mem}}$  discharging for the the different simulations shown in Fig. 4.6. Note that the longest  $t_{\text{SET}}$  (600 ns) corresponds to the least amount of



**Figure 4.8:** (a) Time-evolution of  $V_{\text{mem}}$  and  $V_{\text{cathode}}$  that establish a voltage drop on the CBRAM to enable reset operation. Larger M3 increase the voltage drop, since  $V_{\text{cathode}}$  builds up more.  $V_{\text{mem}}$  corresponding to a  $t_{\text{SET}}$  of 300 ns is considered. (b) Pulse applied to M3 [163].

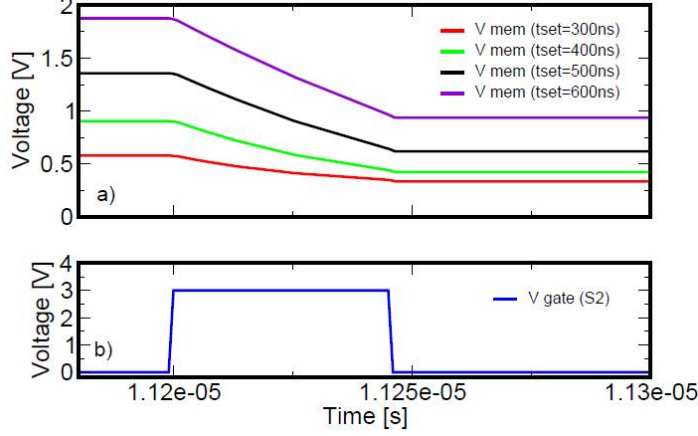
$C_{\text{mem}}$  discharge, and vice-versa. To simulate the reset operation, a pulse of 45 ns with an amplitude of 3 V was applied at M2 and M3, while keeping M1 off. Such high voltage on M3 is required to build up  $V_{\text{cathode}}$ . Fig. 4.8 shows the time evolution of  $V_{\text{cathode}}$  and  $V_{\text{mem}}$  when the initial value of  $V_{\text{mem}}$  was generated by a  $t_{\text{SET}}$  of 300 ns for two different width of M3. The actual voltage drop on the CBRAM can be increased increasing the size of the nMOS as shown in Fig. 4.8. Moreover, during the reset, an additional discharge of  $V_{\text{mem}}$  is possible depending on the size of M3, since M2, that is directly connected to  $V_{\text{mem}}$ , is turned on by S2 (Fig. 4.9(a)).

#### 4.4.2 Parameter Constraints

Due to the intrinsic physics of CBRAM device, some constraints in implementing the proposed circuit should be considered. In particular,  $V_{\text{th}}$  has to be greater than the minimum value of the voltage-drop required to set the CBRAM device for a given pulse-width. The amplitude of S1 should be sufficient to turn on the gate of M1, while the pulse-width of S1 depends on the  $V_{\text{th}}$  and the spread on  $t_{\text{set}}$ . If S1 pulse-width is very long it would always lead to a complete discharge of  $C_{\text{mem}}$  and the  $t_{\text{SET}}$  stochasticity cannot be exploited. However S1 cannot be arbitrarily small, it has to be greater than the minimum  $t_{\text{SET}}$  value at a given voltage applied on the anode of



## 4. USING RRAM FOR NEURON DESIGN



**Figure 4.9:** (a) Time-evolution of  $V_{\text{mem}}$  during the reset operation for  $t_{\text{SET}}$  in the range 300 ns - 600 ns. Different residual voltages are obtained. (b) Pulse applied to M3 [163].

the CBRAM device. The dependence of applied pulse-width and the amplitude of  $V_a$  for CBRAM set-operation is shown in [164]. Thus, by tuning the characteristics of S1, the stochastic response of the neuron can be controlled. The amplitude of S1 would determine the amount of current flowing through M1 (compliance current) and thus the final value of the CBRAM resistance in the set state. The set state resistance would determine the programming conditions for the consecutive reset-operation [165]. Thus, the characteristics of S2 can be tuned based on the final CBRAM resistance obtained after the set-operation.

### 4.4.3 Energy Consumption

For the proposed S-IF, additional energy consumption per spiking cycling of the neuron will be devoted to perform set and reset operation. The extra-energy consumption is dependent on the ratio of  $R_{\text{Off}}$  and  $R_{\text{On}}$ ; in particular on  $R_{\text{On}}$  since hundreds of  $\mu\text{A}$  can flow before M1 would be turned off, if the low resistance state is  $\simeq 10^4 \Omega$ . We estimated the energy consumption during the set operation using:  $E_{\text{set}} = V_{\text{set}} I_{\text{set}} t_{\text{set}}$ . In our simulations we used  $V_{\text{set}} = 3.5 \text{ V}$  (i.e.  $V_{\text{th}}$ ),  $I_{\text{set}} = 350 \mu\text{A}$ ,  $t_{\text{set}}$  in a range between 300 ns and 600 ns that gives an energy mean value of 55 nJ. The energy devoted to reset the CBRAM is negligible. For a real system, this 55 nJ could be strongly reduced increasing the resistance of the low resistive value, since for the proposed application the ratio  $R_{\text{Off}}/R_{\text{On}}$  is not a major constraint.

## 4.5 Conclusion

In this chapter, we showed how CBRAM used in an unexpected fashion may allow designing stochastic neurons with low area overheads. We showed how CBRAM physics naturally leads to a stochastic behavior in the programming time, which may be exploited in a circuit. SPICE simulations validate the concept on a simple Integrate and Fire neuron. The concept could be extended to more complex neuron designs like [50]. These results highlight the benefits of novel non memory technologies, whose impact may go beyond traditional memory markets.

#### 4. USING RRAM FOR NEURON DESIGN

---

## 5

# Conclusions and Perspective

## 5.1 Conclusion

During this research activity we explored how some of the emerging RRAM devices (PCM, CBRAM and OXRAM) may be used inside neuromorphic systems and applications. We mainly emphasized on the emulation of synaptic plasticity effects such as long-term potentiation/depression (LTP/LTD) and learning rules like STDP. We also proposed a methodology to design compact stochastically firing neuron circuits that exploit intrinsic physical effects of RRAM devices.

In the case of PCM, we fabricated and studied devices with different stacks (GST, GeTe and GST+HfO<sub>2</sub>). We showed that while LTP-like conductance can be emulated with identical programming pulses, the same is not true for LTD-like effects. The difference arises due to the underlying physics of phase-change devices. To overcome the limitation imposed by abrupt LTD, we developed a unique low-power methodology known as the “2-PCM Synapse”. It uses 2-PCM devices (one device responsible for LTP and the other for LTD) connected in a complementary architecture. We developed a detailed programming methodology (Read, Write and Refresh schemes) for the “2-PCM Synapse” architecture. We showed that this approach has many advantages such as low-power and high tolerance to PCM-resistance drift as it’s predominantly based on crystallization. We showed that by engineering PCM devices with a HfO<sub>2</sub> dielectric layer the performance of both individual synapses, and the overall system can be improved. We also presented a methodology to strongly mitigate the impact of resistance-drift in neuromorphic systems based on PCM synapses. Using the “2-PCM

## 5. CONCLUSIONS AND PERSPECTIVE

---

Synapse” we showed complex-visual pattern extraction application and performed a detailed energy/power analysis of our system.

In the case of CBRAM, we fabricated and studied (1R, 1T-1R and 8 x 8 matrix) devices with Ag/GeS<sub>2</sub> stack. We found that for CBRAM, both LTP and LTD cannot be emulated with identical pulses. While it was possible to emulate LTP with varying (non-identical) programming conditions, LTD was always abrupt due to uncontrolled filament dissolution. Thus we adopted an alternative approach for neuromorphic systems based on CBRAM synapses. We used the synapses in a probabilistic and binary manner (as opposed to deterministic and multi-level). We used an optimized stochastic STDP rule to compensate for the fewer conductance states. We showed that stochasticity can be implemented either intrinsically (by using weak programming conditions), or extrinsically (by using external PRNG circuits). Using our binary stochastic CBRAM synapse approach we showed complex auditory and visual pattern extraction applications. Similar analysis was also performed for OXRAM 1R devices based on HfO<sub>2</sub> layer. However the study on OXRAM was preliminary and not as detailed as in the case of PCM and CBRAM. In case of all the three technologies that we used, there are some important observations that should be pointed out-

- **Robustness to variability:** In all the learning simulations, it was observed that the neural network is highly tolerant to synaptic variability (see Fig.2.26) For the PCM simulations, a 20% standard deviation dispersion was applied to each synapse parameter described in Tab.2.1. For the CBRAM/OXRAM simulations actual dispersion extracted from experimental data (for example see Fig.3.9, Fig.3.30) was used.
- **Impact of the Neurons:** In all the simulations we used optimized neuron parameters (example Tab.2.3) for high quality learning obtained through genetic evolution algorithm. It is possible that certain shortcomings of the synaptic characteristics were absorbed by the optimization of the neuron parameters. However this point is still under investigation and needs further detailed analysis.
- **Equivalence of binary and multi-level synapses:** In terms of learning performance the results obtained with both binary-stochastic and multi-level deterministic

synapses were equivalent for the cars learning experiment. However some applications might need higher synaptic redundancy in the case of binary synapses to compensate for fewer resistance levels.

## 5.2 Which one is better

In order to choose a technology that defines the ideal synapse, it becomes very important to first fix the overall system constraints, in terms of the final application, the learning rule and the neuron structures. A comparison of state-of-the art RRAM technology parameters, reported in literature, is provided in Tab.1.1.

Direct comparison of the three RRAM technologies discussed in this manuscript for synaptic application is not a straightforward task, due the numerous options and possibilities available at different levels in the design of a neuromorphic system (see Fig.1.28). However some parameters on the basis of which classifications can be done are-

### 5.2.1 Intermediate Resistance States

A technology with a large number of intermediate resistance states is very desirable, however if a stochastic learning rule is used, the number of intermediate resistance states become insignificant and just binary switching is suffice. Some learning applications may require several intermediate synaptic weights, while satisfactory learning can be achieved for others with just binary or few weights. Low number of synaptic weights may also be compensated by increasing the total number of synapses in the network or the per/neuron synaptic redundancy. For the 3 different devices that we fabricated PCM showed the maximum number of intermediate states (60 in the case of GST + HfO<sub>2</sub> devices). Programming scheme used to implement intermediate states is also an important factor, as neuron action potentials are identical. For our devices, OXRAM offered the least number of intermediate states, while CBRAM showed decent RON modulation but at the cost of variable programming pulses.

### 5.2.2 Energy/Power

The energy analysis requires more careful consideration as it is closely linked to the type of learning rule used in the system, and the final application being addressed.

## 5. CONCLUSIONS AND PERSPECTIVE

---

It is always desirable that the programming energy for both SET/RESET events be minimized. However even if one of the two energies (SET or RESET) is higher, it may still minimally impact the total system power dissipation, if a learning rule is chosen such that it pre-dominantly favors the less energy consuming synaptic event. For instance, our simplified STDP learning rules were predominantly LTD dependent, so RESET energy of the synaptic devices was more significant compared to SET energy. This can be further complicated if the architecture of the synaptic devices is changed. As in the case of “2-PCM Synapse”, where LTD is also obtained through crystallization, thus the RESET energy doesn’t play a significant role. Among our 3 technologies, PCM was the most energy consuming, followed by CBRAM and then OXRAM. Our PCM devices consumed more energy as they were large analytical structures and not ultra-scaled. High programming currents would require larger driving transistors or selector devices and would also limit the size of the synaptic arrays.

### 5.2.3 Endurance

As far as device endurance is considered, from Tab.2.4 we can see that for each type of event (LTP/LTD) the programming frequency/device/s is always less than 0.5 Hz. This means that any RRAM technology with an endurance of  $10^8$  cycles would function for a learning time of  $10^8$  s (or  $> 3$  years of continuous learning), assuming a worst case programming frequency of 1 event/dev/sec. From Tab.1.1 we can see that all the three RRAM technologies (PCM,CBRAM,OXRAM) easily satisfy this criteria.

### 5.2.4 Speed

Device switching response time is not an issue for any of the technology, as biological timescales for neuro-synaptic modifications are in the range of ms. All the 3 RRAM technologies respond at least  $10^3$  -  $10^6$  times faster ( $\mu$ s–ns) than the biological timescales. Among our devices, PCM were the fastest in terms of programming pulses, followed by OXRAM and then CBRAM. A fast switching device also brings down the energy dissipated per RESET/SET event.

### 5.2.5 Resistance Window

Generally a large resistance programming window is desirable as it increases the noise margin for the intermediate resistance states. However the exact RON and ROFF ranges may impact the energy dissipation, particularly while reading the resistance of the synaptic array. To minimize read-power dissipation a high RON and ROFF is preferred. However very high RON values increase the parasitic charging times (RC delay) of the memory crossbar. The range of RON and ROFF values will also impact the choice of the driving transistors in the memory control circuitry and the size of largest possible crossbar/matrix.

In terms of device area and scalability, state-of-the art devices for all the 3 RRAM technologies offer highest possible integration density with a cell area of  $4F^2$  (Tab.1.1). In our case the CBRAM devices were the most compact, followed by OXRAM and the PCM.

## 5.3 On-Going and Next Steps

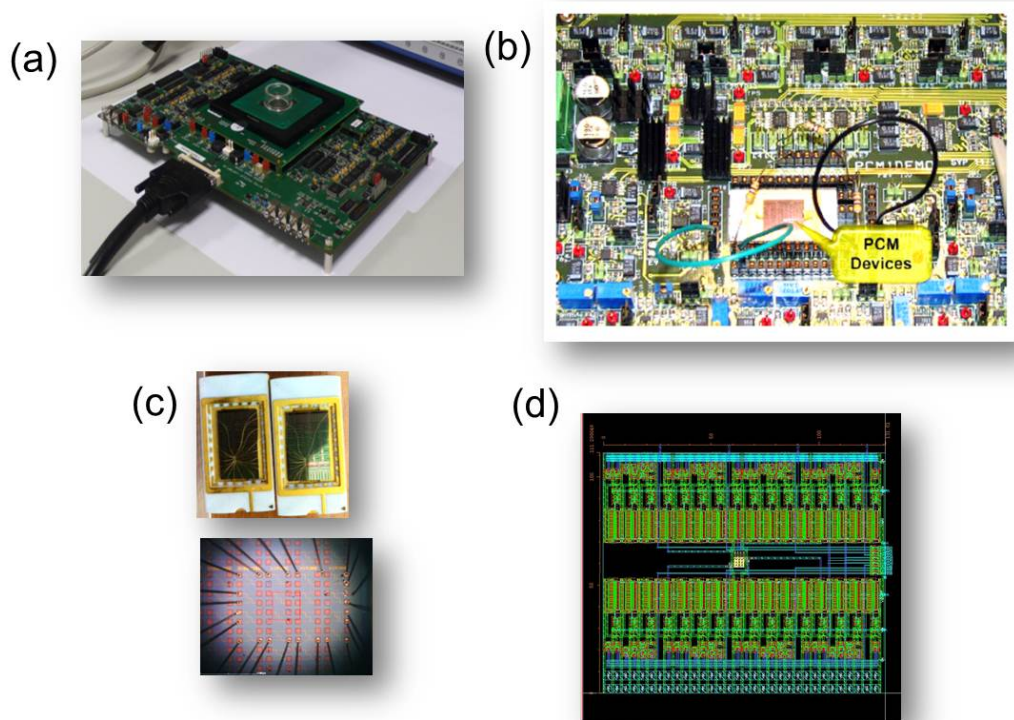
In this section I would like to mention some on-going and future activities that couldn't be completed in the time frame of the PhD, but are very relevant to take the work described in the thesis to the next level-

- OXRAM: We only investigated 1R structures and HfO<sub>2</sub> stack. It would be interesting to see if more intermediate resistance states can be obtained by engineering the material stack, as also shown in literature. As an ongoing activity we are testing 1T-1R OXRAM devices.
- Full test-systems: At CEA-LETI, we are packaging multiple PCM, CBRAM and OXRAM devices using wire-bonding (Fig.5.1c). The idea is to test these packages on special test boards designed by our colleagues at CEA-LIST. These boards (Fig.5.1b) emulate the functionality of neurons and are driven by a PC and FPGA.
- Fully co-integrated chips: We are also designing prototype CMOS-RRAM fully integrated circuits (Fig.5.1d), that contain RRAM arrays and CMOS driving circuits.



## 5. CONCLUSIONS AND PERSPECTIVE

---



**Figure 5.1:** Snapshot of current on-going activities (a) Multi-electrode Array (MEA) of the NeuroPXI system to collect neuron signals [166]. (b) Neuromorphic test-board developed by CEA-LIST for testing packaged RRAM. (c) Wire-bonded and packaged PCM and CBRAM devices. (d) Layout of a CBRAM synapse array + CMOS control circuit designed in collaboration with CEA-LIST.

- **Interfacing with Biology:** Recently we started an activity which involves the use of a real neuron signal acquisition system called NeuroPXI [166]. In this activity, which is in a very early phase, we plan to collect in-vitro/ in-vivo neuron signals (Fig.5.1a) from rat retina and process them with our RRAM based neural network. The motivation behind this activity deals with better understanding/processing of real neuron signals and development of future generation neuro-prosthetic systems.

## 5.4 The road ahead...

Generally speaking the present time is very favorable and interesting for neuromorphic research due to some large funding support and recent projects launched worldwide (see Sec.1.1.1). However, for a sustained and meaningful progress in the field, some shortcomings that need to be addressed are-

- **Standardization:** There is a strong need for standardizing different blocks of a neuromorphic processing core. These may include neuron-models, synaptic-models, communication protocols, input and output data formats (ex-AER), learning rules etc. Standardization of the key blocks would also enable development of neuromorphic specific design and simulation tools
- **Application Benchmarking:** There is a need to create an application database to correlate neuromorphic system learning performance and specifications of building blocks such as synapses and neurons.
- **Learning Rules:** As far as the learning rules are considered, the field requires a lot of development. New learning rules which are more adapted to devices (synaptic or neuron related) should be developed. Just relying on STDP is not sufficient for a broad application base.

## 5. CONCLUSIONS AND PERSPECTIVE

---

## Appendix A

# List of Patents and Publications

### A.1 Patents

1. **Manan Suri** and Barbara DeSalvo, *Utilizing Drift in PCM for Delay-Generation/Timing Application*, DD.12750.
2. Olivier Bichler, **Manan Suri**, Barbara DeSalvo and Christian Gamrat, *Complementary PCM Synapses for STDP*, DD.12749.
3. **Manan Suri**, Olivier Bichler, Damien Querlioz, Christian Gamrat and Barbara DeSalvo, *Design and Optimization of Neuromorphic Systems*, DD.13709.
4. **Manan Suri** and Giorgio Palma, *Stochastic Neuron Design Using RRAM*, DD.14473.

### A.2 Book Chapter

1. **Manan Suri** and Barbara DeSalvo, *Phase Change Memory and Chalcogenide Materials for Neuromorphic Applications: Emphasis on Synaptic Plasticity*, in *Advances in Neuromorphic Memristor Science and Applications*, edited by R. Kozma, R. Pino and G. Kozma, G. Pazienza, Springer Series in Cognitive and Neural Systems, 2012, Vol. 4, Part 2, pp. 155-178.

### A.3 Conference and Journal Papers

1. **M. Suri**, O. Bichler, D. Querlioz, D. Garbin, V. Sousa, L. Perniola, D. Vuillaume, C. Gamrat, and B. DeSalvo, *Phase Change Memory Synapses for Large Scale*

## A. LIST OF PATENTS AND PUBLICATIONS

---

- Energy Efficient Neuromorphic Systems, European Phase Change and Ovonic Symposium, EPCOS, (2013), **invited**.
2. **M. Suri**, D. Garbin, O. Bichler, D. Querlio, D. Vuillaume, C. Gamrat, and B. DeSalvo, Impact of PCM Resistance-Drift in PCM based Neuromorphic Systems and Drift Mitigation Strategy, IEEE/ACM NanoArch, (2013).
  3. **M. Suri**, D. Querlio, O. Bichler, G. Palma, E. Vianello, D. Vuillaume, C. Gamrat, and B. DeSalvo, Bio-Inspired Stochastic Computing Using Binary CBRAM Synapses, IEEE Transactions on Electron Devices, (2013).
  4. D. Garbin, **M. Suri**, D. Querlio, O. Bichler, C. Gamrat, and B. DeSalvo, Probabilistic Neuromorphic System Using Binary Phase-Change Memory (PCM) Synapses: Detailed Power Consumption Analysis, IEEE International Conference on Nanotechnology,(2013).
  5. G.Palma, **M. Suri**, D. Querlio, E. Vianello and B. DeSalvo, Stochastic Neuron Design using Conductive Bridge RAM, IEEE/ACM NanoArch, (2013).
  6. **M. Suri**, O. Bichler, D. Querlio, 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 , IEEE International Electron Devices Meeting, IEDM, p.10.3, (2012).
  7. **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 Performance in Neuromorphic Systems, IEEE International Memory Workshop, IMW, (2012).
  8. **M. Suri**, O. Bichler, D. Querlio, 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, IEEE International Electron Devices Meeting, IEDM, pp. 4.4.1-4.4.4, (2011).

9. **M. Suri**, O. Bichler, D. Querlio, B. Traore, O. Cueto, L. Perniola, V. Sousa, D. Vuillaume, C. Gamrat, and B. DeSalvo, Physical Aspects of PCM Devices for Neuromorphic Applications, *Journal of Applied Physics*, Vol. 112, 5, p. 054904, (2012).
10. **M. Suri**, V. Sousa, L. Perniola, D. Vuillaume, and B. DeSalvo, Phase Change Memory for Synaptic Plasticity Application in Neuromorphic Systems, *IEEE International Joint Conference on Neural Networks, IJCNN*, pp. 619-624, (2011).
11. **M. Suri**, O. Bichler, Q. Hubert, L. Perniola, V. Sousa, C. Jahan, D. Vuillaume, C. Gamrat, and B. DeSalvo, Addition of HfO<sub>2</sub> Interface Layer for Improved Synaptic Performance of Phase Change Memory (PCM) Devices, *Journal of Solid-State Electronics*, (2012).
12. O. Bichler, **M. Suri**, D. Querlio, D. Vuillaume, B. DeSalvo and C. Gamrat, Visual Pattern Extraction using Energy Efficient 2-PCM Synapse Neuromorphic Architecture, *IEEE Transactions on Electron Devices*, Vol. 59, 8, pp. 2206-2214, (2012).
13. Toffoli, **M. Suri**, L. Perniola, A. Persico, C. Jahan, J.F. Nodin, V. Sousa, B. De Salvo and G. Remibold, Phase Change Memory Advanced Electrical Characterization for Conventional and Alternative Applications, *IEEE International Conference on Microelectronic Test Structures, ICMTS*, pp. 114-118, (2012).
14. Navarro, N. Pashkov, **M. Suri**, V. Sousa, L. Perniola, S. Maitrejean, A. Persico, A. Roule, A. Toffoli, B. De Salvo, Electrical Performances of Tellurium-rich GexTe1-x Phase Change Memories, *IEEE International Memory Workshop, IMW*, (2011).
15. N. Pashkov, G. Navarro, J.C. Bastien, **M. Suri**, L. Perniola, V. Sousa, S. Maitrejean, et al., Physical and Electrical Characterization of Germanium or Tellurium Rich GexTe1x for Phase Change Memories, *European Solid State Device Research Conference, ESSDERC*, pp. 91-94, (2011).
16. V. Sousa, L. Perniola, G. Navarro, N. Pashkov, **M. Suri**, A. Persico, E. Henaff, F. Fillot, F. Pierre, A. Roule, et al., GeTe-based Phase Change Memories: Effect

## **A. LIST OF PATENTS AND PUBLICATIONS**

---

of stoichiometric variations and N or C addition, European Phase Change and Ovonic Symposium, EPCOS, (2011).

## Appendix B

# Résumé en Français

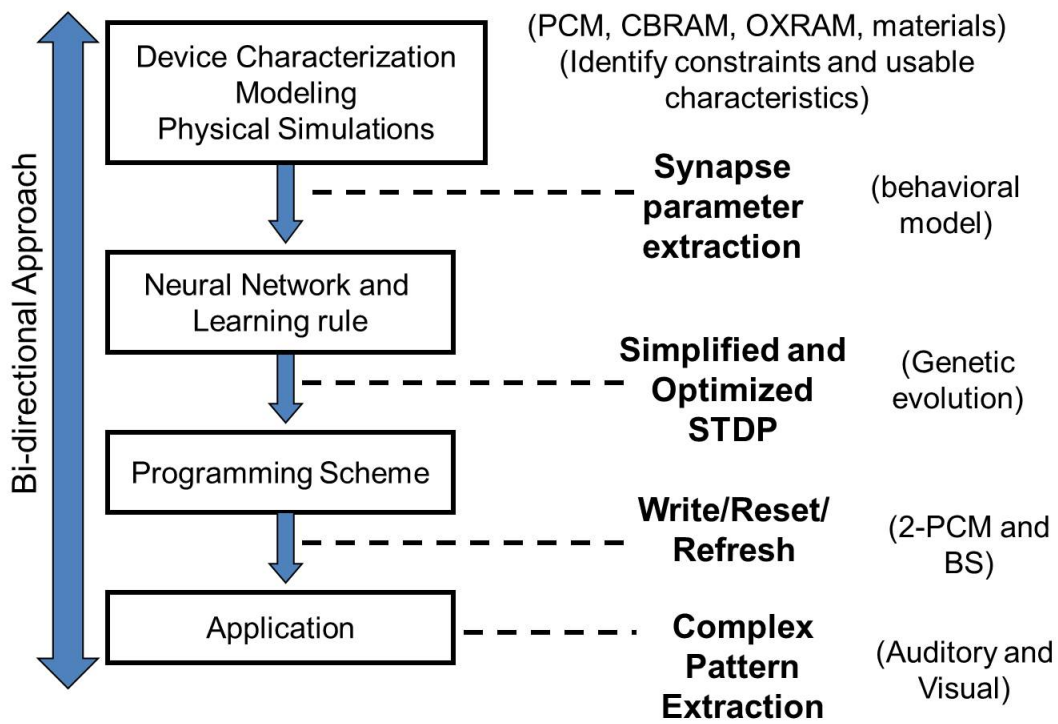
### B.1 Chapitre I: Découverte

Ce chapitre décrit les motivations pour poursuivre la R&D dans le domaine des systèmes neuromorphiques. Nous nous concentrons alors sur quelques concepts de base de la neurobiologie. Un état de l'art sur l'implémentation matériel de synapses biologiques est présenté et leurs limites sont discutées. Le concept de mémoire résistive non-volatile issu de technologies émergentes est introduit. A la fin du chapitre, nous résumons brièvement la portée et la stratégie globale adoptée pour la recherche menée au cours de cette thèse.

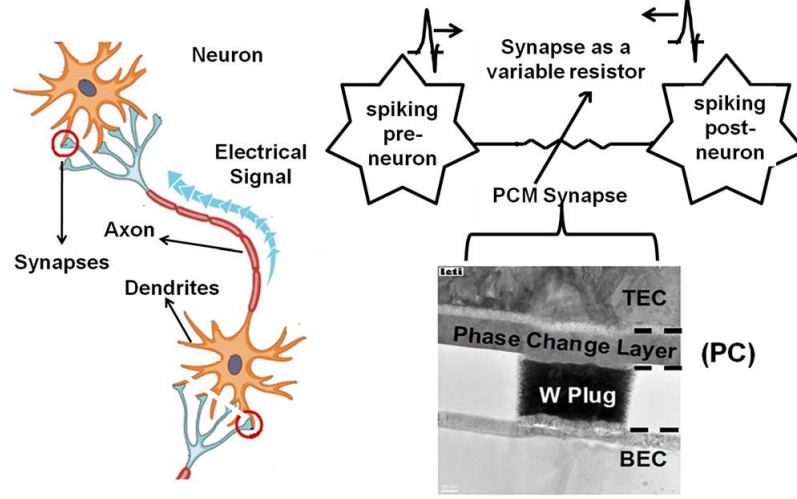
### B.2 Chapitre II: Synapses avec des Mémoires à Changement de Phase

Dans ce chapitre, nous avons démontré que les dispositifs PCM peuvent être utilisés pour imiter les effets de plasticité synaptique en simulant à la fois la LTP et la LTD. Nous avons montré que, bien que la LTP progressive puisse être obtenu par l'application d'impulsions de potentialisation identiques (cristallisation), la nature de la LTD est abrupte lorsque des impulsions de dépression identiques (amorphisant) sont utilisées. La raison pour laquelle le comportement de la LTD est brutal a été expliquée par des expériences et des simulations multi-physiques. Nous avons étudié le rôle de la cinétique de cristallisation (taux de croissance et de nucléation) en émulation LTP, utilisant des dispositifs PCM fabriqués avec deux matériaux chalcogénures différents:





**Figure B.1:** Bi-directional strategy (Top-down + Bottom-up) adopted for the work presented in this PhD thesis. To develop the ideal "synapse-solution" optimization and fine-tuning was performed at different levels such as architectures, learning-rules and programming-schemes.(BS: Binary Stochastic synapses).

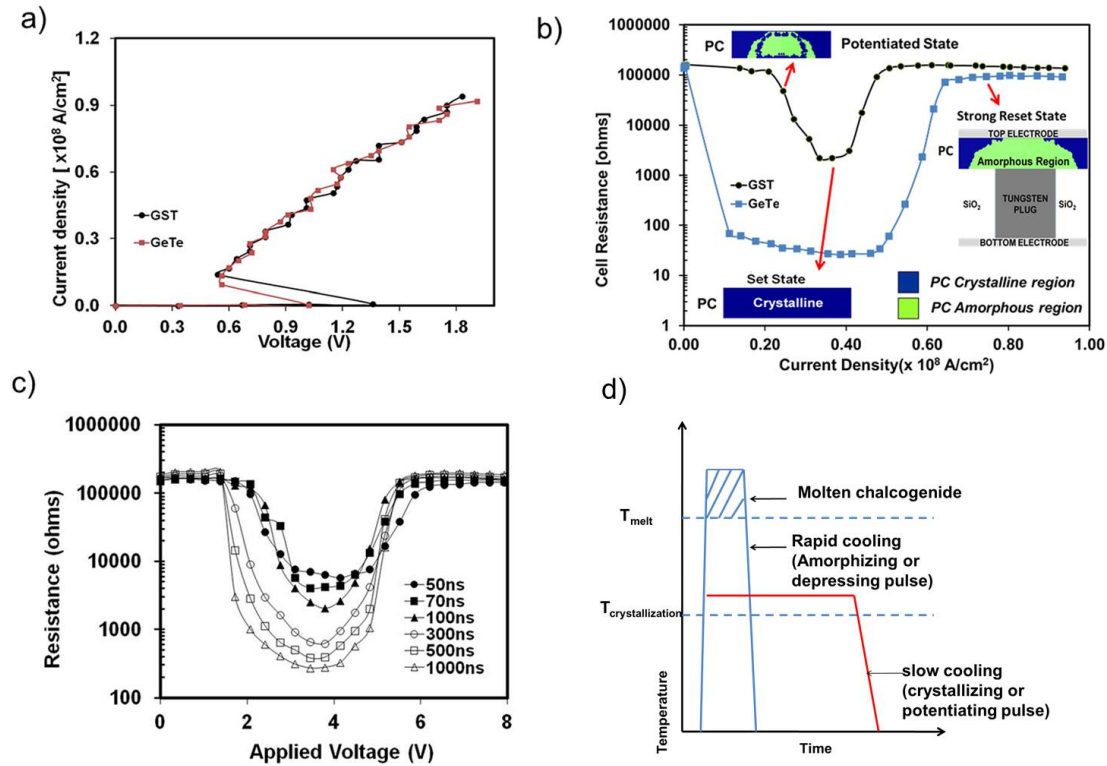


**Figure B.2:** Illustration of biological synapse and the equivalent PCM synapse in a neural circuit connecting a spiking pre- and post- neuron [81]. TEM cross-section image of the GST PCM devices fabricated for this study is shown.

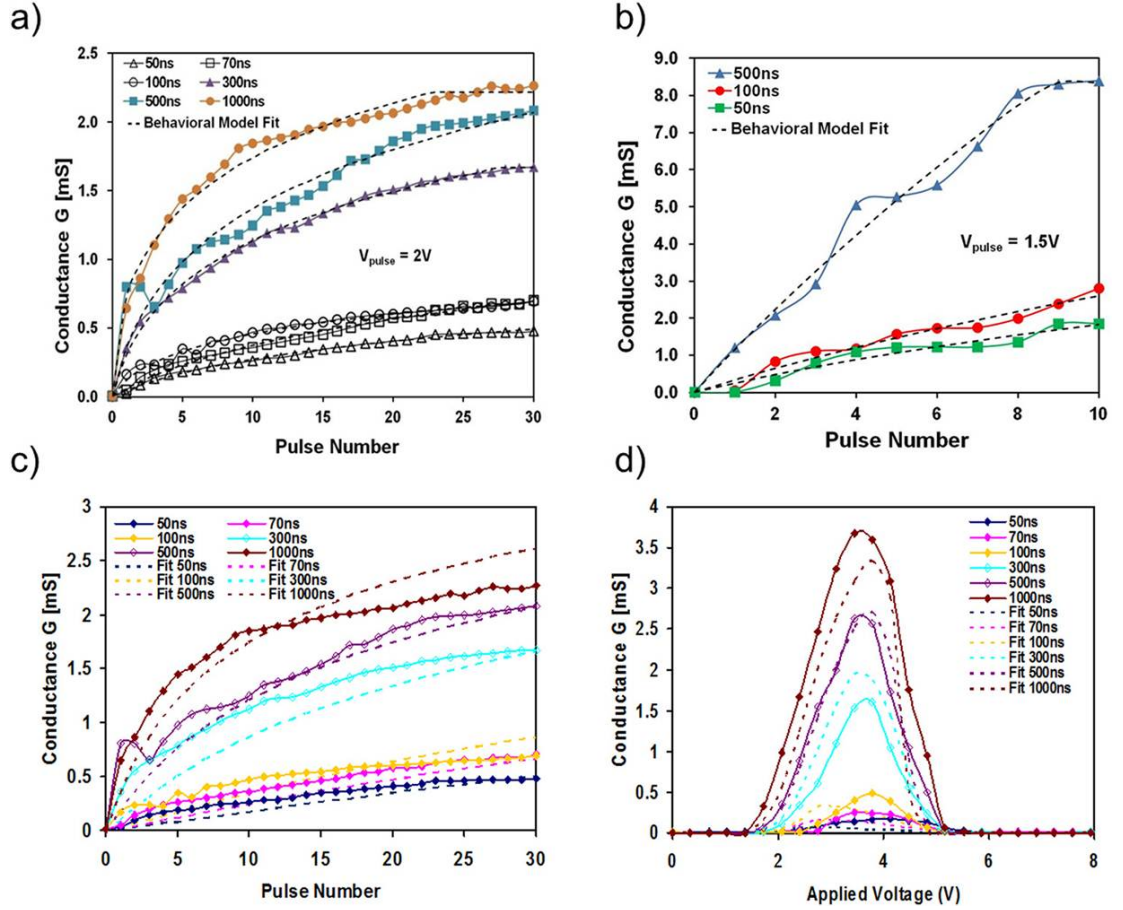
le GST où la nucléation est dominante et le GeTe où la croissance domine. Un modèle polyvalent (i) et un modèle comportemental (ii) compatibles circuit, utiles pour des simulations de réseaux de neurones à grandes échelles avec des dispositifs PCM, ont été développés. Pour surmonter les limitations abrupt de LTD, nous avons développé une nouvelle architecture faible consommation (“2-PCM Synapse”) et une méthodologie de programmation détaillée (en lecture, en écriture et un protocole de rafraîchissement) pour les architectures (i) avec dispositifs de sélection (1T-1R) et (ii) sans dispositif de sélection (1R).

Nous avons simulé un réseau de neurones impulsionnels (SNN, Spiking Neural Network) à 2 couches, spécialement conçu pour les applications d’extraction de motif visuel complexe, composé d’environ 4 millions de dispositifs PCM et simulant une règle d’apprentissage STDP simplifiée. Notre SNN a pu extraire l’orientation du déplacement et les formes des voitures circulant sur une autoroute avec un taux de détection moyen très élevé ( $>90\%$ ) et une très faible consommation d’énergie nécessaire pour la programmation des synapses,  $112\ \mu\text{W}$ . Nous avons montré qu’en modifiant l’interface des dispositifs GST-PCM (par ajout d’une couche de  $\text{HfO}_2$  de  $2\ \text{nm}$ ), l’efficacité énergétique de notre système neuromorphique peut être améliorée tant au niveau des synapses individuelles que pour l’ensemble du système. Avec cette couche d’interface, la consom-

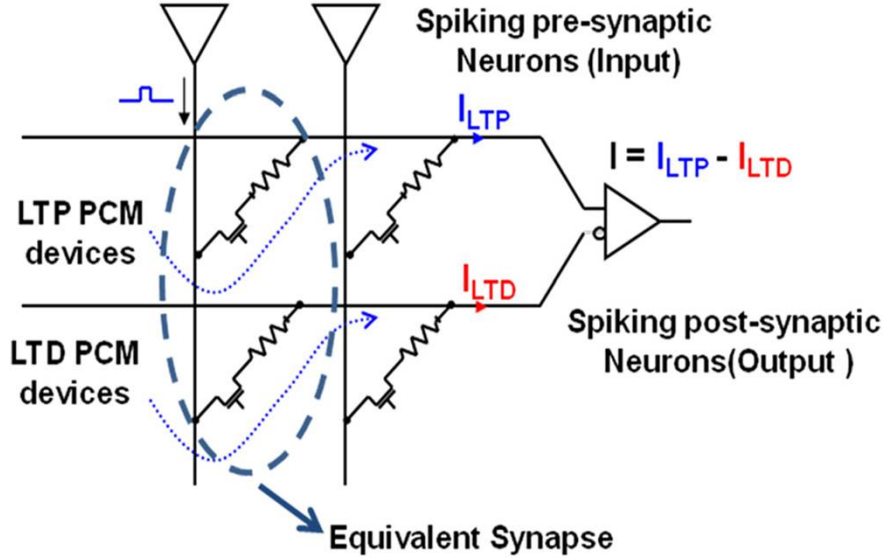
## B. RÉSUMÉ EN FRANÇAIS



**Figure B.3:** (a) IV characteristics for PCM devices with 100 nm thick GST and GeTe layer starting from initially amorphous phase. (b) R-I characteristics of GST and GeTe PCM devices, with inset showing the PCM phase of intermediate resistance states. (c) R-V curves for GST devices with six different pulse widths. Read pulse = 0.1 V, 1 ms. Legend indicates applied pulse widths. (d) Temperature Vs Time profile for PCM programming pulses [81].



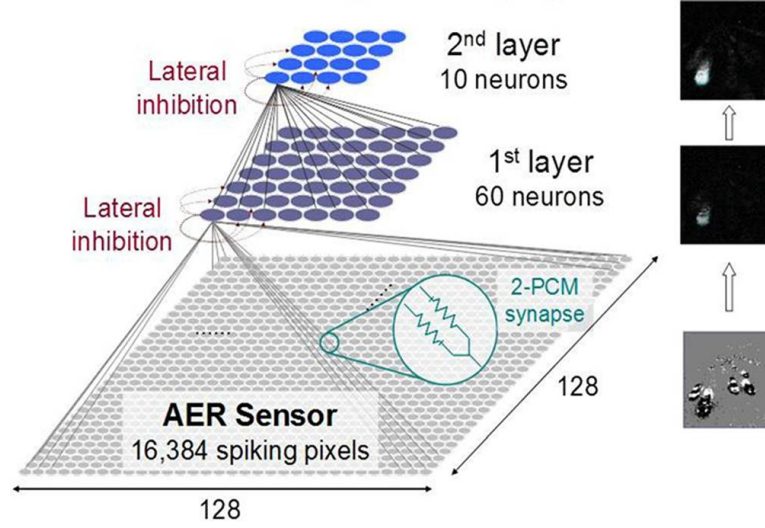
**Figure B.4:** (a) Experimental LTP characteristics of GST PCM devices. For each curve, first a reset pulse (7 V, 100 ns) is applied followed by 30 consecutive identical potentiating pulses (2 V). Dotted lines correspond to the behavioral model fit described in Eq.2.3 and eq.2.4. (b) Experimental LTP characteristics of GeTe PCM devices. (c) LTP simulations for GST devices using circuit-compatible model. (d) Conductance evolution as a function of the applied voltage for GST devices with six different pulse widths, using circuit-compatible model (sec.2.5.2). Legends in Figs.2.6(a–d) indicate pulse widths [81].



**Figure B.5:** The "2-PCM Synapse" concept schematic. The contribution of the current flowing through the LTP device is positive, while that of the LTD device is negative, towards the integration in the output neuron [110].

mation d'énergie a été réduite à  $60 \mu\text{W}$ , tandis que la consommation individuelle des programmations synaptiques est diminué de plus de 50 %.

Nous avons ensuite étudié en détail l'impact de la dérive temporelle de la valeur de résistance des dispositifs PCM dans notre système neuromorphique. Nous montrons que l'architecture "2-PCM Synapse" a une grande tolérance vis-à-vis de la perte de l'information synaptique due à la dérive de la résistance. Pour atténuer davantage les effets de la dérive de la résistance, nous introduisons une autre architecture ainsi qu'une méthodologie (appelé "binaire PCM Synapse programmation") avec une règle d'apprentissage de type STDP stochastique. Des simulations au niveau système ont confirmées que l'utilisation de l'approche "binaire PCM Synapse" n'affecte pas les performances d'apprentissage. La consommation d'énergie par les synapses en fonction des valeurs  $R_{OFF}$  et  $R_{ON}$  a été étudiée. Les résultats montrent que la consommation d'électricité en mode d'apprentissage peut être minimisée en réduisant la valeur de  $R_{OFF}$ , tandis que la consommation d'énergie en mode de lecture peut être optimisée en augmentant à la fois les valeurs de  $R_{ON}$  et  $R_{OFF}$ . La consommation d'énergie synaptique peut être fortement réduite à quelques 100 nW lorsque les meilleurs dispositifs PCM de l'état de l'art sont utilisés. Pour aller plus loin, la nouvelle topologie du



**Figure B.6:** 2-Layer Spiking Neural Network (SNN) topology used in simulation. The network is fully connected and each pixel of the  $128 \times 128$  pixel AER dynamic vision sensor (DVS-retina) is connected to every neuron of the 1st layer through two synapses, receiving positive and negative change in illumination events respectively. Lateral inhibition is also implemented for both layers [110].

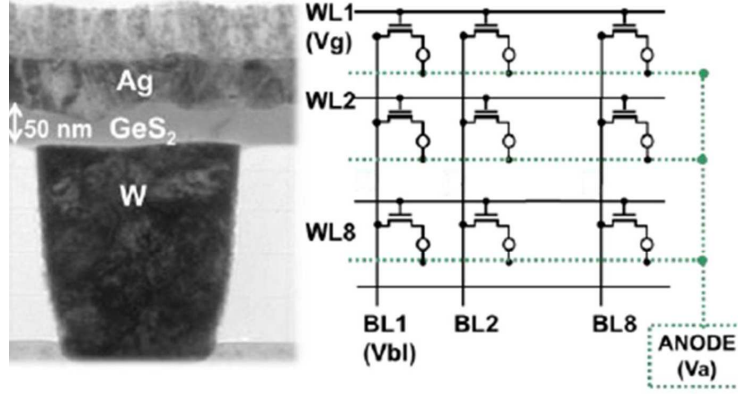
réseau mis en place dans [134] peut être exploitée, avec des neurones spatialement localisées, offrant des performances d'apprentissage similaire avec seulement un dixième des synapses. Ainsi, le besoin d'utiliser seulement 130000 synapses rend la perspective d'une réalisation pratique du matériel encore plus proche. Le chapitre démonte le potentiel de l'utilisation de la technologie PCM pour les futurs systèmes de neuromorphiques embarqués intelligents et omniprésents.

### B.3 Chapitre III: Synapses avec des Mémoires à ‘Filamentary-switching’

Nous avons proposé pour la première fois un système bio-inspirée avec des synapses CBRAM binaires et une règle d'apprentissage STDP stochastique capables de traiter des flux de données analogiques asynchrones pour la reconnaissance et l'extraction de motifs répétitifs d'une manière totalement non supervisée. Les applications démontrées présentent des performances très élevées (sensibilité aux motifs auditifs  $> 2.5$ , taux de détection vidéo  $> 95\%$ ) et une dissipation de puissance synaptique ultra-faible

## B. RÉSUMÉ EN FRANÇAIS

---

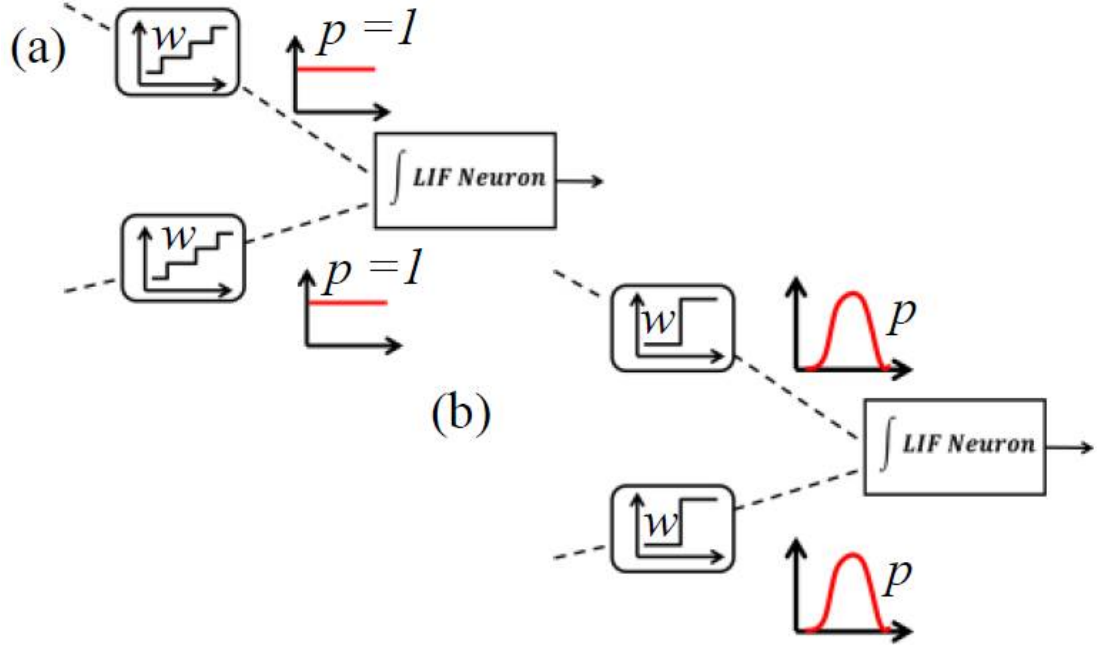


**Figure B.7:** (Left) TEM of the CBRAM resistor element. (Right) Circuit schematic of the 8 X 8 1T-1R CBRAM matrix. (note: the devices used in this study had a GeS<sub>2</sub> layer thickness of 30 nm. The 50 nm TEM is for illustrative purpose only [130].)

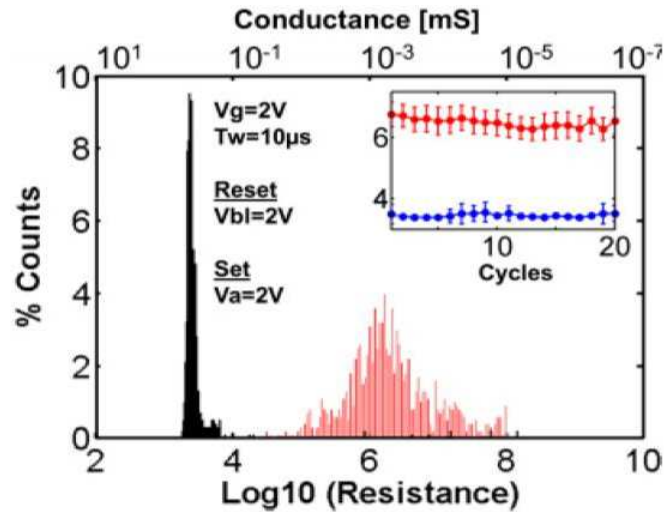
(audio 0.55  $\mu$ W, vidéo 74.2  $\mu$ W) dans le mode d'apprentissage. Nous montrons des stratégies de programmation différentes pour les configurations CBRAM 1R et 1T-1R. La méthodologie de programmation intrinsèque et extrinsèque pour des synapses CBRAM est également discutée.

Nous avons brièvement étudié la possibilité de mettre en oeuvre le comportement synaptique avec des dispositifs OXRAM en configuration 1R (sans dispositif de sélection). La modulation de  $R_{OFF}$  en mode quasi-statique et en mode impulsion a été confirmée. La modulation de  $R_{ON}$  a été démontrée uniquement en mode quasi-statique en raison de la difficulté de contrôler le courant limitant en mode impulsion avec des structures 1R. Les résultats suggèrent que la technologie étudiée pourrait être utilisée avec succès dans les réseaux de neurones impulsionnels en exploitant l'augmentation progressive de la valeur  $R_{OFF}$ , au prix toutefois d'impulsions de programmation non identiques ou variables. Nous avons également étudié les caractéristiques de commutation binaires des dispositifs. La difficulté de contrôler le taux d'échecs avec des conditions de programmation de commutation faibles rendent impossible l'exploitation de la probabilité de commutation intrinsèque. Cependant, notre technologie OXRAM peut être utilisée avec succès dans des réseaux probabilistes en utilisant des conditions de programmation fortes et en introduisant une stochasticité extrinsèque.





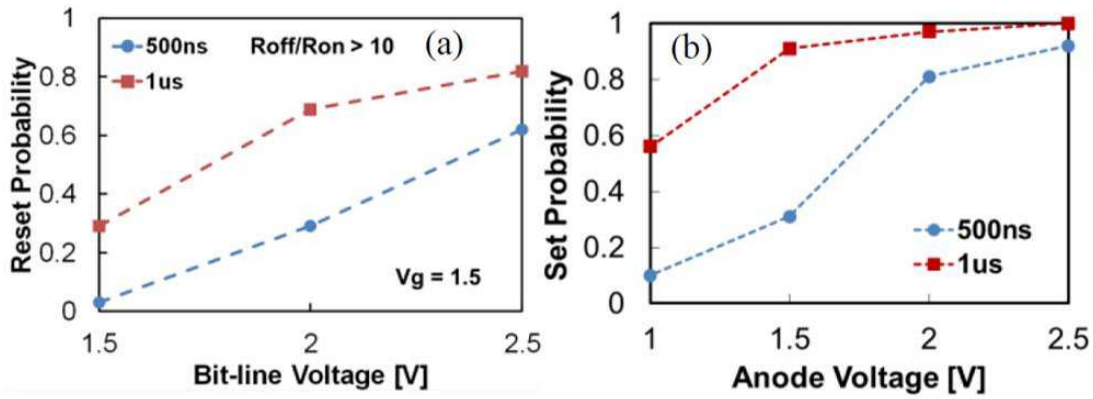
**Figure B.8:** Illustration depicting functional equivalence of deterministic multi-level and stochastic binary synapses.  $p$  indicates probability of change in conductance or switching [130].



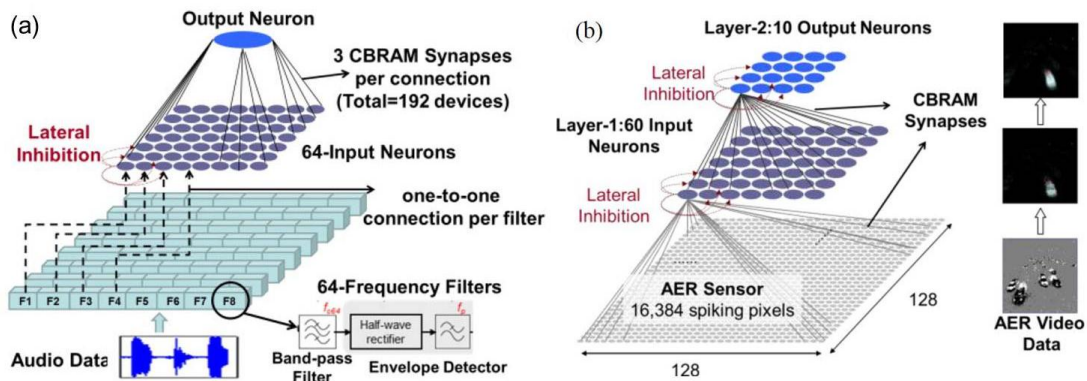
**Figure B.9:** On/Off resistance distributions of the 64 devices of the 8x8 matrix cycled 20 times. Inset shows  $R_{on}$  and  $R_{off}$  values in log scale with dispersion for each cycle [130].



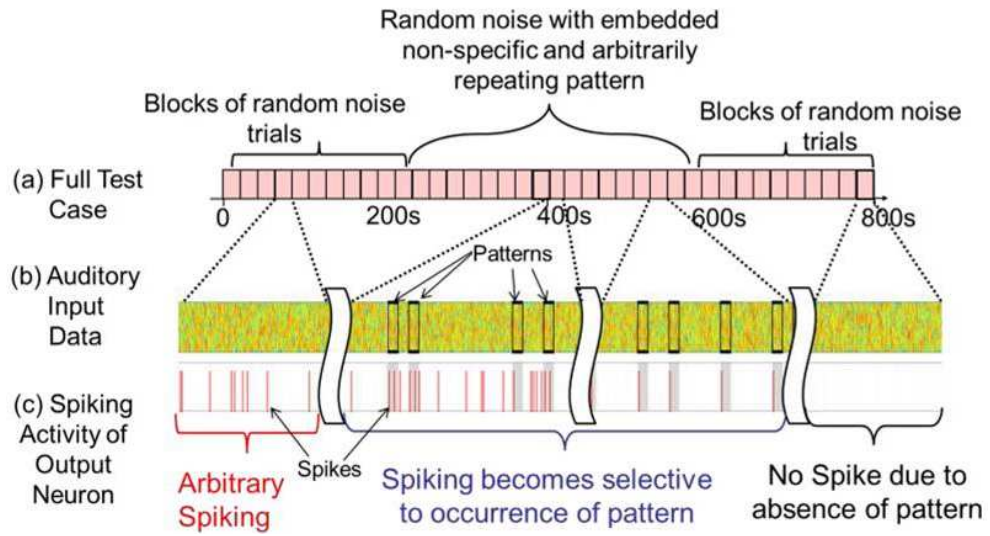
## B. RÉSUMÉ EN FRANÇAIS



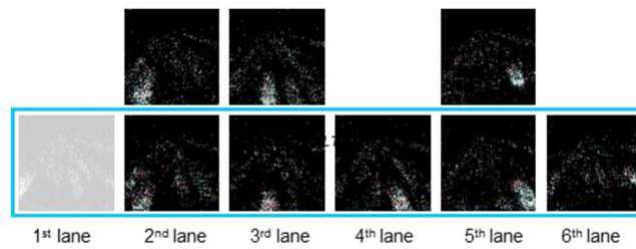
**Figure B.10:** Overall switching probability for the 64 devices of the matrix (switching being considered successful if  $R_{off}/R_{on} > 10$ ) using (a) weak-reset conditions and (b) weak-set conditions.  $V_g$  of 1.5V was used in both experiments [130].



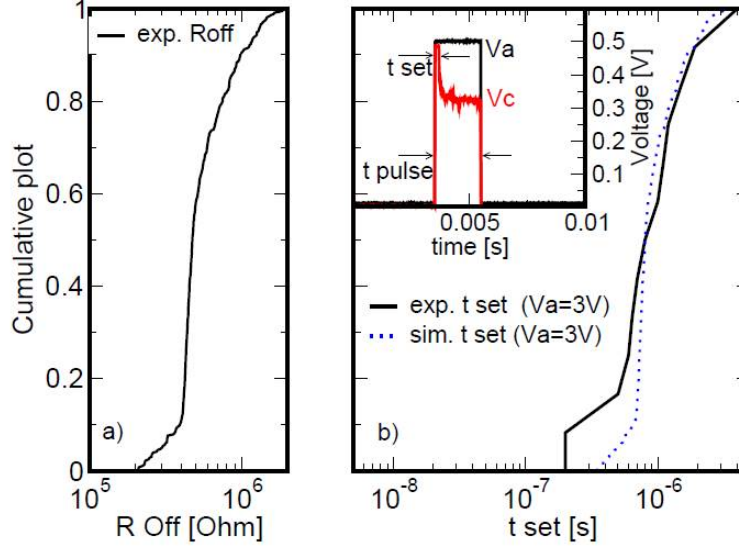
**Figure B.11:** (a) Single-layer SNN simulated for auditory processing.(b) 2-layer SNN for visual processing.(Right) AER video data snapshot with neuron sensitivity maps [130].



**Figure B.12:** (a) Full auditory-data test case with noise and embedded repeated patterns. (b) Auditory input data and (c) spiking activity for selected time intervals of the full test case of the output neuron (shown in Fig.16b) [130].



**Figure B.13:** Final sensitivity map of 9 output neurons from the 1st layer of the neural network shown in Fig.17b. Average detection rate for 5 lanes was 95% [130].



**Figure B.14:** (a)  $R_{\text{off}}$  distribution obtained in  $\text{GeS}_2$  based 1R CBRAM devices. (b) Experimental (line) and simulated (dotted)  $t_{\text{SET}}$  distribution obtained cycling the CBRAM cell with a pulse amplitude  $V_a = 3\text{ V}$ . (b in the inset) Example of a typical oscilloscope trace tracking the voltage on the CBRAM ( $V_c$ ) and the applied pulse ( $V_a$ ). Between every set operation a reset operation was performed (not shown) [163].

### B.4 Chapitre IV: Utiliser des RRAM pour la conception de neurones

Dans ce chapitre, nous avons montré comment la CBRAM utilisé de façon inattendue peut permettre la conception de neurones stochastiques avec une forte densité d'intégration. Nous avons montré comment le comportement physique de la CBRAM conduit naturellement à un comportement stochastique pour le temps de programmation, ce qui peut être exploitée dans un circuit. Des simulations SPICE valident le concept d'un simple neurone "Integrate and Fire". Le concept pourrait être étendu à la conception de neurones plus complexes comme reference [50]. Ces résultats mettent en évidence les avantages des nouvelles technologies de mémoires non-volatiles, dont l'impact peut aller au-delà des marchés traditionnels de la mémoire.

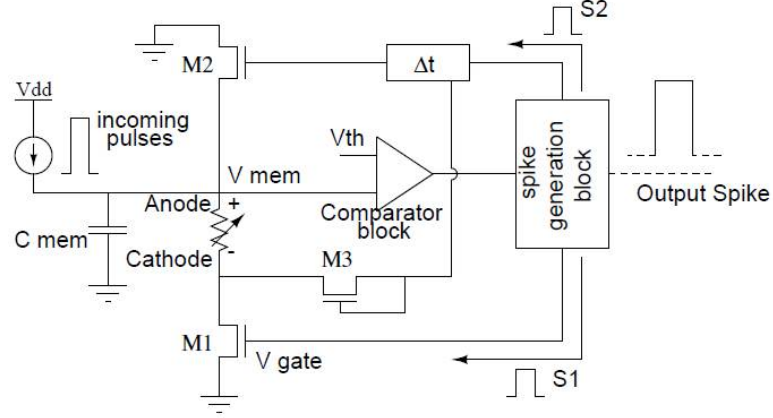


Figure B.15: Proposed circuit-equivalent of the S-IF neuron [163].

## B.5 Chapitre V: Conclusions et Perspectives

Au cours de ce travail de recherche, nous avons exploré la façon dont certains dispositifs RRAM émergents (PCM, CBRAM et OXRAM) peuvent être utilisés dans des systèmes et des applications neuromorphiques.

Nous avons insisté principalement sur l'émulation de la plasticité synaptique comme la potentialisation/dépression à long terme (LTP / LTD) et les règles d'apprentissage comme STDP. Nous avons également proposé une méthodologie pour concevoir des circuits de neurones stochastiques. Dans le cas des PCM, nous avons fabriqué et étudié des dispositifs avec des empilements différents (GST, GeTe et GST + HfO<sub>2</sub>). Nous avons montré que, bien que la conductance LTP peut être émulée avec des impulsions de programmation identiques, il n'en est pas de même pour la LTD. La différence provient de la physique sous-jacente des dispositifs à changement de phase.

Pour surmonter la limitation imposée par la LTD trop abrupte, nous avons développé une méthodologie de faible puissance unique connue sous le nom "2-PCM Synapse". Elle utilise une configuration à 2 PCM (un dispositif en charge de la LTP et l'autre pour la LTD) raccordés dans une architecture complémentaire. Nous avons développé une méthodologie de programmation détaillée (lecture, écriture et actualisation) pour l'architecture "2-PCM Synapse". Nous avons montré que cette approche présente de nombreux avantages telles que la faible consommation et la grande tolérance vis-à-vis de la dérive temporelle de la résistance des PCM car son fonctionnement est principalement basé sur la cristallisation. Nous avons montré que des dispositifs PCM avec

## B. RÉSUMÉ EN FRANÇAIS

---

une couche diélectrique de  $\text{HfO}_2$  améliorent les performances à la fois des synapses individuelles mais aussi de l'ensemble du système.

Nous avons également présenté une méthodologie pour atténuer fortement l'impact de la dérive de la résistance dans des systèmes neuromorphiques basés sur des synapses PCM. Avec l'utilisation de la "2-PCM Synapse" nous avons montré des applications d'extraction de schémas visuels complexes et nous avons aussi effectué une analyse énergétique détaillée de notre système.

Dans le cas des CBRAM, nous avons fabriqué et étudié (1R, 1T-1R et 8 x 8 matrice) des dispositifs avec des empilements Ag/GeS<sub>2</sub>. Nous avons constaté que pour la CBRAM, aussi bien la LTP que la LTD ne peuvent pas être émulées avec des impulsions identiques. Alors qu'il était possible d'émuler la LTP avec des conditions de programmation variables (Pulses non identiques), la LTD a toujours été brutale en raison de la dissolution incontrôlée des filaments. Ainsi, nous avons adopté une approche alternative pour les systèmes neuromorphiques basé sur les synapses CBRAM. Nous avons utilisé les synapses de manière probabiliste et binaire (par opposition à déterministe et multi-niveau). Nous avons utilisé une règle STDP stochastique optimisée pour compenser les états de conductance les plus bas. Nous avons montré que la stochasticité peut être mise en oeuvre soit intrinsèquement (en utilisant des conditions de programmation faibles), ou de façon extrinsèque (en utilisant des circuits de PRNG externes). Grâce à notre approche stochastique et binaire des synapses CBRAM, nous avons montré des applications d'extraction de motifs visuels et auditifs complexes. Une analyse similaire a également été réalisée pour les composants OXRAM 1R basé sur une couche de  $\text{HfO}_2$ .

D'une manière générale, nous sommes à l'heure actuelle dans des conditions très favorables pour la recherche dans le domaine du neuromorphique en raison d'importantes aides financières et de récents projets lancés dans le monde entier (Voir Sec.1.1.1). Cependant, pour un progrès soutenu et significatif dans le domaine, certaines pistes qui doivent être abordées. Ce sont:

- La Normalisation: Il y a un fort besoin de standardisation des blocs différents d'un noyau de traitement neuromorphique. Il peut s'agir de modèles de neurone, de synapse, des protocoles de communication, du format des données de sortie et d'entrée (ex-AER), des règles d'apprentissages etc... La normalisation

des blocs principaux permettrait également le développement d'architecture neuromorphique spécifique et d'outils de simulation.

- Analyse comparative de l'application: Il est nécessaire de créer une base de données d'applications pour corréler les performances d'apprentissage de systèmes neuromorphiques et des blocs de base spécifiques comme les synapses et les neurones.
- Les règles d'apprentissage: Pour les règles d'apprentissage considérées, le domaine nécessite beaucoup de développement. De nouvelles règles d'apprentissages plus adaptées aux composants (synapses ou neurones liés) devraient être développées. S'appuyer seulement sur la STDP n'est pas suffisant pour une large base d'application.

## **B. RÉSUMÉ EN FRANÇAIS**

---

# List of Figures

1.1	Proposed future computing roadmap with emerging beyond-moore technologies (adapted from IBM research colloquia-2012, Madrid, M. Ritter et. al.). . . . .	2
1.2	Data obtained from the Web of Knowledge using the search expressions: Topic=(neuromorphic and memristor) OR Topic=(neuromorphic and RRAM) OR Topic=(neuromorphic and PCM) OR Topic=(neuromorphic and Phase change) OR Topic=(neuromorphic and resistive switching) OR Topic=(neuromorphic and magnetic) OR Topic=(phase change memory and synapse) OR Topic=(conductive bridge memory and synapse) OR Topic=(PCM and synapse) OR Topic=(CBRAM and synapse) OR Topic=(RRAM and synapse) OR Topic=(OxRAM and synapse) OR Topic=(OxRAM and neuromorphic) for the time period Jan 2007- April 2013 (a) Publications, (b) Citations. . . . .	4
1.3	(a)Number of synapses Vs number of processing cores required for cat scale brain simulations using IBM Blue Gene supercomputer, (b) Growth of Top 500 supercomputers overlaid with recent IBM results and projection for realtime human-scale cortical simulation. Green line (lower) shows the 500th fastest supercomputer, dark blue line (middle) shows the fastest supercomputer, and light blue line (upper) shows the summed power of the top 500 machines [12]. . . . .	6
1.4	(a) System power scaling for IBM Watson supercomputer w.r.t human brain, (adapted from IBM research colloquia-2012, Madrid, Ritter et. al.). (b) Biological and silicon neurons have much better power and space efficiencies than digital computers [14]. . . . .	7



## LIST OF FIGURES

---

1.5	Species with increasing intelligence, number of neurons and synapses. Neuron and synapse numbers extracted from [22], [23], [24], [25], [26]. . . . .	9
1.6	Illustration showing the basic structure of a neuron cell. Inset shows a zoom of the biological synapse. Adapted and modified from [27]. . . . .	9
1.7	Illustration showing the chemical synaptic transmission process. Adapted and modified from [27]. . . . .	11
1.8	(a) Illustration of neuron Action-Potential (spike). (b) EPSP and IPSP, adapted from [29]. . . . .	12
1.9	Experimentally observed, classical anti-symmetric STDP rule, in cultured hippocampus neurons. $\Delta t < 0$ implies LTD while $\Delta t > 0$ implies LTP [37]. Change in EPSP amplitude is indicative of change in synaptic strength. . . . .	14
1.10	(a) Illustration showing different types of cells in the retina (b) Anatomical diagram of visual stimuli signal pathway starting at the photoreceptors and ending at the optic-nerve. Adapted from [32]. . . . .	15
1.11	Pathway from the retina through the LGN of the thalamus to the primary visual cortex in the human brain [32]. . . . .	16
1.12	(a) Illustration of the human ear and (b) cross-section of the cochlea, adapted from [29]. . . . .	18
1.13	(a) Illustration of uncoiled basilar membrane with different frequency sensitive regions, adapted from [47](b) inner hair cell, movement of the stereocilium leads to generation of receptor potentials, adapted from [29]	18
1.14	(a) Illustration showing the organ of corti in the cochlear cross-section. (b) zoomed view of the organ of corti showing location of the inner hair cells, adapted from [21] . . . . .	19
1.15	(a) Simplified circuit equivalent of the Hodgkin-Huxley (HH) neuron model. (b) Circuit model with synapses as variable programmable resistors [49].	21
1.16	Intel ETANN synapse structure implemented using EEPROM floating-gate devices [8]. . . . .	23
1.17	(a) Layout of poly silicon floating-gate synaptic device [52]. (b) circuit schematic of floating-gate synapse with transconductance amplifier [52]. (c) layout of floating-gate pFET synaptic device [53]. . . . .	23

1.18	(a) circuit schematic of analog DRAM based synapse with three additional transistors [55]. (b) circuit schematic of DRAM based synapse with transconductance amplifier [56]. . . . .	24
1.19	Capacitor based synapse with additional learning and weight update circuit blocks [58]. . . . .	25
1.20	IBM's 45 nm node neurosynaptic core and 8-T transposable modified SRAM cell [60]. . . . .	26
1.21	Synapse schematic comprising of 4-bit SRAM cells for the wafer-scale FACETS neuromorphic project [61]. . . . .	26
1.22	(a) Physical structure of the NOMFET. It is composed of a p+ doped bottom-gate covered with silicon oxide (200 nm). Source and drain electrodes are made of gold and Au NPs (20 nm diameter) are deposited on the interelectrode gap (5 $\mu$ m), before the pentacene deposition [64]. (b) The carbon nanotube synapse circuit [65]. (c) Neural Network Crossbar with OG-CNTFET synapse [66]. (d) Schematic representation of a nanotube network-based OG-CNTFET [67]. . . . .	29
1.23	(a) Schematics of a Ag <sub>2</sub> S atomic switch inorganic synapse. Application of input pulses causes the precipitation of Ag atoms from the Ag <sub>2</sub> S electrode, resulting in the formation of a Ag atomic bridge between the Ag <sub>2</sub> S electrode and a counter metal electrode. When the precipitated Ag atoms do not form a bridge, the inorganic synapse works as STP. After an atomic bridge is formed, it works as LTP [68]. (b) SEM image of complex Ag networks produced by reaction of aqueous AgNO <sub>3</sub> with (inset) lithographically patterned Cu seed posts [70]. . . . .	30
1.24	Classification of the resistive switching effects which are considered for non-volatile memory applications [72]. . . . .	31
1.25	(a) Photo of the constituents of the copper-sulphate based memistor device. (b) Characteristic programming curves showing hysteresis loop in the memistor devices, adapted from [76]. . . . .	34
1.26	(a) Photo of the ADALINE architecture with 1 neuron and 3x3 memistor synapses [76]. (b) Recent photo of the ADALINE system containing memistors taken at IJCNN-2011. . . . .	34

## LIST OF FIGURES

---

1.27	(a) Cross-section schematic of the tungsten oxide based 3-terminal memistor. (b) Programming characteristics of the solid state memistor device, adapted from [80]. . . . .	35
1.28	Bi-directional strategy (Top-down + Bottom-up) adopted for the work presented in this PhD thesis. To develop the ideal "synapse-solution" optimization and fine-tuning was performed at different levels such as architectures, learning-rules and programming-schemes.(BS: Binary Stochastic synapses). . . . .	36
2.1	Illustration of biological synapse and the equivalent PCM synapse in a neural circuit connecting a spiking pre- and post- neuron [81]. TEM cross-section image of the GST PCM devices fabricated for this study is shown. . . . .	40
2.2	(a) Cross-section of the GST-PCM device showing amorphous and crystalline regions. (b) Potentiation and depression using voltage pulse trains with changing amplitude, adapted from [82]. . . . .	41
2.3	(a) Spike scheme with set and reset pulse trains. (b) Spikes for different forms of STDP [83]. . . . .	42
2.4	PCM conductance increases or decreases in response to negative or positive pulses, respectively. The pulse amplitudes vary with identical 30 ns widths [85]. . . . .	42
2.5	(a) IV characteristics for PCM devices with 100 nm thick GST and GeTe layer starting from initially amorphous phase. (b) R-I characteristics of GST and GeTe PCM devices, with inset showing the PCM phase of intermediate resistance states. (c) R-V curves for GST devices with six different pulse widths. Read pulse = 0.1 V, 1 ms. Legend indicates applied pulse widths. (d) Temperature Vs Time profile for PCM programming pulses [81]. . . . .	44

2.6	(a) Experimental LTP characteristics of GST PCM devices. For each curve, first a reset pulse (7 V, 100 ns) is applied followed by 30 consecutive identical potentiating pulses (2 V). Dotted lines correspond to the behavioral model fit described in Eq.2.3 and eq.2.4. (b) Experimental LTP characteristics of GeTe PCM devices. (c) LTP simulations for GST devices using circuit-compatible model. (d) Conductance evolution as a function of the applied voltage for GST devices with six different pulse widths, using circuit-compatible model (sec.2.5.2). Legends in Figs.2.6(a–d) indicate pulse widths [81]. . . . .	45
2.7	(a) Illustration showing a single test sequence applied to the PCM device, consisting of one potentiating and two depressing events of varying intensity. (b) Variation of PCM resistance for all the 30 test sequences. The circle indicate resistance after application of the potentiating pulse, while the square and triangle indicate the reading values after application of 1st and 2nd depressing pulses respectively. (c) Variation of PCM resistance with every individual pulse [92]. . . . .	49
2.8	(a) Illustration showing a single test sequence, consisting of two identical depressing pulses and one potentiating pulse. (b) Variation of PCM resistance for all the 30 test sequences. The circle indicate resistance after application of the potentiating pulse, while the square and triangle indicate the reading values after application of 1st and 2nd depressing pulses respectively. (c) Variation of PCM resistance with every individual pulse [92]. . . . .	50
2.9	Experimental LTD characteristics of GST and GeTe PCM devices. Inset shows simulated phase morphology of GST layer after the application of consecutive depressing pulses [81]. . . . .	51
2.10	(a) 2D Axi-symmetrical half cell description used for physical simulations. (b) Simulated time evolution of applied voltage pulse and drop across the device for a potentiating pulse. (c) Simulated maximum temperature in GST layer with the applied pulse. (d) Simulated current passing through the device during the applied pulse. (e) Simulated resistance of the device with the applied pulse [81]. . . . .	52

## LIST OF FIGURES

---

2.11	(a) Simulated depressing (reset) pulse indicating the instance of time snapshot. (b) Time snapshot of the simulated phase morphology of the GST phase change layer [81]. . . . .	53
2.12	(a) Simulated LTP curves while fixing the nucleation rate (NR) and varying the growth rate GR compared to GST (taken as reference: GR = 1, NR = 1). Corresponding simulations of GST layer morphology are shown (0th pulse: reset; 1st-5th: potentiating). (b) Simulated LTP curves while fixing the growth rate (GR = 1) and varying the nucleation rate (NR) compared to GST (taken as reference material: NR = 1, GR = 1). Corresponding simulation of GST layer morphology are also shown [81]. . . . .	54
2.13	(a) TEM of the GST PCM device with HfO <sub>2</sub> interface layer. (b) Resistance versus voltage applied curve, for the interface layer (2nm thick HfO <sub>2</sub> ) PCM device, during the initial forming step[98]. . . . .	61
2.14	(a) LTP curves for the GST devices with 2 nm thick HfO <sub>2</sub> layer. Inset shows the test procedure (Reset pulse: 4V/100ns and potentiating pulse: 2.1 V. (b) Current values for each pulse of the LTP test shown in (a). [98].	62
2.15	(a) Experimentally acquired waveforms for the applied voltage pulse and the actual voltage drop across the PCM devices. The waveforms were acquired for the first 5 pulses of the 200 ns LTP curve shown in Fig.2.14a. (b) Graph showing a plot of effective pulse width and the normalized differential resistance (for 200 ns LTP curve shown in Fig.2.14). The resistance values were normalized with respect to maximum value of $\Delta R$ , [98]. . . . .	63
2.16	The "2-PCM Synapse" concept schematic. The contribution of the current flowing through the LTP device is positive, while that of the LTD device is negative, towards the integration in the output neuron [110]. .	64
2.17	(a) Biological STDP (from [37]) and simplified STDP used in the proposed PCM implementation. In the simplified rule, a synapse receiving a postsynaptic spike with no presynaptic spike in the LTP window undergoes an LTD regardless of the existence of a presynaptic spike [111]. (b) Write, Reset and Read pulses for the programming scheme proposed in sec.2.7.2, [110]. . . . .	65

2.18	Read Operations. Current from both LTP and LTD PCM devices is integrated in the output neuron, with a positive and negative contribution, respectively [111]. . . . .	67
2.19	Write operations based on the simplified-STDP rule. For a specific PCM, $G \nearrow$ denotes an increase in conductance (thus, partial crystallization of the device), while $G \rightarrow$ denotes no change in conductance [111]. . . . .	67
2.20	Refresh principle: The two devices forming a synapse are reset, and the one that had the higher conductance is reprogrammed such that the equivalent weight of the synapse stays unchanged [111]. . . . .	69
2.21	Refresh-operation flowchart [111]. . . . .	69
2.22	Refresh-operation: RESET pulses generation to re-initialize the LTP and LTD devices conductance to the minimum when $V_{\text{RESET}} > 2.V_{\text{SET}}$ . [111]. . . . .	70
2.23	Refresh-operation without selectors: RESET pulses generated to re-initialize the LTP and LTD devices conductance to the minimum when $V_{\text{RESET}} < 2.V_{\text{SET}}$ . The LTD (respectively LTP) device is reset when the negative part $-V_{\text{ER}}$ (respectively positive part $V_{\text{ER}}$ ) of the erase pulse coming from the input neuron overlaps with the post-synaptic erase pulse [111]. . . . .	71
2.24	2-Layer Spiking Neural Network (SNN) topology used in simulation. The network is fully connected and each pixel of the $128 \times 128$ pixel AER dynamic vision sensor (DVS-retina) is connected to every neuron of the 1st layer through two synapses, receiving positive and negative change in illumination events respectively. Lateral inhibition is also implemented for both layers [110]. . . . .	73
2.25	AER video data snapshot. Cars passing on a freeway recorded with the DVS-sensor described in [117]. The original video has no lane demarcations, yellow demarking lines were drawn later for lane-clarity [110]. . .	74

## LIST OF FIGURES

---

2.26	How synapses actually look inside the neural network: Simulated variability for (a) GST- (b) GeTe- and (c)Interface- PCM devices. The plots show the increase in conductance as a function of the number of SET pulses (LTP-curves), for 100 different sets of parameters, obtained by applying 20% dispersion (standard deviation of the mean value) from values extracted from fitting [98], [111]. Inset of (c) shows the experimentally obtained LTP-curve for a 200 ns potentiating pulse. . . . .	76
2.27	Learning Results for GST, GeTe and Interface synapses: (a) Final output neuron sensitivity patterns for the 6 lanes and (b) Lane-specific car detection rates, [110], [98]. Lane-6 is not learnt by any neuron. . . . .	77
2.28	PCM programming current scaling trend [110]. . . . .	80
2.29	(a) Resistance drift with time for different initial programming conditions. Measurement was carried out on GST PCM devices. Equation governing the drift dynamics is also shown. (b) Experimental Resistance-Voltage curves for different programming pulse widths on GST PCM devices [122]. . . . .	81
2.30	Illustration depicting functional equivalence of deterministic multi-level and stochastic binary synapses. $p$ indicates probability of change in conductance or switching [130]. . . . .	83
2.31	(a) Simplified stochastic STDP learning rule. On corresponds to set and Off to reset of the PCM synapse. (b) Schematic of the Binary-PCM Synapse architecture and the proposed programming-scheme [122]. . . .	83
2.32	Schematic of the Binary-PCM Synapse architecture and the proposed programming-scheme for selector-free configuration. . . . .	85
2.33	(a) Distribution of synaptic resistance states for the "2-PCM Synapse" architecture at the end of the visual learning simulation. (b) Distribution of synaptic resistance states for the "Binary-PCM Synapse" architecture with 20 k $\Omega$ mean $R_{off}$ [122]. . . . .	86
2.34	(a) Distribution of synapses in off-state for the "Binary-PCM Synapse" and (b) Distribution of synapses in on-state, for the PCM synapses with mean $R_{off}$ values of 20 k $\Omega$ , 30 k $\Omega$ and 123 k $\Omega$ [122]. . . . .	87

2.35	(Left) Comparison of learning statistics for the "2-PCM Synapse" and "Binary-PCM Synapse" architectures. (Right) Car detection rates for the "Binary-PCM Synapse" architecture [122]. For both statistics two cases of "Binary-PCM Synapse" are shown (with mean $R_{off} = 20 \text{ k}\Omega$ and $30 \text{ k}\Omega$ ). . . . .	88
2.36	(a) Ratio between the number of RESET and SET events as a function of the resistance window $R_{OFF}/R_{ON}$ . (b) Programming power as a function of decreasing $R_{OFF}$ - red line (keeping $R_{ON} = 110 \text{ }\Omega$ constant) and increasing $R_{ON}$ - blue line, (keeping $R_{OFF} = 1.06 \text{ M}\Omega$ constant). (c) Read power as a function of decreasing $R_{OFF}$ - red line (keeping $R_{ON} = 110 \text{ }\Omega$ constant), and increasing $R_{ON}$ - blue line (keeping $R_{OFF} = 1.06 \text{ M}\Omega$ constant) [133]. . . . .	90
2.37	SET and RESET events as functions of resistance window [133]. . . . .	91
3.1	(a) Schematic of the CBRAM device showing the formation of the conductive-filament (b) Simulated modulation of $R_{off}$ , adapted from [137]. . . . .	94
3.2	(a) Shapes of pre- and post- neuron spikes used to emulate STDP on Ag/Ge <sub>0.3</sub> Se <sub>0.7</sub> CBRAM devices (b) Simulated STDP-like curve for the Ag/Ge <sub>0.3</sub> Se <sub>0.7</sub> CBRAM devices, adapted from [137]. . . . .	95
3.3	(a) Incremental increase and decrease of device conductance on application of potentiating and depressing pulses (b) Demonstration of STDP in the Ag+Si/Ag devices, inset shows a SEM image of the fabricated synapse array, adapted from [138]. . . . .	96
3.4	(Left) TEM of the CBRAM resistor element. (Right) Circuit schematic of the 8 X 8 1T-1R CBRAM. matrix. (note: the devices used in this study had a GeS <sub>2</sub> layer thickness of 30 nm. The 50 nm TEM is for illustrative purpose only [130].) . . . . .	97
3.5	(a) On-state resistance modulation using current compliance. Fitting using model [136] is also shown (extracted filament radius are indicated). (b) Resistance dependence on gate voltage during the set-to-reset transition [130]. . . . .	98
3.6	On/Off resistance distribution of an isolated 1T-1R device during 400 cycles when strong programming is used [130]. . . . .	99



## LIST OF FIGURES

---

3.7	Computed distributions (generated using Roff data from fig. 3.6 and model [136], of: (a) Tset and (b) Vset (Inset) values for consecutive successful set operation (mean and sigma are indicated). For computing (a) the applied voltage is 1 V and for (b) a ramp rate of 1 V/s is used in the quasi-static mode [130]. . . . .	100
3.8	Stochastic switching of 1T-1R device during 1000 cycles using weak-conditions (switch-probability=0.49) [130]. . . . .	100
3.9	On/Off resistance distributions of the 64 devices of the 8x8 matrix cycled 20 times. Inset shows Ron and Roff values in log scale with dispersion for each cycle [130]. . . . .	101
3.10	Overall switching probability for the 64 devices of the matrix (switching being considered successful if $R_{off}/R_{on} > 10$ ) using (a) weak-reset conditions and (b) weak-set conditions. Vg of 1.5V was used in both experiments [130]. . . . .	101
3.11	(a) Circuit schematic with CBRAM synapses without selector devices, LIF neurons, in the external probability case. (b) Circuit schematic with CBRAM synapses with selector devices, LIF neurons, in the external probability case. In both cases, the presented voltages waveforms implement the simplified STDP learning rule for the CBRAMs [130]. . .	103
3.12	Probabilistic STDP learning rule (used for audio application). X-axis shows the time difference of post-and pre-neuron spike [130]. . . . .	104
3.13	Schematic showing two different approaches for implementing stochasticity with binary synapses [130]. . . . .	105
3.14	Tunable Pseudo-random-number generator (PRNG) circuit [131], the output being tuned according to STDP in Fig.3.12. . . . .	106
3.15	Concept and data-flow of the unsupervised learning simulations [130]. .	106
3.16	(a) Single-layer SNN simulated for auditory processing.(b) 2-layer SNN for visual processing.(Right) AER video data snapshot with neuron sensitivity maps [130]. . . . .	107
3.17	(a) Full auditory-data test case with noise and embedded repeated patterns. (b) Auditory input data and (c) spiking activity for selected time intervals of the full test case of the output neuron (shown in Fig.16b) [130]. . . . .	108

3.18	(a) Pattern Sensitivity ( $d'$ ) for the test case shown in fig. 3.17. The system reaches a very high sensitivity ( $d' > 2$ ). (b) Number of false detections by the output neuron during the auditory learning [130]. . . . .	108
3.19	Final sensitivity map of 9 output neurons from the 1st layer of the neural network shown in Fig.17b. Average detection rate for 5 lanes was 95% [130]. . . . .	109
3.20	Illustration depicting OXRAM working principle and underlying physics [74]. . . . .	112
3.21	(a) LTD obtained by increasing RESET pulse, LTP by increasing SET pulse. [146] (b) LTD obtained by increasing the RESET pulse amplitude, LTP by increasing compliance current [147]. (c) LTD only, obtained by identical RESET pulses [148]. (d) and (e): LTD and LTP obtained by identical RESET and SET voltage [149], [150]. . . . .	113
3.22	(a) Schematic view of the tested OXRAM memory device. (b) Typical I-V characteristic for tested OXRAM devices (10 nm $\text{HfO}_x$ ). . . . .	114
3.23	(a) Schematic description of test case for $R_{OFF}$ modulation in quasi-static mode with SET 'refresh' operation. (b) Experimental data confirming $R_{OFF}$ modulation in quasi-static mode (10 nm $\text{HfO}_x$ ). . . . .	115
3.24	(left) Schematic description of test for $R_{OFF}$ modulation in pulsed-mode with SET 'refresh' operation. (right) Experimental data confirming $R_{OFF}$ modulation in pulsed-mode (10 nm $\text{HfO}_x$ ). . . . .	115
3.25	(left) Schematic description of test for $R_{OFF}$ modulation in pulse mode without SET 'refresh' operation. (right) Experimental data confirming $R_{OFF}$ modulation in pulse mode (10 nm $\text{HfO}_x$ ). . . . .	116
3.26	(left) Schematic description of test for $R_{OFF}$ modulation in pulse mode with identical reset-pulses. (right) Experimental data for the described test with identical reset pulses (10 nm $\text{HfO}_x$ ). . . . .	117
3.27	Cumulative device response in conductance to identical RESET pulse train for (a) 10 nm thick and (b) 15 nm thick $\text{HfO}_x$ layer devices. . . . .	117
3.28	Schematic description of test for $R_{ON}$ modulation in quasi-static mode. (b) Experimental data showing the trend in $R_{ON}$ . . . . .	118
3.29	Device response to a pulse train of identical SET pulses. . . . .	118

## LIST OF FIGURES

---

3.30	Binary switching operation for our OXRAM devices with 'strong' programming conditions. . . . .	119
3.31	Binary switching operation for our OXRAM devices with 'weak' RESET programming conditions. . . . .	120
3.32	Binary switching operation for our OXRAM devices with 'weak' SET programming conditions. . . . .	121
4.1	(a) $R_{\text{Off}}$ distribution obtained in $\text{GeS}_2$ based 1R CBRAM devices.(b) Experimental (line) and simulated (dotted) $t_{\text{SET}}$ distribution obtained cycling the CBRAM cell with a pulse amplitude $V_a=3\text{V}$ . (b in the inset) Example of a typical oscilloscope trace tracking the voltage on the CBRAM ( $V_c$ ) and the applied pulse ( $V_a$ ). Between every set operation a reset operation was performed (not shown) [163]. . . . .	125
4.2	(a) Schematic image shown the basic concept of a Integrate and Fire neuron [50]. (b) Schematic showing the basic concept of our proposed Stochastic Integrate-Fire neuron (S-IF) [163]. . . . .	127
4.3	(a)-(d) Schematic of output neuron firing patterns for different example test cases [163]. . . . .	128
4.4	Proposed circuit-equivalent of the S-IF neuron [163]. . . . .	129
4.5	Circuit used to demonstrate the concept of a S-IF effect when the CBRAM is in the set state [163]. . . . .	129
4.6	Full evolution of $V_{\text{mem}}$ simulating the circuit shown in Fig. 4.5. (a) Pre-neuron incoming pulses are used to build up $V_{\text{mem}}$ . (b) Initially $V_{\text{mem}}$ builds up as consequence of incoming currents (charging phase). Set operation lead to different discharge of $C_{\text{mem}}$ ( $t_{\text{dsc}}$ ). During the recharging phase a different number of incoming pulses will raise $V_{\text{mem}}$ till $V_{\text{th}}$ . (c) Expected different inter-spike intervals depending on the $t_{\text{set}}$ [163]. . . . .	131
4.7	(a) Pre-neuron incoming pulses are used to build up $V_{\text{mem}}$ . (b) Zoom on $V_{\text{mem}}$ during the discharging phase for different $t_{\text{SET}}$ in the range 300 ns - 600 ns. Lower $t_{\text{SET}}$ leads to lower residual membrane voltage $V_{\text{mem}}$ [163].	132

4.8	(a) Time-evolution of $V_{\text{mem}}$ and $V_{\text{cathode}}$ that establish a voltage drop on the CBRAM to enable reset operation. Larger M3 increase the voltage drop, since $V_{\text{cathode}}$ builds up more. $V_{\text{mem}}$ corresponding to a $t_{\text{SET}}$ of 300 ns is considered. (b) Pulse applied to M3 [163]. . . . .	133
4.9	(a) Time-evolution of $V_{\text{mem}}$ during the reset operation for $t_{\text{SET}}$ in the range 300 ns - 600 ns. Different residual voltages are obtained. (b) Pulse applied to M3 [163]. . . . .	134
5.1	Snapshot of current on-going activities (a) Multi-electrode Array (MEA) of the NeuroPXi system to collect neuron signals [166]. (b) Neuromorphic test-board developed by CEA-LIST for testing packaged RRAM. (c) Wire-bonded and packaged PCM and CBRAM devices. (d) Layout of a CBRAM synapse array + CMOS control circuit designed in collaboration with CEA-LIST. . . . .	142
B.1	Bi-directional strategy (Top-down + Bottom-up) adopted for the work presented in this PhD thesis. To develop the ideal "synapse-solution" optimization and fine-tuning was performed at different levels such as architectures, learning-rules and programming-schemes.(BS: Binary Stochastic synapses). . . . .	150
B.2	Illustration of biological synapse and the equivalent PCM synapse in a neural circuit connecting a spiking pre- and post- neuron [81]. TEM cross-section image of the GST PCM devices fabricated for this study is shown. . . . .	151
B.3	(a) IV characteristics for PCM devices with 100 nm thick GST and GeTe layer starting from initially amorphous phase. (b) R-I characteristics of GST and GeTe PCM devices, with inset showing the PCM phase of intermediate resistance states. (c) R-V curves for GST devices with six different pulse widths. Read pulse = 0.1 V, 1 ms. Legend indicates applied pulse widths. (d) Temperature Vs Time profile for PCM programming pulses [81]. . . . .	152

## LIST OF FIGURES

---

B.4	(a) Experimental LTP characteristics of GST PCM devices. For each curve, first a reset pulse (7 V, 100 ns) is applied followed by 30 consecutive identical potentiating pulses (2 V). Dotted lines correspond to the behavioral model fit described in Eq.2.3 and eq.2.4. (b) Experimental LTP characteristics of GeTe PCM devices. (c) LTP simulations for GST devices using circuit-compatible model. (d) Conductance evolution as a function of the applied voltage for GST devices with six different pulse widths, using circuit-compatible model (sec.2.5.2). Legends in Figs.2.6(a–d) indicate pulse widths [81]. . . . .	153
B.5	The "2-PCM Synapse" concept schematic. The contribution of the current flowing through the LTP device is positive, while that of the LTD device is negative, towards the integration in the output neuron [110]. .	154
B.6	2-Layer Spiking Neural Network (SNN) topology used in simulation. The network is fully connected and each pixel of the 128×128 pixel AER dynamic vision sensor (DVS-retina) is connected to every neuron of the 1st layer through two synapses, receiving positive and negative change in illumination events respectively. Lateral inhibition is also implemented for both layers [110]. . . . .	155
B.7	(Left) TEM of the CBRAM resistor element. (Right) Circuit schematic of the 8 X 8 1T-1R CBRAM. matrix. (note: the devices used in this study had a GeS <sub>2</sub> layer thickness of 30 nm. The 50 nm TEM is for illustrative purpose only [130].) . . . . .	156
B.8	Illustration depicting functional equivalence of deterministic multi-level and stochastic binary synapses. $p$ indicates probability of change in conductance or switching [130]. . . . .	157
B.9	On/Off resistance distributions of the 64 devices of the 8x8 matrix cycled 20 times. Inset shows $R_{on}$ and $R_{off}$ values in log scale with dispersion for each cycle [130]. . . . .	157
B.10	Overall switching probability for the 64 devices of the matrix (switching being considered successful if $R_{off}/R_{on} > 10$ ) using (a) weak-reset conditions and (b) weak-set conditions. $V_g$ of 1.5V was used in both experiments [130]. . . . .	158

B.11 (a) Single-layer SNN simulated for auditory processing.(b) 2-layer SNN for visual processing.(Right) AER video data snapshot with neuron sensitivity maps [130]. . . . .	158
B.12 (a) Full auditory-data test case with noise and embedded repeated patterns. (b) Auditory input data and (c) spiking activity for selected time intervals of the full test case of the output neuron (shown in Fig.16b) [130]. . . . .	159
B.13 Final sensitivity map of 9 output neurons from the 1st layer of the neural network shown in Fig.17b. Average detection rate for 5 lanes was 95% [130]. . . . .	159
B.14 (a) $R_{\text{OFF}}$ distribution obtained in $\text{GeS}_2$ based 1R CBRAM devices.(b) Experimental (line) and simulated (dotted) $t_{\text{SET}}$ distribution obtained cycling the CBRAM cell with a pulse amplitude $V_a=3\text{V}$ . (b in the inset) Example of a typical oscilloscope trace tracking the voltage on the CBRAM ( $V_c$ ) and the applied pulse ( $V_a$ ). Between every set operation a reset operation was performed (not shown) [163]. . . . .	160
B.15 Proposed circuit-equivalent of the S-IF neuron [163]. . . . .	161

## LIST OF FIGURES

---

# List of Tables

1.1	Comparison of emerging RRAM technology with Standard VLSI technologies. Adapted from ITRS-2012. (Values indicated for PCM and Redox are the best demonstrated.) Redox includes both CBRAM and OXRAM devices. . . . .	32
2.1	Fitting parameters of the behavioral model for 300 ns GST LTP curve and 100 ns GeTe LTP curve shown in Fig.2.6a and Fig.2.6b respectively.	57
2.2	Parameters used for the GST compact model simulations shown in Fig.2.6c.	58
2.3	Neuron parameters for the learning. A different set of parameters is used depending on the PCM materials. See [114] for a detailed explanation of the parameters. . . . .	75
2.4	Learning statistics, over the whole learning duration ( $8 \times 85 = 680$ s). The SET pulses number includes both the write pulses for the learning and the additional pulses to reprogram the equivalent synaptic weight during refresh operations (Fig.2.20). . . . .	78
2.5	Energy statistics and synaptic power for the test case described in table 2.4, by using voltage and current values extracted from literature. . . . .	79
3.1	Network statistics for auditory and visual learning simulations with stochastic binary CBRAM synapses [130]. . . . .	110
3.2	Energy/Power statistics for auditory and visual learning simulations with stochastic binary CBRAM synapses [130]. . . . .	111
3.3	Recent OXRAM based synaptic emulations . . . . .	112
3.4	Learning statistics, over the whole learning duration for binary OXRAM synapses( $8 \times 85 = 680$ s). . . . .	121



## LIST OF TABLES

---

3.5	Energy/Power statistics for visual learning simulations with stochastic binary OXRAM synapses (Total learning duration = 680 s ). . . . .	121
4.1	Parameters used in the simulations . . . . .	126

# Bibliography

- [1] W. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *Bulletin of Mathematical Biology*, vol. 5, pp. 115–133, Dec. 1943. 3
- [2] F. Rosenblatt, “The perceptron: a probabilistic model for information storage and organization in the brain,” pp. 89–114, 1988. 3
- [3] M. Minsky and S. Papert, *Perceptrons: An introduction to computational geometry*. M.I.T. Press, 1969. 3
- [4] P. J. Werbos, *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. PhD thesis, Harvard University, 1974. 3
- [5] D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, *Parallel Distributed Processing, Vol. 1: Foundations*. The MIT Press, jul 1987. 3
- [6] C. Mead, “Neuromorphic electronic systems,” *Proceedings of the IEEE*, vol. 78, pp. 1629–1636, Oct. 1990. 3
- [7] J.-Y. Boulet, D. Louis, C. Godefroy, P. Tannhof, and G. Paillet, “Neuron circuit,” *United States Patent*, vol. 5621863, 1997. 3
- [8] M. Holler, S. Tam, H. Castro, and R. Benson, “An electrically trainable artificial neural network (etann) with 10240 ‘floating gate’ synapses,” in *Neural Networks, 1989. IJCNN., International Joint Conference on*, pp. 191–196 vol.2, 1989. 3, 22, 23, 166
- [9] U. Ramacher, W. Raab, J. Hachmann, J. Beichter, N. Bruls, M. Wesseling, E. Sicheneder, J. Glass, A. Wurz, and R. Manner, “Synapse-1: a high-speed general purpose parallel neurocomputer system,” in *Parallel Processing Symposium, 1995. Proceedings., 9th International*, pp. 774–781, 1995. 3

## BIBLIOGRAPHY

---

- [10] C. Gamrat, A. Mougin, P. Peretto, and O. Ulrich, “The architecture of mind neurocomputers,” in *MicroNeuro Int. Conf. on Microelectronics for Neural Networks, Munich, Germany*, pp. 463–469, 1991. 3
- [11] S. Thorpe, D. Fize, and C. Marlot, “Speed of processing in the human visual system,” *Nature*, no. 381, pp. 520–522, 1996. 3, 17
- [12] R. Ananthanarayanan, S. K. Esser, H. D. Simon, and D. S. Modha, “The cat is out of the bag: cortical simulations with 109 neurons, 1013 synapses,” *SC Conference*, vol. 0, pp. 1–12, 2009. 6, 165
- [13] A. Muthuramalingam, S. Himavathi, and E. Srinivasan, “Neural network implementation using fpga: issues and application,” *International journal of information technology*, vol. 4, no. 2, pp. 86–92, 2008. 5
- [14] C.-S. Poon and K. Zhou, “Neuromorphic silicon neurons and large-scale neural networks: challenges and opportunities,” *Frontiers in Neuroscience*, vol. 5, no. 108, 2011. 7, 165
- [15] J. Arthur, P. Merolla, F. Akopyan, R. Alvarez, A. Cassidy, S. Chandra, S. Esser, N. Imam, W. Risk, D. Rubin, R. Manohar, and D. Modha, “Building block of a programmable neuromorphic substrate: A digital neurosynaptic core,” in *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pp. 1–8, june 2012. 7
- [16] M. Paliwal and U. A. Kumar, “Neural networks and statistical techniques: A review of applications,” *Expert Systems with Applications*, vol. 36, no. 1, pp. 2–17, 2009. 7
- [17] Q. V. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. S. Corrado, J. Dean, and A. Y. Ng, “Building high-level features using large scale unsupervised learning,” *arXiv preprint arXiv:1112.6209*, 2011. 7
- [18] T. Chen, Y. Chen, M. Duranton, Q. Guo, A. Hashmi, M. Lipasti, A. Nere, S. Qiu, M. Sebag, and O. Temam, “Benchmn: On the broad potential application scope of hardware neural network accelerators,” in *IEEE International Symposium on Workload Characterization (IISWC)*, November 2012. 8

- [19] M. V. Srinivasan and M. Ibbotson, “Biologically inspired strategies, algorithms and hardware for visual guidance of autonomous helicopters,” tech. rep., DTIC Document, 2011. 8
- [20] T. W. Berger, M. Baudry, R. D. Brinton, J.-S. Liaw, V. Z. Marmarelis, A. Yoon-dong Park, B. J. Sheu, and A. R. Tanguay Jr, “Brain-implantable biomimetic electronics as the next era in neural prosthetics,” *Proceedings of the IEEE*, vol. 89, no. 7, pp. 993–1012, 2001. 8
- [21] E. R. Kandel, J. H. Schwartz, and T. M. Jessell, *Principles of neural science*. McGraw-Hill, Health Professions Division, New York, 2000. 8, 10, 19, 166
- [22] B. Schierwater, “My favorite animal, *Trichoplax adhaerens*,” *BioEssays : news and reviews in molecular, cellular and developmental biology*, vol. 27, pp. 1294–1302, Dec. 2005. 9, 166
- [23] J. G. White, E. Southgate, J. N. Thomson, and S. Brenner, “The structure of the nervous system of the nematode *caenorhabditis elegans*,” *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, vol. 314, no. 1165, pp. 1–340, 1986. 9, 166
- [24] R. Menzel and M. Giurfa, “Cognitive architecture of a mini-brain: the honeybee,” *Trends in Cognitive Sciences*, vol. 5, pp. 62–71, Feb. 2001. 9, 166
- [25] R. W. Williams, “Mapping genes that modulate mouse brain development: a quantitative genetic approach,” in *Mouse brain development*, pp. 21–49, Springer, 2000. 9, 166
- [26] S. Herculano-Houzel, “The human brain in numbers: a linearly scaled-up primate brain,” *Frontiers in Human Neuroscience*, vol. 3, no. 31, 2009. 9, 166
- [27] J. Karey, L. Ariniello, and M. McComb, *Brain Facts: A primer on the brain and nervous system*. The Society for Neuroscience, Washington DC, 2002. 9, 11, 166
- [28] M. V. Bennett and R. Zukin, “Electrical coupling and neuronal synchronization in the mammalian brain,” *Neuron*, vol. 41, no. 4, pp. 495 – 511, 2004. 10

## BIBLIOGRAPHY

---

- [29] A. Borisyuk, G. Ermentrout, A. Friedman, and D. Terman, *Tutorials in Mathematical Biosciences I, Mathematical Neuroscience*, vol. 1860. 12, 18, 166
- [30] B. Walmsley, F. R. Edwards, and D. J. Tracey, “The probabilistic nature of synaptic transmission at a mammalian excitatory central synapse.,” *Neuroscience*, vol. 7, p. 1037, Apr 1987. 12
- [31] J. H. Byrne and J. L. Roberts, *From molecules to networks: An introduction to cellular and molecular neuroscience*. Elsevier Science, 2004. 13
- [32] P. Dayan and L. F. Abbott, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. The MIT Press, 2001. 13, 14, 15, 16, 166
- [33] R. C. Malenka and M. F. Bear, “LTP and LTD: an embarrassment of riches.,” *Neuron*, vol. 44, pp. 5–21, Sept. 2004. 13
- [34] H. D. Abarbanel, L. Gibb, R. Huerta, and M. I. Rabinovich, “Biophysical model of synaptic plasticity dynamics.,” *Biol Cybern*, vol. 89, pp. 214–226, Sept. 2003. 13
- [35] D. H. O’Connor, G. M. Wittenberg, and S. S.-H. Wang, “Graded bidirectional synaptic plasticity is composed of switch-like unitary events,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 27, pp. 9679–9684, 2005. 13
- [36] H. D. I. Abarbanel, S. S. Talathi, L. Gibb, and M. I. Rabinovich, “Synaptic plasticity with discrete state synapses,” *Phys. Rev. E*, vol. 72, p. 031914, Sep 2005. 13
- [37] 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. 10464–10472, Dec. 1998. 13, 14, 65, 102, 166, 170
- [38] H. Markram, W. Gerstner, and P. J. Sjstrm, “A history of spike-timing-dependent plasticity,” *Frontiers in Synaptic Neuroscience*, vol. 3, no. 4, 2011. 13

- [39] L. F. Abbott and S. B. Nelson, “Synaptic plasticity: taming the beast.,” *Nature neuroscience*, vol. 3 Suppl, pp. 1178–1183, Nov. 2000. 13
- [40] D. E. Feldman, “Timing-based LTP and LTD at vertical inputs to layer II/III pyramidal cells in rat barrel cortex.,” *Neuron*, vol. 27, pp. 45–56, July 2000. 13
- [41] C. D. Holmgren and Y. Zilberter, “Coincident spiking activity induces long-term changes in inhibition of neocortical pyramidal cells.,” *J Neurosci*, vol. 21, pp. 8270–8277, Oct. 2001. 13
- [42] H. Hayashi and J. Igarashi, “LTD windows of the STDP learning rule and synaptic connections having a large transmission delay enable robust sequence learning amid background noise,” *Cognitive Neurodynamics*, vol. 3, pp. 119–130, June 2009. 14
- [43] J. G. Daugman, “Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters,” *J. Opt. Soc. Am. A*, vol. 2, pp. 1160–1169, July 1985. 17
- [44] A. Anzai, X. Peng, and D. C. Van Essen, “Neurons in monkey visual area V2 encode combinations of orientations,” *Nat Neurosci*, vol. 10, pp. 1313–1321, Oct. 2007. 17
- [45] T. Masquelier and S. J. Thorpe, “Unsupervised Learning of Visual Features through Spike Timing Dependent Plasticity,” *PLoS Comput Biol*, vol. 3, pp. e31+, Feb. 2007. 17, 110
- [46] A. Delorme, L. Perrinet, S. Thorpe, and M. Samuelides, “Network of integrate-and-fire neurons using Rank Order Coding B: spike timing dependant plasticity and emergence of orientation selectivity.,” *Neurocomputing*, vol. 38–40, no. 1–4, pp. 539–45, 2001. 17
- [47] G. M. Shepherd, *The Synaptic Organization of the Brain*. Oxford University Press, 2004. 18, 166
- [48] L. Abbott, “Lapicques introduction of the integrate-and-fire model neuron (1907),” *Brain Research Bulletin*, vol. 50, no. 56, pp. 303 – 304, 1999. 21

## BIBLIOGRAPHY

---

- [49] X. Zhang, “A mathematical model of a neuron with synapses based on physiology,” *nature precedings*, 2008. 21, 22, 166
- [50] G. Indiveri, B. Linares-Barranco, T. J. Hamilton, A. van Schaik, R. Etienne-Cummings, T. Delbruck, S.-C. Liu, P. Dudek, P. Hafliger, S. Renaud, J. Schemmel, G. Cauwenberghs, J. Arthur, K. Hynna, F. Folowosele, S. SAIGHI, T. Serrano-Gotarredona, J. Wijekoon, Y. Wang, and K. Boahen, “Neuromorphic silicon neuron circuits,” *Frontiers in Neuroscience*, vol. 5, no. 73, 2011. 21, 25, 66, 126, 127, 135, 160, 176
- [51] A. Hodgkin and A. Huxley, “A quantitative description of membrane current and its application to conduction and excitation in nerve,” *Bulletin of Mathematical Biology*, vol. 52, no. 1-2, pp. 25–71, 1990. 22
- [52] B. Lee, B. Sheu, and H. Yang, “Analog floating-gate synapses for general-purpose vlsi neural computation,” *Circuits and Systems, IEEE Transactions on*, vol. 38, no. 6, pp. 654–658, 1991. 23, 24, 166
- [53] P. Hasler and J. Dugger, “Correlation learning rule in floating-gate pfet synapses,” *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on*, vol. 48, no. 1, pp. 65–73, 2001. 23, 24, 166
- [54] S. Ramakrishnan, P. Hasler, and C. Gordon, “Floating gate synapses with spike-time-dependent plasticity,” *Biomedical Circuits and Systems, IEEE Transactions on*, vol. 5, no. 3, pp. 244–252, 2011. 24
- [55] J. Lont and W. Guggenbhl, “Analog cmos implementation of a multilayer perceptron with nonlinear synapses,” *Neural Networks, IEEE Transactions on*, vol. 3, no. 3, pp. 457–465, 1992. 24, 167
- [56] B. Lee and B. Sheu, “General-purpose neural chips with electrically programmable synapses and gain-adjustable neurons,” *Solid-State Circuits, IEEE Journal of*, vol. 27, no. 9, pp. 1299–1302, 1992. 24, 167
- [57] T. Watanabe, K. Kimura, M. Aoki, T. Sakata, and K. Ito, “A single 1.5-v digital chip for a 106 synapse neural network,” *Neural Networks, IEEE Transactions on*, vol. 4, no. 3, pp. 387–393, 1993. 25

- [58] S. Mitra, S. Fusi, and G. Indiveri, “A vlsi spike-driven dynamic synapse which learns only when necessary,” in *Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on*, pp. 4 pp.–, 2006. 25, 167
- [59] G. Indiveri, E. Chicca, and R. Douglas, “A VLSI reconfigurable network of integrate-and-fire neurons with spike-based learning synapses,” in *Proceedings of 12th European Symposium on Artificial Neural Networks (ESANN04)*, pp. 405–410, 2004. 25
- [60] J. Seo, B. Brezzo, Y. Liu, B. Parker, S. Esser, R. Montoye, B. Rajendran, J. Tierno, L. Chang, D. Modha, and D. Friedman, “A 45nm cmos neuromorphic chip with a scalable architecture for learning in networks of spiking neurons,” in *Custom Integrated Circuits Conference (CICC), 2011 IEEE*, pp. 1–4, 2011. 25, 26, 167
- [61] J. Schemmel, D. Bruderle, A. Grubl, M. Hock, K. Meier, and S. Millner, “A wafer-scale neuromorphic hardware system for large-scale neural modeling,” in *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, pp. 1947–1950, 2010. 25, 26, 167
- [62] A. Rast, S. Yang, M. Khan, and S. Furber, “Virtual synaptic interconnect using an asynchronous network-on-chip,” in *Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, pp. 2727–2734, 2008. 27
- [63] F. Alibart, S. Pleutin, D. Gurin, C. Novembre, S. Lenfant, K. Lmimouni, C. Gamrat, and D. Vuillaume, “An organic nanoparticle transistor behaving as a biological spiking synapse,” *Advanced Functional Materials*, vol. 20, no. 2, pp. 330–337, 2010. 28
- [64] O. Bichler, W. Zhao, F. Alibart, S. Pleutin, S. Lenfant, D. Vuillaume, and C. Gamrat, “Pavlov’s dog associative learning demonstrated on synaptic-like organic transistors,” *Neural Computation*, vol. 25, no. 2, pp. 549–566, 2013. 28, 29, 167



## BIBLIOGRAPHY

---

- [65] A. K. Friesz, A. C. Parker, C. Zhou, K. Ryu, J. M. Sanders, H.-S. P. Wong, and J. Deng, “A biomimetic carbon nanotube synapse circuit,” in *Biomedical Engineering Society (BMES) Annual Fall Meeting*, 2007. 28, 29, 167
- [66] J.-M. Retrouvey, J.-O. Klein, S.-Y. Liao, and C. Maneux, “Electrical simulation of learning stage in og-cntfet based neural crossbar,” in *Design and Technology of Integrated Systems in Nanoscale Era (DTIS), 2010 5th International Conference on*, pp. 1–5, 2010. 28, 29, 167
- [67] G. Agnus, W. Zhao, V. Derycke, A. Filoramo, Y. Lhuillier, S. Lenfant, D. Vuillaume, C. Gamrat, and J.-P. Bourgoin, “Two-terminal carbon nanotube programmable devices for adaptive architectures,” *Advanced Materials*, vol. 22, no. 6, pp. 702–706, 2010. 28, 29, 167
- [68] T. T. K. T. J. K. G. M. A. Takeo Ohno, Tsuyoshi Hasegawa, “Short-term plasticity and long-term potentiation mimicked in single inorganic synapses,” *Nature Materials*, no. 8, p. 591595, 2011. 28, 30, 167
- [69] A. Nayak, T. Ohno, T. Tsuruoka, K. Terabe, T. Hasegawa, J. K. Gimzewski, and M. Aono, “Controlling the synaptic plasticity of a cu<sub>2</sub>s gap-type atomic switch,” *Advanced Functional Materials*, vol. 22, no. 17, pp. 3606–3613, 2012. 28
- [70] A. V. Avizienis, H. O. Sillin, C. Martin-Olmos, H. H. Shieh, M. Aono, A. Z. Stieg, and J. K. Gimzewski, “Neuromorphic atomic switch networks,” *PLoS ONE*, vol. 7, p. e42772, 08 2012. 28, 30, 167
- [71] K. Cantley, A. Subramaniam, H. Stiegler, R. Chapman, and E. M. Vogel, “Spike timing-dependent synaptic plasticity using memristors and nano-crystalline silicon tft memories,” in *Nanotechnology (IEEE-NANO), 2011 11th IEEE Conference on*, pp. 421–425, 2011. 28
- [72] R. Waser, R. Dittmann, G. Staikov, and K. Szot, “Redox-based resistive switching memories—nanoionic mechanisms, prospects, and challenges,” *Advanced Materials*, vol. 21, no. 25-26, pp. 2632–2663, 2009. 30, 31, 33, 167
- [73] H.-S. 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, no. 12, pp. 2201–2227, 2010. 33, 39, 81

- [74] H.-S. Wong, H.-Y. Lee, S. Yu, Y.-S. Chen, Y. Wu, P.-S. Chen, B. Lee, F. Chen, and M.-J. Tsai, “Metal oxide rram,” *Proceedings of the IEEE*, vol. 100, no. 6, pp. 1951–1970, 2012. 33, 111, 112, 175
- [75] C. Gopalan, Y. Ma, T. Gallo, J. Wang, E. Runnion, J. Saenz, F. Koushan, P. Blanchard, and S. Hollmer, “Demonstration of conductive bridging random access memory (cbram) in logic {CMOS} process,” *Solid-State Electronics*, vol. 58, no. 1, pp. 54 – 61, 2011. Special Issue devoted to the 2nd International Memory Workshop (IMW 2010). 33, 96
- [76] B. Widrow, *An adaptive ADALINE neuron using chemical memistors*. Technical Report No 1553-2, october 1960. 33, 34, 167
- [77] B. Widrow and M. E. Hoff, “Adaptive Switching Circuits,” in *1960 IRE WESCON Convention Record, Part 4*, (New York), pp. 96–104, IRE, 1960. 33
- [78] L. Chua, “Memristor-the missing circuit element,” *Circuit Theory, IEEE Transactions on*, vol. 18, no. 5, pp. 507–519, 1971. 33
- [79] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, “The missing memristor found,” *Nature*, vol. 453, pp. 80–83, May 2008. 33
- [80] S. Thakoor, A. Moopenn, T. Daud, and A. P. Thakoor, “Solid-state thin-film memistor for electronic neural networks,” *Journal of Applied Physics*, vol. 67, pp. 3132–3135, mar 1990. 33, 35, 168
- [81] M. Suri, O. Bichler, D. Querlioz, B. Traoré, 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. 40, 44, 45, 51, 52, 53, 54, 109, 151, 152, 153, 168, 169, 170, 177, 178
- [82] D. Kuzum, R. G. D. Jeyasingh, B. Lee, and H.-S. P. Wong, “Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing,” *Nano Letters*, vol. 12, no. 5, pp. 2179–2186, 2012. 40, 41, 47, 64, 65, 85, 168

## BIBLIOGRAPHY

---

- [83] D. Kuzum, R. Jeyasingh, S. Yu, and H.-S. Wong, “Low-energy robust neuromorphic computation using synaptic devices,” *Electron Devices, IEEE Transactions on*, vol. 59, no. 12, pp. 3489–3494, 2012. 40, 42, 168
- [84] M. J. Breitwisch, R. Cheek, L. C.L., D. Modha, and B. Rajendran, “System for electronic learning synapse with spike-timing dependent plasticity using phase change memory,” *United States Patent Application Publication*, vol. US2010/0299297 A1, 2010. 40, 70
- [85] Y. Li, Y. Zhong, L. Xu, J. Zhang, X. Xu, H. Sun, and X. Miao, “Ultrafast synaptic events in a chalcogenide memristor,” *Science Reports*, vol. 3, no. 1619, 2013. 41, 42, 168
- [86] S. R. Ovshinsky, “Optical cognitive information processing a new field,” *Japanese Journal of Applied Physics*, vol. 43, no. 7B, p. 4695, 2004. 43
- [87] C. D. Wright, Y. Liu, K. I. Kohary, M. M. Aziz, and R. J. Hicken, “Arithmetic and biologically-inspired computing using phase-change materials,” *Advanced Materials*, vol. 23, no. 30, pp. 3408–3413, 2011. 43
- [88] A. Fantini, L. Perniola, M. Armand, J.-F. Nodin, V. Sousa, A. Persico, J. Cluzel, C. Jahan, S. Maitrejean, S. Lhostis, A. Roule, C. Dressler, G. Reimbold, B. De Salvo, P. Mazoyer, D. Bensahel, and F. Boulanger, “Comparative assessment of gst and gete materials for application to embedded phase-change memory devices,” in *Memory Workshop, 2009. IMW '09. IEEE International*, pp. 1–2, 2009. 43
- [89] A. Toffoli, M. Suri, L. Perniola, A. Persico, C. Jahan, J.-F. Nodin, V. Sousa, B. DeSalvo, and G. Reimbold, “Phase change memory advanced electrical characterization for conventional and alternative applications,” in *Microelectronic Test Structures (ICMTS), 2012 IEEE International Conference on*, pp. 114–118, 2012. 43
- [90] T. V. P. Bliss and G. L. Collingridge, “A synaptic model of memory: long-term potentiation in the hippocampus,” *Nature*, vol. 361, no. 6407, pp. 31–39, 1993. 43

- [91] A. Pirovano, A. Lacaita, A. Benvenuti, F. Pellizzer, and R. Bez, “Electronic switching in phase-change memories,” *Electron Devices, IEEE Transactions on*, vol. 51, no. 3, pp. 452–459, 2004. 43
- [92] M. Suri, V. Sousa, L. Perniola, D. Vuillaume, and B. DeSalvo, “Phase change memory for synaptic plasticity application in neuromorphic systems,” in *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pp. 619–624, 2011. 49, 50, 169
- [93] C. Sandhya, A. Bastard, L. Perniola, J. C. Bastien, A. Toffoli, E. Henaff, A. Roule, A. Persico, B. Hyot, V. Sousa, B. De Salvo, and G. Reimbold, “Analysis of the effect of boron doping on gete phase change memories,” in *Reliability Physics Symposium (IRPS), 2012 IEEE International*, pp. 6C.3.1–6C.3.5, 2012. 51
- [94] R. M. Shelby and S. Raoux, “Crystallization dynamics of nitrogen-doped ge<sub>2</sub>sb<sub>2</sub>te<sub>5</sub>,” *Journal of Applied Physics*, vol. 105, no. 10, p. 104902, 2009. 51
- [95] M. Boniardi, D. Ielmini, A. Lacaita, A. Redaelli, A. Pirovano, I. Tortorelli, M. Allegra, M. Magistretti, C. Bresolin, D. Erbetta, A. Modelli, E. Varesi, F. Pellizzer, and R. Bez, “Impact of material composition on the write performance of phase-change memory devices,” in *Memory Workshop (IMW), 2010 IEEE International*, pp. 1–4, 2010. 51
- [96] L. van Pieterse, M. H. R. Lankhorst, M. van Schijndel, A. E. T. Kuiper, and J. H. J. Roosen, “Phase-change recording materials with a growth-dominated crystallization mechanism: A materials overview,” *Journal of Applied Physics*, vol. 97, no. 8, p. 083520, 2005. 51
- [97] G. Navarro, N. Pashkov, M. Suri, V. Sousa, L. Perniola, S. Maitrejean, A. Persico, A. Roule, A. Toffoli, B. De Salvo, P. Zuliani, and R. Annunziata, “Electrical performances of tellurium-rich ge<sub>2</sub>sb<sub>2</sub>te<sub>5</sub> phase change memories,” in *Memory Workshop (IMW), 2011 3rd IEEE International*, pp. 1–4, 2011. 51
- [98] 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

## BIBLIOGRAPHY

---

- performance in neuromorphic systems,” in *Memory Workshop (IMW), 2012 4th IEEE International*, pp. 1–4, 2012. 51, 61, 62, 63, 76, 77, 170, 172
- [99] S. Braga, N. Pashkov, L. Perniola, A. Fantini, A. Cabrini, G. Torelli, V. Sousa, B. De Salvo, and G. Reimbold, “Effects of alloy composition on multilevel operation in self-heating phase change memories,” in *Memory Workshop (IMW), 2011 3rd IEEE International*, pp. 1–4, 2011. 51
- [100] A. Fantini, V. Sousa, L. Perniola, E. Gourvest, J. C. Bastien, S. Maitrejean, S. Braga, N. Pashkov, A. Bastard, B. Hyot, A. Roule, A. Persico, H. Feldis, C. Jahan, J.-F. Nodin, D. Blachier, A. Toffoli, G. Reimbold, F. Fillot, F. Pierre, R. Annunziata, D. Benshael, P. Mazoyer, C. Vallee, T. Billon, J. Hazart, B. De Salvo, and F. Boulanger, “N-doped gete as performance booster for embedded phase-change memories,” in *Electron Devices Meeting (IEDM), 2010 IEEE International*, pp. 29.1.1–29.1.4, 2010. 51
- [101] A. Gliere, O. Cueto, and J. Hazart, “Coupling the level set method with an electrothermal solver to simulate gst based pcm cells,” in *Simulation of Semiconductor Processes and Devices (SISPAD), 2011 International Conference on*, pp. 63–66, 2011. 51, 53, 55
- [102] G. Bruns, P. Merkelbach, C. Schlockermann, M. Salinga, M. Wuttig, T. D. Happ, J. B. Philipp, and M. Kund, “Nanosecond switching in gete phase change memory cells,” *Applied Physics Letters*, vol. 95, no. 4, p. 043108, 2009. 55
- [103] D. Querlioz, P. Dollfus, O. Bichler, and C. Gamrat, “Learning with memristive devices: How should we model their behavior?,” in *Nanoscale Architectures (NANOARCH), 2011 IEEE/ACM International Symposium on*, pp. 150–156, 2011. 56
- [104] J. Rubin, D. D. Lee, and H. Sompolinsky, “Equilibrium properties of temporally asymmetric Hebbian plasticity,” *Physical review letters*, vol. 86, pp. 364–367, Jan. 2001. 56
- [105] K. Sonoda, A. Sakai, M. Moniwa, K. Ishikawa, O. Tsuchiya, and Y. Inoue, “A compact model of phase-change memory based on rate equations of crystallization

- and amorphization,” *Electron Devices, IEEE Transactions on*, vol. 55, no. 7, pp. 1672–1681, 2008. 56, 57
- [106] I. Karpov, M. Mitra, D. Kau, G. Spadini, Y. A. Kryukov, and V. G. Karpov, “Fundamental drift of parameters in chalcogenide phase change memory,” *Journal of Applied Physics*, vol. 102, no. 12, pp. 124503–124503–6, 2007. 57
- [107] Q. Hubert, C. Jahan, A. Toffoli, L. Perniola, V. Sousa, A. Persico, J.-F. Nodin, H. Grampeix, F. Aussenac, and B. De Salvo, “Reset current reduction in phase-change memory cell using a thin interfacial oxide layer,” in *Solid-State Device Research Conference (ESSDERC), 2011 Proceedings of the European*, pp. 95–98, 2011. 61, 62
- [108] W. K. Njoroge, H. Dieker, and M. Wuttig, “Influence of dielectric capping layers on the crystallization kinetics of  $\text{ag}_{51}\text{in}_{65}\text{sb}_{59}\text{te}_{30}$  films,” *Journal of Applied Physics*, vol. 96, no. 5, pp. 2624–2627, 2004. 61
- [109] R. Pandian, B. J. Kooi, J. T. De Hosson, and A. Pauza, “Influence of capping layers on the crystallization of doped  $\text{sb}_{x}\text{te}_{100-x}$  fast-growth phase-change films,” *Journal of Applied Physics*, vol. 100, no. 12, pp. 123511–123511–9, 2006. 61
- [110] 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,” in *Electron Devices Meeting (IEDM), 2011 IEEE International*, pp. 4.4.1–4.4.4, 2011. 63, 64, 65, 73, 74, 77, 80, 102, 109, 154, 155, 170, 171, 172, 178
- [111] O. Bichler, M. Suri, D. Querlioz, D. Vuillaume, B. DeSalvo, and C. Gamrat, “Visual pattern extraction using energy-efficient 2-pcm synapse neuromorphic architecture,” *Electron Devices, IEEE Transactions on*, vol. 59, no. 8, pp. 2206–2214, 2012. 65, 67, 69, 70, 71, 76, 109, 170, 171, 172
- [112] J. Lisman and N. Spruston, “Questions About STDP as a General Model of Synaptic Plasticity,” *Frontiers in Synaptic Neuroscience*, vol. 2, 2010. 66
- [113] G. M. Wittenberg and S. S.-H. Wang, “Malleability of Spike-Timing-Dependent Plasticity at the CA3-CA1 Synapse,” *The Journal of Neuroscience*, vol. 26, no. 24, pp. 6610–6617, 2006. 66

## BIBLIOGRAPHY

---

- [114] O. Bichler, D. Querlioz, S. Thorpe, J. Bourgoïn, and C. Gamrat, “Unsupervised features extraction from asynchronous silicon retina through Spike-Timing-Dependent Plasticity,” in *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pp. 859–866, 2011. 66, 68, 72, 74, 75, 181
- [115] A. Pirovano, F. Pellizzer, I. Tortorelli, R. Harrigan, M. Magistretti, P. Petruzza, E. Varesi, D. Erbetta, T. Marangon, F. Bedeschi, R. Fackenthal, G. Atwood, and R. Bez, “Self-aligned  $\mu$ Trench phase-change memory cell architecture for 90nm technology and beyond,” in *Solid State Device Research Conference, 2007. ESSDERC 2007. 37th European*, pp. 222–225, 2007. 70, 79
- [116] D. Querlioz, O. Bichler, and C. Gamrat, “Simulation of a memristor-based spiking neural network immune to device variations,” in *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pp. 1775–1781, 2011. 72, 81, 102, 105, 110
- [117] P. Lichtsteiner, C. Posch, and T. Delbruck, “A 128 times; 128 120 dB 15  $\mu$ s Latency Asynchronous Temporal Contrast Vision Sensor,” *Solid-State Circuits, IEEE Journal of*, vol. 43, no. 2, pp. 566–576, 2008. 72, 73, 74, 109, 171
- [118] G. W. Burr, M. J. Breitwisch, M. Franceschini, D. Garetto, K. Gopalakrishnan, B. Jackson, B. Kurdi, C. Lam, L. A. Lastras, A. Padilla, B. Rajendran, S. Raoux, and R. S. Shenoy, “Phase change memory technology,” *Journal of Vacuum Science Technology B Microelectronics and Nanometer Structures*, vol. 28, no. 2, pp. 223–262, 2010. 75
- [119] J. Liang, R. Jeyasingh, H.-Y. Chen, and H.-S. Wong, “A 1.4  $\mu$ A reset current phase change memory cell with integrated carbon nanotube electrodes for cross-point memory application,” in *VLSI Technology (VLSIT), 2011 Symposium on*, pp. 100–101, 2011. 79
- [120] F. Xiong, A. D. Liao, D. Estrada, and E. Pop, “Low-Power Switching of Phase-Change Materials with Carbon Nanotube Electrodes,” *Science*, vol. 332, no. 6029, pp. 568–570, 2011. 79
- [121] D. H. Im, J. Lee, S. Cho, H. G. An, D. H. Kim, I. S. Kim, H. Park, D. Ahn, H. Horii, S. Park, U.-i. Chung, and J. Moon, “A unified 7.5nm dash-type confined

- cell for high performance pram device,” in *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*, pp. 1–4, 2008. 79
- [122] M. Suri, D. Garbin, O. Bichler, D. Querlioz, C. Gamrat, D. Vuillaume, and B. Desalvo, “Impact of pcm resistance-drift in neuromorphic systems and drift-mitigation strategy,” in *Nanoscale Architectures (NANOARCH), 2013 IEEE/ACM International Symposium on*, 2013. 81, 83, 86, 87, 88, 172, 173
- [123] D. Ielmini, S. Lavizzari, D. Sharma, and A. Lacaita, “Physical interpretation, modeling and impact on phase change memory (pcm) reliability of resistance drift due to chalcogenide structural relaxation,” in *Electron Devices Meeting, 2007. IEDM 2007. IEEE International*, pp. 939–942, 2007. 81, 82, 86
- [124] N. Papandreou, H. Pozidis, T. Mittelholzer, G. F. Close, M. Breitwisch, C. Lam, and E. Eleftheriou, “Drift-tolerant multilevel phase-change memory,” in *Memory Workshop (IMW), 2011 3rd IEEE International*, pp. 1–4, 2011. 81, 82, 86
- [125] D. H. Goldberg, G. Cauwenberghs, and A. G. Andreou, “Probabilistic synaptic weighting in a reconfigurable network of vlsi integrate-and-fire neurons,” *Neural Networks*, vol. 14, no. 6-7, pp. 781–793, 2001. 82
- [126] W. Senn and S. Fusi, “Convergence of stochastic learning in perceptrons with binary synapses,” *Phys. Rev. E*, vol. 71, p. 061907, Jun 2005. 82
- [127] J. H. Lee and K. K. Likharev, “Defect-tolerant nanoelectronic pattern classifiers,” *Int. J. Circuit Theory Appl.*, vol. 35, pp. 239–264, May 2007. 82
- [128] Y. Kondo and Y. Sawada, “Functional abilities of a stochastic logic neural network,” *Neural Networks, IEEE Transactions on*, vol. 3, pp. 434 –443, may 1992. 82
- [129] P. Appleby and T. Elliott, “Stable competitive dynamics emerge from multispikes interactions in a stochastic model of spike-timing-dependent plasticity,” *Neural computation*, vol. 18, no. 10, pp. 2414–2464, 2006. 82
- [130] 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



## BIBLIOGRAPHY

---

- Processing Applications,” in *Electron Devices Meeting (IEDM), 2012 IEEE International*, p. 10.3, 2012. 83, 97, 98, 99, 100, 101, 103, 104, 105, 106, 107, 108, 109, 110, 111, 156, 157, 158, 159, 172, 173, 174, 175, 178, 179, 181
- [131] F. Brglez, C. Gloster, and G. Kedem, “Hardware-based weighted random pattern generation for boundary scan,” in *Test Conference, 1989. Proceedings. Meeting the Tests of Time., International*, pp. 264–274, 1989. 82, 104, 106, 174
- [132] N. Papandreou, H. Pozidis, A. Pantazi, A. Sebastian, M. Breitwisch, C. Lam, and E. Eleftheriou, “Programming algorithms for multilevel phase-change memory,” in *Circuits and Systems (ISCAS), 2011 IEEE International Symposium on*, pp. 329–332, 2011. 86
- [133] D. Garbin, M. Suri, O. Bichler, D. Querlioz, C. Gamrat, and B. Desalvo, “Probabilistic neuromorphic system using binary phase-change memory (pcm) synapses: Detailed power consumption analysis,” in *Nanotechnology (NANO), 2013 IEEE International Conference on*, 2013. 90, 91, 173
- [134] 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*, vol. 32, pp. 339–348, 2012. 92, 102, 105, 155
- [135] R. Waser and M. Aono, “Nanoionics-based resistive switching memories,” *Nat. Mater.*, vol. 6, p. 833, Nov. 2007. 93
- [136] G. Palma, E. Vianello, C. Cagli, G. Molas, M. Reyboz, P. Blaise, B. De Salvo, F. Longnos, and F. Dahmani, “Experimental investigation and empirical modeling of the set and reset kinetics of ag-ges2 conductive bridging memories,” in *Memory Workshop (IMW), 2012 4th IEEE International*, pp. 1–4, 2012. 93, 97, 98, 100, 173, 174
- [137] S. Yu and H.-S. Wong, “Modeling the switching dynamics of programmable-metallization-cell (pmc) memory and its application as synapse device for a neuromorphic computation system,” in *Electron Devices Meeting (IEDM), 2010 IEEE International*, pp. 22.1.1–22.1.4, 2010. 94, 95, 173

- [138] 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, no. 4, pp. 1297–1301, 2010. PMID: 20192230. 94, 95, 96, 173
- [139] E. Vianello, C. Cagli, G. Molas, E. Souchier, P. Blaise, C. Carabasse, G. Rodriguez, V. Jousseau, B. De Salvo, F. Longnos, F. Dahmani, P. Verrier, D. Bretegnier, and J. Liebault, “On the impact of ag doping on performance and reliability of ges2-based conductive bridge memories,” in *Solid-State Device Research Conference (ESSDERC), 2012 Proceedings of the European*, pp. 278–281, sept. 2012. 96
- [140] M. Kund, G. Beitel, C.-U. Pinnow, T. Rohr, J. Schumann, R. Symanczyk, K.-D. Ufert, and G. Muller, “Conductive bridging ram (cbram): an emerging non-volatile memory technology scalable to sub 20nm,” in *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International*, pp. 754–757, dec. 2005. 97
- [141] S. Yu and H.-S. Wong, “Compact modeling of conducting-bridge random-access memory (cbram),” *Electron Devices, IEEE Transactions on*, vol. 58, pp. 1352–1360, may 2011. 97
- [142] S. Choi, S. Ambrogio, S. Balatti, F. Nardi, and D. Ielmini, “Resistance drift model for conductive-bridge (cb) ram by filament surface relaxation,” in *Memory Workshop (IMW), 2012 4th IEEE International*, pp. 1–4, 2012. 98
- [143] D. Ielmini, F. Nardi, and C. Cagli, “Resistance-dependent amplitude of random telegraph-signal noise in resistive switching memories,” *Applied Physics Letters*, vol. 96, no. 5, p. 053503, 2010. 98
- [144] V. Chan, S.-C. Liu, and A. van Schaik, “Aer ear: A matched silicon cochlea pair with address event representation interface,” *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 54, no. 1, pp. 48–59, 2007. 107
- [145] T. R. Agus, S. J. Thorpe, and D. Pressnitzer, “Rapid formation of robust auditory memories: Insights from noise,” *Neuron*, vol. 66, no. 4, pp. 610–618, 2010. 107

## BIBLIOGRAPHY

---

- [146] S. Yu, Y. Wu, R. Jeyasingh, D. Kuzum, and H.-S. Wong, “An electronic synapse device based on metal oxide resistive switching memory for neuromorphic computation,” *Electron Devices, IEEE Transactions on*, vol. 58, no. 8, pp. 2729–2737, 2011. 112, 113, 175
- [147] Y. Wu, S. Yu, H.-S. 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,” in *Memory Workshop (IMW), 2012 4th IEEE International*, pp. 1–4, 2012. 112, 113, 175
- [148] S. Yu, B. Gao, Z. Fang, H. Yu, J. Kang, and H.-S. Wong, “A neuromorphic visual system using rram synaptic devices with sub-pj energy and tolerance to variability: Experimental characterization and large-scale modeling,” in *Electron Devices Meeting (IEDM), 2012 IEEE International*, pp. 10.4.1–10.4.4, 2012. 112, 113, 175
- [149] 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, no. 25, p. 254023, 2011. 112, 113, 175
- [150] S. Park, H. Kim, M. Choo, J. Noh, A. Sheri, S. Jung, K. Seo, J. Park, S. Kim, W. Lee, J. Shin, D. Lee, G. Choi, J. Woo, E. Cha, J. Jang, C. Park, M. Jeon, B. Lee, B. Lee, and H. Hwang, “Rram-based synapse for neuromorphic system with pattern recognition function,” in *Electron Devices Meeting (IEDM), 2012 IEEE International*, pp. 10.2.1–10.2.4, 2012. 112, 113, 175
- [151] J. Alspector, B. Gupta, and R. B. Allen, “Performance of a stochastic learning microchip,” in *NIPS* (D. S. Touretzky, ed.), pp. 748–760, Morgan Kaufmann, 1988. 123, 124
- [152] M. van Daalen, P. Jeavons, and J. Shawe-Taylor, “A stochastic neural architecture that exploits dynamically reconfigurable fpgas,” in *FPGAs for Custom Computing Machines, 1993. Proceedings. IEEE Workshop on*, pp. 202–211, 1993. 123

- [153] S. Fusi, P. J. Drew, L. Abbott, *et al.*, “Cascade models of synaptically stored memories,” *Neuron*, vol. 45, no. 4, pp. 599–612, 2005. 123
- [154] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, vol. 313, no. 5786, pp. 504–507, 2006. 123
- [155] R. Salakhutdinov and G. Hinton, “An efficient learning procedure for deep boltzmann machines,” *Neural Computation*, vol. 24, no. 8, pp. 1967–2006, 2012. 123
- [156] K. Wiesenfeld, F. Moss, *et al.*, “Stochastic resonance and the benefits of noise: from ice ages to crayfish and squids,” *Nature*, vol. 373, no. 6509, pp. 33–36, 1995. 123
- [157] D. Querlioz and V. Trauchesse, “Stochastic resonance in an analog current-mode neuromorphic circuit,” *to be published in Procs. of IEEE ISCAS*, 2013. 123
- [158] J. Alspector, J. Gannett, S. Haber, M. Parker, and R. Chu, “A vlsi-efficient technique for generating multiple uncorrelated noise sources and its application to stochastic neural networks,” *Circuits and Systems, IEEE Transactions on*, vol. 38, no. 1, pp. 109–123, 1991. 124
- [159] K. Cameron, T. Clayton, B. Rae, A. Murray, R. Henderson, and E. Charbon, “Poisson distributed noise generation for spiking neural applications,” in *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, pp. 365–368, IEEE, 2010. 124
- [160] T.-J. Chiu, J. Gong, Y.-C. King, C.-C. Lu, and H. Chen, “An octagonal dual-gate transistor with enhanced and adaptable low-frequency noise,” *Electron Device Letters, IEEE*, vol. 32, no. 1, pp. 9–11, 2011. 124
- [161] T. Oya, T. Asai, and Y. Amemiya, “Stochastic resonance in an ensemble of single-electron neuromorphic devices and its application to competitive neural networks,” *Chaos, Solitons & Fractals*, vol. 32, no. 2, pp. 855–861, 2007. 124
- [162] S. Yu and H.-S. Wong, “Compact modeling of conducting-bridge random-access memory (cbram),” *Electron Devices, IEEE Transactions on*, vol. 58, no. 5, pp. 1352–1360, 2011. 125

## BIBLIOGRAPHY

---

- [163] G. Palma, M. Suri, D. Querlioz, E. Vianello, and B. Desalvo, “Stochastic neuron design using conductive bridge ram,” in *Nanoscale Architectures (NANOARCH), 2013 IEEE/ACM International Symposium on*, 2013. 125, 127, 128, 129, 131, 132, 133, 134, 160, 161, 176, 177, 179
- [164] E. Vianello, G. Molas, F. Longnos, P. Blaise, E. Souchier, C. Cagli, G. Palma, J. Guy, M. Bernard, M. Reyboz, G. Rodriguez, A. Roule, C. Carabasse, V. Delaye, V. Jousseau, S. Maitrejean, G. Reimbold, B. De Salvo, F. Dahmani, P. Verrier, D. Bretegnier, and J. Liebault, “Sb-doped ges2 as performance and reliability booster in conductive bridge ram,” in *Electron Devices Meeting (IEDM), 2012 IEEE International*, pp. 31.5.1–31.5.4, 2012. 134
- [165] F. Longnos, E. Vianello, G. Molas, G. Palma, E. Souchier, C. Carabasse, M. Bernard, B. DeSalvo, D. Bretegnier, and J. Liebault, “On disturb immunity and p/e kinetics of sb-doped ges2/ag conductive bridge memories,” in *to be published in proceedings of IEEE International Memory Workshp*, 2013. 134
- [166] S. Bonnet, J.-F. Bêche, S. Gharbi, O. Abdoun, F. Bocquelet, S. Joucla, V. Agache, F. Sauter, P. Pham, F. Dupont, *et al.*, “Neuropxi: A real-time multi-electrode array system for recording, processing and stimulation of neural networks and the control of high-resolution neural implants for rehabilitation,” *IRBM*, vol. 33, no. 2, pp. 55–60, 2012. 142, 143, 177