Resistance level modulation in PCM memory for neuromorphic applications

Ahmed Trabelsi

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : EEATS - Electronique, Electrotechnique, Automatique, Traitement du Signal (EEATS) Spécialité : Nano électronique et Nano technologies
Unité de recherche : Laboratoire d'Electronique et de Technologie de l'Information (LETI)

Modulation des niveaux de résistance dans une mémoire PCM pour des applications neuromorphiques

Resistance level modulation in PCM memory for neuromorphic applications

Présentée par :
Ahmed TRABELSI

Direction de thèse :

Veronique SOUSA Ingénieur HDR, directeur de recherche CEA LETI	Directrice de thèse
Carlo CAGLI Ingénieur, CEA Leti actuellement STMicroelectronics Crolles (France)	Co-encadrant de these

Rapporteurs :

Marc BOCQUET
Professeur à l'université Aix-Marseille
Damien DELERUYELLE
Professeur à l'institut National des Sciences Appliquées (INSA) de Lyon

Thèse soutenue publiquement le **2 avril 2024**, devant le jury composé de :

Veronique SOUSA Ingénieur HDR, directeur de recherche CEA LETI	Directrice de thèse
Abdelkader SOUFI Professeur, Institut National des Sciences Appliquées (INSA) de Lyon	Président
Marc BOCQUET Professeur à l'université Aix-Marseille	Rapporteur
Damien DELERUYELLE Professeur, Institut National des Sciences Appliquées (INSA)	Rapporteur
Quentin RAFHAY Maître de conférences à Phelma - Grenoble INP	Examinateur

Invités :

Carlo CAGLI
Ingénieur, CEA Leti actuellement STMicroelectronics Crolles (France)
Elisa Vianello
Directeur du programme IA embarqué | CEA-Leti

Acknowledgments

I extend my heartfelt gratitude to those who have played crucial roles in the completion of this thesis. Special thanks to my Supervisor Carlo Cagli, whose unwavering support, mentorship, and invaluable guidance have significantly shaped the trajectory of this work. His expertise and selfless contribution to my professional development have been invaluable and will forever be part of my personal and professional life. I am also deeply appreciative of my thesis director, Veronique Sousa, for her unwavering support, mentorship, and invaluable guidance throughout the research process. Her expertise has been instrumental in shaping the trajectory of this work. To my beloved family, words cannot adequately express my gratitude for your unwavering love, steadfast support, and boundless encouragement. Throughout this academic journey, your enduring presence has been my guiding light, providing comfort in moments of challenge and celebration in moments of triumph. Your sacrifices and belief in my endeavors have been a constant source of motivation, shaping the foundation upon which this thesis stands. Thank you for being the pillars of strength that have allowed me to pursue my academic aspirations with confidence and resilience. Your love has been an invaluable treasure, and I am profoundly grateful for each member's unique contribution to the tapestry of my life. I extend my deepest thanks to my incredible girlfriend, for her unwavering and steadfast support,

profound understanding, and continuous encouragement during the numerous challenges and triumphant moments of this demanding academic journey. To my friends, your camaraderie has been a source of strength. I appreciate the willingness of the research participants to share their experiences, enriching the depth of this study. My colleagues and peers, including Gabriele Navarro from the PCM team, have provided valuable feedback and engaging discussions, contributing to the intellectual growth of this work. I am thankful for the resources and facilities provided by CEA Leti. To any additional contributors, your support has not gone unnoticed and is deeply appreciated. In conclusion, this thesis is a culmination of the collective efforts of these individuals and entities, and I am grateful for their integral role in my academic journey.

Ahmed Trabelsi

Contents

Abstract - Résumé

TITLE: Resistance level modulation in PCM memory for neuromorphic applications

Abstract:

The exponential growth of data in recent years has led to a significant increase in energy consumption, creating a pressing need for innovative memory technologies to overcome the limitations of conventional solutions. This data deluge has resulted in a forecasted consumption surge in data centers, with an expected fourfold increase in data by 2025 compared to the present volume. To address this challenge, emerging memory technologies such as RRAM (Resistive RAM), PCM (Phase-Change Memory), and MRAM (Magnetoresistive RAM) are being developed to offer high density, fast access times, and non-volatility, thereby revolutionizing storage and memory solutions (Molas & Nowak, 2021).

One promising technique to address the need for innovative memory technologies is the use of frequency modulation to modulate resistance in PCM which is a crucial aspect of its use in neuromorphic computing. PCM is a non-volatile memory technology based on the reversible phase transition between amorphous and crystalline phases of certain materials. The ability to alter conductance levels makes PCM well-suited for synaptic realizations in neuromorphic computing. The progressive crystallization of the phase-change material and the subsequent increase in device conductance enable PCM to be used in neuromorphic applications. Additionally, PCM-based memristor neural networks have been developed, and the resistance drift effect in PCM has been quantified, opening up new paths for the development of PCM-based memristor neuromorphic accelerators. Furthermore, frequency modulation has been identified as a promising technique to modulate resistance in PCM. This approach can be applied to PCM as well as RRAM, and it is expected to yield improved learning effects in more complex networks using multi-level cells (J. Wang et al., 2011). The primary aim of this thesis is to explore innovative methods for controlling resistance levels in PCM devices with a focus on their application in neuromorphic systems. The research involves a comprehensive understanding of the mechanisms underlying PCM devices and an identification of parameters that may influence the reliability of these devices.

Additionally, the thesis aims to propose a novel approach to effectively modulate resistance levels in PCM devices, contributing to advancements in this field.

Speciality: Nanoelectronics and Nanotechnology

Keywords: Phase-Change Memory, Neuromorphic computing, resistance level modulation, reliability, phase-change materials.

Thesis work prepared at: CEA, LETI, MINATEC Campus, 17 rue des Martyrs, 38054 Grenoble Cedex 9, France.

TITRE : Modulation du niveau de résistance dans la mémoire PCM pour les applications neuromorphiques

Résumé :

La croissance exponentielle des données au cours des dernières années a entraîné une augmentation significative de la consommation d'énergie, créant ainsi un besoin urgent de technologies de mémoire innovantes pour surmonter les limitations des solutions conventionnelles. Cette inondation de données a entraîné une augmentation prévue de la consommation dans les centres de données, avec une multiplication par quatre des données d'ici 2025 par rapport au volume actuel. Pour relever ce défi, des technologies de mémoires émergentes telles que la RRAM (RAM résistive), la PCM (mémoire à changement de phase) et la MRAM (RAM magnéto-résistive) sont en cours de développement pour offrir une haute densité, des temps d'accès rapides et une non-volatilité, révolutionnant ainsi les solutions de stockage et de mémoire (Molas & Nowak, 2021).

Une technique prometteuse pour répondre au besoin de technologies de mémoire innovantes est l'utilisation de la modulation de fréquence pour moduler la résistance dans la PCM, qui est un aspect crucial de son utilisation en informatique neuromorphique. La PCM est une technologie de mémoire non volatile basée sur la transition de phase réversible entre les phases amorphe et cristalline de certains matériaux. La capacité de modifier les niveaux de conductance rend la PCM bien adaptée aux réalisations synaptiques en informatique neuromorphique. La cristallisation progressive du matériau à changement de phase et l'augmentation subséquente de la conductance du dispositif permettent à la PCM d'être utilisée dans des applications neuromorphiques. De plus, des réseaux neuronaux

basés sur la mémoire PCM ont été développés, et l'effet de dérive de la résistance dans la PCM a été quantifié, ouvrant de nouvelles voies pour le développement d'accélérateurs neuromorphiques à base de memristors PCM. De plus, la modulation de fréquence a été identifiée comme une technique prometteuse pour moduler la résistance dans la PCM. Cette approche peut être appliquée à la PCM ainsi qu'à la RRAM, et on s'attend à ce qu'elle produise des effets d'apprentissage améliorés dans des réseaux plus complexes utilisant des cellules multi-niveaux (Wang et al., 2011). L'objectif principal de cette thèse est d'explorer des méthodes innovantes pour contrôler les niveaux de résistance dans les dispositifs PCM en mettant l'accent sur leur application dans les systèmes neuromorphiques. La recherche implique une compréhension approfondie des mécanismes sous-jacents aux dispositifs PCM et une identification des paramètres susceptibles d'influencer la fiabilité de ces dispositifs. De plus, la thèse vise à proposer une nouvelle approche pour moduler efficacement les niveaux de résistance dans les dispositifs PCM, contribuant ainsi aux avancées dans ce domaine.

Spécialité : Nanoelectronique et Nanotechnologie

Mots-clés : Mémoire à changement de phase, Informatique neuromorphique, Modulation du niveau de résistance, Fiabilité, Matériaux à changement de phase.

Travail de thèse réalisé à : CEA, LETI, Campus MINATEC, 17 rue des Martyrs, 38054 Grenoble Cedex 9, France.

Introduction

"Technology is best when it brings people together."

Matt Mullenweg CEO of Tumblr

Context

The exponential growth of data in recent years has created a pressing need for innovative memory technologies to overcome the limitations of conventional solutions. The sheer volume of data being generated, processed, and stored has surpassed the capabilities of traditional memory technologies, leading to challenges in terms of capacity, speed, power consumption, and reliability. To tackle these obstacles, researchers and industry experts have embarked on a quest to explore new frontiers in memory technology.

Von Neumann architecture is the traditional computer architecture where the central processing unit for data computation and the main memory for data storage are separated. This separation leads to a bottleneck known as the von Neumann bottleneck, which limits the system performance. In contrast, near-memory computing, in-memory computing, and neuromorphic computing are emerging computational paradigms that aim to overcome this bottleneck. In near-memory computing (Singh et al., 2018), the processor is placed in close proximity to the memory, reducing the data transfer distance and latency. This architecture allows for more efficient data processing and can improve system performance("Beyond von Neumann," 2020). In-memory computing (Sebastian et al., 2020), the memory is not just used for data storage, but also for performing computational operations. This approach merges storage and computational operations, allowing for parallel processing and reducing the need to move data between the processor and memory. Finally, neuromorphic computing (S. Furber, 2016), is inspired by the structure and function of the brain. In a neuromorphic computer, both processing and memory are governed by neurons and synapses, allowing for highly parallel operation and efficient processing of spiking neural network-based algorithms (Schuman et al., 2022). In the quest to overcome the limitations of the von Neumann architecture and unlock more efficient and powerful computing systems, emerging computational paradigms are being explored. One notable avenue in this exploration is phase-change memory (PCM) (Stanisavljevic et al., 2015), which leverages materials capable of transitioning between amorphous and crystalline states to symbolize the binary data of 0 and 1. This technology offers several advantages, including high capacity, fast read and write speeds, and non-volatility (data retention even without power). PCM has the potential to significantly enhance memory performance and storage density, making it a promising candidate for next-generation memory solutions.

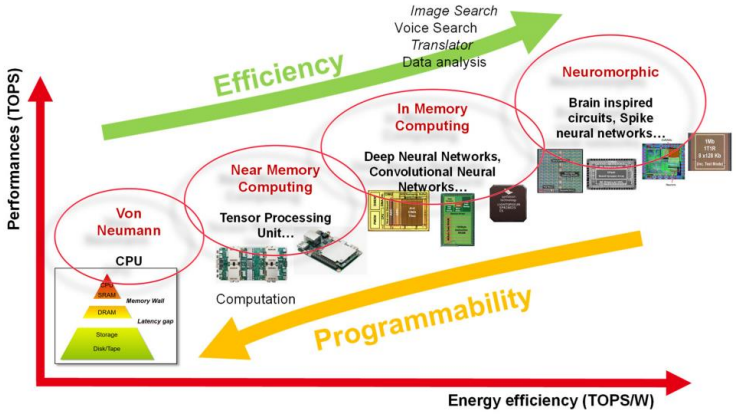


Figure. Performance vs. energy efficiency in computing systems.

Another emerging technology is resistive random-access memory (RRAM) (Wong et al., 2012). RRAM relies on the manipulation of resistance states within a memory cell to store and retrieve data. This technology offers advantages such as high density, low power consumption, and excellent scalability. RRAM's ability to retain data even when powered off and its fast read and write speeds make it an attractive option for future memory systems. Magnetic random-access memory (MRAM) (Khvalkovskiy et al., 2013) is another promising alternative. MRAM utilizes the magnetic properties of materials to store data. It offers advantages such as fast access times, high endurance, and non-volatility. MRAM has the potential to bridge the gap between volatile and non-volatile memory, providing a unique combination of speed and persistence. These type of memory named also “memristors” (which are devices that can change their resistance based on the history of applied voltage, allowing them to store information) offers advantages such as high density, fast operation, and low power consumption. Its ability to function as both storage and logic elements opens up new possibilities for memory architectures. By investigating and developing these next-generation memory technologies, researchers aim to overcome the challenges posed by the exponential growth of data. As these technologies continue to advance, they have the potential to reshape the landscape of memory technology and enable new innovations across various industries.

In this thesis, inspired by neuromorphic research and taking into account the latest advances in artificial intelligence, we have demonstrate a new way to send synaptic weights into PCM implementing algorithms for neuromorphic computing.

The first chapter of the thesis, we explored Neuromorphic Computing, comparing biological and artificial neural networks, delving into synapses and learning algorithms. We discussed network topologies, focusing on feed-forward and convolutional neural networks. The chapter concluded with current and future challenges in Neuromorphic Computing, addressing hardware limitations and ethical concerns.

The second chapter, we delved into semiconductor memory technologies for neuromorphic computing applications, including CBRAM, OXRAM, FeRAM, STT-RAM, and PCRAM. Each technology unique features were highlighted, forming a basis for comparison. Our analysis aimed to identify strengths and

weaknesses, aiding in the selection of the most suitable non-volatile memory (NVM) for specific neuromorphic applications.

The third chapter, we explored the integration of neuromorphic computing with non-volatile memory (NVM) technologies. We discussed Spike-timing-dependent-plasticity (STDP) as a crucial mechanism for mimicking learning in biological neural networks. The core operation of neuromorphic computing, vector-matrix multiplication, was highlighted for tasks like pattern recognition. Phase-Change Memory (PCM) was identified as a versatile option, serving as both synapse and neuron.

The fourth chapter focus on the exploration of phase-change memory (PCM) technology beyond the traditional von Neumann architecture offers insights into evolving computing. Exploring the phase-change mechanism unveils PCM's versatility for complex computing tasks. Challenges like resistance drift are acknowledged, emphasizing ongoing research for stability. Comparisons between phase-change materials, particularly $Ge_2Sb_2Te_5$ and Ge-rich GST, underscore material importance.

Chapter 5 present the frequency modulation model within phase-change systems. The primary goal is to introduce a model for programming synaptic weights in PCM devices. By concentrating on frequency modulation, we find a practical and effective approach to manipulating conductance levels. The effectiveness of frequency modulation is validated through simulations, experiments, and measurements.

Finally, the sixth chapter introduces the application of the frequency modulation in real case scenario on 28nm FDSOI technology for transferring synaptic weights. We convert pre-calculated weights into conductance values and then into frequency. We validate our algorithm by assessing the accuracy of a CNN trained with PCM-based weights, achieving up to 90% accuracy. Our experiments confirm the effectiveness of frequency modulation, positioning our work as a strong proof of concept for a CNN model based on this approach.

Chapter 1:

Fundamentals of Neuromorphic Computing

"Neuromorphic computing is not just about building brain-like computers; it's a quest to understand the principles that make our brains so powerful and efficient."

Karlheinz Meier, Professor of Physics, University of Heidelberg

1. Overview of Neuromorphic Computing

Computers have become vital in our modern lives, being present everywhere around the world. As data-intensive applications become more prevalent, there is a growing need for hardware that can provide fast access, high capacity, large bandwidth, low cost, and the ability to handle artificial intelligence tasks. However, the increasing demand for big data poses additional challenges. Firstly, energy consumption has become a significant issue, driven by the advancement of complex algorithms and architectures. Currently, a substantial amount of global energy, ranging from 5% to 15%(Dayarathna et al., 2016), is consumed by data-related activities like transmission and processing. This percentage is expected to rise rapidly due to the exponential growth of data generated by ubiquitous sensors in the era of the Internet of Things. Secondly, the traditional von Neumann computer architecture, which separates processing and memory units physically, is becoming a bottleneck for data processing (R. Nair, 2015). This architecture, while immensely influential in the field of technology for many years, suffers from inefficiencies in performance, primarily due to the slow and energy-intensive movement of data (Yang et al., 2019).

Conventional von Neumann computers, utilizing complementary metal oxide semiconductor (CMOS) technology, lack the inherent capacity to learn from and handle complex data in a manner similar to the human brain. To overcome these limitations, extensive research is being conducted globally to explore alternative computing approaches that draw inspiration from biological principles. One such approach involves the development of neuromorphic systems, which aim to mimic the information processing mechanisms observed in the human brain. These systems strive to replicate the brain's ability to process and interpret data in a highly parallel and energy-efficient manner (Yang et al., 2019).

The term 'neuromorphic' was first introduced by Carver Mead in the 1990s (Mead, 1990). It was coined to describe computing systems that incorporate mixed signal analog/digital very large scale integration (VLSI) technology and draw inspiration from the neurobiological architectures of the brain. These neuromorphic systems aim to replicate the unique characteristics of the brain, such as parallelism, fault tolerance, and low power consumption, in order to achieve efficient and brain-like information processing capabilities(Indiveri, 2021). By combining analog and digital components, these systems seek to emulate the intricate and efficient computational capabilities observed in biological neural networks (Mead, 1990). The field of "neuromorphic engineering" has emerged as an interdisciplinary research field with a primary focus on constructing electronic neural processing systems that directly emulate the bio-physics of real neurons and synapses (Chicca et al., 2014). Initially, the term "neuromorphic" referred to the replication of the physical characteristics and behaviors of biological neural systems in electronic form. However, more recently, the definition of the term has been expanded in two additional directions (Chicca & Indiveri, 2020). Firstly, it has been used to describe spike-based processing systems that are engineered to explore large-scale computational neuroscience models. These systems aim to replicate the behavior of neurons and synapses in electronic form, focusing on spike-based communication and neural dynamics. Secondly, neuromorphic computing involves dedicated electronic neural architectures that implement circuits for

neurons and synapses. It should be noted that this concept is distinct from AI machine learning approaches, which rely on software algorithms to minimize recognition errors in pattern recognition tasks (LeCun et al., 2015).

However, there is ongoing debate regarding the precise definition of neuromorphic computing. It can range from strict, high-fidelity mimicking of neuroscience principles, where detailed synaptic chemical dynamics are mandatory, to more loosely brain-inspired principles, such as simple vector-matrix multiplication. Generally, there is a consensus that neuromorphic computing should involve some form of time, event, or data-driven computation. Spiking neural networks (SNNs), often referred to as the third generation of neural networks (Maass, 1997), exemplify these principles. However, there is significant cross-fertilization between the technologies required to develop efficient SNNs and those used for more traditional non-SNNs, known as artificial neural networks (ANNs), which are typically driven by discrete time steps.

Nature serves as a crucial inspiration for the development of a more sustainable computing landscape, where neuromorphic systems demonstrate significantly lower power consumption compared to conventional processors. This reduction in power consumption is achieved through the integration of non-volatile memory, analog/digital processing circuits, and dynamic learning capabilities when dealing with complex data. One of the remaining challenges in computing is to construct artificial neural networks (ANNs) that closely mimic their biological counterparts. If the fundamental technical issues associated with neuromorphic computing can be resolved in the coming years, the market for neuromorphic computing is projected to experience substantial growth. It is estimated that the market will expand from $0.2 billion in 2025 to $20 billion in 2035. This growth will be driven by the advancements in neuromorphic computers, which offer ultra-low power consumption and high-speed processing capabilities, thereby fueling the demand for neuromorphic devices.

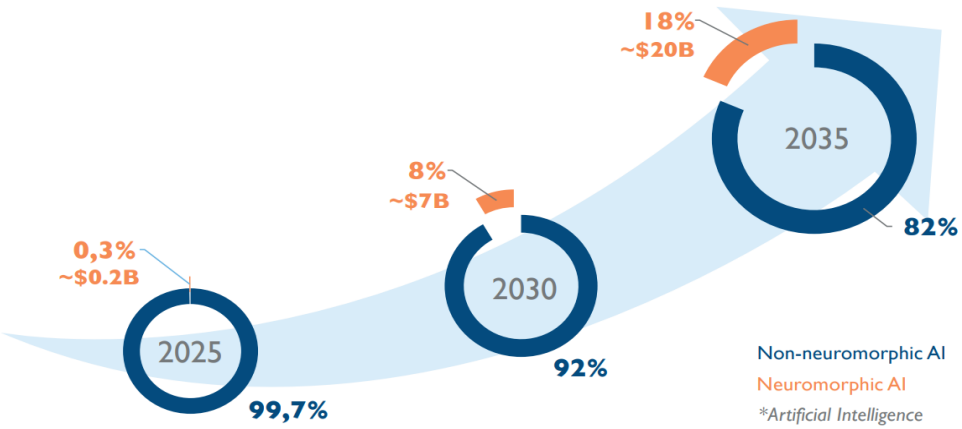


Figure 1. 1 Neuromorphic into AI computing & sensing - 2025 - 2030 - 2035 revenue evolution (Yole Développement, May 2021)

1.1 Biological Neural Network

Over the years, there has been extensive research in the field of neuroscience aimed at understanding the human brain. Initially, the focus was primarily on investigating the anatomical structure of the brain. However, attempts to comprehend the functional workings of its intricate neural network were often clouded by speculative theories masquerading as knowledge for many centuries.

It was around the mid-18th century that a more comprehensive understanding of the brain's functionality began to emerge. During this period, scientific studies demonstrated that nerve impulses, previously attributed to "animal spirits," were actually electrical signals akin to charges in an electrical circuit. This realization marked a significant step forward in comprehending the mechanisms underlying neural communication. Furthermore, advancements in neuroscience research and microscopy techniques have shed light on the structure and characteristics of neurons. This has led to the conceptualization of the brain as a network of interconnected neurons that communicate with each other through chemical signals. These findings have helped unravel the intricate ways in which information is processed and transmitted within the brain.

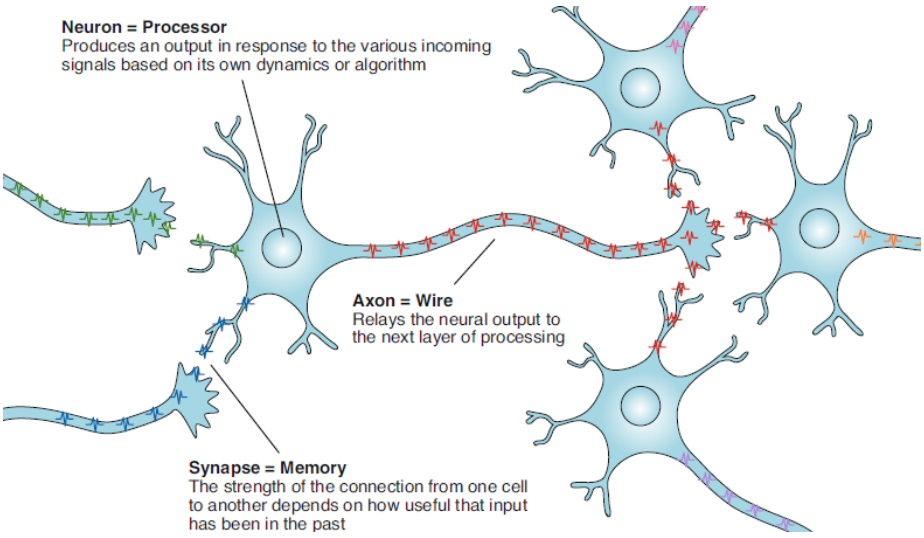


Figure 1. 2 In both biology and computing, neural networks involve the processing of incoming signals (voltage spikes). These signals undergo weighting, where they can be amplified or reduced before reaching the neuron. Once the signals reach the neuron, they are combined and processed. The neuron responds to the collective presence of these signals and may generate its own signal, commonly known as firing. It is worth noting that the timing of signals plays a crucial role. When multiple signals arrive simultaneously, they have a greater likelihood of influencing the neuron to fire. Subsequently, the axon transmits these signals to other neurons within the network. (Bains, 2020)

The human brain is estimated to consist of approximately 100 billion neurons, and each neuron is believed to have around 1,000 to 10,000 connections, resulting in a total of trillions to quadrillions of connections in the brain (Herculano-Houzel, 2009). The complexity of this system surpasses that of the entire global mobile network. Neurons in the brain constantly form and dissolve connections, sometimes within a matter of seconds (Kolb & Gibb, 2011). Despite advancements in understanding

the brain, the processes underlying thought remain a mystery. Artificial intelligence (AI) has made significant progress in attempting to unravel the nature of thought processes within the brain (Lin, 2017). Researchers are utilizing Artificial Neural Networks (ANNs) in computational tools to simulate the biological brain (Fukushima, 1980). AI aims to address questions about how networks of neurons in the visual processing areas of the brain convert optical images and how they can be emulated to create intelligent devices. However, explaining these phenomena often requires the language of mathematics, which is a central focus of computational neuroscience (Lin, 2017).

1.2 Biological Neurons versus Artificial Neurons

Biological neurons are the fundamental components of the human brain, while artificial neurons form the basis of Artificial Neural Networks (ANNs). ANNs are computer algorithms that draw inspiration from biological neurons and are designed to perform specific computational tasks (Bush & Sejnowski, 1996) (Izhikevich, 2006) (Maex & Schutter, 2003).

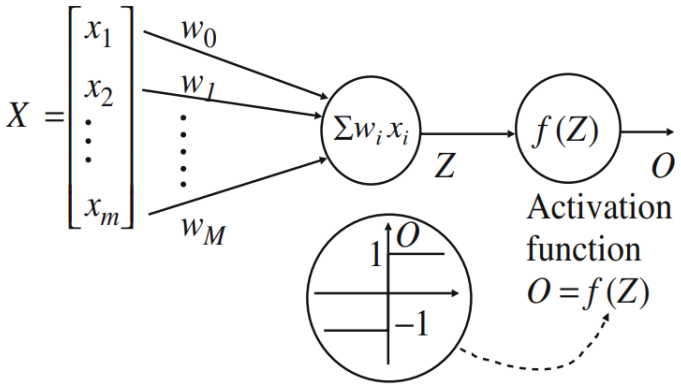


Figure 1. 3 The artificial model proposed by McCulloch and Pitts in 1943 represents a simplified version of a biological neuron. One notable aspect of this model is that the relationship between the input and output is governed by a non-linear function known as the activation function. In the depicted model, the activation function takes the form of a hard threshold, meaning that the neuron's output is binary, either firing or not firing, based on whether the weighted sum of the inputs exceeds a specific threshold value (Fukushima, 1980).

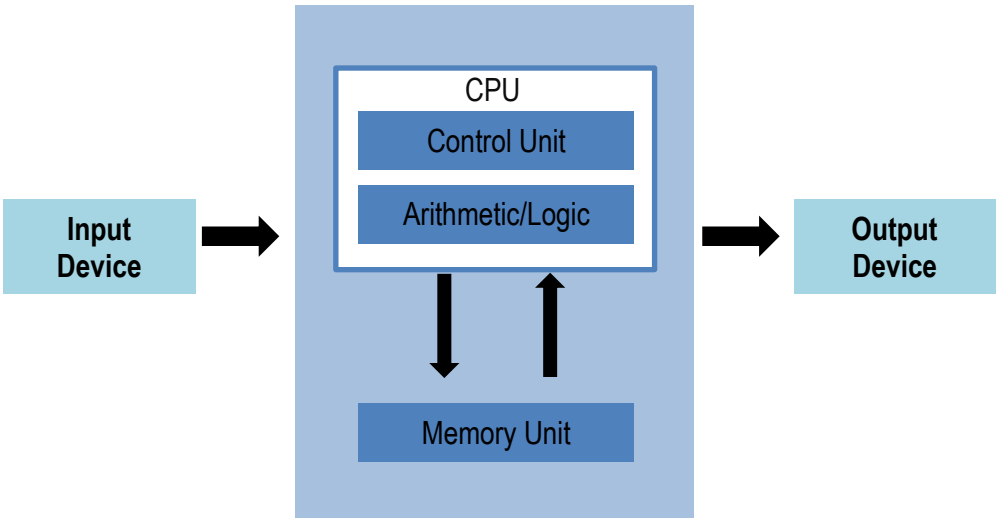
Figure 1.3 depicts a model of an artificial neuron that was described by McCulloch and Pitts in 1943. It consists of inputs that can be represented as either an electrical impulse (1) or the absence of an electrical impulse (0) (McCulloch & Pitts, 1943). Each input has an associated weight, which is multiplied by the activation function. If the weighted sum of the inputs is equal to or greater than a certain threshold value (θ), the neuron fires and returns a value of 1. Otherwise, if the sum is below the threshold, the neuron does not fire and returns a value of 0 (McCulloch & Pitts, 1943). Additionally, an artificial neuron has an activation threshold and establishes weighted connections with neighboring neurons (Lin, 2017) (Maass, 1997). When the combined activation received from its neighbors exceeds the activation threshold, the neuron fires and transmits this signal intensity to the neighboring neurons. Training the network involves modifying the weights associated with the connections to perform specific tasks, enabling learning.

During the 1950s, artificial neural networks (ANNs) emerged as a subject of significant interest in computer science research. This was driven by the recognition that computers had unmatched capabilities in terms of memory and processing speed, while humans demonstrated remarkable skills in complex actions and reasoning. Researchers sought to bridge this gap by studying the underlying principles of the human brain and developing computational models inspired by biological neurons.

Neurons, whether artificial or biological, can be interconnected to form networks. Artificial neural networks (ANNs) are computer algorithms that simulate the behavior of biological neural networks. They are designed to process and interpret complex data patterns, learn from experience, and perform tasks such as pattern recognition, classification, and decision-making. ANNs consist of interconnected artificial neurons that exchange information and computations to collectively solve problems.

Biological neural networks (BNNs), on the other hand, refer to the intricate network of neurons in living organisms, such as the human brain. These networks are responsible for complex cognitive functions, including perception, memory, learning, and decision-making. BNNs exhibit remarkable capabilities in terms of adaptability, fault tolerance, and efficiency, which have inspired the development of artificial neural networks.

a :



b :

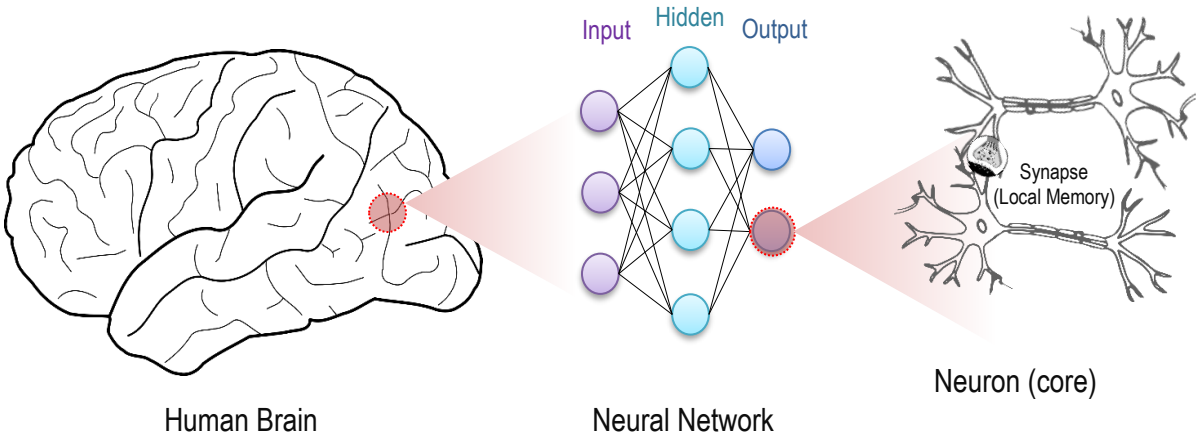


Figure 1. 4 a) The Von Neumann architecture employs a design where faster and more expensive memory is positioned closer to the cores in multiprocessor platforms. This includes caches and local memory. On the other hand, slower and less expensive memory, such as magnetic memory, is located in other layers to reduce the cost of the CPU. This arrangement forms a memory hierarchy. b) The neuromorphic architecture draws inspiration from the neural networks found in biological brains. It overcomes the bottleneck issue of the Von Neumann architecture by enabling parallel and cognitive computing. In this architecture, synapses act as local memories connected to each neuron, functioning as computational cores.

By studying the structure and function of biological neural networks and emulating them in artificial neural networks Figure 1.4, researchers aim to unlock the potential for machine intelligence that mimics human-like cognitive abilities. The interdisciplinary field of computational neuroscience plays a crucial role in understanding the principles of neural information processing and translating them into computational models and algorithms.

The development and advancement of artificial neural networks have led to various applications across numerous domains, including image and speech recognition, natural language processing, robotics, medical diagnosis, and financial forecasting. As researchers continue to explore and refine the capabilities of ANNs, they strive to bridge the gap between human and artificial intelligence, leveraging the strengths of both to create intelligent systems that can tackle complex problems and enhance our understanding of cognitive processes.

In 2014, two influential articles were published that garnered significant attention from scientists and sparked interest in neuromorphic platforms as innovative computing architectures.

One of these articles, authored by (Merolla et al., 2014), detailed research conducted at IBM and sponsored by DARPA. They demonstrated a computing hardware system comprised of compact modular cores for a large-scale neuromorphic architecture. These cores combined digital neurons with a vast synaptic array. The general-purpose neuromorphic processor they developed utilized thousands of neuro-synaptic cores, incorporating one million neurons and 256 million reconfigurable synapses. The purpose of the system is to mimics the functioning of the human brain and is designed to be energy-efficient and highly scalable

The second notable work published in 2014 was the SpiNNaker (Spiking Neural Network Architecture) project (S. B. Furber et al., 2014). This project, which had been underway for a decade, presented a comprehensive description of its goals and progress. SpiNNaker aimed to create massively parallel architectures with millions of cores, taking inspiration from the connectivity properties of the mammalian brain. The hardware platform they developed was specifically designed for modeling large-scale spiking neural networks in real time, resembling biological processes.

The field of neuromorphic and neuro-inspired computing has gained momentum since then, with an increasing number of academic and industrial research teams embracing these approaches. In recent years, numerous valuable publications have emerged, explaining the use of novel materials capable of emulating some of the properties observed in biological synapses (Benjamin et al., 2014; Jo et al., 2010; Rajendran & Alibart, 2016).

1.3 Synapse and Learning

A synapse is a specialized structure that facilitates the exchange of spike signals and the adjustment of connection strength between two neurons. It plays a crucial role in learning within neural network systems. Physiologically, a synapse connects the axon of a presynaptic neuron (before the synapse) to the dendrite of a postsynaptic neuron (after the synapse). There are two main types of biological synapses: chemical and electrical.

In a chemical synapse, neurotransmitters are the key players in signal transmission between the presynaptic and postsynaptic neurons. When an action potential occurs in the presynaptic neuron, it triggers the release of a chemical substance into the synaptic cleft, which is the space between the two neurons. The neurotransmitter diffuses across the synaptic cleft and induces a change in the voltage of the postsynaptic neuron's membrane. In the biological neural system, a synapse is considered excitatory if the neurotransmitter increases the voltage of the postsynaptic neuron, and inhibitory if it decreases the voltage (Purves et al., 2001).

On the other hand, an electrical synapse consists of gap junctions, which are small channels that directly connect the cytoplasms of two cells (Hu & Bloomfield, 2003). These synapses allow for rapid electrical communication between neurons. The fundamental mechanism of synaptic transmission involves the depolarization of the synaptic terminal by a presynaptic spike, leading to calcium influx through presynaptic calcium channels. This calcium flow triggers the release of neurotransmitter vesicles into the synaptic cleft. The neurotransmitter then temporarily binds to postsynaptic channels, opening them and enabling ionic current to flow across the membrane.

1.4 Synaptic Behavior and plasticity

Neurons perform computations by integrating inputs received from other neurons and generating spikes based on various connections. The computation process is influenced by synaptic plasticity, which refers to the ability of synapses to modify their connection strength in response to neuronal activity. This synaptic plasticity is considered fundamental for adaptation and learning, even in

traditional neural network models. Many synaptic weight updating rules in these models are derived from Hebb's law (Morris, 1999).

The primary importance of neural networks lies in their ability to learn from the environment and enhance their performance. Various learning algorithms exist to facilitate this process. While the interconnection configuration of a neural network is significant for learning, the specific way in which synapse weights are adjusted distinguishes different learning algorithms.

Simon Haykin's book "*Neural Networks: A Comprehensive Foundation Subsequent*" outlines five fundamental learning algorithms: memory-based, Hebbian, error-correction, competitive, and Boltzmann learning. Memory-based learning involves explicitly memorizing the training data. Hebbian and competitive learning draw inspiration from neurobiology. Error-correction learning utilizes an optimum filtering rule, and Boltzmann learning is rooted in concepts derived from statistical mechanics. In more details:

- Memory-Based Learning: Memory-based learning, often referred to as instance-based learning, involves storing and memorizing the entire training dataset. When a new input is presented, the algorithm finds the most similar training example(s) and uses them to make predictions. It is a form of lazy learning, as it does not build a model during training but relies on the stored examples during inference.

- Hebbian Learning: Hebbian learning is inspired by the Hebbian theory of synaptic plasticity in neuroscience. It posits "cells that fire together, wire together." In the context of artificial neural networks, this learning rule strengthens the connections between neurons when they are simultaneously active, promoting associative learning.

- Error-Correction Learning: Error-correction learning, often associated with the backpropagation algorithm, is a widely used method in training artificial neural networks. It involves computing the error between the network's output and the target output and then adjusting the network's weights to minimize this error. It is a supervised learning approach and plays a crucial role in training feedforward and deep neural networks.

- Competitive Learning: Competitive learning is another biologically inspired learning algorithm. In competitive learning, neurons or nodes within a network compete to become active based on their input and weights. The winner, the neuron with the most significant response, is the one that "learns" by adjusting its weights to better recognize the presented input.

- Boltzmann Learning: Boltzmann learning is based on concepts from statistical mechanics, particularly the Boltzmann distribution. It is used in training Boltzmann machines, a type of neural network that can be used for various tasks, including optimization and probabilistic modeling. Boltzmann learning involves adjusting the network's weights to move towards a state with lower energy, where lower energy states correspond to more probable configurations.

In general, learning algorithms can be categorized into three types: supervised or teacher-guided learning, semi-supervised learning, and unsupervised or self-guided learning. Supervised learning relies on labeled training data with explicit teacher signals. Semi-supervised learning combines labeled and unlabeled data. Unsupervised learning algorithms, in contrast, aim to uncover patterns and structures within the data without explicit teacher signals.

- In supervised learning algorithms, a teacher possesses knowledge about the environment and shares it with the neural network through examples of inputs and their corresponding outputs. The network undergoes a training process where synapse weights are adjusted using a modification rule until the desired computation is achieved. The supervision continues until the network is capable of producing similar outputs to specific inputs observed during the training phase. Examples of supervised algorithms include error-correction algorithms like back-propagation using gradient descent, as well as well-known algorithms such as support vector machines (SVM) and Bayesian learning algorithms. Supervised learning uses labeled data for training, where input data and corresponding target values are used to teach the algorithm to make predictions (regression) or classifications (classification). During testing, the algorithm applies what it learned to new, unlabeled data, evaluating its performance by comparing predictions to true labels or target values. It is fundamental in tasks like image recognition and medical diagnosis, where past data informs accurate predictions with new data.

- Semi-supervised learning algorithms lie between supervised learning and unsupervised learning approaches. Obtaining labeled data can be challenging, expensive, and time-consuming as it often requires the expertise of human annotators. On the other hand, collecting unlabeled data is relatively easier. Semi-supervised learning leverages a combination of labeled and unlabeled data to develop more effective classifiers. In the context of semi-supervised learning, the learning process can be seen as an exam, where the labeled data represents a few example problems that the teacher has solved during the course. Additionally, the teacher provides a set of unsolved problems. By utilizing both labeled and unlabeled data, semi-supervised learning requires less human effort while achieving higher accuracy. This makes it highly valuable in both theoretical and practical applications.

- Unsupervised learning algorithms operate in the absence of a teacher, and the network has no knowledge about the environment. In unsupervised learning, there is no labeled output data available. Instead, the focus is on discovering patterns and structures within the data that go beyond what can be considered pure unstructured noise. Clustering is a classic and straightforward example of unsupervised learning. It involves grouping data points into clusters based on their similarities or proximity to each other. Hebbian plasticity is another form of unsupervised learning, which is particularly useful for clustering input data. However, it may be less suitable when a desired outcome for the network is already known in advance.

1.5 Network topology

The arrangement of connections between neurons in an artificial neural network is referred to as the topology, architecture, or graph of the network. The specific structure of the interconnections is closely

related to the learning algorithms used to train the neural network. Different ways of structuring the interconnections lead to various topologies, which can be broadly categorized into two main classes: Feed-Forward Neural Networks (FFNN) and Recurrent Neural Networks (RNN), also known as feedback neural networks. These classes are illustrated in Figure 1.5.

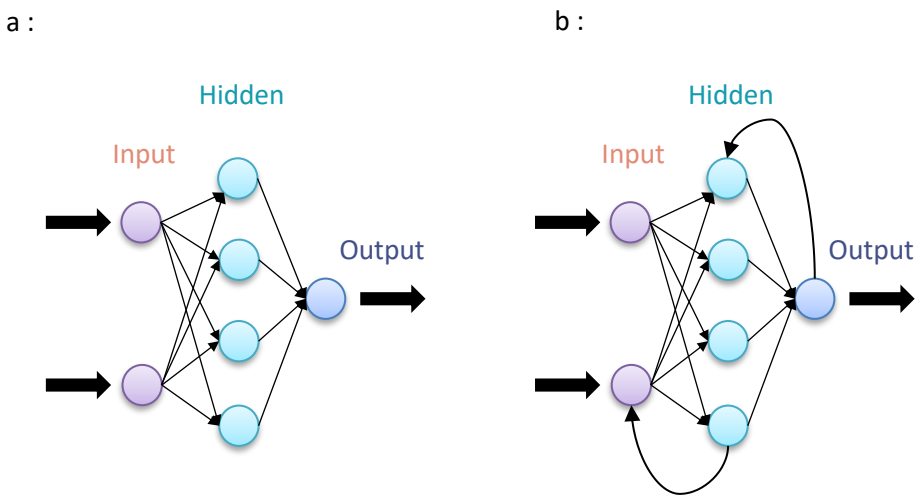


Figure 1. 5 Two example types of artificial neural network architectures: a) Feed-Forward Neural Networks have a structure where information flows in a single direction, without any feedback connections. b) Recurrent Neural Networks are characterized by their feedback connections, which allow information to circulate within the network in cycles.

1.5.1. Feed-Forward Neural Networks

The Feed-Forward Neural Network (FFNN) can be categorized into two distinct structures: single-layer FFNN and multilayer FFNN. The single-layer FFNN consists of an input layer and an output layer, forming a strictly feed-forward or acyclic graph. The input layer is not considered in the calculations as no computations are performed within its nodes (neurons). On the other hand, the multilayer FFNN incorporates one or more hidden layers between the input and output layers, as depicted in Figure 1.5.a where there is one hidden layer. By introducing hidden layers, the neural network can extract higher-order statistics, which is particularly useful when dealing with large input layer sizes. Among the various types of neural networks, feed-forward neural networks are widely employed due to their simplicity, flexible structure, good representation capabilities, and the ability for universal approximation. In terms of interconnectivity between nodes (neurons), there are two types of feed-forward architectures:

- Fully connected: In this configuration, every node in each layer of the network is connected to every other node in the subsequent layer. These networks can be referred to as globally connected networks, and the Restricted Boltzmann Machine (RBM) serves as an example of a fully connected FFNN.

- Partially connected: In this configuration, certain communication links are absent or missing. Convolutional neural networks are a notable example of partially connected FFNN. Partially connected topologies provide an alternative with reduced redundancy, offering potential efficiency improvements in neural networks.

1.5.2. Convolutional Neural Networks (CNN)

A Convolutional Neural Network (CNN) is a multi-layer feed-forward network architecture where neurons in one layer receive inputs from multiple neurons in the previous layer and produce an output based on a weighted sum of its inputs, which is typically passed through a threshold or sigmoidal function. The connectivity pattern between nodes in one layer and the subsequent layer forms the convolution kernel, which performs the weighted sum operation. In each layer of a CNN, there are typically one or a few convolution kernels that connect a group of neurons from one layer to the target neuron in the next layer (Lecun et al., 1998). This architecture has been extensively studied in the neuromorphic community for visual processing tasks (Krizhevsky et al., 2012). Traditionally, CNNs have been implemented on CPUs and GPUs, which consume significant power. However, in recent years, there has been a growing interest in utilizing System-on-Chip (SoC) solutions and FPGA platforms to implement CNNs. These alternative platforms offer improved performance and reduced power consumption compared to traditional CPU and GPU implementations. This shift towards more efficient hardware implementations aims to enhance the overall efficiency and scalability of CNNs for various applications.

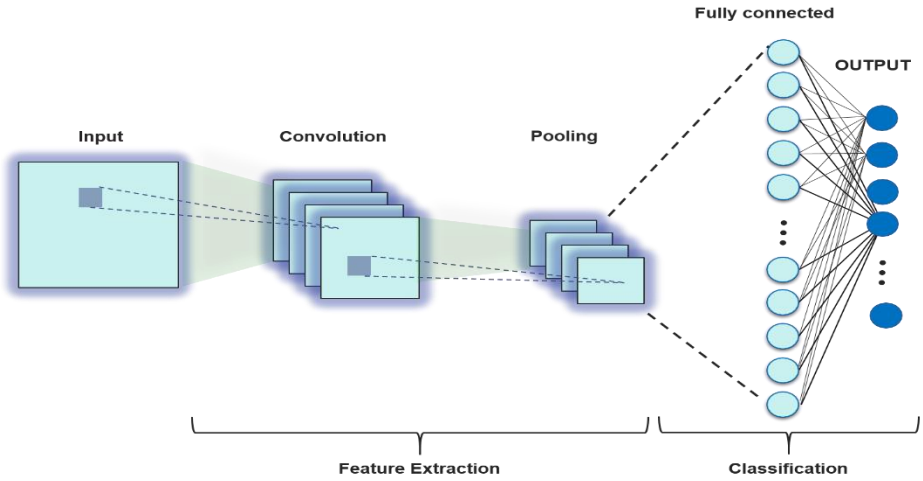


Figure 1. 6 : A simplified diagram illustrating the architecture of a basic convolutional neural network (CNN). The diagram depicts the sequential flow of information within the network.

Convolutional neural networks (CNNs) are specifically designed to process data in the form of multidimensional arrays, making them well-suited for tasks such as image recognition, video understanding, speech recognition, and natural language understanding. Figure 1.6 shows an example for CNN. Inspired by the human visual system (W. Choi et al., 2018; R. G. Kim et al., 2018), CNNs employ convolution as a linear operator to aggregate information from neighboring pixels. In a CNN, neurons are arranged in three dimensions (width, height, depth), and each layer transforms the 3D

input volume into a corresponding 3D output volume. The input layer of a CNN typically represents an image, with its width and height corresponding to the image dimensions, and the depth representing the color channels (e.g., RGB) (Ciregan et al., 2012). CNNs commonly consist of four layers: convolutional, activation, pooling, and fully connected layers (Ciregan et al., 2012). The output feature maps of the final convolution or pooling layer are often flattened into a 1D array and connected to fully connected layers, which apply learnable weights. These layers are followed by nonlinear functions (e.g., ReLU), (Romanuke, 2017) and the final fully connected layer typically outputs the probabilities for each class in classification tasks. CNNs can have different architectures, such as AlexNet (Alom et al., 2018), VGGNet (Muhammad et al., 2018), GoogleNet (Al-Qizwini et al., 2017), and ResNet (Mukti & Biswas, 2019, p. 50). Convolution in CNNs involves element-wise multiplication of a filter (weight) with parts of the image, resulting in feature maps. This process is repeated until the entire image is scanned, and it often leads to a reduction in the image size.

1.6 Current and Future Challenges in Neuromorphic Computing

Neuromorphic computing, inspired by the structure and function of the human brain, aims to develop computer systems that can perform complex cognitive tasks efficiently (Schuman et al., 2017). While there have been significant advancements in this field, researchers and engineers face several challenges in both the current and future development of neuromorphic computing. Designing and fabricating efficient and scalable hardware architectures for neuromorphic computing is a major challenge (Benjamin et al., 2014). Current neuromorphic hardware, such as neuromorphic chips, often struggle to match the complexity and computational power of the human brain while maintaining low power consumption (Chicca et al., 2014). Researchers are exploring new materials, device technologies, and architectures to enable efficient and large-scale implementation of neuromorphic systems (Kuzum et al., 2013; Schemmel et al., 2010). Achieving neuroplasticity and learning capabilities similar to the human brain is another challenge(Pfeil et al., 2012). Developing algorithms and mechanisms for unsupervised learning, reinforcement learning, and lifelong learning in neuromorphic systems is an ongoing research area. Scaling up neuromorphic systems to handle complex real-world problems requires addressing issues related to scalability, connectivity, and integration of neural components. Additionally, programming models and software tools need to be developed to exploit the unique capabilities of neuromorphic hardware (Appeltant et al., 2011). Evaluating and benchmarking the performance of neuromorphic systems is a challenge due to the lack of standardized evaluation metrics and test datasets. Ethical and privacy concerns related to data security, user privacy, and potential misuse of neuromorphic technologies also need to be addressed.

Conclusion

In this opening chapter, we embarked on a journey into the captivating world of Neuromorphic Computing. Our exploration began with a fundamental understanding of biological neural networks, shedding light on the complex web of interconnected neurons that drive our cognitive processes. By comparing biological neurons with their artificial counterparts, we uncovered the remarkable efforts in mimicking the brain's computational power. The chapter delved deeper into the fascinating realm of

synapses, where learning and adaptation take place. Synaptic behavior and plasticity were discussed, highlighting the foundation for many learning algorithms that power artificial neural networks.

We navigated through the diverse landscape of network topologies, recognizing the significance of feed-forward and convolutional neural networks in various applications. These network architectures play a vital role in enabling machines to process and understand data in ways that resemble human cognitive processes.

Finally, we contemplated the current and future challenges that Neuromorphic Computing faces. The field holds immense promise, yet it is not without its hurdles. As we look ahead, we must address these challenges, which range from hardware limitations to ethical and societal concerns. Solving these issues will pave the way for a more advanced and ethically responsible era of Neuromorphic Computing. This chapter serves as a solid foundation for our journey into Neuromorphic Computing, offering insights into the biological inspirations, the tools at our disposal, and the road ahead. It is a reminder of the exciting potential and the responsibility that comes with harnessing the power of brain-inspired computational systems. As we continue our exploration, we will delve deeper into the intricacies of Neuromorphic Computing, building upon this knowledge to unlock its full potential.

Résumé

- Explored Neuromorphic Computing fundamentals.
- Compared biological and artificial neurons.
- Examined synaptic learning and algorithms.
- Discussed network topologies, including CNNs.
- Highlighted current and future challenges.

Chapter 2

Emerging Nonvolatile Memories for Neuromorphic Computing

"As we venture into the era of neuromorphic computing the choice of nonvolatile memories is pivotal — shaping not just the storage, but the very foundation of cognitive computing architectures."

Dr. Krisztian Flautner, CEO of AImotive.

2. The Market for Semiconductor Memory

Today's computing systems are designed based on the Von Neumann architecture, which distinguishes the roles of the Central Processing Unit (CPU) and the Memory Unit. In this architecture, the CPU handles arithmetic operations, logic functions, control tasks, and input/output operations as instructed by a set of stored instructions, which make up a computer program stored in the Memory Unit. The Memory Unit holds both the code of computer programs and the data, including information to be processed by the CPU and the computation results (Godfrey, 1993).

Nonetheless, as the size of the memory array increases, the time and power required to access information also increase. Therefore, when a large storage capacity is needed for computations, the power consumption and access time of the memory become the dominant factors affecting overall power and performance. In fact, there exists a performance disparity between processors and memory, where the speed at which data can be accessed from memory, constrained by latency and bandwidth, typically limits the computation performance. This performance gap is commonly known as the memory bottleneck. In the pursuit of advanced non-volatile memory solutions, several technologies have developed in research over the past 15 years (Parkin, 2004; Qureshi et al., 2009). These technologies offer advantages over Flash memory by eliminating its limitations, such as low endurance (limited write operations), the need for high voltage supply during programming, lengthy write times, and complex erase procedures (Wong & Salahuddin, 2015). Additionally, Flash memory presents challenges as a Front-End-Of-Line (FEOL) technology, making it difficult to co-integrate with sub-32 nm CMOS (Skotnicki, 2007).

Phase-Change Random Access Memory (PCRAM or PCM), Conductive-Bridging Random Access Memory (CBRAM), and metal Oxide resistive Random Access Memory (OxRAM) and other more are the leading emerging non-volatile memory technologies. These emerging memory technologies utilize physical mechanisms that differ from traditional methods of storing charge in a capacitor or the floating gate of a transistor, as seen in SRAM, DRAM, and Flash. They are integrated in the Back-End-Of-Line. It is important to note that the term resistive RAM (RRAM or ReRAM) is commonly used in literature to encompass both OxRAM and CBRAM. The following sections will provide an overview of the primary emerging non-volatile memory technologies, stressing the key properties and performance aspects.

2.1 CBRAM

In figure 2.1a, the diagram illustrates the structure of Conductive-Bridging Random Access Memory (CBRAM) devices. These devices consist of a Metal-Insulator-Metal setup, where the top electrode becomes electrochemically active or oxidized under a positive bias, while the bottom electrode remains electrochemically inert. The insulating materials positioned between the top and bottom electrodes can be solid electrolytes (Vianello et al., 2012) or metal oxides (Guy et al., 2013).

When a positive voltage is applied to the top electrode, mobile metal ions from the top electrode move through the solid electrolyte or oxide, driven by the electric field. These ions then deposit and reduce on the inert of the bottom electrode, creating conductive filaments (CF) composed of elements from the top electrode (typically Cu or Ag). These filaments bridge the top and bottom electrodes, causing the device to transition into a Low Resistance State (LRS), known as the SET operation.

Conversely, when the voltage is reversed, the metal ions migrate back to the top electrode, dissolving the conductive filaments and returning the device to a High Resistance State (HRS), known as the RESET operation (Barci et al., 2014). The CBRAM's current-voltage characteristics, as shown in Figure 2.1b, exhibit distinct behavior during SET and RESET operations.

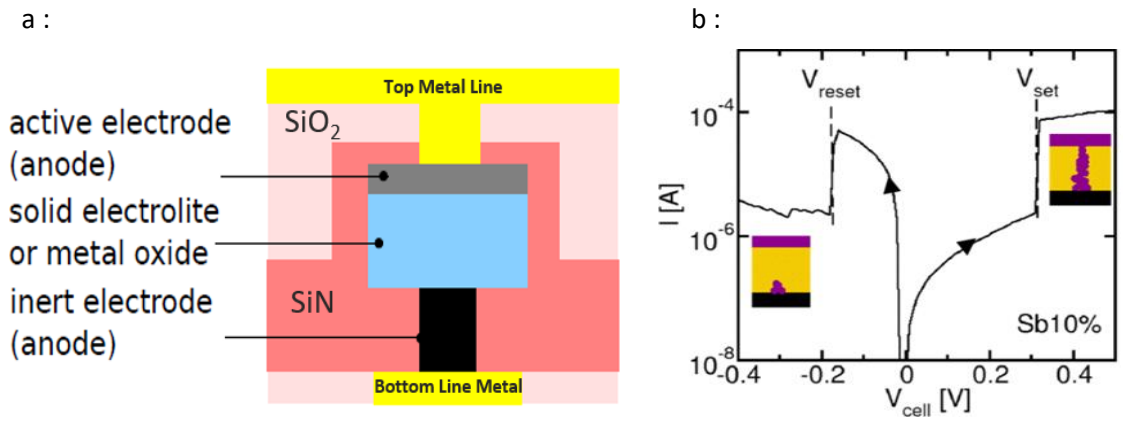


Figure 2. 1 (a) presents a schematic representation of the CBRAM device, as described in reference (Palma et al., 2012). (b) The current-voltage characteristics of CBRAM are shown, as outlined in (Vianello et al., 2012). This diagram illustrates the typical behavior of the CBRAM device

By employing interface engineering of chalcogenide CBRAM with a dual-layer electrolyte stack, researchers have achieved a resistance ratio between HRS and LRS exceeding 10^6 (one million). However, due to the stochastic nature of ion migration, each SET and RESET operation results in a different configuration of the conductive filaments. This variability leads to significant resistance fluctuations, particularly in the HRS. One significant limitation of CBRAM technology is its retention issues, primarily attributed to the high mobility of ions within the matrix. This results in relatively short-lived ON states, diminishing its overall performance and reliability.

2.2 OxRAM

Oxide-based resistive RAM (OxRAM) devices, similar to CBRAM, feature a straightforward MIM (metal-insulator-metal) structure, as illustrated in Figure 2.2a. This structure comprises a metal oxide layer positioned between a top and a bottom electrode. When an electric field is applied to the device, it initiates the creation and migration of oxygen vacancies (Vo) within the oxide layer. These oxygen vacancies are essentially oxygen atoms missing from their lattice positions within the oxide material. They arise due to various factors, including defects in the manufacturing process, the composition of the oxide material, and the influence of external factors, such as temperature and electrical stress. The presence and controlled movement of these oxygen vacancies play a vital role in the resistive switching

behavior of OxRAM devices, enabling data storage and retrieval functions. This allows for the formation and destruction of conductive filaments (CFs) in oxygen vacancies, resulting in a change in the device's resistance. The device can be switched between a Low Resistance State (LRS) and a High Resistance State (HRS) through SET and RESET operations, respectively. Figure 2.3 provides a schematic illustration of these switching processes. Initially, a forming or electroforming process is required to create a conductive filament in the oxide layer for fresh samples in their pristine resistance state (Wong et al., 2012). During the forming process, oxygen ions migrate towards the top electrode interface under the influence of the electric field. If the top electrode material is oxidizable, an interface oxide layer forms. Otherwise, if the top electrode material is inert, oxygen accumulates as nonlattice atoms. As a result, the top electrode/oxide interface acts as an oxygen reservoir (Fujimoto et al., 2006) for subsequent SET and RESET operations. The operation of OxRAM can be classified into two switching modes, namely unipolar and bipolar, based on the polarity of the voltage required for SET and RESET. Figures 8b and 8c depict schematic current-voltage characteristics for these two switching modes.

In the unipolar switching mode (Fig. 2.2b), the SET and RESET operations are determined solely by the magnitude of the applied voltage. As a result, both operations can be achieved using the same polarity of programming voltage. During the RESET operation, when current passes through the conductive filament (CF), Joule heating occurs, causing the temperature to rise. This elevated temperature activates the thermal diffusion of oxygen ions, which then move away from the CF due to the concentration gradient (H. D. Lee et al., 2010). As a result, the device transitions to the High Resistance State (HRS). If both positive and negative voltage polarities can be used for both SET and RESET operations, the unipolar switching mode is also referred to as the nonpolar mode.

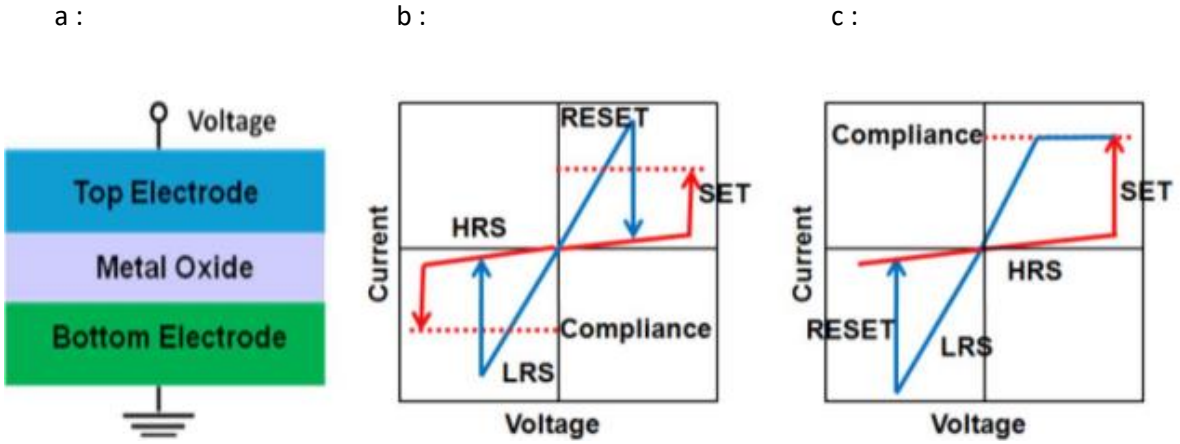


Figure 2. 2 (a) Schematic of MIM structure of OxRAM devices and (b) Schematic unipolar and (c) bipolar current-voltage characteristics (Wong et al., 2012)

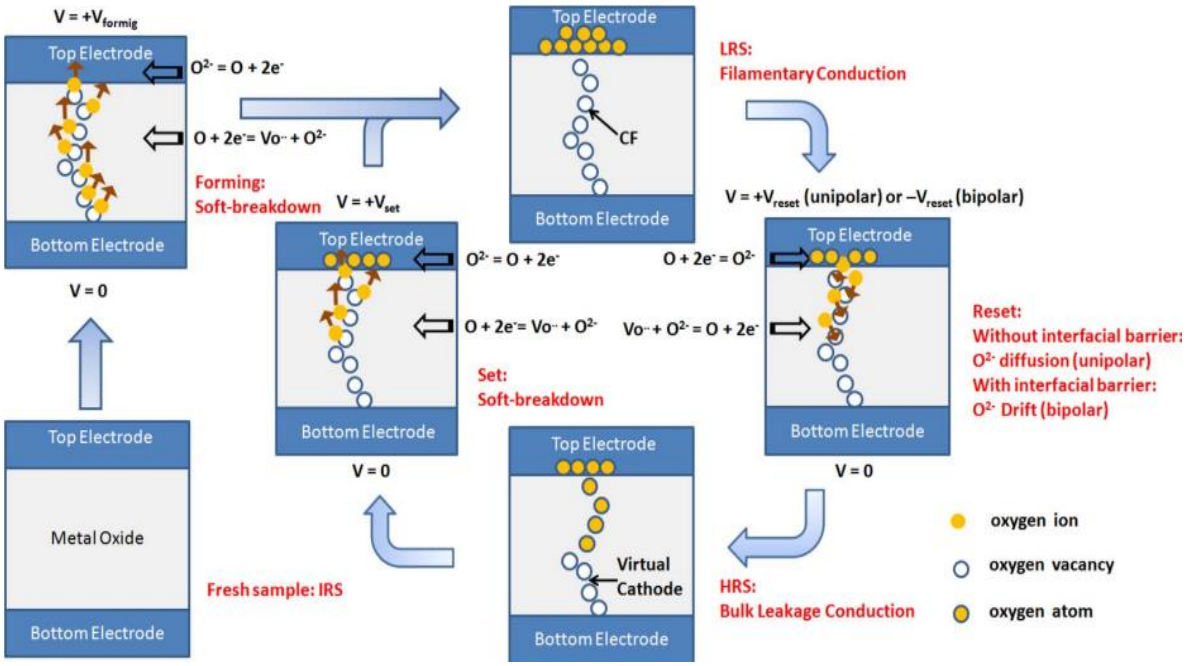


Figure 2. 3 A diagram depicting the operational mechanism of OxRAM, sourced from (Wong et al., 2012).

In the bipolar switching mode (Fig. 2.2c), the SET and RESET operations are carried out using reverse voltage polarities. The presence of an interfacial oxide layer at the top electrode can create a significant diffusion barrier. In such cases, the thermal diffusion caused by Joule heating and the concentration gradient alone may not be sufficient for the RESET process. Therefore, a reverse electric field is required to enhance the migration of oxygen ions and facilitate the RESET operation.

2.3 Ferroelectric Random-Access Memory (FeRAM)

FeRAM memories represent one of the earliest emerging memory technologies that have been brought into production. These memories are primarily constructed using a combination of lead, zirconium, and titanium (PZT) (A. Chen, 2016; Meena et al., 2014) or more recently, HfO_2 which stands for hafnium dioxide (Böscke et al., 2011). FeRAM employs a ferroelectric material, where an electric dipole remains present even without an external electric field. By applying an electric field, the memory can be toggled between different polarization states (see Figure 2.4) by altering the positions of zirconium and titanium atoms. Notably, this change in atomic configuration persists even after the electric field is removed, providing the FeRAM device with its non-volatile nature.

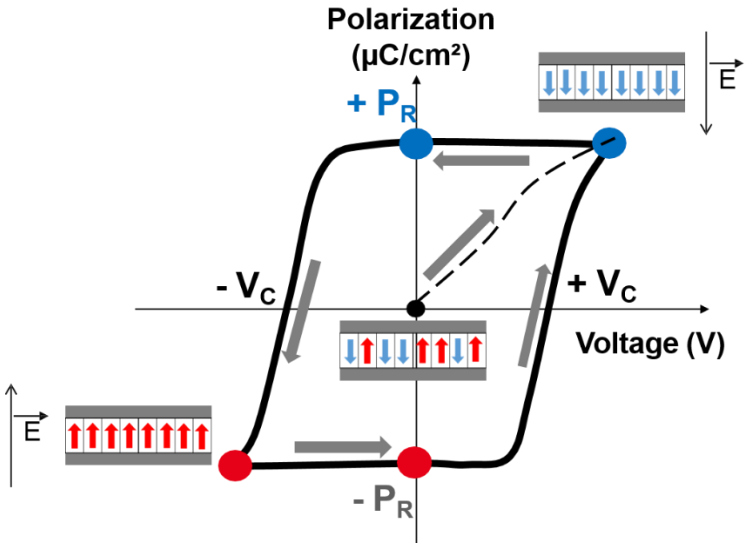


Figure 2. 4 The hysteresis curve was generated using silicon-doped hafnium oxide devices (Fujisaki, 2013). The structure of a FeRAM closely resembles that of a DRAM, except that the dielectric material in a DRAM is substituted with this ferroelectric material. This ferroelectric material is integrated with a MOS transistor (source: Courtesy of L. Grenouillet) (see Figure 2.5).

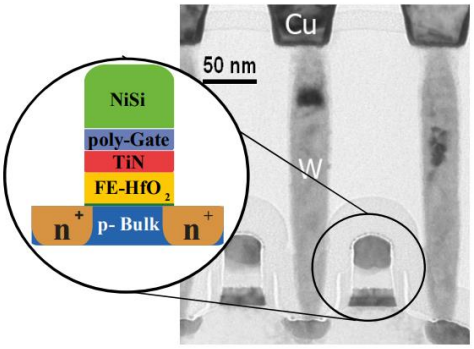


Figure 2. 5 TEM cross section illustrates implementation of FE-HfO2 with schematization of a FeRAM memory (Mikolajick et al., 2014)

To determine the memory state, a write operation is performed, and if a current pulse is detected, it signifies that the memory was in the OFF state (A. Chen, 2016). However, this read operation has a disadvantage as it causes data loss. On the positive side, FeRAMs exhibit excellent characteristics such as high speed, low energy consumption, low operating voltage (approximately 2V), and exceptional endurance with a high number of cycles for the 16Kbit arrays with 10^9 cycles. The FeFET devices (Ferroelectric Field Effect Transistors) follow a similar operational principle. These non-volatile ferroelectric transistors are currently the subject of extensive research due to their potential for faster performance compared to DRAMs and higher data density than Flash memories.

2.4 STT-RAM

In Spin-Transfer-Torque Magnetic Random Access Memory (STT-MRAM) devices, data is stored by controlling the magnetization orientation of a nano-scale ferromagnetic layer. Figure 2.6a provides a schematic view of a typical STT-MRAM bit-cell.

The core component of STT-MRAM is the Magnetic Tunnel Junction (MTJ), which consists of two magnetic layers separated by an insulating MgO tunneling barrier. The orientation of the magnetization in the Free Layer (FL) can be switched between two states to store information, while the Reference Layer (RL) has a fixed magnetic orientation acting as a stable reference (Khvalkovskiy et al., 2013). If RL and FL have the same orientation, the device is in the Parallel (P) state. If their orientations are opposite, the device is in the Anti-Parallel (AP) state. The STT-MRAM operates based on two phenomena: the Tunnelling Magneto-Resistance (TMR) effect (Julliere, 1975), responsible for the resistance difference between the P and AP states, and the Spin-Transfer Torque (STT) effect (Berger, 1996; Slonczewski, 1989), used to switch the magnetization of FL.

During a read operation, the device's resistance is sensed to determine the magnetic state of FL, thereby retrieving the stored information. In the write operation, a current flow generates a torque on FL's magnetization. If the torque is sufficient, the magnetic state of FL is switched, and data is written. The direction of the current flow determines whether a positive voltage leads to AP-to-P transition, while a negative voltage leads to P-to-AP transition (Ad & Dc, 2015).

STT-MRAM is expected to have excellent endurance because its switching mechanism does not involve any magnetic degradation or atomic movement during write operations, unlike other memory technologies like PCRAM, CBRAM, or OxRAM. However, the dielectric breakdown of the MgO tunnel barrier may occur if the voltage across it exceeds approximately 400mV (Min et al., 2010). The complexity of the MRAM stack is a notable factor contributing to the challenges faced by this technology, potentially impeding its wider adoption in the market.

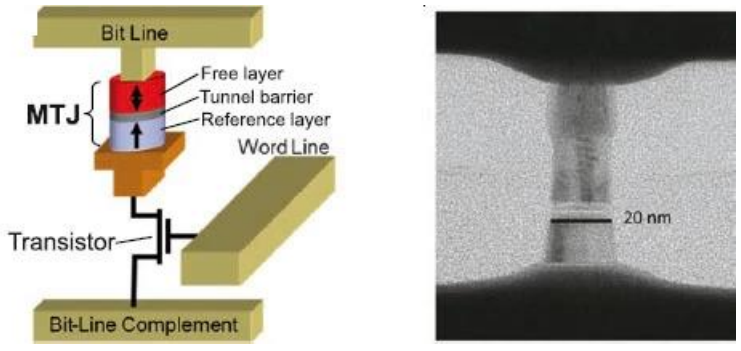


Figure 2. 6 (a) Diagram of a bit cell in spin-transfer-torque magnetoresistive random-access memory, (b) Image obtained through cross-sectional transmission electron microscopy depicting a perpendicular magnetic tunnel junction (p-MTJ) with a diameter of 20 nanometers. (Yuasa et al., 2018)

2.5 PCRAM

The fundamental principle governing the operation of Phase-Change Random Access Memory (PCRAM or PCM) hinges on the distinctive electrical traits of phase-change materials. These materials distinctly showcase a pronounced contrast in resistivity between their amorphous and crystalline phases. Specifically, the amorphous phase is marked by elevated electrical resistivity, while the crystalline phase exhibits notably diminished resistivity (Raoux et al., 2008). Within PCRAM, the alteration of the phase-change material between its amorphous and crystalline states is accomplished through Joule heating. As depicted in Figure 2.7a, typical current-voltage characteristics illustrate the crystalline and amorphous phases of phase-change materials. The process of crystallization is executed by elevating the material's temperature beyond its crystallization threshold (referred to as the SET operation). Conversely, the amorphous state is achieved by melting the material into a liquid form and rapidly cooling it into a disordered amorphous phase (known as the RESET operation). These operations are facilitated by electrical current pulses: high-power pulses are essential for the RESET operation, while moderately powered but lengthier pulses are utilized for the SET operation. To retrieve stored information, low-power pulses are employed to gauge the device's resistance (Raoux et al., 2008). Figure 2.7b schematically illustrates a mushroom-shaped PCRAM cell. A particularly appealing feature of PCRAM is its capacity for achieving multi-level cell (MLC) storage. This implies that the device can be programmed to possess resistance states spanning various levels, beyond the complete SET and RESET resistance levels. Accomplishing this involves adjusting the ratio between the sizes of the crystalline and amorphous regions within the active region. The MLC functionality serves as an effective approach to curbing memory expenses, as it permits the storage of more data within a given silicon area (Stanisavljevic et al., 2015).

a : b :

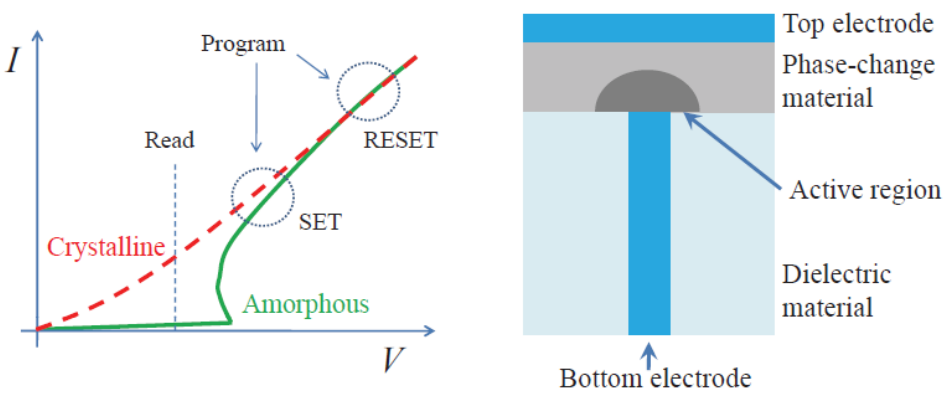


Figure 2. 7 (a) The standard current-voltage behaviors of crystalline and amorphous phases of phase-change materials. (b) An illustrative cross-sectional representation of a phase-change memory cell.(Pershin & Di Ventra, 2011)

2.6 Comparison of NVM technologies

The ongoing quest for advanced Non-Volatile Memory (NVM) technologies has ignited a dynamic landscape of options, each boasting distinctive attributes that cater to various application demands. The comparative analysis of these emerging NVM technologies sheds light on their respective strengths and challenges.

Comparing the performance of emerging nonvolatile memory (NVM) technologies reveals a spectrum of capabilities tailored to diverse applications. Among these options, Phase Change Memory (PCM) emerges as a standout choice for various scenarios, particularly in neuromorphic applications. PCM's notable attributes, including rapid read and write speeds, commendable endurance, and non-volatility, position it as a compelling contender. Its compatibility with existing semiconductor manufacturing processes further enhances its appeal. While PCM exhibits some limitations in terms of write endurance and power consumption during writes, its blend of attributes makes it suitable for applications demanding efficient learning and inference capabilities, aligning well with the objectives of neuromorphic computing. One of the main challenges associated with PCM, however, is its retention at high temperatures. This issue can impact the reliability of data storage in elevated-temperature environments. Regarding endurance and power consumption, PCM generally performs satisfactorily in these aspects. It's worth noting that even though a single bit in PCM consumes just picoWatts (pW) of power, it is essential to consider the overall energy expenditure in the process of charging and discharging the bitlines, as well as data transfer operations. While the power consumption of individual bits may be low, a significant amount of energy can be consumed in the context of memory array operations. Nevertheless, PCM technology typically operates effectively with supply voltages below 2 or 3 volts and current in the range of a few hundred microamperes (μA), making it a competitive option for various applications. The performance evaluation, however, underscores the importance of considering application-specific requirements and staying attuned to the ever-evolving technological advancements in the NVM landscap

Parameter	PCM	ReRAM	STT-MRAM	FeRAM	OxRAM	CBRAM	3D XPoint	NAND Flash
Endurance	Good	Good	Good	Good	Good	Good	Good	Average
Retention	Good	Good	Good	Good	Average	Average	Good	Good
Write Speed	Good	Good	Good	Good	Good	Good	Good	Average
Read Speed	Good	Good	Good	Good	Good	Good	Good	Good
Power Efficiency	Good	Good	Good	Good	Good	Good	Average	Good
Scalability	Good	Good	Good	Average	Good	Good	Good	Good
Non-volatility	Good	Good	Good	Good	Good	Good	Good	Good
Fabrication Complexity	Average	Good	Average	Good	Average	Average	Average	Good

Table 1: Comparison of the performance of the different emerging nonvolatile memory technologies

2.7 Emerging memories for neuromorphic application

Emerging memories have gained significant attention in the realm of neuromorphic computing due to their unique alignment with the brain's information processing mechanisms. Neuromorphic applications aim to emulate the brain's efficient and parallel processing abilities. Emerging memories, such as resistive random-access memory (RRAM) and phase-change memory (PCM), offer promising avenues for achieving this goal. These memories can store and process information in a non-volatile manner, allowing synaptic weights to be stored directly within the memory cells themselves, mimicking the synaptic connections between neurons. This eliminates the need to transfer data constantly between separate memory and processing units, resulting in faster and more energy-efficient computations. The analog behavior of emerging memories also simulates the graded nature of neural communication, enabling more accurate representation of synaptic strength. This potential to leverage emerging memories for neuromorphic applications could revolutionize artificial intelligence and edge computing, providing a brain-like approach to computation and enabling advanced cognitive capabilities in various applications.

Furthermore, integrating emerging memories in neuromorphic systems opens possibilities for efficient learning and adaptation. The inherent plasticity of these memories mirrors the concept of synaptic plasticity in biological neural networks. This adaptability allows for the implementation of learning algorithms like spike-timing-dependent plasticity (STDP) (Feldman, 2012), enabling the system to learn from its environment and improve performance over time. Organizing emerging memories in crossbar arrays replicates the connectivity seen in biological neural networks, facilitating parallelism and supporting the required connectivity for complex neural computations(Liu & Zeng, 2022; Wan et al., 2022).

Practically, using emerging memories in neuromorphic applications offers reduced power consumption, accelerated processing speeds, and improved scalability. With the rise of edge computing and the need for efficient AI implementations in resource-constrained environments, emerging memories offer a tantalizing solution. Continued innovation in this field is likely to yield even more advanced functionalities. Challenges include developing specialized hardware architectures and algorithms that harness the capabilities of emerging memories, resulting in neuromorphic processors rivaling traditional architectures in performance while consuming less power.

In the broader scope, the synergy between emerging memories and neuromorphic computing aligns with the broader goals of AI research building machines that can emulate the cognitive abilities of the human brain. Achieving such sophisticated neuromorphic systems will involve contributions from material scientists, hardware engineers, neuroscientists, and computer scientists. The potential impacts span industries, from revolutionizing autonomous systems to enabling new human-computer interaction modes. The integration of emerging memories into neuromorphic applications stands at the forefront of technological innovation, promising a future where machines can operate with the elegance and efficiency of biological neural networks.

In conclusion, the convergence of emerging memories and neuromorphic computing holds the promise of revolutionizing the design and deployment of intelligent systems, leading to more energy-efficient, adaptable, and brain-inspired computational technologies.

Conclusion

In this chapter, we explored the landscape of semiconductor memory technologies with a focus on their potential applications in the field of neuromorphic computing. We covered CBRAM, OXRAM, FeRAM, STT-RAM, and PCRAM, highlighting the unique features and advantages of each.

A crucial part of our investigation involved a thorough comparison of these non-volatile memory (NVM) technologies. This comparative analysis allowed us to assess their strengths and weaknesses, facilitating informed decision-making in selecting the most suitable NVM technology for specific neuromorphic applications. Our exploration also extended into emerging memory technologies poised to play a pivotal role in future neuromorphic computing developments. This section underlines the ever-evolving nature of memory technologies and emphasizes the importance of ongoing research and development. By acquiring this knowledge, we have equipped ourselves with the essential understanding needed to harness the potential of memory technologies in creating brain-inspired computational systems. This chapter provides a solid foundation for our ongoing exploration of neuromorphic computing and its exciting potential in the realm of cognitive computing and machine learning.

Résumé

- Explored a range of NVM technologies (CBRAM, OXRAM, FeRAM, STT-RAM, PCRAM).
- Conducted a comparative analysis to assess their suitability for neuromorphic applications.
- Highlighted emerging memory technologies for future advancements.
- Gained insights into the landscape of semiconductor memory options for neuromorphic computing.
- Prepared to make informed decisions regarding memory technology adoption in future chapters.

Chapter 3

Applications and Future Directions

"Nonvolatile memories are not just data repositories; they are the keystones of innovation, unlocking doors to a future where our technologies seamlessly blend with the fabric of our daily lives."

Dr. Wei Lu, Professor of Electrical Engineering and Computer Science, University of Michigan.

3. Neuromorphic computing using non-volatile memory

For over twenty years, the adaptable nature of the 'stored program' Von Neumann architecture has been the driving force behind remarkable advancements in system performance. Nonetheless, with the deceleration of device miniaturization owing to concerns about power consumption and voltage considerations, the resources expended on transferring data through the bottleneck known as the 'Von Neumann bottleneck' between the memory and the processor have become problematic. This is especially pronounced in applications that focus on handling data, like tasks such as real-time image recognition and natural language processing. In these scenarios, modern Von Neumann systems face significant challenges in matching the performance levels of an average human. The human brain presents a captivating alternative to the conventional Von Neumann computing paradigm, known as Non-Von Neumann computing, for upcoming computing systems. This innovative approach is marked by its massively parallel architecture, which interconnects a multitude of energy-efficient computing units referred to as neurons, as well as adaptable memory components known as synapses (Figure 3.1). The brain's design enables it to excel in comparison to contemporary processors across numerous tasks that entail the classification of unstructured data and the identification of patterns. This suggests that by emulating the brain's architecture, future computing systems could potentially achieve remarkable advancements in tasks that involve complex data analysis and recognition. The effort to duplicate the complex structure of the human brain requires embracing the artificial intelligence paradigm, notably exemplified by neural networks.

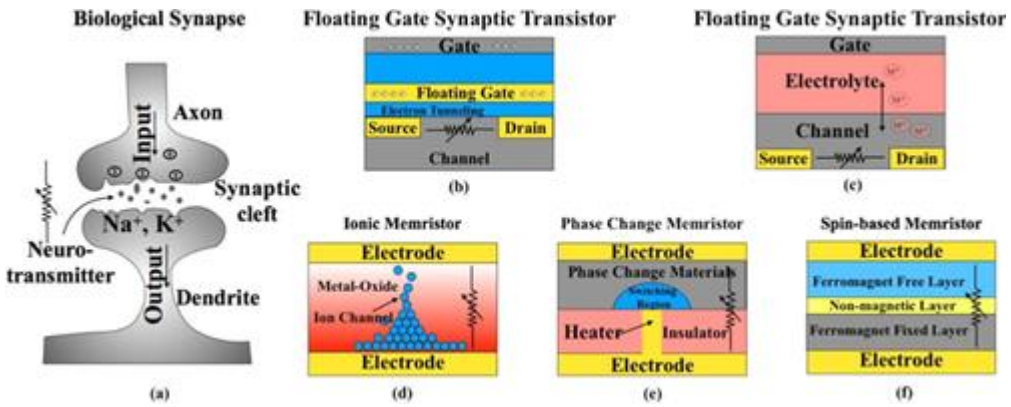


Figure 3. 1 : A biological synapse can be analogously conceptualized as a form of resistive switching. In terms of artificial synaptic devices, the floating gate transistor and ionic transistor utilize the conductance properties of the source and drain electrodes to emulate the behavior of synapses. Conversely, the memristor, characterized by its two-terminal nature, bears a closer resemblance to the biological synapse. It models the synapse by representing the conductance of its top and bottom electrodes.(X. Zhang et al., 2018)

These networks, acknowledged as robust machine learning models, exhibit a degree of intelligence on a technological platform, akin to the way a biological nervous system operates.

Enabling efficient neural networks necessitates the integration of high-density and parallel synaptic storage and computation. These networks, composed of interconnected neurons and synapses, possess computational capabilities and learning capacities akin to those observed in the human brain (Maass & Markram, 2004; Siegelmann & Sontag, 1995). Figure 3.2 illustrates three distinct categories into which neural networks can be grouped, based on the types of neuron models they employ (Maass, 1997).

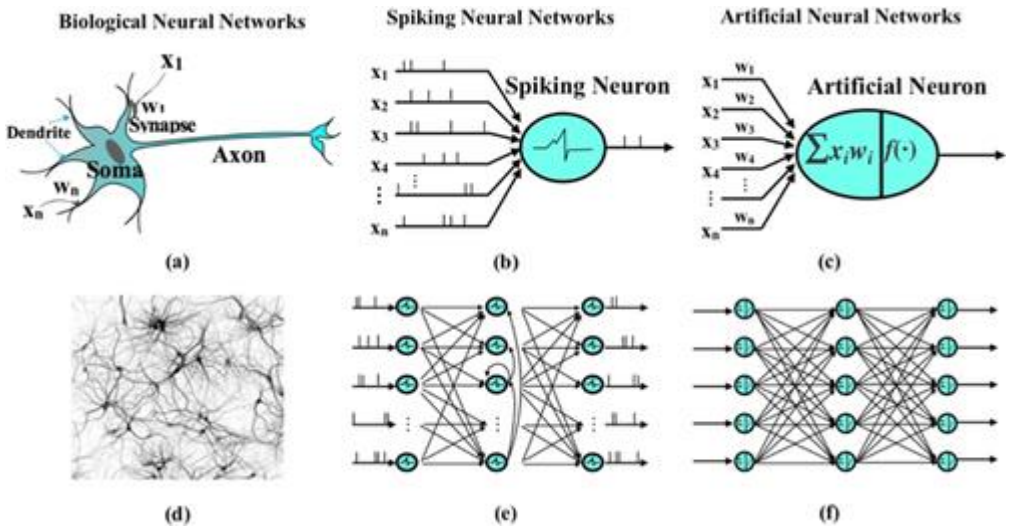


Figure 3. 2 : Schematic of the biological neurons networks, spiking neural networks, and artificial neural networks.

Deep neural network (DNN) has showcased exceptional performance in tasks like object, image, and speech recognition (Indiveri et al., 2013). Nonetheless, DNNs demand substantial datasets and significant energy resources during the training phase to establish the network model, often employing the backpropagation with gradient descent algorithm. This training process can be viewed as a mathematical technique to iteratively enhance the alignment with existing data by adjusting the synaptic weights connecting neuron pairs, however, this lacks a biological parallel in actual nervous systems. In contrast, spiking neural networks (SNNs) seek to emulate certain biological mechanisms, aiming for a learning model that functions more closely to biology. (Table 2 offers a comparative analysis of diverse biological neural networks, SNNs, and ANNs, considering aspects such as synapse models, neuron models, network topology, learning algorithms, implementation, applications, and other features (Andrew, 2003; Esser et al., 2016).

	Biological neural networks	Spiking neural networks	Artificial neural networks
Synapses	Diverse	Short term plasticity (STP), Long term plasticity (LTP), etc.	Numerical Matrix
Neurons	Diverse	Integrate & Fire, Hodgkin–Huxley, etc.	Sigmoid, Tanh, ReLU, Leaky ReLU
Topology	Complex	Hopfield Network, Liquid State Machines, etc.	FNN, CNN, RNN, LSTM, DNC, etc.
Learning algorithm	–	Spike timing dependent plasticity, etc.	Gradient Descent Backpropagation, etc.
Application	Cognition, Inference, Imagination, etc.	Realtime recognition camera, Brain-like Neuromorphic Chip, etc.	Autonomous driving, Voice control system, Medical Dignosis, etc.
Implementation	Brain	TrueNorth, SpiNNaker, Neurogrid, Darwin, etc.	Tensorflow, PyTorch, MXNet, GPU, TPU, Cambrian, etc.
Features	The most complex and powerful computing system and learning system	Biological Close; Realtime; Online Learning; Low power; Noise input; Spatio-Temporal	Multilayer; Feasible and practical with current computing system; Data/computation intensive

Table 2 : Comparison of biological neurons networks, spiking neural networks, and artificial neural networks about synapse models, neuron models, network topology, learning algorithms, and developments (adapted X. Zhang et al., 2018)

This approach requires less data and operates with minimal power by relying heavily on neuronal spike exchanges for information processing (Serrano-Gotarredona et al., 2013). This represents the neuromorphic direction, wherein systems endeavor to integrate biologically relevant architecture, information encoding, and learning mechanisms akin to the human brain. In neuromorphic SNNs, spikes play a crucial role in coordinating learning, governed by Hebbian rules such as spike-timing-dependent plasticity (STDP) and spike-rate-dependent plasticity (SRDP). For the integration of DNNs and SNNs into hardware circuits and systems, the conventional choice has been the adoption of

complementary metal-oxide semiconductor (CMOS) technology, encompassing both digital and analog (or mixed) circuitry (Querlioz et al., 2015). CMOS circuits offer the advantage of flexible design, scalability, and the ability to operate transistors efficiently in the power-conserving subthreshold domain. However, CMOS circuits encounter a deficiency in possessing a nonvolatile memory component capable of sustaining synaptic weights in a desired long-term and multistate fashion.

In contrast, emerging nonvolatile memory technologies, such as resistive switching random access memory (RRAM) (Saïghi et al., 2015), spin-transfer torque magnetic random access memory (STT-MRAM) (Senn & Fusi, 2005) and phase-change memory (PCM) (Boybat et al., 2018), inherently offer a more suitable synaptic element for hardware DNNs and SNNs. These memory variants are compact, scalable, and built upon a two-terminal resistive structure, with their resistance modifiable through electrical pulses. These memory devices have been scaled down to dimensions of approximately 10 nm in height and hundreds of nanometers in width. Notably, PCM exhibit analog switching behavior, permitting gradual increases or decreases in resistance through appropriate pulses. The integration of these emerging memory technologies into CMOS circuits is facilitated through back-end-of-line (BEOL) integration techniques (Eryilmaz et al., 2015). Furthermore, PCM crossbars and matrices facilitate swift and energy-efficient in-memory computing , harnessing the principles of Ohm's law and Kirchhoff's laws, as well as the non-iterative resolution of linear algebra problems. The amalgamation of physical, architectural, and scaling benefits positions nonvolatile resistive memories as a promising technology for embedding synaptic elements within high-density neuromorphic systems.

3.1 Spike-timing-dependent-plasticity

Synaptic plasticity is a fundamental process in the nervous system that involves the modification of synaptic strength, playing a central role in the brain's ability to learn and remember. One key mechanism of synaptic plasticity is spike-timing-dependent plasticity (STDP), which is often considered a manifestation of Hebbian learning. As Donald Hebb succinctly put it, 'Cells that fire together, wire together ' (Keysers & Gazzola, 2014). STDP relies on the precise timing of action potentials in pre-synaptic and post-synaptic neurons. When a pre-synaptic spike precedes a post-synaptic spike, it leads to Long-Term Potentiation (LTP), effectively strengthening the synaptic connection. Conversely, when the spiking is asynchronous or the post-synaptic spike precedes the pre-synaptic one, it results in Long-Term Depression (LTD), leading to a weakening of the synaptic connection (Bi & Poo, 1998).

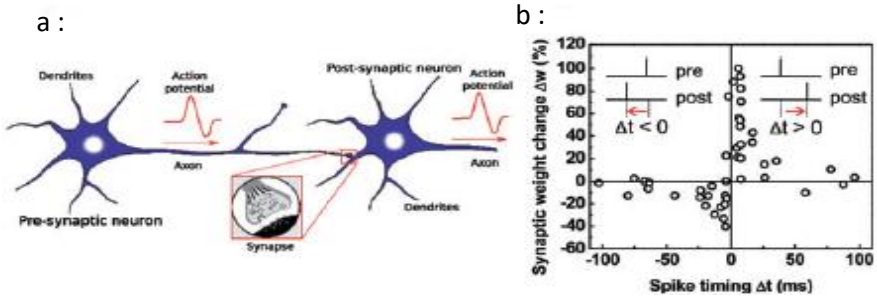


Figure 3. 3: (a) Within the realm of biology, the transmission of excitatory and inhibitory postsynaptic potentials occurs from one neuron to another via synapses, employing a combination of chemical and electrical communication. This dynamic process contributes to the initiation of fresh 'action potentials.' (b) In the biological context, alterations in synaptic strength have been observed to be influenced by the relative timing of spikes

from the pre-synapse and post-synapse (denoted as $\Delta t = t_{post} - t_{pre}$), a phenomenon exemplified in glutamatergic synapses within the hippocampus.(Burr et al., 2017).

This intricate interplay of neural activity and synaptic plasticity allows the brain to adapt and store information, ultimately contributing to the complex processes of learning and memory (Morrison et al., 2008).

Artificial adaptations of this spike-dependent synaptic plasticity, employing asynchronous spikes with consistent amplitude and duration, are often denoted as Spiking Neural Networks (SNNs) (Gruning & Bohte, 2014). Regrettably, there are instances where hardware implementations of traditional Deep Neural Networks (DNNs), which also use asynchronous spikes solely for energy-efficient node-to-node communication without relying on a global clock, are inaccurately labeled as SNNs. It is worth noting that SNNs, as per the definition provided here, encompass the energy-efficient advantages of spike-based communication approaches, particularly when spike occurrences remain adequately sparse. In the realm of SNNs, akin to the presumed operation of the brain, data is encoded into the temporal pattern and frequency of spikes (Gruning & Bohte, 2014). In an SNN, the spiking activity of a pre-synaptic neuron (carrying electrical current) modulates the membrane potential (electrical voltage) of a post-synaptic neuron. This modulation is determined by the 'synaptic weight' (electrical conductance), eventually leading to a post-synaptic spike through mechanisms like leaky-integrate-and-fire or similar neuron models.

The adaptation of Spike-Timing-Dependent Plasticity (STDP) as a localized learning principle for Non-Volatile Memory (NVM) arrays is remarkably straightforward (depicted in Figure 3.4). In this arrangement, one axis of the crossbar array denotes pre-synaptic neurons, while the orthogonal axis signifies post-synaptic neurons. The voltage applied to the wiring leading to the latter neurons corresponds to their membrane potential. Consequently, the implementation of the STDP learning rule becomes the primary task, involving the adjustment of NVM conductance according to the timing of pulses within the pre- and post-synaptic neurons. Intriguingly, non-volatile memory (NVM) devices, even when their conductance change is limited to just one direction, offer a promising avenue for implementing STDP in artificial neural networks. This is made possible by strategically partitioning the functions of Long-Term Potentiation (LTP) and Long-Term Depression (LTD) across two distinct NVM devices, as depicted in Figure 3.4, a concept originally proposed by (Dr. M. Suri et al., 2013).

The key question that arises is why NVM devices should either increase or decrease their conductivity. In the context of simulating synaptic plasticity, NVM devices serve as analogs for biological synapses, adjusting their conductance to mirror the strengthening (LTP) and weakening (LTD) of synaptic connections. The direction of conductance change, whether an increase or decrease, is a critical factor, as it accurately represents the modification in the synaptic connection strength. This decision hinges on the specific characteristics of the NVM technology in use. Moreover, the feasibility of implementing STDP in a simple crossbar array is contingent on the choice of NVM technology. Some NVM technologies align seamlessly with this purpose, allowing the straightforward crossbar architecture to effectively accommodate the division of LTP and LTD functionalities. However, it is important to recognize that the suitability of a particular NVM technology for STDP in crossbar arrays may vary. For instance, phase-change memory (PCM) may pose challenges due to its inherent properties and

limitations. Consequently, the selection of the NVM technology becomes a pivotal determinant in the successful realization of STDP within a given system. Nonetheless, it is important to recognize that STDP functions solely as a localized learning rule and not as a comprehensive computational framework. Therefore, substantial research is still required to explore domains and system architectures where STDP can be effectively harnessed to offer robust and highly valuable non-Von Neumann computational capabilities.

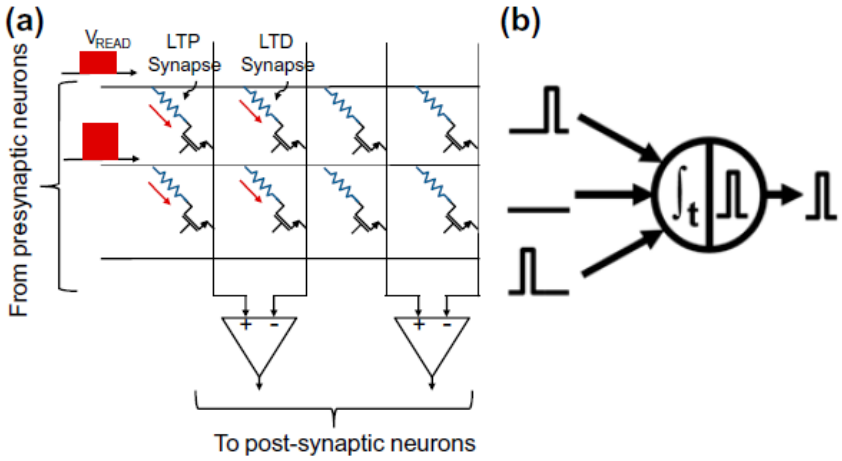


Figure 3. 4 : (a) The incorporation of Spike-Timing-Dependent Plasticity (STDP) through a two-NVM-per-synapse arrangement (adapted from (Dr. M. Suri et al., 2013)) is demonstrated. Given the inherent abrupt RESET in phase-change memory (PCM) devices, Long-Term Depression (LTD) and Long-Term Potentiation (LTP) are executed through SET switching in distinct devices. The overall weight of the synapse is contingent upon the disparity between these two conductance levels. (b) The vital spiking characteristics intrinsic to spiking neural networks are elucidated: downstream spikes are influenced by the temporal accumulation of continuous inputs, with adjustments in synaptic weight contingent upon the relative timing of spikes.

3.2 Vector-matrix multiplication for neuromorphic application

In contrast to Spiking Neural Networks (SNNs), which leverage the biologically realistic Spike-Timing-Dependent Plasticity (STDP) but often lack a robust learning framework, Deep Neural Networks (DNNs) have achieved remarkable breakthroughs in recent times. DNNs encompass various architectures, including Convolutional Neural Networks (CNNs) (Li et al., 2022), specialized for processing visual data, such as images and videos, excelling in tasks like image recognition. Deep Belief Networks (DBNs) (Hua et al., 2015) combine supervised and unsupervised learning, with multiple layers of stochastic variables for capturing complex patterns, applied in domains like speech recognition and recommendation systems. Additionally, Multilayer Perceptrons (MLPs) (Baum, 1988), featuring interconnected layers of neurons, serve as foundational models for diverse tasks, from classification to regression. DNNs, including these architectures, are typically trained through supervised learning and the error backpropagation technique, showcasing the power of data-driven

approaches in solving complex problems. Unlike the asynchronous, single-valued spikes of SNNs, the neuron outputs within a DNN are continuous numerical values processed using synchronous time steps. While these techniques are not directly observed in biological systems, the remarkable success of DNNs can be attributed to a fortuitous combination of several key elements. These include objective function minimization through gradient descent, made possible by the backpropagation algorithm. Additionally, the availability of substantial labeled datasets has provided essential training material. Moreover, the high parallelism achieved with contemporary Graphics Processing Units (GPUs), particularly in matrix multiplications, has significantly propelled the achievements of DNNs in a variety of commercially relevant domains (LeCun et al., 2015).

The suitability of analog resistive memory arrays for the fundamental multiply-accumulate operations in DNNs during both forward-inference and training has been recognized for a considerable period (Shibata & Ohmi, 1997). At each crosspoint, the multiplication is executed by adhering to Ohm's law, while current summation along rows or columns is accomplished through Kirchhoff's current law (illustrated in Figure 3.5). Consequently, these multiply-accumulate tasks can be concurrently conducted at the data's location via local analog computing. This approach conserves power by circumventing the time and energy expenses associated with transferring weight data (Burr, Narayanan, et al., 2015; Burr, Shelby, et al., 2015). Incorporating device read-currents along columns enables the simultaneous computation of the sums of $\Sigma\omega_{ij}x_i$, essential for forwarding the excitation of neurons xi. Correspondingly, integrating along rows achieves the parallel computation of sums of $\Sigma\omega_{ij}\delta_j$, pivotal for error term backpropagation δ_j (Gokmen & Vlasov, 2016).

In 2005, Senn and Fusi (Senn & Fusi, 2005) delved into the realm of pattern classification within simplistic perceptron networks incorporating binary synapses. Alibart et al. showcased a small-scale pattern classification task by employing a 2×9 crossbar array along with the delta learning rule (Alibart et al., 2013). Notably, the nonlinear conductance response of Non-Volatile Memory (NVM) emerged as a significant hindrance to online learning in these setups. Other researchers introduced a read-before-write strategy, which involved sensing the device's conductance before selecting the suitable programming pulse (Alibart et al., 2012; P.-Y. Chen et al., 2015; Jang et al., 2015). While this strategy might be well-suited for the idiosyncrasies of real NVM devices, its potential to scale efficiently to arrays featuring millions of NVM synapses remains uncertain due to its inherently sequential nature.

Novel weight-update strategies compatible with crossbar architectures were proposed (Gao et al., 2015; M. V. Nair & Dudek, 2015), involving the independent firing of programming pulses by upstream and downstream neurons. By aligning these pulses at shared crosspoints, the intended weight adjustments were achieved. Remarkably, no decrease in test accuracy was observed for the MNIST (Deng, 2012) benchmark when this approach was applied. Various learning rules were also explored, such as gradient descent (M. V. Nair & Dudek, 2015), winner-take-all (Tymoshchuk & Wunsch, 2019), and Sanger's learning rule (S. Choi et al., 2015). These rules found applications in tasks like image classification using the MNIST dataset (M. V. Nair & Dudek, 2015), cancer dataset classification (S. Choi et al., 2015), and image compression and reconstruction. The stochastic Restricted Boltzmann Machine neural network, featuring contrastive divergence, could also benefit from accelerated implementations.

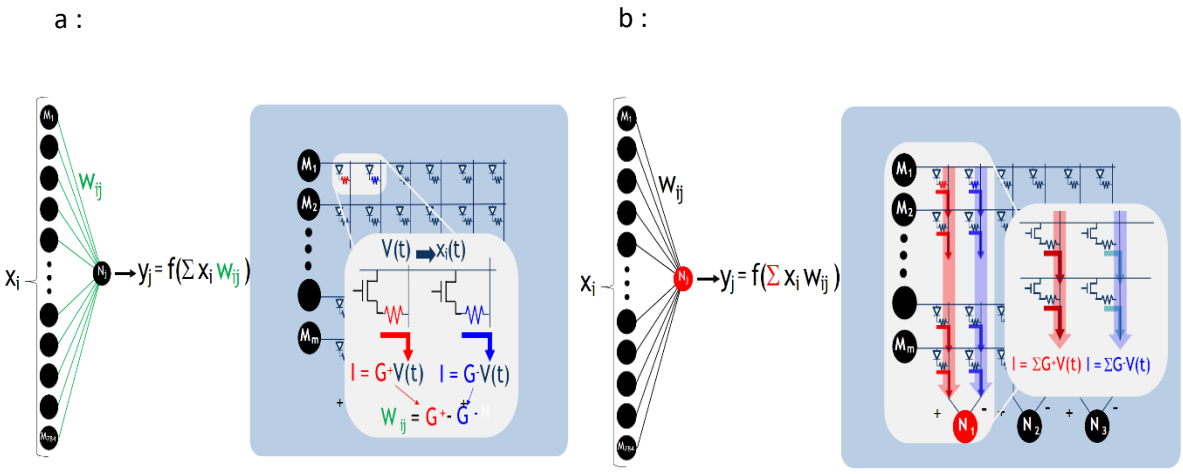


Figure 3. 5 : a): NVM Crossbar Array Configuration it illustrates the configuration of a non-volatile memory (NVM) crossbar array for mapping a neural network. Each intersecting point of horizontal and vertical lines represents a junction where NVM devices are situated. These devices serve as synaptic connections in the neural network. The Ohm's Law principle, represented by the product x_i w_{ij}, is applied here to determine the conductance values of these connections. (b): Signal Flow in NVM Crossbar Array In this figure, we visualize the signal flow within the NVM crossbar array designed for neural network mapping. The sum of signals from various sources is performed in this array using Kirchhoff's Law, symbolized by the sum operation. It showcases how signals are aggregated and distributed across the NVM crossbar, crucial for the proper functioning of the neural network. The multiply-accumulate (MAC) operation, which is integral to both the forward-inference and backpropagation processes of DNNs, can be realized by performing vector-matrix multiplication using large arrays of Non-Volatile Memory (NVM) devices. Even when NVM devices exhibit asymmetric conductance responses, where, for instance, one transition (e.g., RESET) is more abrupt than the other (e.g., SET), these devices can still be utilized for network training. This is feasible by assigning two conductance values to represent signed synaptic weights, enabling effective network training despite the asymmetry. (Solomon, 2019; Yusiong, 2012) (Tsai et al., 2019).

However, issues such as non-linear conductance response, limited dynamic conductance range, and variability required careful consideration. Research indicated that a certain level of non-linearity could be tolerated, as long as the range of conductance where the non-linear response occurs is a small fraction (around 10%) of the overall conductance range (Sidler et al., 2016). The impact of non-linearity on recognition accuracy was studied in the context of a sparse-coding algorithm as well. In terms of dynamic range, a proposal by Burr et al. recommended using around 20 to 50 programming steps between minimum and maximum conductance values, with each conductance pair representing a single weight. This approach echoed findings from Yu et al. who advocated for 6-bit resolution (Yu et al., 2015).

Furthermore, researchers delved into assessing the influence of parameter variations and device reliability on DNN training (Chabi et al., 2014).Gamrat et al. demonstrated the utilization of 'spike-coding' for inference purposes, showcasing competitive performance on the MNIST dataset using pre-trained weights stored on ideal devices (Gamrat et al., 2015). In a different study, Garbin et al. proposed a strategy involving the parallel combination of 10-20 HfO2-based RRAM devices, which offered increased robustness against device variability (Garbin et al., 2015). They managed to train and perform forward inference on a compact one-layer deep neural network (DNN) using a 12×12 memristor crossbar, eliminating the need for a separate selection device (Prezioso et al., 2015). Through modeling that relied on the programming dynamics of oxide memristor devices (Bayat et al., 2015), they achieved impressive results in MNIST digit classification with high accuracy (Kataeva et al., 2015). Lastly, non-volatile memory (NVM)-based DNN implementations were pitted against GPUs in terms of power consumption and speed, revealing the potential for a remarkable 25X speedup and significantly reduced power usage in DNN training (Burr, Narayanan, et al., 2015).

3.3 PCM as a synapse

Phase Change Memory (PCM), (explained in details in the next chapter), relies on the significant difference in electrical resistivity between the amorphous (low-conductance) and crystalline (high-conductance) phases of materials known as phase change materials. In the context of Non-Volatile Memory (NVM) devices, the process of transitioning to the high-conductance state is referred to as 'SET,' while the transition to the low-conductance state is termed 'RESET.' PCM finds relevance in neuromorphic applications where maintaining a 'device history' is beneficial. However, only the SET process can be performed incrementally, wherein repetitive pulses gradually crystallize a high-resistance amorphous region within the device. On the other hand, the RESET process involves melting and rapid cooling, making it a more abrupt transition, especially within an array of devices that might not exhibit perfect homogeneity.

In the past, a dual-phase change memory (PCM) strategy was introduced (Bichler et al., 2012) to implement Spike-Timing-Dependent Plasticity (STDP), utilizing distinct devices for Long-Term Potentiation (LTP) and Long-Term Depression (LTD) mechanisms (refer to Figure 3.4). In this approach, when an input neuron generates a spike, it emits a read pulse and enters the 'LTP mode' for a duration denoted as t_{LTP}. During this period, if the post-synaptic neuron fires a spike, the LTP synapse receives a partial SET pulse; if not, the LTD synapse is programmed.

Suri et al. made enhancements to the synaptic performance of conventional $Ge_2Sb_2Te_5$ (GST)-based PCM devices by introducing a thin layer of HfO2 (M. Suri, Bichler, Hubert, et al., 2012). The improved dynamic range was attributed to the impact of this interface layer on crystallization kinetics, influenced by activation energies associated with growth and nucleation sites. Similar to the subsequent application of PCM arrays for vector-matrix computations in the backpropagation algorithm (Burr, Shelby, et al., 2015), the two-PCM approach necessitates a complex refresh protocol. This protocol involves disabling inputs, reading effective weights, and performing RESET operations on synapses that have fully transitioned to the SET state, ensuring weight retention with reduced conductance values.

Suri et al. employed a behavioral model (derived from actual device data) of GST and GeTe PCM devices to conduct an event-based simulation using Xnet. The purpose was to extract features from a dynamic vision sensor and count cars in six highway traffic lanes (M. Suri, Bichler, et al., 2013). In subsequent work, they developed a model compatible with circuits, incorporating the electrical and thermal characteristics of both top and bottom electrodes alongside phase change material parameters (M. Suri, Bichler, Querlioz, et al., 2012). The authors noted that achieving maximum conductance required fewer pulses if either growth or nucleation rates were increased. Additionally, they found that growth and nucleation rates significantly influenced the shape of the amorphous plug after RESET pulses, though not its size. Furthermore, during partial-SET processes, the conductance was more sensitive to the nucleation rate than the growth rate. Due to its dominance in growth, GeTe saturated in conductance more rapidly than the nucleation-dominated GST, implying that GST could offer a greater range of intermediate conductance states compared to GeTe.

Furthermore, a single PCM cell per synapse was employed to implement both symmetric and asymmetric Spike-Timing-Dependent Plasticity (STDP). This was achieved through the application of RESET pulses with varying amplitudes and staircase down pulses of different amplitudes for partial SET operations. By employing short pulse timings, the overall energy consumption remained low despite the use of high programming currents. Simulation results demonstrated the feasibility of associative and sequential learning. The authors subsequently highlighted that by adjusting the energy of the spikes, the total energy expenditure in neuromorphic implementations could be further reduced (Kuzum et al., 2012). In another study, the same researchers utilized molecular dynamics to model the physical changes occurring within phase change materials during STDP potentiation and depression. This was achieved through a stepwise enhancement in material order in response to heat pulses (as opposed to electrical pulses) of varying heights and durations (Skelton et al., 2015).

Eryilmaz et al. conducted an experimental demonstration of array-level learning using a 10×10 array of transistor-selected Phase Change Memory (PCM) cells. They showcased Hebbian Spike-Timing-Dependent Plasticity (STDP) learning with the ability to learn simple patterns (Eryilmaz et al., 2013). The study revealed that higher initial resistance variation required longer training durations. Ambrogio et al. utilized measurements from a few transistor-selected PCM cells (at the 45 nm node) to simulate larger networks (Ambrogio et al., 2016). By constructing a two-layer network with 28×28 pre-neurons and one post-neuron, they achieved 33% recognition probability for MNIST digit recognition, with a corresponding error rate of 6%. With a three-layer network comprising 256 neurons, recognition probability reached an impressive 95.5% (with an error rate of 0.35%). The authors also discussed the network's capability to both forget previous information and learn new information, whether in parallel or in a sequential manner.

Li and Zhong et al. examined four variants of STDP updates, encompassing asymmetric Hebbian and anti-Hebbian updates with potentiation, as well as symmetric updates involving depression and potentiation. They applied pulses within different time windows to implement these forms, studying both physical devices and simulations (Zhong et al., 2015). Jackson et al. introduced two STDP schemes. One involved generating STDP-encoded neuronal firing delays within the electronic pulses reaching the synapse, while the other tracked the delay using a basic RC circuit within each neuron (Jackson et al., 2013). The latter approach demonstrated feasibility using programming energies under

5 pJ in phase change devices with 10 nm pores (19 nm actual dimensions). The authors subsequently simulated 100 leaky integrate-and-fire neurons to successfully learn a simple task involving predicting the next item in a sequence of four stimuli.

3.4 PCM as a neuron

In biological neurons, a delicate lipid-bilayer membrane acts as a barrier, separating the internal electrical charge of the cell from the external environment. This membrane, working alongside various electrochemical processes, maintains an equilibrium membrane potential. When excitatory and inhibitory postsynaptic potentials are received via the neuron's dendrites, this equilibrium potential can change. Upon reaching a certain level of excitation, an action potential, often termed "neuronal firing" (refer to Figure 3.5). Replicating these intricate neuronal behaviors, which encompass the stable potential equilibrium, transient dynamics, and the neurotransmission process, is considered pivotal for creating biologically feasible neuromorphic computing systems.

While the Hodgkin-Huxley model and other threshold-based neuronal models capture the complex neuronal dynamics, they often require simplification for practical hardware implementation (Camuñas-Mesa et al., 2019). In this simplification, the integration of postsynaptic potentials (related to the neuronal soma) and subsequent firing events (associated with the axon) remain the two most crucial dynamic elements. A growing number of research aims to leverage the properties of non-volatile memory (NVM) devices to emulate these intricate neuronal dynamics (as illustrated in Figure 3.6). The ultimate goals encompass achieving substantial efficiency in terms of area and power consumption, along with ensuring smooth integration with densely packed synaptic arrays.

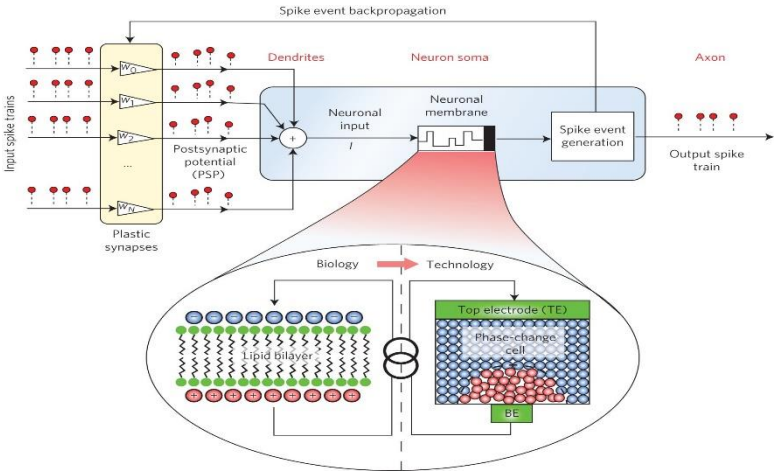


Figure 3. 6: A diagram of an artificial neuron includes three main components: the input section known as dendrites, the central body referred to as the soma which encompasses the neuronal membrane and the mechanism for generating spike events, and finally the output part called the axon. The dendrites are linked to adaptable synapses, connecting the neuron with other neurons in a network. The crucial computational aspect lies in the neuronal membrane, which holds the membrane potential using a nanoscale phase-change device. Thanks to their natural nanosecond-scale dynamics, dimensions at the nanometer level, and inherent randomness, these devices allow the simulation of extensive and densely packed groups of neurons. This simulation mimics biological processes, allowing for the representation and processing of signals in a way inspired by nature.(Tuma et al., 2016)

When employing a Nonvolatile memory as a substitute for a neuron, it is not a strict requirement for the NVM to exhibit a continuous range of conductance states. Instead, the crucial aspect is its ability to exhibit an accumulating behavior (explained in details in chapter 5), resulting in a "firing" event after a specific number of input pulses have been received. These pulses might trigger changes in an internal state that might not be immediately evident in the external conductance, but become relevant once the firing threshold is reached and the neuron undergoes a "firing" event.

Ovshinksy and Wright were among the pioneers who initially proposed utilizing phase-change memory (PCM) devices for creating artificial neurons (Ovhsinsky, 2006). In other work, Tuma and colleagues demonstrated the integration of post-synaptic inputs using PCM-based neurons (Tuma et al., 2016). They experimentally showcased how the evolution of the neuronal membrane potential could be encoded using the phase configuration within the PCM device. This innovation allowed for the detection of temporal correlations within extensive event-based data streams.

3.5 Applications

This ability to mimic the human brains functionality makes the PCM well suited the neuromorphic realization and neuromorphic devices have demonstrated significant effectiveness in their ability to receive and process information from their surroundings (Tuchman et al., 2020).

Furthermore, there is potential for the integration of neuromorphic computer technology into prostheses. This technology's notable advantage lies in its efficient acceptance and processing of external signals. For individuals with prosthetic limbs, opting for neuromorphic devices over conventional ones could lead to a more natural and fluid experience (Tuchman et al., 2020).

Other domain, which is imaging much like the human eye, neuromorphic vision sensors operate by creating images, but they do so in a unique way as event-based imaging devices (Sherif & Ahmed, 2022). Instead of capturing images in a traditional frame-by-frame manner, they respond to changes in light intensity, which are external signals (Khaled Ahmed et al., 2023). This event-based approach enables them to operate at a faster pace independently of conventional frame rates.

In a neuromorphic sensor, each pixel functions independently and almost instantly communicates changes to its neighboring pixels. This decentralized and rapid processing at the pixel level contributes to highly efficient data utilization. Similar to traditional vision sensors, neuromorphic sensors do not suffer from issues like motion blur or delayed responses to changes in the environment.

Given these qualities, integrating neuromorphic vision sensors into virtual and augmented reality systems could be highly advantageous. Their ability to capture and process visual information in a way that mimics the human eye's response to external signals could lead to more immersive and responsive experiences in these technologies.

Neuromorphic computing holds promise for a wide range of applications, including large-scale initiatives and product customization. One notable application is its potential to handle vast amounts of data generated by environmental sensors more efficiently. These sensors can monitor various environmental parameters such as water content, temperature, radiation levels, and other

characteristics. The neuromorphic computing framework can analyze this data, identify patterns, and facilitate the extraction of valuable insights. This could be particularly beneficial for sectors that rely on environmental monitoring, such as agriculture, climate science, and industrial processes.

Moreover, neuromorphic devices, due to the unique properties of the materials used in their construction, have the potential to facilitate product customization. These materials can be transformed into controllable fluids, making them adaptable for various purposes. They can then be processed through additive manufacturing in liquid form, allowing for the creation of devices that can be customized to meet the specific requirements of individual users or applications. This capability could have significant implications for industries like healthcare, where personalized medical devices and treatments are becoming increasingly important, and for other sectors seeking tailored solutions.

Finally neuromorphic computing is well-suited for edge computing due to its low energy consumption (Greengard, 2020). The edge refers to the boundary of a network where devices can connect to a cloud platform. This property makes neuromorphic computing particularly relevant in scenarios where fast and energy-efficient processing is essential, such as in autonomous vehicles like driverless cars. Arnaud et al in 2018 introduced a 28nm Fully Depleted Silicon-on-Insulator (FDSOI is a semiconductor technology that uses a very thin layer of silicon, separated from the underlying substrate by an insulating layer) e-NVM (embedded Non-Volatile Memory) solution designed specifically for automotive micro-controller applications, utilizing PCM technology. This advancement marks a significant milestone in semiconductor technology tailored to the automotive industry's needs. (Arnaud et al., 2018)

In the context of driverless cars, neuromorphic computing can enable vehicles to react more quickly to their surroundings, even when they are not connected to a reliable internet source. This capability could significantly enhance the safety and environmental suitability of driverless cars, as they would have the ability to make critical decisions independently, without relying solely on cloud-based processing (Schuman et al., 2017).

Additionally, the superior sensory capabilities of neuromorphic computing can enhance various "smart technologies" (Sharifshazileh et al., 2021), extending its applicability to a broader range of situations, similar to its role in driverless cars. This modification can improve the effectiveness of smart devices in various contexts by providing them with faster and more efficient processing capabilities.

Furthermore, neuromorphic computing has the potential to expand communication channels, which can lead to innovative applications and improved connectivity in various industries and IoT (Internet of Things) scenarios. This could result in more efficient and responsive communication systems in areas like telecommunications, industrial automation, and beyond (Schuman et al., 2017).

Conclusion

In this chapter, we have explored the fascinating realm of neuromorphic computing and its integration with non-volatile memory (NVM) technologies. We have covered essential concepts and applications,

highlighting the immense potential within this field. We began by delving into Spike-timing-dependent-plasticity (STDP), a critical mechanism that mimics the learning and adaptation processes in biological neural networks. This mechanism is instrumental in emulating the brain's ability to modify neural connections based on their firing patterns. Our attention then shifted to the fundamental operation of neuromorphic computing: vector-matrix multiplication. This process serves as the backbone for numerous neural network computations, enabling efficient tasks like pattern recognition and data analysis. We discussed the versatile use of Phase-Change Memory (PCM) as both a synapse and a neuron, making it a compelling candidate for neuromorphic applications. PCM offers the potential to emulate both synaptic and neuronal functions within a single technology. Finally, we explored a wide range of applications for neuromorphic computing, spanning cognitive computing, artificial intelligence, robotics, and sensory processing. These applications emphasize the transformative potential of neuromorphic computing in various domains.

In summary, this chapter has provided insights into the synergy between non-volatile memory technologies and neuromorphic computing. We have examined crucial mechanisms, core operations, and the adaptable use of PCM in synaptic and neuronal roles. With this foundational knowledge, we are well-prepared to explore in detail the PCM devices, their characteristics, and their specific applications within the broader landscape of neuromorphic computing. This exploration holds the potential to reshape the fields of artificial intelligence, cognitive computing, and more, opening up new horizons for innovative and efficient computing solutions.

Résumé

- Explored Spike-timing-dependent-plasticity (STDP) for emulating learning in neural networks.
- Investigated vector-matrix multiplication, a core operation in neuromorphic computing.
- Explored the versatile use of Phase-Change Memory (PCM) as both synapses and neurons.
- Discussed a wide range of applications, from cognitive computing to robotics.
- Prepared to explore PCM devices in greater detail, offering transformative potential for artificial intelligence and cognitive computing.

Chapter 4

The Phase-Change Memory Technology

"In the dance of electrons and the rhythm of rapid transformations, phase-change memory orchestrates a symphony of data, ushering in a new era where storage is not just a memory but an experience."

Dr. Thomas Mikolajick, Professor of Nanoelectronics, Technical University of Dresden

4. Beyond von Neumann architecture

The von Neumann architecture, developed by John von Neumann, forms the foundation of modern computer design. It comprises a central processing unit (CPU), memory, input/output devices, and a control unit. In this architecture, instructions and data are stored in separate memory units, and the CPU sequentially fetches and executes instructions (illustrated in Figure 4.1a). However, to go beyond the limitations of the von Neumann architecture, researchers are exploring innovative approaches such as in-memory computing and neuromorphic computing utilizing resistive memories (illustrated in Figure 4.1b).

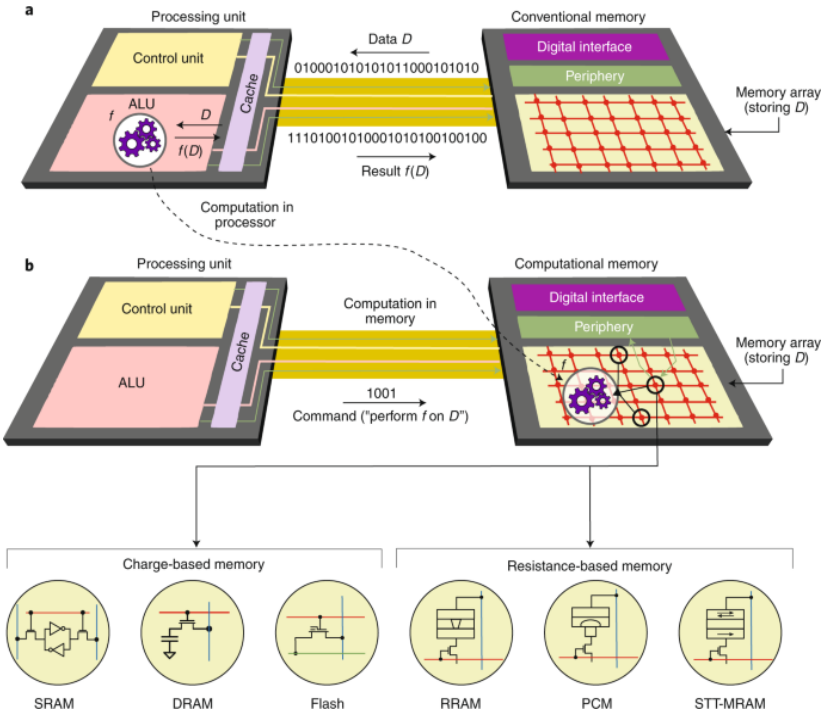


Figure 4. 1 : a. In a conventional computing setup, when a operation f is executed on data D, the data D needs to be transferred to a processing unit. This leads to significant delays and energy consumption. b. However, in the context of in-memory computing, the operation f(D) is carried out within a computational memory unit, leveraging the inherent physical properties of memory devices. This eliminates the necessity to move the data D to the processing unit. The computational tasks take place within the boundaries of the memory array and its associated circuitry, all without revealing the specific content of individual memory elements. Both memory technologies that rely on charges, like SRAM, DRAM, and flash memory, as well as those based on resistance, such as RRAM, PCM, and STT-MRAM, have the potential to function as components of such computational memory units.(Sebastian et al., 2020)

In-memory computing leverages the properties of non-volatile memory (NVM) technologies, such as resistive random-access memory (RRAM) or phase-change memory (PCM), to perform computations

directly on the data. By integrating computing and storage, in-memory computing eliminates the need for data movement between storage and processing units, significantly reducing latency and improving energy efficiency. This approach is particularly beneficial for data-intensive tasks, such as large-scale analytics and machine learning, where the proximity of computation to data leads to faster processing and improved performance.

Neuromorphic computing, inspired by the architecture and principles of the human brain, seeks to develop computational systems that can perform tasks efficiently and emulate cognitive functions. Resistive memories, with their ability to store and process data simultaneously, offer potential advantages for neuromorphic computing. These memories can mimic the synaptic connections in the brain, enabling the development of artificial neural networks. By exploiting the analog and parallel computing capabilities of resistive memories, neuromorphic computing can achieve efficient and intelligent processing for applications such as pattern recognition, deep learning, and cognitive computing.

Within the context of advancing computing paradigms, such as in-memory computing and neuromorphic computing, leveraging resistive memories offers opportunities to transcend the conventional von Neumann architecture. In-memory computing represents a paradigm shift characterized by the seamless integration of computational and storage elements, culminating in accelerated and energy-efficient processing. On the other hand, neuromorphic computing exploits the potential of resistive memories to emulate brain-inspired functionality, potentially enabling intricate cognitive processes within computing systems. These groundbreaking innovations bear the potential to reshape diverse domains, spanning from data analytics to artificial intelligence, and are poised to herald a new era of more potent and efficient computational systems.

In this chapter, we shall commence with the electrical characterization of PCM (Phase-Change Memory) devices, where we shall present the preliminary findings as the first outcomes of our study. This experimental endeavor aims to provide crucial insights into the performance and behavior of PCM devices.

4.1 Brief history of PCM technology

Phase-change memory (PCM) capitalizes on phase-change materials, which can be switched between amorphous and crystalline phases with varying electrical resistivity. The amorphous phase typically demonstrates high electrical resistivity, while the crystalline phase showcases significantly lower resistivity, sometimes differing by three to four orders of magnitude. This substantial difference in resistance is harnessed for information storage within PCM. (In PCM, the state with high resistance signifies a logical "0," while the state with low resistance signifies a logical "1"). As such, a PCM device essentially comprises a layer of phase-change material positioned between two metal electrodes.

During the mid-1950s, researchers Kolomiets and Goryunova at the Ioffe Physical-Technical Institute made a significant discovery regarding the semiconducting properties of chalcogenide-based glasses (Bogoslovskiy & Tsendin, 2012). Subsequently, in 1968, Stanford R. Ovshinsky from Energy Conversion Devices observed a rapid and reversible switching effect within the Si12Te48As30Ge10

(STAG) composition (Ovshinsky, 1968). Notably, he also witnessed a memory effect for the first time by making slight alterations to the STAG material composition. This effect led to the preservation of the low-resistance state achieved through switching, even in the absence of a voltage source (Ovshinsky, 1968). Ovshinsky recognized the potential commercial utility of these materials as a key element in electronic switches and memory cells (Ovshinsky, 1970). In 1970, an array of 256-bit amorphous semiconductor memory cells was developed by R. G. Neale, D. L. Nelson, and Gordon E. Moore (Neale & Aseltine, 1973).

From the 1970s through the early 2000s, further efforts aimed at creating dependable PCM cells faced substantial challenges. These difficulties primarily stemmed from issues such as device degradation and operational instability. Consequently, the enthusiasm for developing electrical memory cells utilizing phase-change materials gradually waned.

Nevertheless, phase-change materials found significant utility starting in the 1990s within the realm of optical memory devices. Remarkably, these materials continue to serve as the primary medium for storing information on CDs, DVDs, and Blu-Ray disks (Wuttig & Yamada, 2007). In the context of optical memory, a laser source is used to heat the phase-change material. The key to information storage lies in the contrast in optical reflectivity between the amorphous and crystalline phases of the material.

The success and research findings related to optical storage utilizing phase-change materials prompted a resurgence of interest in PCM during the early 2000s. Leading companies like Intel, Samsung, STMicroelectronics, and SKHynix licensed the technology from Ovonyx, who initially possessed the exclusive PCM technology invented by Ovshinsky. Ovonyx was later acquired by Micron in 2012. Subsequently, these companies embarked on creating their own PCM chips of varying sizes, some reaching up to 8 Gb (Y. Choi et al., 2012).

In 2008, Numonyx, a memory company formed by Intel and STMicroelectronics, introduced the first PCM product consisting of 128-Mbit memories produced through a 90-nm process. Micron later acquired Numonyx in 2010. In 2012, Micron unveiled a 45-nm 1-Gbit PCM chip designed for mobile phones, particularly for Nokia, although it was subsequently withdrawn in 2014.

In 2015, a significant milestone in the realm of Phase-Change Memory (PCM) technology was attained, as Intel and Micron jointly unveiled the 3D Xpoint memory architecture. This groundbreaking development was predicated upon the utilization of a phase-change alloy as the primary storage component within the memory element. The year 2018 saw the commercialization of this technological innovation under the Intel Optane brand. Intel Optane introduced a range of nonvolatile memory modules characterized by their diminutive capacities, encompassing 16 GB to 64 GB, designed to ameliorate prevailing storage systems' performance through low-latency attributes. Notably, it is essential to append that subsequent to this progression, Intel Optane faced challenges in the market, eventually leading to its withdrawal from commercial availability.

4.2 How the PCM cell is build

The characteristics of phase-change memory (PCM) depend on the design of memory cells. An important aspect is the programming current. This study exclusively focuses on the "Wall structure" (Burr et al., 2016). A visual representation of this design is depicted in Figure 4.2.b. Each memory cell consists of a core made of phase-change material surrounded by two electrodes. The Wall structure relies on an upper electrode that comes into contact with the entire phase-change material, and a slender lower electrode referred to as the heater with a dimension of 5nm. This heater electrode is crafted from a resistive substance, TiSiN, which possesses higher resistivity than the phase-change material. Due to its increased resistance, the majority of the electrical energy is absorbed by the lower electrode. Consequently, the dynamic area that undergoes a phase transition during writing procedures is the dome-shaped section in contact with this electrode (Gallo & Sebastian, 2020). The primary advantage of this arrangement mainly stems from its capability to diminish write currents: the thin lower electrode encounters a dense current flow and consequently becomes significantly heated. This substantial heating facilitates the thermal triggering of state changes within the cell. The magnitude of these current intensities typically ranges in the hundreds of microamperes (Navarro, Thesis 2013).

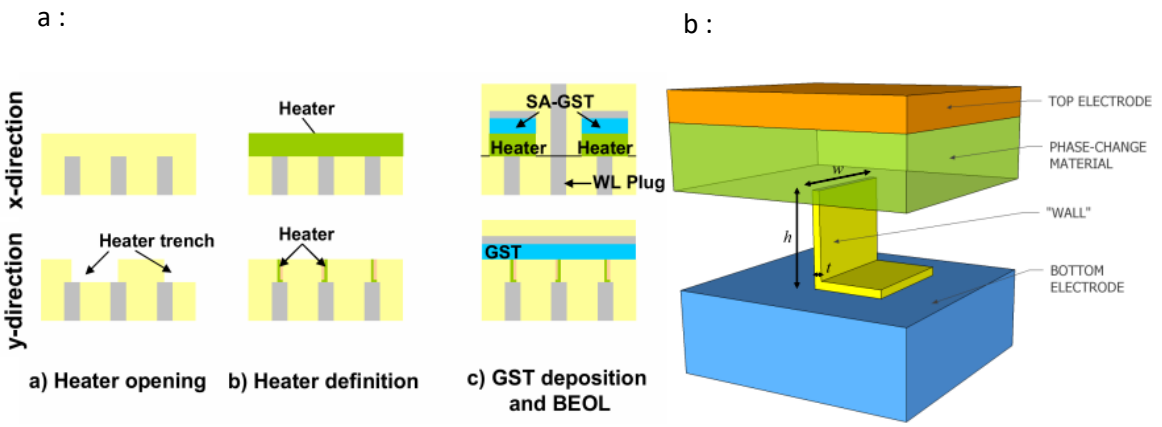


Figure 4. 2 a) PCM “Wall” structure process flow as described in (Servalli, 2009). b) The basic scheme of the “Wall” structure, evidences the structure of the plug (Navarro, Thesis 2013).

The Wall structure is achieved by a sequence of technological processes, some of which involve raising the temperature. Starting from a flat substrate, different layers of metals and oxides are applied and etched. Notably, material heating takes place during the deposition phase (Servalli, 2009). Temperatures reaching around 400°C are attained within roughly a minute.

The integration of the PCM cell with other components, like transistors, is undertaken. As these components require a final annealing process, the PCM cell undergoes annealing at approximately

400°C for several tens of minutes. The term "thermal budget" encompasses the information about the time and temperature associated with a thermal treatment.

Particular attention is needed during manufacturing stages where the phase-change material experiences temperatures surpassing its glass transition point. Throughout the previously mentioned temperature increases, the initially deposited amorphous GeSbTe enriched with germanium undergoes crystallization, potentially resulting in significant microstructure changes. Consequently, it has been observed that a strong forming current is essential before the cell can be used. The current is designed to perform a deep reset of the cell, aiming to mitigate microstructural alterations induced by the cell fabrication procedures. Nonetheless, it is worth noting that this process has been observed to induce degradation. The application of a substantial current, reaching levels of up to 500 µA, exerts notable effects on the cell, particularly with regard to the integrity of the heater electrode (Thesis C. Pigot, 2019). Therefore, for accurate forming current adjustment, understanding maturation linked to a specific thermal budget holds paramount importance.

It is evident that investigating microstructure evolution holds relevance not just during memory writing procedures, but also during uniform temperature annealing linked to various phases of the manufacturing process.

4.3 The PCM cell

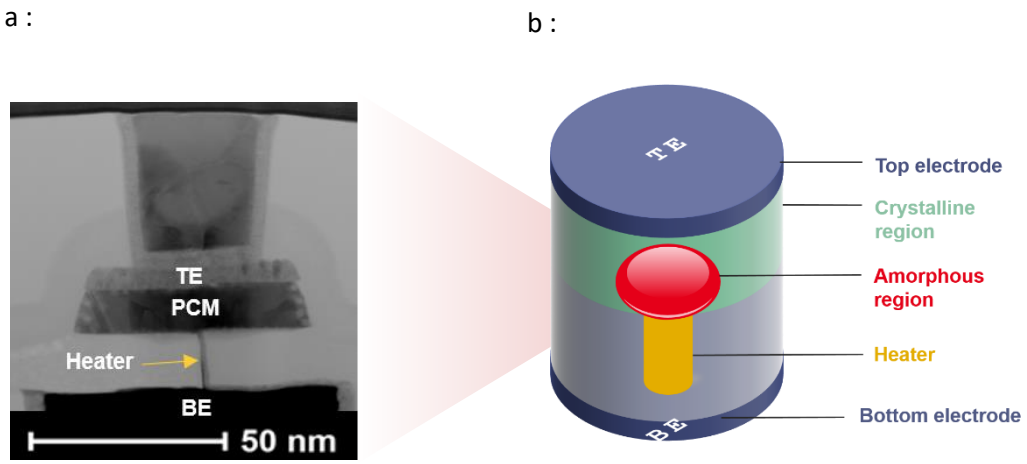


Figure 4. 3 : a) TEM image of PCM cell. b) Schematic of the PCM

A fundamental requirement for a memory device is its capability to store and retrieve data. In PCM, (showing in Figure 4.3a, a TEM image) data is recorded through the transformation of a phase-change material within the memory device, switching between crystalline (ordered) and amorphous (disordered) phases, and vice versa (showing in Figure 4.3b). This change comes with a substantial alteration in electrical and optical properties. The amorphous phase exhibits high electrical resistivity and low optical reflectivity, while the crystalline phase has low electrical resistivity and high optical

reflectivity. This disparity in optical characteristics has been effectively utilized in optical data storage devices like DVDs and Blu-Ray discs.

For electrical data storage using PCM, however, it is the difference in electrical resistivity between these two phases that serves as the basis for information storage. Consequently, a WRITE operation in PCM involves the transition between the amorphous and crystalline states by applying an electrical pulse. On the other hand, a READ operation typically entails measuring the electrical resistance of the PCM device, thereby determining whether it resides in the amorphous state (high resistance, representing logical "0") or the crystalline state (low resistance, representing logical "1").

Following the discovery of the memory effect, it quickly became evident that this phenomenon is linked to a transition in materials from an amorphous phase to a crystalline phase. The amorphous phase represents an energetically unstable glass, yet its crystallization process at room temperature occurs over an extended period. However, by subjecting the amorphous material to elevated temperatures below the melting point, it rapidly undergoes crystallization. To revert the material back to its amorphous state, it must be heated beyond its melting temperature and then rapidly cooled down. This rapid cooling solidifies the atomic arrangement into a disordered configuration.

In PCM, the necessary heat is generated by passing an electric current through the phase-change material, a phenomenon known as the Joule heating effect (is the process in which the passage of an electric current through a conductor produces heat due to the resistance of the material) (Xuan, 2008). The electrical pulse utilized to transition the device into the high-resistance amorphous state is termed the RESET pulse, while the pulse employed to switch the device back to the low-resistance crystalline state is termed the SET pulse (Figure 4.4).

The key advantage of PCM is its non-volatile nature. Non-volatile memory retains data even when power is disconnected. Unlike volatile memory like DRAM, PCM does not require constant power to retain its stored information. This makes PCM ideal for applications where power loss or disruptions are a concern. However, it is important to note that PCM is not without its challenges. One significant issue is related to data retention endurance and resistance drift. Over time and with repeated write and erase cycles, PCM cells can experience deterioration in their ability to reliably store and retain data. This phenomenon is often referred to as endurance degradation. Additionally, resistance drift can occur, leading to changes in the electrical characteristics of the PCM cells, further impacting their performance and reliability.

PCM also offers fast read and write speeds compared to traditional storage technologies. It can perform read and write operations at a similar speed to DRAM, which is significantly faster than flash memory. This makes PCM suitable for applications that require high-speed data access.

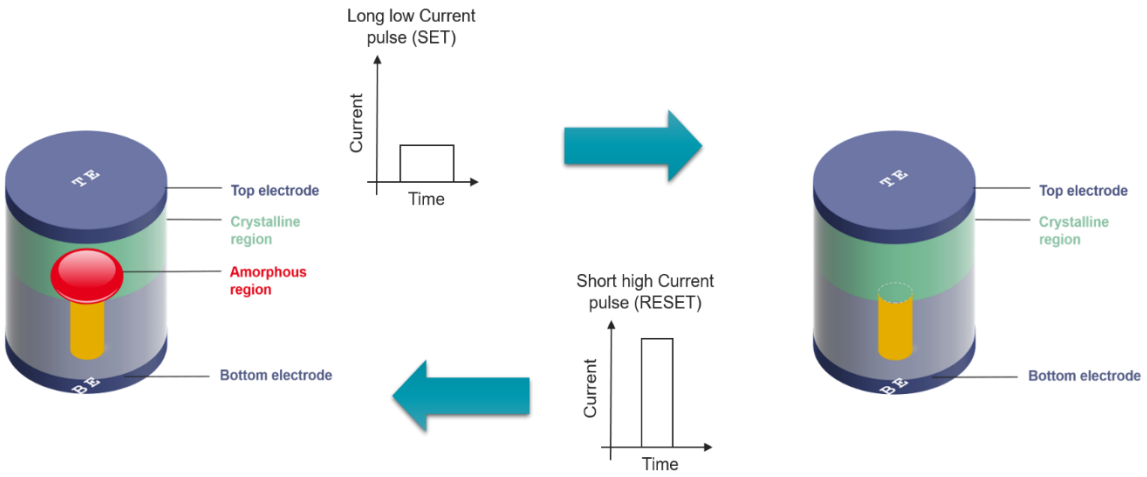


Figure 4. 4 The operational principle of PCM involves a mushroom-type PCM device, which comprises a layer of phase-change material positioned between a top electrode (TE) and a narrower bottom electrode (BE). The process involves two distinct electrical pulses. Starting from amorphous state, a lengthy low current pulse, known as the SET pulse, is applied. This SET pulse is designed to transition the PCM device into the low-resistance crystalline state. Conversely, a short high current pulse, referred to as the RESET pulse, is then administered. This RESET pulse serves the purpose of converting the PCM device into the high-resistance amorphous state.

Another benefit of PCM is its high endurance. PCM cells can endure a large number of read and write cycles without degrading their performance, making them more durable than flash memory. This endurance is due to the ability of PCM cells to switch between the crystalline and amorphous states reliably.

PCM technology has matured significantly, progressing beyond its early stages of development, and it holds great promise for future memory systems. It has the potential to replace or complement existing memory technologies, offering higher capacity, faster performance, and improved reliability. Researchers and industry experts continue to explore and optimize PCM to unlock its full potential and make it commercially viable for a wide range of applications.

Phase-change memory (PCM) has gained significant interest in the field of neuromorphic computing, which aims to develop computer systems that mimic the structure and functionality of the human brain. PCM's unique characteristics make it well-suited for neuromorphic applications. Here's an overview of PCM's relevance and potential in neuromorphic computing:

- **Analog behavior:** PCM cells exhibit analog behavior, meaning they can exist in multiple resistance states between the crystalline and amorphous phases. This property allows PCM to represent and process information in a continuous and graded manner, similar to the synapses in biological neural networks. By adjusting the resistance levels, PCM can simulate the strengths and weights of synaptic connections, enabling more accurate and efficient neural network simulations. (Antolini et al., 2023; W. Kim et al., 2019)

- **Energy efficiency:** Neuromorphic systems aim to replicate the brain's energy efficiency. PCM offers low power consumption during both read and write operations, making it an attractive option for energy-efficient neuromorphic architectures. Compared to traditional digital memory technologies, PCM's analog nature allows for more efficient computation, reducing the overall energy consumption of neuromorphic systems.(Bichler et al., 2012; Song & Das, 2020)

- **Parallel processing:** PCM enables parallel processing, a fundamental aspect of neuromorphic computing. The ability to perform simultaneous read and write operations in PCM arrays allows for efficient and parallelized synaptic weight updates, facilitating the parallel processing of neural network operations. This feature can lead to significant improvements in computational speed and efficiency.(Zhou et al., 2016)

- **Adaptability and plasticity:** The synaptic connections in biological neural networks exhibit adaptability and plasticity, allowing them to learn, adapt, and store information. PCM's ability to modify its resistance levels enables it to emulate synaptic plasticity, making it suitable for implementing learning and memory functions in neuromorphic systems. The re-configurability and non-volatility of PCM also allow for the retention of learned information even in the absence of power. (Nandakumar & Rajendran, 2017; Sung et al., 2022)

- **Integration with conventional circuits:** PCM can be integrated with conventional CMOS circuits, enabling seamless integration of memory and computation in neuromorphic systems. This integration allows for closer proximity between the memory and processing units, minimizing data transfer and latency issues associated with separate memory and processing components.(Kumar & Suri, 2023)

- **Compact and scalable:** PCM has the potential to offer high-density memory arrays, allowing for the construction of compact and scalable neuromorphic systems. The small cell size of PCM enables a large number of synapses to be packed within a small area, facilitating the implementation of large-scale neural networks.(Qureshi et al., 2009; Raoux et al., 2008)

While PCM shows promise for neuromorphic applications, there are still challenges to overcome. Improving the endurance and reliability of PCM cells, reducing variability in resistance levels, and developing efficient programming algorithms for synaptic plasticity are among the ongoing research areas. However, PCM's unique properties make it a compelling option for enabling energy-efficient and high-performance neuromorphic computing systems that can emulate the computational capabilities of the human brain.

4.4 Operation principals of PCM

4.4.1 SET/RESET operation

The fundamental processes of crystallization and amorphization that underlie the WRITE operation of PCM are depicted in Figure 4.5. To amorphize the phase-change material within the PCM device during a RESET operation, a high-voltage or high-current pulse with sharp edges is applied.

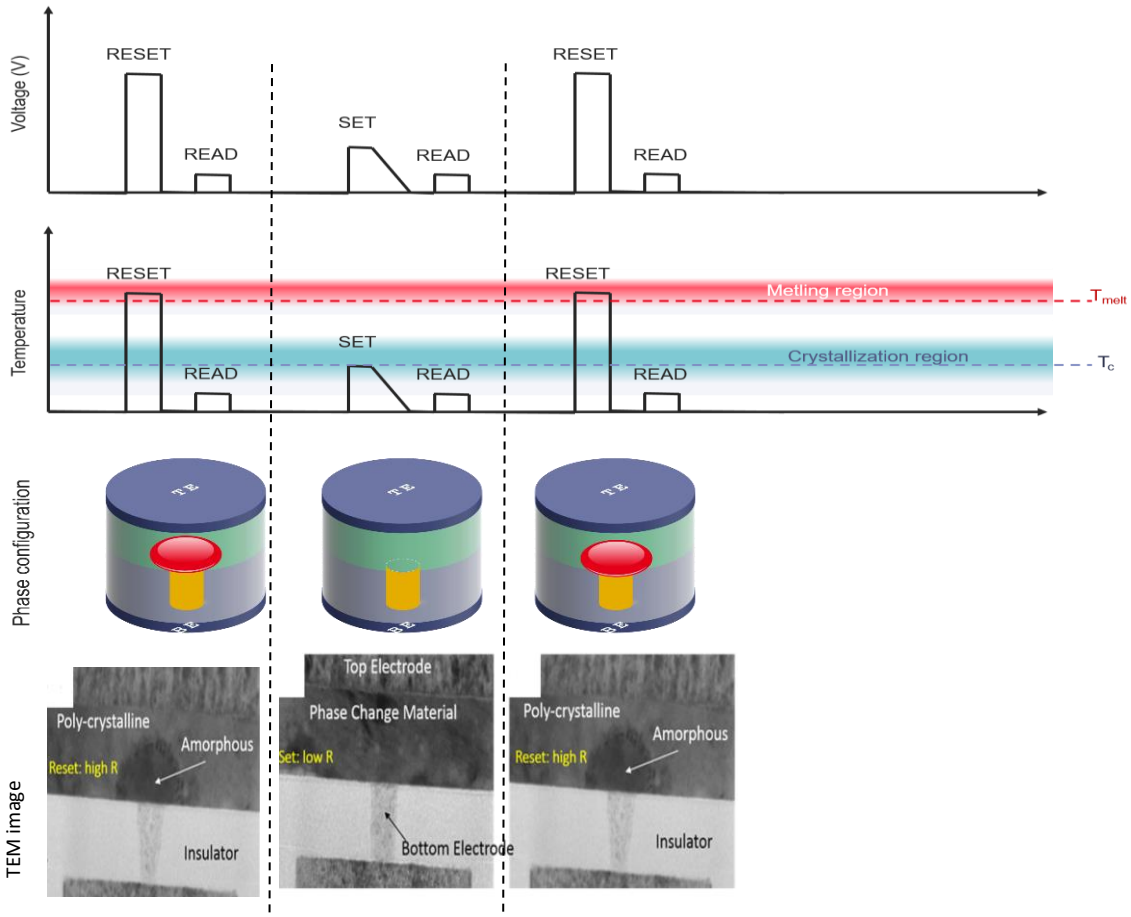


Figure 4. 5 The underlying principles of the WRITE operation in PCM are as follows: RESET Operation: A RESET operation transitions the PCM device into a high-resistance state by inducing amorphization in the phase-change material. This is achieved by heating the material above its melting temperature, known as Tmelt, followed by rapidly cooling it. As a result of this process, the atomic arrangement becomes disordered, leading to a high-resistance state in the device. SET Operation: Conversely, a SET operation shifts the PCM device to a low-resistance state by crystallizing a region that was previously amorphous. The extent of this amorphous region can be controlled by adjusting the amplitude of the pulse power. In summary, the WRITE operation in PCM involves RESET to achieve a high-resistance state through amorphization and SET to achieve a low-resistance state through controlled crystallization. The process is influenced by manipulating factors such as the pulse shape parameters, including rise time, width, and fall time.

The resulting power dissipation must be significant enough to induce Joule heating, raising the temperature within the PCM device above the phase-change material's melting temperature, denoted as T_{melt}. The ensuing melting disrupts any previously established periodic atomic arrangement.

Once the phase-change material is in its molten state, it must undergo rapid cooling (quenching) to effectively "freeze" the atomic structure into a disordered state. Should the rapid crystallization regime (as shown in Figure 4.5) be swiftly surpassed by speedy quenching, the atomic mobility at temperatures below this regime becomes extremely limited. This causes the atoms to remain immobilized and prevents them from reorganizing into their energetically preferred configuration during the cooling process. As a result, they become trapped in a no equilibrium or "glassy" amorphous state. This phenomenon is commonly referred to as the glass transition, leading to the creation of the amorphous state characterized by high resistance during a RESET operation.

4.4.2 The amorphous and the crystalline phases

The amorphization process can occur remarkably quickly, often taking just a few tens of picoseconds, thanks to the rapid melting kinetics of PCM (Sonoda et al., 2008). During this process, the phase-change material is typically heated to temperatures exceeding approximately 1000 K (Yu et al., 2014).

To transition from the amorphous state to the crystalline state (SET) in PCM, a voltage or current pulse is applied. This pulse serves to elevate the temperature within the PCM device to a range where rapid crystallization occurs. Additionally, the duration of the pulse must be sufficiently extended to facilitate the complete crystallization of any amorphous region that might have been previously formed.

As a result of this process, a crystalline state is generated, characterized by low resistance during a SET operation. The crystallization procedure generally takes longer than the amorphization process, spanning tens to hundreds of nanoseconds. Crystallization occurs at temperatures typically ranging from around 500°C to 600°C, positioned above (T_C) but still below the melting temperature (T_{melt}) of the material (W. Zhang et al., 2019).

The speed at which PCM crystallizes is influenced by the amount of initially amorphous material targeted for crystallization and the crystallization kinetics inherent to the phase-change material being employed (Jeyasingh et al., 2014). These kinetics are significantly reliant on temperature. Crystallization kinetics within PCM at elevated temperatures can be either nucleation-driven or growth-driven, and this topic has garnered extensive research interest (Orava et al., 2015). In nucleation (Karpov et al., 2007), a stochastic process unfolds where a crystalline nucleus eventually attains a critical size that renders it stable and permits growth instead of dissolution. The development of this nucleus to a critical size involves an incubation time (Orava & Greer, 2017). The specific critical size hinges on temperature and is influenced by factors such as the free-energy difference between amorphous and crystalline phases (diminishing the critical size with an increase) and the interfacial energy density between these phases (augmenting the critical size with an increase). Subsequent to reaching the critical size, crystal growth ensues as a deterministic process. The rate of crystal growth

is highly sensitive to temperature, contingent on factors like the free-energy disparity between amorphous and crystalline phases (boosting growth velocity with an increase) and viscosity (reducing growth velocity with an increase). Under conventional conditions, such as those in optical disks, it is been demonstrated that crystallization in certain materials, such as PCM (Ag,In)-doped Sb2Te (AIST),(Pries et al., 2021) is driven by growth (with slow nucleation), while in GST (Germanium-Antimony-Telluride) (Boniardi et al., 2014; Guo et al., 2019; Ielmini & Lacaita, 2011), it is driven by nucleation (with fast nucleation) (W. Zhang et al., 2019). Nonetheless, it has been postulated that in nanoscale PCM devices, nucleation's role might be diminished, and crystallization might be predominantly dictated by crystal growth (Sebastian et al., 2014). This is because, after a RESET operation, a substantial population of nuclei already exists within the melt-quenched amorphous phase (B.-S. Lee et al., 2014), and an interface between the amorphous and crystalline states is present. Consequently, considerable growth of the existing nuclei and at the amorphous-crystalline interface might dominate over additional nucleation, even in materials driven by nucleation.

4.5 Phase change mechanism and phase change transition

4.5.1 Switching process

For the crystallization and amorphization scheme described above to be practically applicable for electrical data storage using PCM, the capacity to rapidly and significantly elevate the temperature within the device is essential, regardless of its resistance state. In optical storage, this can be accomplished by exposing the phase-change material to a laser source with adequate power, regardless of the material's current state. However, in PCM, achieving rapid and substantial power dissipation relies on a key property: a highly nonlinear current/voltage (I-V) characteristic exhibited by the phase-change material. This characteristic, observed at a threshold voltage (Vth) of a few volts, enables the application of a relatively low-amplitude voltage pulse. Importantly, the amplitude of this pulse remains largely unaffected by the resistance state.

The typical I-V characteristics of the amorphous and crystalline states are depicted in Figure 4.6. While the crystalline state demonstrates fairly linear behavior at lower voltages, the amorphous state showcases a highly nonlinear relationship between the applied voltage and the resulting current.

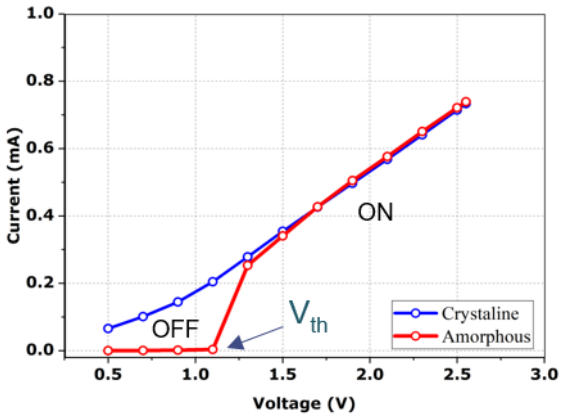


Figure 4. 6 The switching I-V characteristic of a PCM device, initially in the amorphous state. When the voltage surpasses the threshold value (Vth), a threshold switching event transpires, triggering a swift increase in current, leading to a voltage snapback. During memory switching (completing crystallization), the I-V characteristic of the amorphous ON state converges with that of the crystalline state. The blue arrow depicts the continuation of the I-V characteristic, originating from the crystalline state, when higher voltages are applied. In this case, the phase-change material experiences heating up to elevated temperatures, eventually culminating in its melting.

In the state known as the amorphous OFF state (or subthreshold regime) of a PCM device, the current demonstrates various behaviors with increasing applied voltage. These include ohmic, exponential, and super-exponential behavior. Once the applied voltage surpasses a certain threshold called the threshold switching voltage (Vth), the conductivity of the amorphous phase undergoes rapid augmentation due to a feedback-driven mechanism, resulting in a phenomenon known as negative-differential resistance, often leading to voltage snapback.

The state achieved upon threshold switching is typically referred to as the amorphous ON state, as the amorphous phase has not yet undergone crystallization. Once a sufficient amount of current passes through the PCM device in the amorphous ON state for an adequate duration, memory switching occurs, resulting in total crystallization. This causes the I-V characteristic of the amorphous ON state to merge with that of the crystalline state.

The underlying cause of the threshold switching mechanism in PCM has been a subject of extensive debate, remaining unresolved even though the phenomenon was initially observed over 50 years ago by Ovshinsky (Menzel et al., 2015; Ovshinsky, 1968). A plethora of models have been put forth to elucidate threshold switching in PCM (Bogoslovskiy & Tsendin, 2012), generally falling into two broad categories: thermal models, where the switching is connected to an electro thermal instability within the device, and purely electronic models (Eaton, 1964; Mott, 1971). It is important to note that consensus remains elusive. However, it is widely acknowledged that the nature of the mechanism is electronic.

4.5.1 Memory Switching

The concept of the threshold switching mechanism, as elucidated in the preceding section, pertains to a reversible shift between a state of low conductivity and one of high conductivity contingent on the

applied bias. Upon surpassing the threshold voltage of the chalcogenide material, current initiation within the device occurs. However, without sustaining the chalcogenide under a biased state, the low-conductivity condition spontaneously returns. It is important to note that the threshold switching phenomenon by itself is unsuitable for use in nonvolatile memory cells.

Nevertheless, within the class of materials that exhibit threshold switching, certain compounds, like GST, can sustain the high-conductive state without the necessity of an applied bias. For these materials, once a high current commences flowing, the chalcogenide substance elevates its temperature beyond the crystallization threshold, facilitating an eventual phase transition (referred to as memory switching or MS). Given that the performance of PCM cells is closely connected to the kinetics of the phase transition, comprehending the physics of MS becomes a pivotal aspect for advancing this technology.

It is worth underscoring that the transition associated with threshold switching (Krebs et al., 2009) is fundamentally electronic in nature and differs from the memory effect in chalcogenide materials characterized by the phase change transition (Le Gallo et al., 2016). This distinction has been empirically confirmed in PCM devices, where threshold switching can manifest devoid of a shift from the amorphous to crystalline phase. In such instances, achieving the memory phase change demands higher currents or lengthier durations.

4.5.2 Multilevel operation

A crucial characteristic of a phase change memory (PCM) device is its capacity for modifying the size of the disordered, amorphous region in an almost entirely continuous manner. This manipulation is achieved by applying specific electrical pulses due to the uneven distribution of temperature within the PCM device. In the mushroom-shaped PCM device illustrated in Figure 4.5, the point of greatest temperature, caused by Joule heating from an electrical pulse, typically occurs near the bottom electrode. Consequently, by administering a RESET pulse that expends more energy, a larger amorphous region is generated. This is because the temperature required for melting (T_{melt}) is attained farther from the bottom electrode. The expanded amorphous region leads to a heightened resistance within the PCM device. Exploiting this characteristic allows for the encoding of more than a single bit of information within a single phase change memory (PCM) device. This is achievable due to the ability to establish a continuum of resistance states, each capable of representing specific bit patterns (such as "11" and "10") depicted in Figure 4.7.

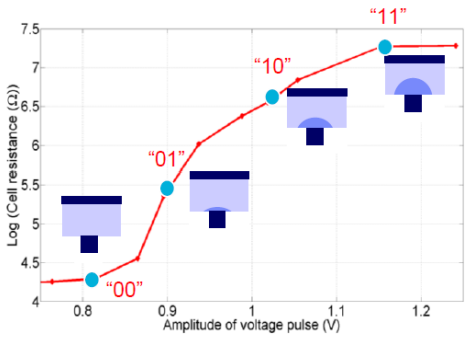


Figure 4. 7 : The relationship between the amplitude of the voltage pulse and the logarithm of cell resistance is evident in the study conducted by Sebastian et al. (Proc. E\PCOS, 2016). As the applied voltage increases, the Phase Change Memory (PCM) cell demonstrates a spectrum of resistance and conductance levels. It is crucial to note, however, that the increase in resistance levels cannot be extended limitlessly; there exists a threshold beyond which further increments are constrained.

Additionally, adjustments to the pulse width (for SET operations) or the duration of the trailing edge enable the programming of multiple resistance levels. The relationship between PCM resistance and the applied programming power is commonly referred to as the programming curve. A typical programming curve, obtained using a mushroom-type PCM device initially in the RESET state, is depicted in Figure 4.8.

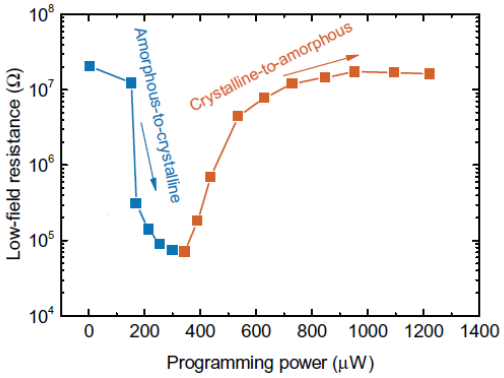


Figure 4. 8 Resistance under low-field conditions in relation to the applied programming power is depicted by the programming curve of a phase change memory (PCM) device that begins in the RESET state.(Le Gallo & Sebastian, 2020)

The programming curve displays distinct characteristics on both sides. The left side is involving the transition from amorphous to crystalline phase (crystalline-to-amorphous phase transition is not feasible in this segment of the curve). On the other hand, the right side of the programming curve is mainly, with the phase transition being influenced by the melt-quench process (allowing for both crystalline-to-amorphous and amorphous-to-crystalline transitions). Reliable storage of multiple levels of information using PCM has been successfully demonstrated, with up to 3 bits (8 levels) per memory cell achieved.

4.6 Resistance drift

Resistance drift refers to the phenomenon of the programmed resistance gradually increasing over time due to a structural relaxation of the material. This relaxation process impacts both the electrical conduction properties and the atomic arrangement within the material.

The drift phenomenon is illustrated in Figure 4.9 for both the RESET and SET states. In the SET state, the resistance remains fairly stable, while in the RESET state, the resistance experiences an increase following a power-law:

$$R(t) = R0\left(\frac{t}{t_0}\right)^{\nu} \qquad Eq.1$$

Where R is the resistance, t is the time, and R0 and t0 are constants. While only the RESET state in GST is affected by drift, there are certain materials, such as Ge-rich GST, where the SET state exhibits a notable drift due to a crystalline disordered phase. The inherent variability in drift poses a significant challenge, particularly in the context of implementing multilevel storage in PCM and for synaptic weight storage in neuromorphic applications. This drift variability varies not only from one material to another but also depends on the chosen material, which, in turn, differs based on the specific application, such as automotive or neuromorphic systems.(Jaguemont et al., 2018; Schuman et al., 2017)

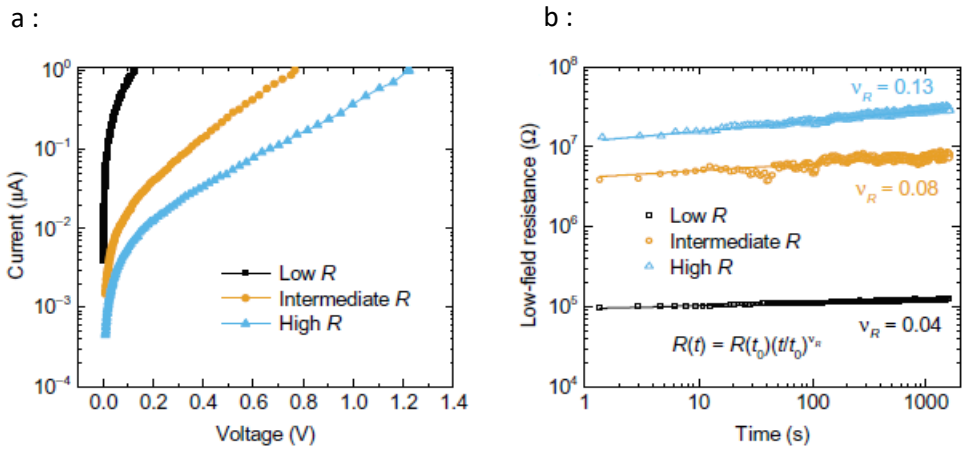


Figure 4. 9 (a) The I-V characteristics of three distinct resistance states are shown: low, intermediate, and high. With a larger amorphous region, the low-field resistance increases, and the slope of log(I) versus V decreases.(b) The low field resistance is plotted against time for the three different resistance states: low, intermediate, and high (Mohseni et al., 2022).

Resistance drift observed in PCM devices can be attributed to the spontaneous structural relaxation of the amorphous phase-change material. This relaxation process is a direct consequence of the amorphization process, as described earlier. When the molten phase-change material is rapidly quenched, its atomic configurations become locked into a highly stressed glass state. Over time, these

atomic configurations tend to relax towards a more energetically favorable "ideal glass" configuration. The observed increase in resistance has been linked to the atomic rearrangements resulting from this relaxation process. It is important to note that in the SET state, where the phase-change material undergoes intentional crystallization, there may be regions that contain some amorphous material.

Various strategies have been attempted to address the impact of resistance drift and ensure the accurate retrieval of stored information from a PCM device. One approach involves capitalizing on the nonlinearity exhibited by the I-V characteristic of the amorphous state (Figure 4.8.a). By measuring the resistance in the high-field regime, a more reliable estimation of the phase configuration can be obtained, which is less susceptible to drift (Sebastian et al., 2011).

As demonstrated in Figure 4.8.a, the slope of log(I) versus V at higher voltages can serve as an indicator of the amorphous region's size (Gallo et al., 2015). This measure is less influenced by drift compared to the low-field resistance. However, without prior knowledge of the programmed state, exploring the high-field regime for every programmed state necessitates applying varying read voltages and detecting the voltage or time at which a specific current threshold (It) is attained. This value (often referred to as the M-metric) (Sebastian et al., 2011) is then used to represent the programmed state. Research has shown that the M-metric can effectively mitigate the impact of drift (Stanisavljevic et al., 2015).

A fascinating method to counteract the impact of resistance drift during the READ operation involves the creation of a "projected PCM device" (Figure 4.10) (Koelmans et al., 2015). This device is constructed with a meticulously designed segment comprising a noninsulating material, referred to as the "projection segment," that runs parallel to the phase-change segment. The resistance of this projection segment is deliberately chosen in such a way that it exerts only a minor influence on the WRITE operation while significantly affecting the READ operation. This is made possible due to the highly nonlinear electrical transport behavior of the amorphous phase.

The concept behind this approach is as follows: During the WRITE operation, the current flows through the phase-change segment because the resistance of the amorphous ON state is lower than that of the projection segment. However, during the READ operation, the current is directed through the projection segment, as its resistance is lower than that of the amorphous OFF state. Consequently, this strategy fully separates information retrieval from information storage, effectively concealing undesirable traits of the amorphous phase like resistance drift, temperature sensitivity, and noise during READ. This innovative approach has demonstrated a reduction in the drift exponent by nearly two orders of magnitude, along with virtually eliminating current noise and temperature dependency in the READ current (Koelmans et al., 2015).

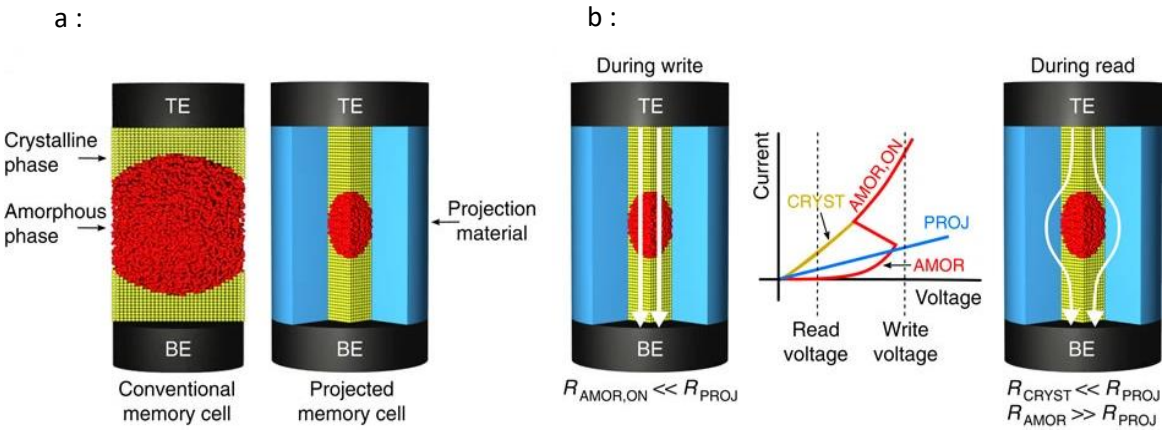


Figure 4. 10 (a) The depiction showcases a conventional phase-change memory cell on the left and a projected phase-change memory cell on the right. In both cases, TE represents the top electrode, while BE stands for the bottom electrode. The crystalline and amorphous phases are illustrated schematically. The projected memory cell features an additional segment containing projection material. (b) The intended I-V characteristics corresponding to the phase-change and projection segments are illustrated schematically. The amorphous phase section is labeled as "AMOR," with the associated resistance RAMOR, and the crystalline phase section is labeled as "CRYST" with resistance RCRYST. The projection segment is marked as "PROJ" with resistance RPROJ. During write mode, the write voltage surpasses the threshold voltage, prompting the amorphous section to enter the ON-state with a resistance RAMOR,ON that is lower than RPROJ. This ensures that the majority of the current flows through the phase-change segment, inducing Joule heating and a subsequent phase transition. During read mode, as RPROJ is substantially lower than RAMOR at low fields, the current predominantly courses through the portion of the projection segment parallel to the amorphous section. Elsewhere, the current primarily flows through the crystalline section (Koelmans et al., 2015).

4.7 The phase-change materials

The material used for phase change memories must possess specific properties (Figure 4.10). One fundamental characteristic of the material is its crystallization temperature. A high crystallization temperature demands a significant amount of energy to induce the state change in the cell. Consequently, a memory utilizing such a material would consume substantial energy and generate significant heat. Conversely, if the crystallization temperature is too low, a cell in its amorphous state might spontaneously crystallize as its environmental temperature increases, potentially even at room temperature.

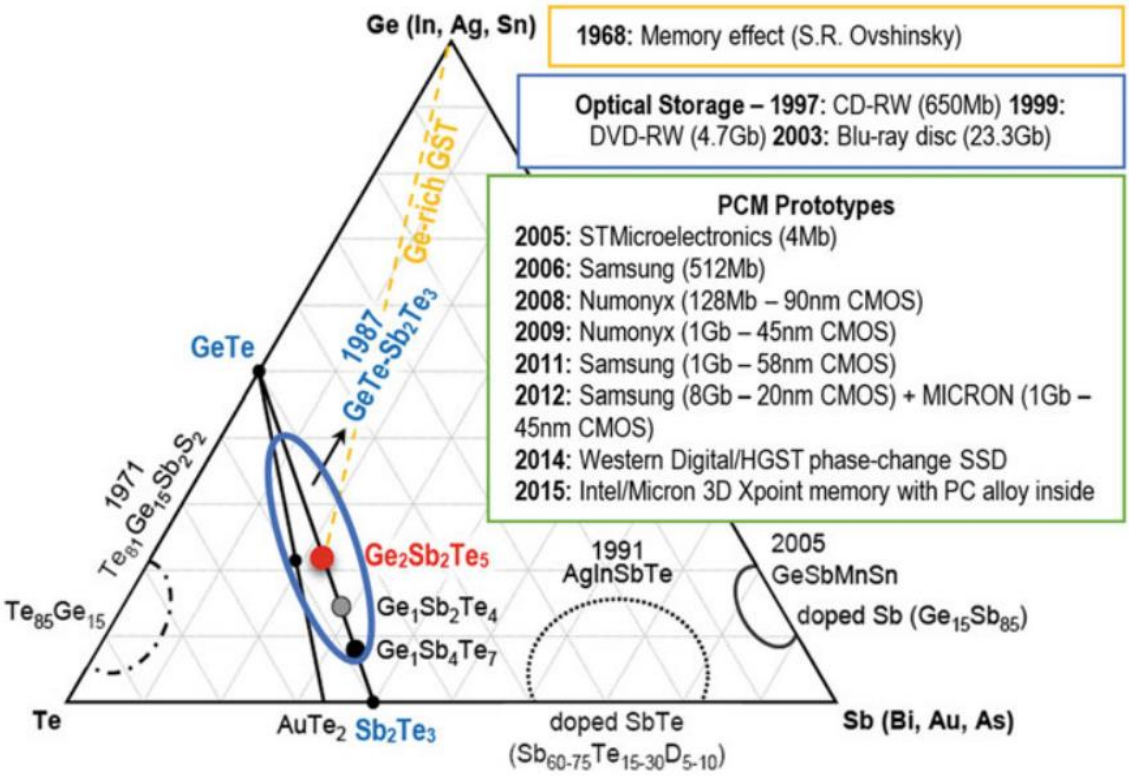


Figure 4. 11 The ternary phase diagram depicted here illustrates various phase change alloys, accompanied by their respective discovery years. This diagram provides a comprehensive overview of the evolution of both optical and resistive phase change memories. A significant body of research has been dedicated to alloys positioned along the pseudo-binary GeTe-Sb2Te3 tie line within the ternary Ge-Sb-Te system. Notably, the well-known Ge2Sb2Te5 alloy falls within this category. More recently, investigations have extended to include GexTe1-x and Ge2Sb2Te5 alloys doped with elements like B, C, N, O, SiC, SiN, SiO2, among others. Additionally, research has encompassed Ge-rich Ge-Sb-Te alloys. These explorations are motivated by their potential utilization in embedded phase change memories, highlighting the continual pursuit of enhancing the capabilities and performance of phase change memory technologies. (Noé et al., 2017)

In such cases, information could be lost. Thus, a compromise is necessary to select a material that enables good information retention while utilizing a reasonable amount of energy for state transition (Navarro et al., 2013).

The crystallization speed is also an important property. The material should be capable of rapid crystallization to ensure competitive write times compared to other existing memory technologies. The glass transition temperature corresponds to the temperature at which a sudden drop in viscosity is observed in the material. Below this temperature, the material can be considered solid and, particularly, its evolution over timeframes of the order of a year is negligible. This glass transition temperature must be higher than the operational temperature of the devices to ensure long-term information storage.

Lastly, the resistivity contrast between the amorphous and crystalline phases must be high, ideally differing by several orders of magnitude, to ensure clear reading of the cell's state.

It becomes apparent that materials suitable for use in PCMs (as indicated in Figure 4.11) must satisfy specific and seemingly opposing criteria: qualitatively, the amorphous and crystalline phases need to have structures that are sufficiently close for rapid reorganization, yet they also need to exhibit significant differences in their electronic properties. Comparative studies of various materials have been conducted to determine which ones are best suited for use in PCMs; D. Lencer, for example, establishes a strong connection between the electronic bonding properties of materials and their phase-change behaviors (Lencer et al., 2008).

4.7.1 $Ge_2Sb_2Te_5$

The germanium-antimony-tellurium alloy was first investigated by Yamada in 1987 (Yamada et al., 1987, 1991). This alloy stands out as one of the most extensively studied alloys in the realm of PCM research. $Ge_2Sb_2Te_5$, often referred to as GST, has been recognized as a promising material since the early 1990s due to its remarkable stability at room temperature and its relatively rapid crystallization when exposed to laser irradiation (within 50 nanoseconds). These attributes positioned GST as an excellent candidate for optical recording purposes. Additionally, its properties led to it being one of the first phase-change materials considered for potential use in PCM applications. GST boasts a crystallization temperature of approximately 150°C, while its melting point is around 660°C (Kalb et al., 2005). It is also acknowledged that the $Ge_2Sb_2Te_5$ alloy exhibits behaviors typical of fragile liquids (Orava et al., 2012; Sosso et al., 2012): whereas the kinetic coefficient of crystallization is usually proportional to the inverse of viscosity, fragile liquids show a slower decrease in their crystallization kinetics as viscosity increases. This characteristic enables this alloy to crystallize rapidly over a broad temperature range spanning from its glass transition temperature to its melting temperature.

Another property of the GeSbTe alloy, often abbreviated as GST, contributing to its success as a PCM material is its electronic switching behavior (Pirovano et al., 2004) amorphous GST, when subjected to a voltage, remains insulating up to a specific threshold voltage, beyond which it becomes conductive. This effect was first observed by Ovshinski in 1968 (Ovshinsky, 1968). This characteristic enables the passage of the current necessary for melting the material while still allowing a significant contrast in resistivity between the amorphous and crystalline phases as long as the applied read voltage is below the threshold voltage of the electronic switching. The current-voltage characteristic is depicted in Figure 4.6. A substantial resistivity contrast between the amorphous and crystalline phases is evident for voltages below the threshold: the amorphous phase is much more resistive than the crystalline phase.

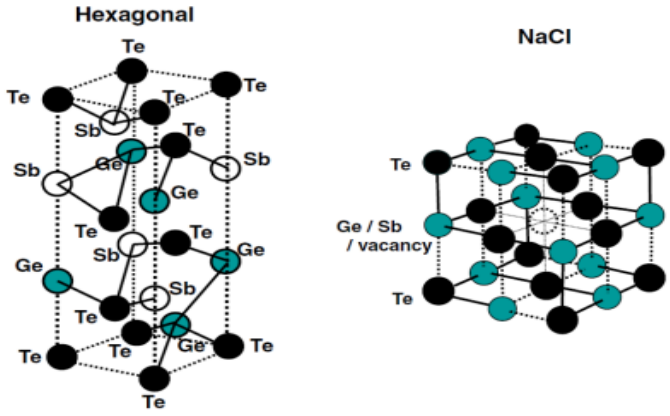


Figure 4. 12 The hexagonal and NaCl-type structures are possible crystalline arrangements that Ge2Sb2Te5 (GST) can adopt, as described in (Terao et al., 2009).

However, the resistivities associated with each phase become comparable when the voltage exceeds the threshold and tend to equalize as the voltage increases. $Ge_2Sb_2Te_5$ can crystallize in two distinct structures: the NaCl-type structure and the hexagonal structure (Wuttig & Yamada, 2007). Both of these structures are illustrated in Figure 4.12. Among these two, the hexagonal structure is the more stable one. In memory applications, the NaCl-type structure because it lead to achieve a rapid crystallization, facilitating efficient memory effects.

4.7.2 Ge-rich GST

A weakness of PCM technology is the information retention quality at high temperatures. For instance, in automotive applications, PCM devices need to retain their information for 10 years at 150°C (Zuliani et al., 2013). The crystallization temperature of $Ge_2Sb_2Te_5$ at 150°C makes it unsuitable for such applications. Other alloys must be developed to prevent the spontaneous crystallization of amorphous cells. One solution that has been studied and adopted is modifying the alloy's stoichiometry. An alloy with higher crystallization temperature is GeSbTe enriched with germanium. This solution has been chosen by companies like STMicroelectronics to address markets where high-temperature information retention is crucial (Sousa et al., 2015). The evolution of crystallization temperature with the addition of germanium is depicted in Figure 4.13.

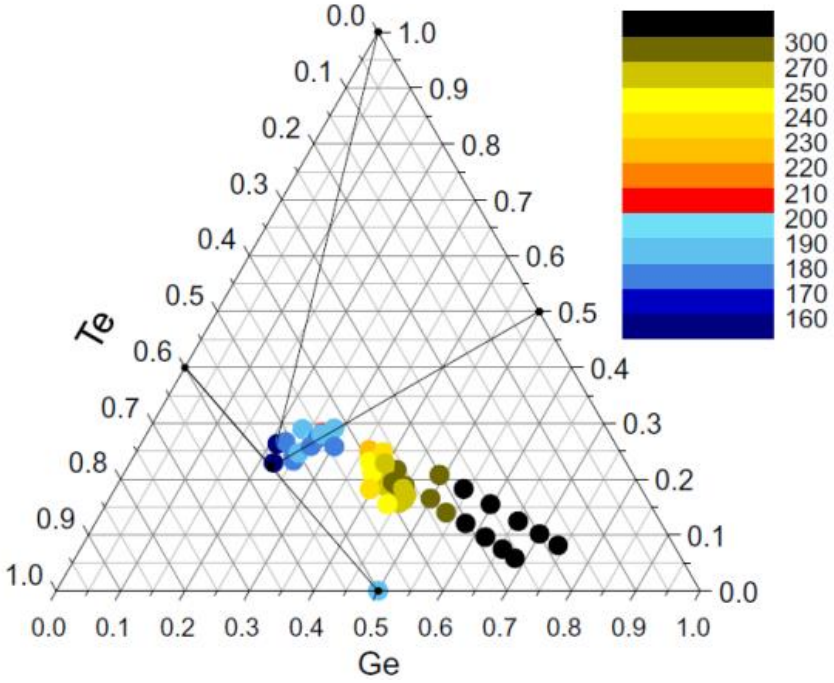


Figure 4. 13 The variation of crystallization temperature of GST with increasing germanium enrichment is depicted in the figure. Each colored point on the simplex corresponds to a crystallization temperature in degrees Celsius.

Back In 2011, where Cheng (Cheng et al., 2011) . initiated the exploration of the GeSbTe ternary system towards compositions enriched in germanium (Cheng et al., 2011). They subsequently improved this approach by introducing nitrogen into the mix (Wong et al., 2012). Focusing on the pseudo-binary line of Ge-Sb2Te3, they identified a high-performance composition known as the "golden" composition. This germanium-rich mixture exhibited heightened thermal stability in the RESET state while maintaining rapid switching speeds. Utilizing these Ge-rich compositions, Zuliani et al. successfully demonstrated data retention even under the typical soldering reflow temperature profile (Zuliani et al., 2013). Meanwhile, by investigating the pseudo-binary line of Ge-Ge2Sb2Te5, an explanation was proposed for the robust stability of both the RESET and SET states(Sousa et al., 2015). In the subsequent sections, we present the outcomes of these investigations. The term "GST + Ge x %" refers to the indicated proportion of Germanium added to the Ge2Sb2Te5 alloy using co-sputtering from two targets namely, Ge and Ge2Sb2Te5.

At the thin film level, X-ray diffraction patterns obtained from Ge-rich GST thin films subjected to annealing at 400°C reveal the presence of two distinct cubic phases, which correspond to the separation of the Ge and Ge2Sb2Te5 phases as indicated by the equilibrium phase diagram (Bordas et al., 1986). The extent of the germanium phase's volume fraction becomes more pronounced with higher germanium percentages, as anticipated. Notably, for this category of germanium-rich GST alloys, unlike the stoichiometric compounds Ge2Sb2Te5, the crystallization process involves a local

alteration in composition alongside the phase separation. It is worth mentioning that when the germanium content is held constant, the volume fraction of the germanium cubic phase diminishes due to the introduction of nitrogen. This trend mirrors the outcomes observed for the GeTeN alloy, suggesting that a portion of the germanium becomes bonded with nitrogen, forming Ge-N bonds during the formation process.

Germanium-rich GST alloys have been integrated into an industrial test vehicle that includes a PCM memory cell structure on a 90 nm CMOS platform. Despite the observed phase separation in the characterization of thin films, the devices display functional resistance-interruption (RI) characteristics (Fig. 4.14). These RI characteristics reveal that the transitions from the SET state to the RESET state are not as sudden as those obtained with stoichiometric compounds like Ge2Sb2Te5. However, the RESET current remains largely unaffected. In comparison to Ge2Sb2Te5 devices, the enhanced thermal stability of the amorphous phase becomes evident through improved retention of devices programmed in the RESET state. For GST alloys enriched with 45% germanium, extrapolated measurements suggest that the RESET state's stability can be maintained for at least 10 years at temperatures up to 200°C. This duration and temperature threshold exceed the specified requirements for automotive applications (10 years at 150°C).

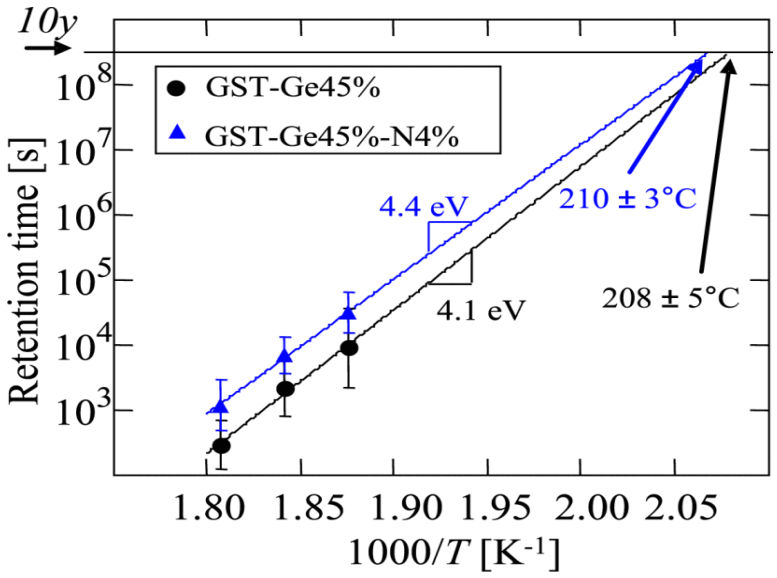


Figure 4. 14: Creating an Arrhenius plot depicting the data retention time for GST-Ge45%, with and without Ndoping, demonstrates the influence of doping on the material's stability. (Navarro, Thesis 2013).

4.8 GST and GE-rich GST comparison

In our detailed study, we closely looked at two important materials: the regular GST and the innovative more Germanium (GE) GE-rich GST variant. We found something interesting these materials behave differently when it comes to drifting over time. Specifically, the regular GST material tends to have less drifting compared to the GE-rich GST variant. This finding has important implications for our focus on applying these materials in neuromorphic engineering.

In this section, we will provide a detailed overview of the setup and protocols employed for the electrical characterization. Subsequently, we will present the comparative results for the two materials mentioned earlier.

4.8.1 Electrical characterization

- **The array**

The investigated PCM device is a 16kbit 1T1R NOR array integrated into the Back-End-Of-Lines (BEOL) fabrication process of the LETI Memory Advanced Demonstrator (MAD), which uses a 130nm CMOS technology with a 300mm FDSOI wafer diameter. Most of the devices feature a 'wall' structure that is heater-based. This structure employs a vertical thin wall heater, with chalcogenide material deposited above it. This confined structure improves thermal efficiency by depositing phase-change material in limited areas surrounded by oxide. The 'wall' devices heater width of 100nm. The array is organized using Bitlines (BL) and Wordlines (WL), decoded by simple pass-gate MUXs that connect the IO ring to the selector gates and the top and bottom electrodes (TE and BE respectively). The selector and current limiter in the well-known 1T1R structure are implemented using a standard logic 28nm FDSOI thin oxide NMOS. The PCM cell comprises a phase-change material (Ge-rich GST or Ge2Sb2Te5) sandwiched between Pt/Ti electrodes and is embedded between metal 5 and metal 6. Additionally, a 5nm wide TiN pillar serves as the heater element, connecting the BE to the underlying W plug."

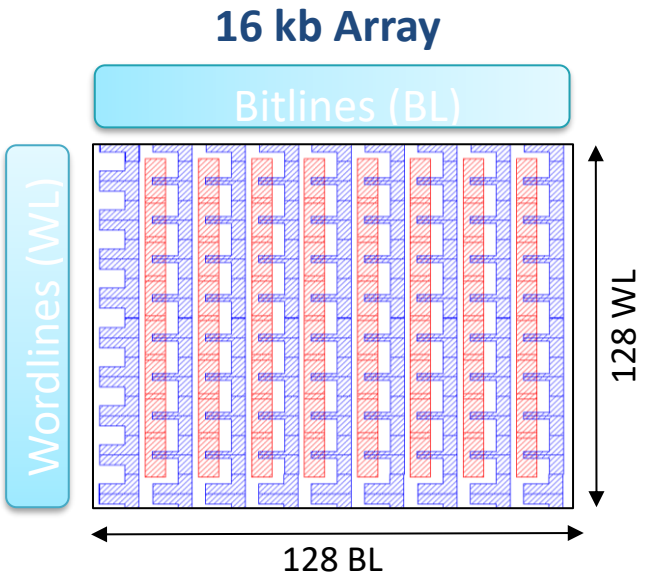


Figure 4. 15 : Schematic of a 16kb array

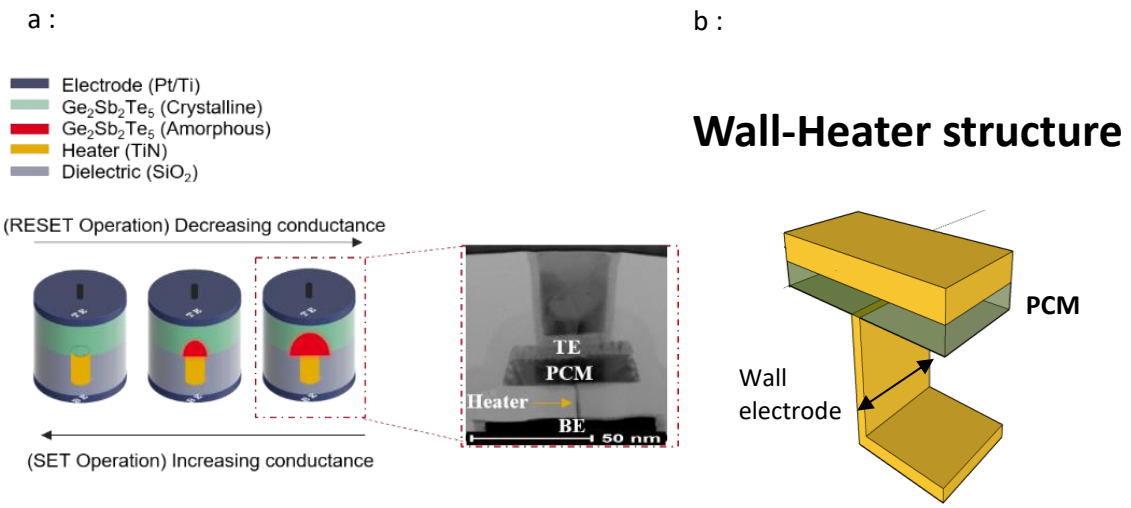


Figure 4. 16: a) Schematic of the PCM. On the right a TEM image of the PCM cell. b) Schematic of a "wall" PCM device

▪ Experimental setup

In phase-change memories (PCMs), the cell resistance serves as the primary parameter for measuring the memory state. Programming in PCMs involves applying voltage pulses, with adjustable rise, width, and fall times (as depicted in Fig. 4.17).

The key parameters extracted for basic electrical characterization include:

- RESET resistance: This represents the maximum resistance of a PCM in the amorphous phase.
- SET resistance: It refers to the resistance of a PCM in the crystalline phase.
- Resistance Window: This is the logarithmic difference between the RESET resistance and the SET resistance.
- RESET current: The current required to transition the resistance from the SET state to a value corresponding to 90% of the resistance window.
- SET current: The current necessary to program the cell into the SET state.
- Threshold voltage (V_{TH}): The voltage at which threshold switching occurs.

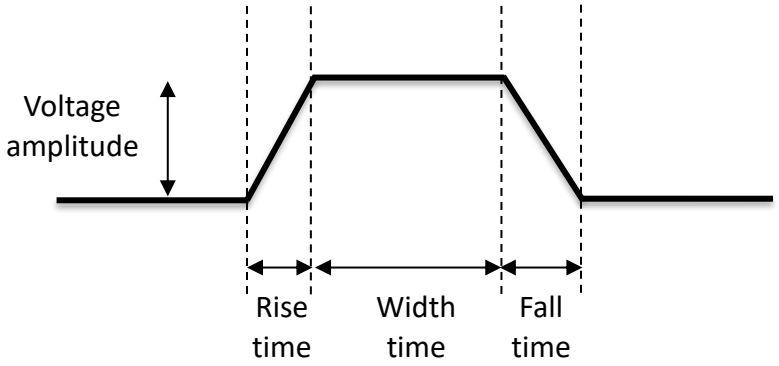


Figure 4. 17 : Example of a voltage pulse used to program our PCM devices

▪ Test setup

The setup can be divided into two components: an analog component responsible for set, reset, and read operations on the memories, and a digital component responsible for addressing the multiplexers. The analog portion is controlled by a Pulse Generator Unit (PGU) Keysight B1500. This device generates voltage pulses and measures current on the BL, SL, or WL lines.

The digital aspect is managed by an Arduino Mega microcontroller (see Figure 4.18). It sends addresses to the memories for writing, erasing, or reading to the various decoders. It's connected to the Trigger output of the B1500. When the B1500 completes its measurement, it sends a trigger signal to the microcontroller, which then handles changing the address.

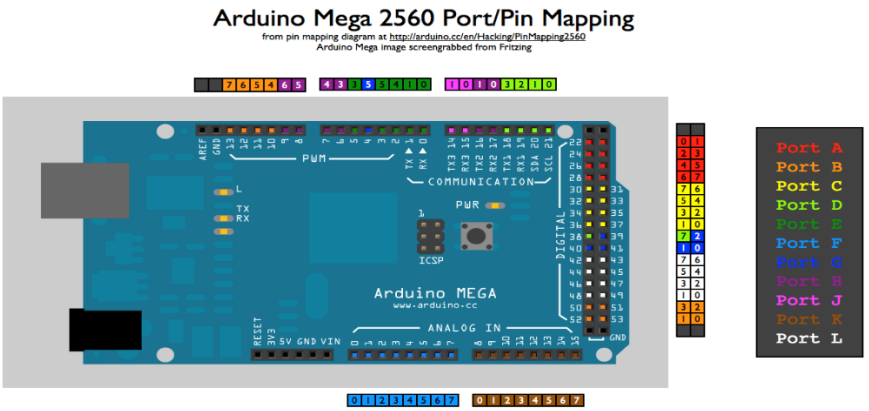


Figure 4. 18 : schematic of an Arduino Mega. (Source : *Fritzing*)

The general operational diagram of the electrical characterization measurements is illustrated in the provided Figure below

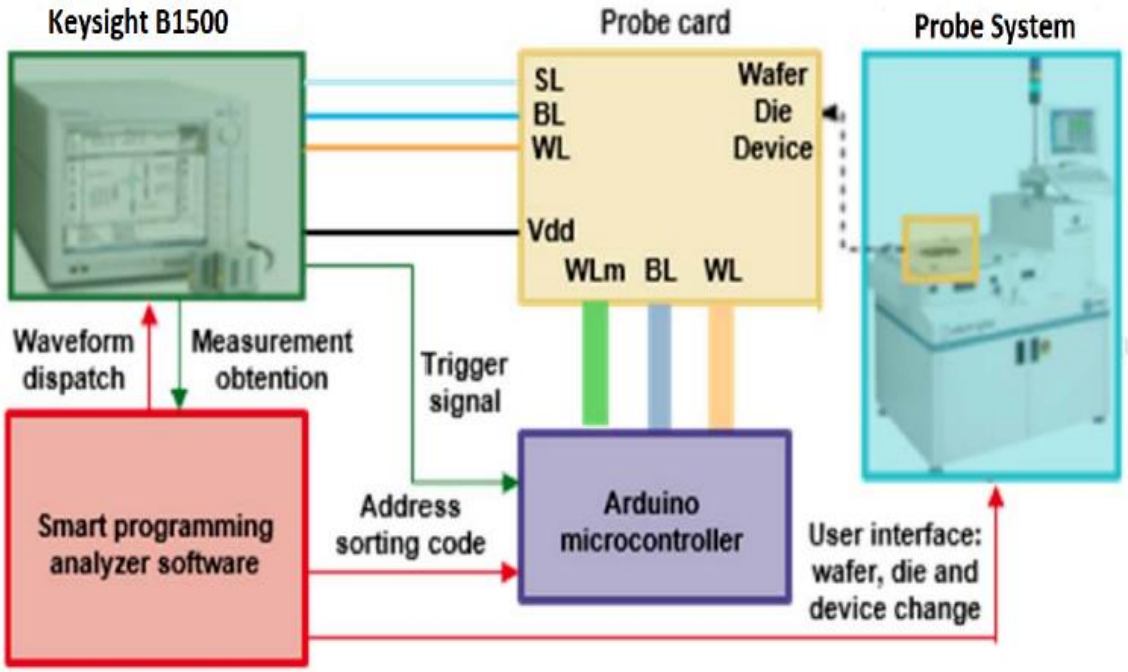


Figure 4. 19 : Experimental setup to characterize 16kb arrays. Adapted from (Nguyen, Thesis.,2018)

As visible in this diagram, a measurement board is employed to link the 25 pads to the probes (refer to Figure 4.20). The specimens for study are positioned on an Cascade S300 test bench. A set of 25 SMA cables is utilized to establish connections from the board to either the microcontroller or the PIV outputs. Lastly, the 2.75V supply voltage (Vdd) is sourced from either the microcontroller or an external power supply.

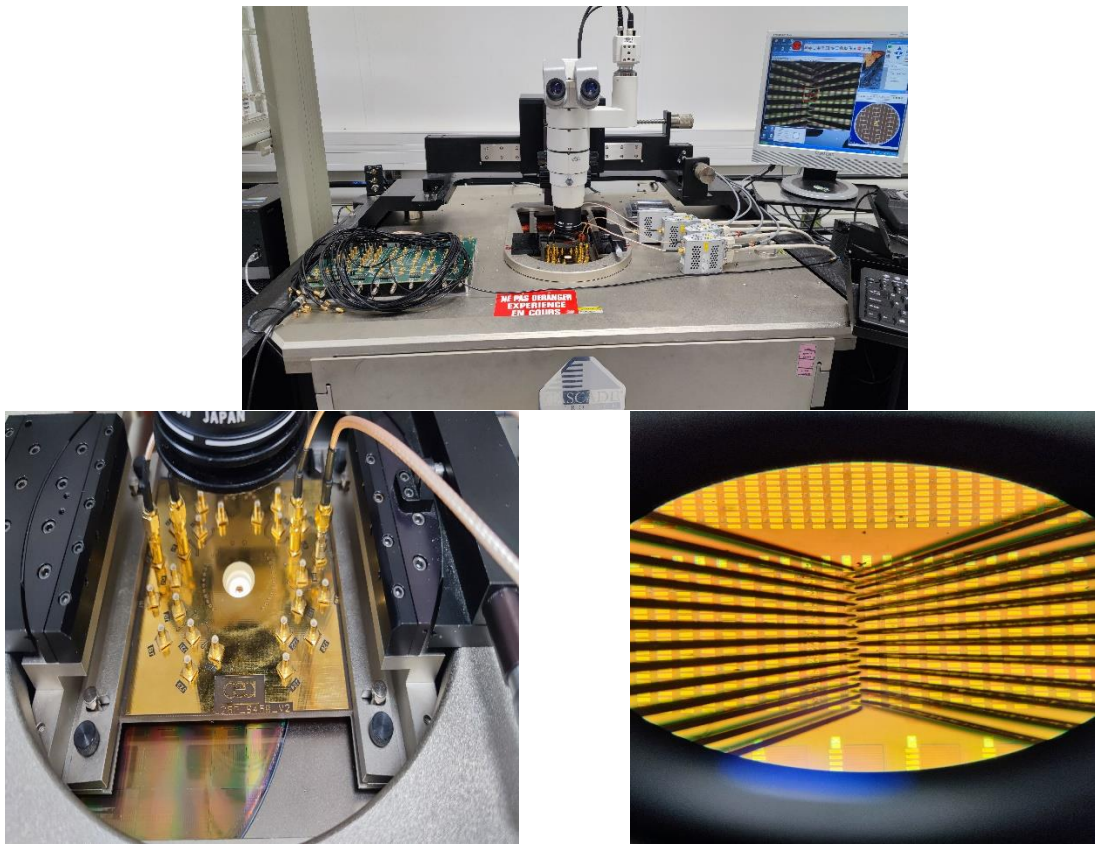


Figure 4. 20 : Image of the measurement setup with the card and the 25 tips connected to the DUT

A key aspect to highlight is the absence of direct communication between the B1500 and the microcontroller. When the B1500 completes its measurement, it transmits a trigger signal to the microcontroller, which subsequently manages the address alteration process. However, it's worth noting that the Arduino does not send a signal back to the B1500 once the address change has been carried out successfully (this approach would be excessively time-consuming). If the timing is not appropriately synchronized, the B1500 might initiate the new measurement sequence before the address change has been fully implemented. Hence, accurately gauging the time interval between the trigger signal dispatched from the B1500 to the Arduino and the subsequent address alteration pulses sent by the Arduino to the multiplexers is of significant importance.

- **Test protocol**

The primary objective of this study is to identify the optimal material that minimizes drift, thus enhancing the efficiency of storing synaptic weights. In the context of synaptic weight storage, drift refers to the undesirable alteration or fluctuation of resistance of the PCM devices over time. Such drift can significantly impact the accuracy and reliability of neural networks or synaptic circuits. By investigating various materials and their properties, this research endeavors to pinpoint the substance that best

preserves synaptic weights, thereby mitigating drift-related concerns. The outcome of this study is poised to not only advance the field of neural network engineering but also contribute to the design of more robust and enduring artificial intelligence systems.

4.8.2 Comparative results

The electrical analysis of the components is conducted under an ambient temperature of 300K. To mitigate any possible impact of recrystallization on our measurements. We first need to characterize the resistance drift of our 16 kb device. The line-cell is programmed by altering the reset pulse, we induce varying reset voltage between 1.0V, 1.2V and 1.4V

After programming, we measure the resistance's evolution over a duration of milliseconds, and this data is then fitted to the standard drift equation, denoted as $R(t) = R0\left(\frac{t}{t_0}\right)^{\nu}$ (Fig. 4.21). In the provided figure, distinct examples are presented for each individual condition. It is essential to reiterate that the conditions being discussed pertain to the varying reset current voltages employed as part of the experimental setup. The figure illustrates the phenomenon of resistance drift towards higher resistance values, underscoring the prevalence of the drift effect.

Subsequent to obtaining these measurements, our focus shifted to the quantification of the drift coefficient ν (refer to Figure 4.22). By employing the established drift equation and referring to Figure 4.21, we derived the coefficients. Our analysis revealed drift coefficients ranging from 0.015 to 0.08 across three discrete conditions for both materials (Ge2Sb2Te5 and germanium-rich GST).

Consequently, it becomes evident that the drift coefficient undergoes alteration irrespective of the voltage magnitude of the applied reset pulse. Notably, the measured drift coefficient ν at 300K is observed to be lower than that of the Ge-rich GST, as depicted in Figure 4.22.

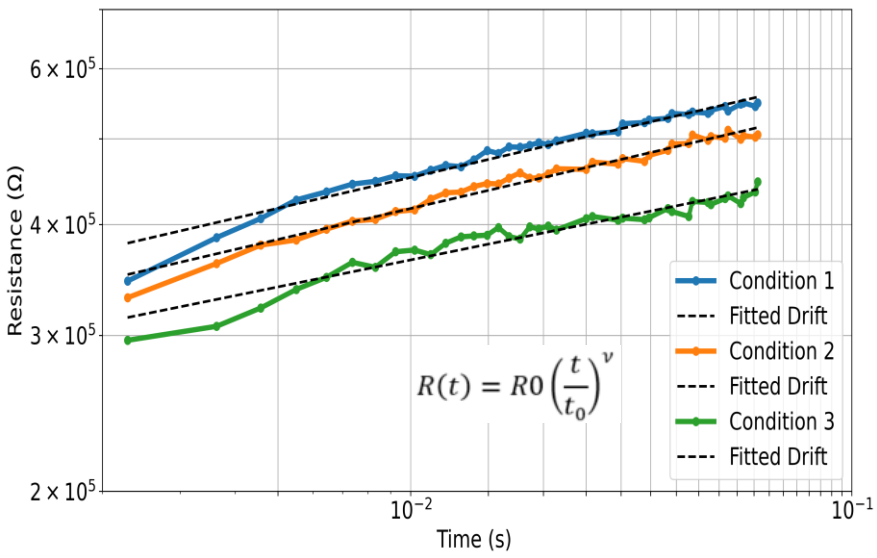


Figure 4. 21 : Resistance as a function of time measured in Ge2Sb2Te5 with three different level of programing current showing the drift phenomena.

Figure 4.22 illustrates a comparison between the materials GST225 and Ge-rich GST under identical conditions, as previously described. The analysis indicates that Ge-rich GST exhibits a higher drift, yet the plot reveals a plateau, suggesting that the drift remains relatively constant across all three conditions, as measured with Rzero (initial resistance before programming). In contrast, GST225 demonstrates a lower drift compared to Ge-rich GST, but the magnitude of drift appears to depend on the specific conditions applied.

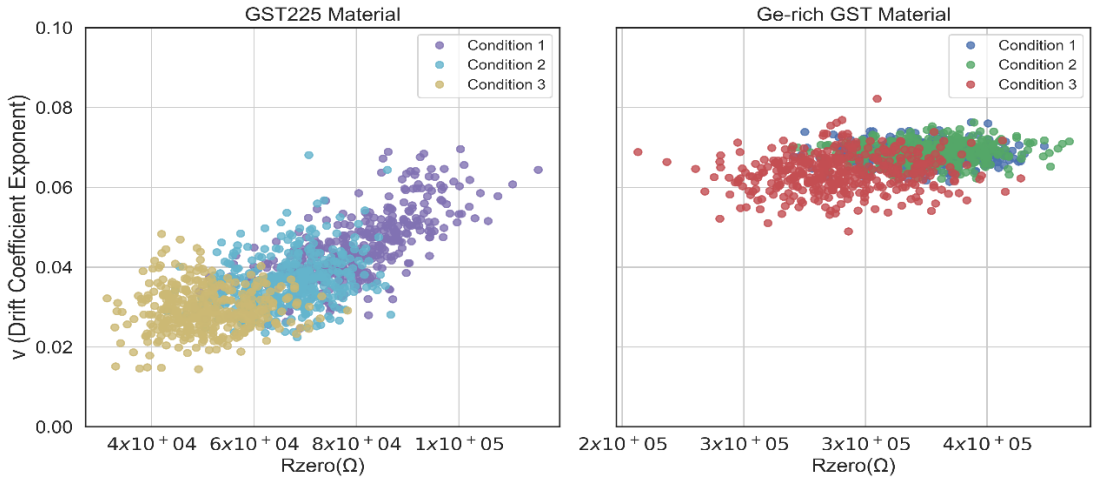


Figure 4. 22 : Drift coefficient ν ranging from 0.015 to 0.08 across three discrete conditions for both materials (Ge2Sb2Te5 and germanium-rich GST).

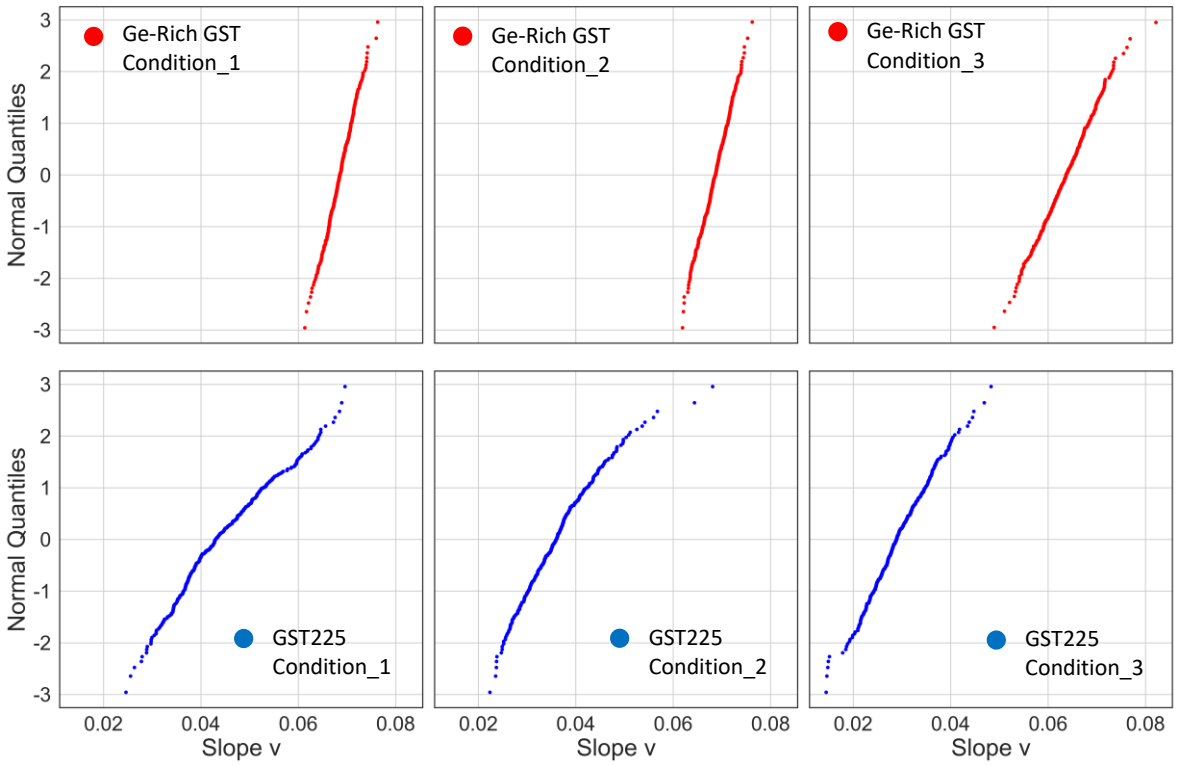


Figure 4. 23 : The distribution of nu (v) for both materials, GST225 and Ge-rich GST, was examined under identical 3 conditions.

In Figure 4.23, the Nu distribution plots provide insights into material behavior. A more vertical plot signifies greater consistency, and higher verticality is indicative of better performance. The first three plots for Ge-rich GST align with the trend observed in the preceding plot, demonstrating a relatively constant nu distribution within the range of ≈ 0.06 to 0.08 but higher from the GST225

Conversely, the Nu distribution plots for GST225 exhibit some dispersion across the three conditions. However, in the case of GST225, the third condition stands out by showing a lower drift ranging between ≈ 0.02 and 0.04 and a somewhat constant trend. While not perfect, this stability is notably improved compared to the other two conditions for GST225.

In the realm of neuromorphic computing, where the goal is to emulate the functionality of biological neural networks through artificial systems, the stability and persistence of materials play a pivotal role. Neuromorphic systems aim to replicate the brain's remarkable ability to process and store information in a highly efficient and adaptive manner. However, achieving this level of complexity and functionality requires materials that can reliably maintain their properties over extended periods.

Stability refers to a material's ability to resist changes in its properties or structure over time and under varying conditions. Persistence, on the other hand, refers to the duration for which a material can retain its intended properties without significant degradation. These factors are critical in the context of neuromorphic applications for several reasons:

- Longevity of Functionality: Neuromorphic systems are designed to perform tasks that involve learning, memory, and information processing, mirroring the plasticity and adaptability of the human brain. For these systems to be effective, the underlying materials must exhibit stability and persistence. Any drift or degradation in material properties could lead to inaccurate computations, compromised memory retention, and reduced overall system performance.

- Reliability: The reliability of neuromorphic systems heavily relies on the consistent behavior of the materials they are built upon. Materials that are prone to significant drift or degradation might introduce errors in computations, leading to unreliable outcomes. This is especially problematic in applications where precision and accuracy are crucial, such as pattern recognition or decision-making tasks.

- Energy Efficiency: One of the advantages of neuromorphic computing is its potential to be highly energy-efficient. This efficiency is partly achieved by emulating the brain's ability to perform computations with minimal power consumption. Stable materials contribute to this efficiency by minimizing the need for frequent recalibration or replacement of components due to drift or degradation, thereby reducing energy overhead.

- Adaptability: The brain's ability to adapt and learn from experience is a hallmark of its functionality. Materials with high stability and persistence enable neuromorphic systems to maintain their learned connections and weights over time, facilitating the retention of knowledge and the gradual refinement of the system's performance.

Conclusion

The exploration of phase-change memory (PCM) technology beyond the traditional von Neumann architecture has provided valuable insights into the evolution of computing paradigms. The brief historical overview of PCM technology highlights its journey from conceptualization to practical implementation, showcasing its potential to redefine the landscape of memory and processing systems. Examining the construction of the PCM cell elucidates the intricate design principles that underlie its functionality. The PCM cell serves as the fundamental building block, embodying the dynamic interplay between amorphous and crystalline states that form the basis of its operation. The SET/RESET operations play a pivotal role in manipulating the phase of the material, enabling the storage and retrieval of data in a non-volatile manner. Delving into the phase-change mechanism and the associated switching process unveils the memory and multilevel operations of PCM. These operations contribute to the versatility of PCM, allowing for efficient data storage and retrieval while accommodating the complexity of modern computing tasks. However, the chapter also addresses challenges such as resistance drift, acknowledging the need for ongoing research and development to enhance the stability and reliability of PCM-based systems. The discussion on phase-change materials, particularly Ge2Sb2Te5 and Ge-rich GST, highlights the significance of material selection in optimizing PCM performance. In the comparison between GST and Ge-rich GST, it becomes evident that the choice of phase-change material plays a crucial role in determining the efficacy of PCM technology. This analysis provides valuable insights for future advancements and optimizations in the field. In essence, the chapter contributes to the understanding of PCM technology, shedding light on its principles, mechanisms, and materials. As we move beyond the traditional von Neumann architecture, PCM emerges as a promising candidate for shaping the future of memory and computing. The continuous exploration and refinement of PCM technology hold the potential to usher in a new era of computational efficiency and data storage capabilities, paving the way for innovative applications in the realm of information technology.

Résumé

- Explored PCM technology evolution beyond von Neumann architecture.
- Provided a concise historical overview of PCM development.
- Examined PCM cell design and amorphous-crystalline interplay.
- Emphasized SET/RESET operations for non-volatile data storage.
- Explored phase-change mechanism and multilevel operations.
- Addressed challenges, including resistance drift considerations.
- Discussed significance of Ge2Sb2Te5 and Ge-rich GST materials.
- Compared GST and Ge-rich GST for PCM performance optimization.
- Provided guidance for future advancements and optimizations

Chapter 5

Frequency modulation of conductance level in PCM device for neuromorphic applications

"Within the crystalline realms of a PCM device, each modulated pulse tells a story—a narrative of neuromorphic possibilities, where memory becomes a canvas for the ever-evolving intelligence of machines."

Dr. Alan Bennett, Professor of Electrical Engineering, Massachusetts Institute of Technology.

5. Frequency modulation of conductance level in PCM device for neuromorphic applications

In recent years, there has been a growing exploration of phase change memory (PCM) devices for the development of more effective neuromorphic applications. This exploration capitalizes on the physical properties of PCM devices, as discussed in previous chapters, to enable certain computational operations to be performed directly within the memory. The exploration of PCM devices in neuromorphic computing is a promising avenue for developing more effective and energy-efficient cognitive systems. Their unique characteristics, including non-volatility, low power consumption, in-memory computing capabilities, and synaptic weight adjustability, make PCM technology a compelling choice for simulating and enhancing the brain's computational processes.

In this chapter, we delve into a comprehensive exploration of the pivotal attributes that render neuromorphic applications viable within PCM. Our focus will be on clarifying the intricacies associated with programming synaptic weights through traditional methods. Furthermore, we will introduce novel approaches to synaptic weight programming. The discourse will not only encompass theoretical foundations but will also feature practical insights, including a presentation of COMSOL simulations and corresponding measurements.

5.1 The key attribute of Phase change memory

As research and development in this field continue to advance, PCM-based neuromorphic systems hold the potential to revolutionize various domains, from artificial intelligence to edge computing and beyond. One of the fundamental properties of PCM that makes it suitable for in-memory computing (IMC) is its ability to store two levels of resistance or conductance values in a non-volatile manner, allowing for reversible switching between these two levels (binary storage capability) Figure 5.1.

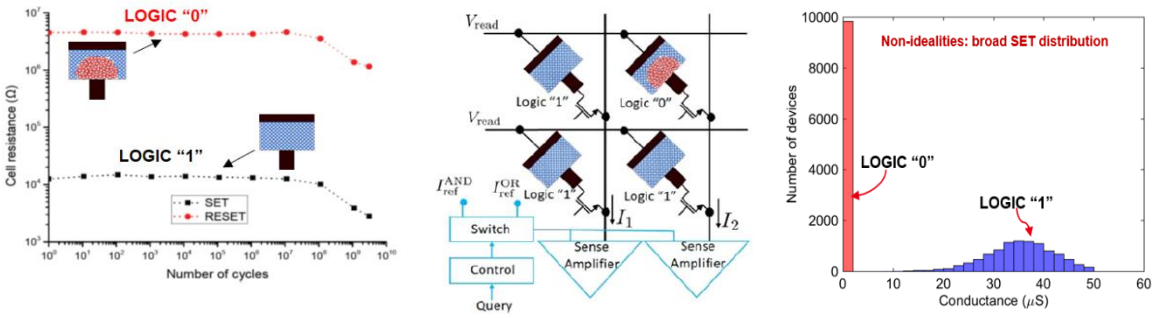


Figure 5. 1 : Essential physical characteristics that support neuromorphic computing include non-volatile binary storage, which enables in-memory logical operations essential for tasks like hyper-dimensional computing. However, there are significant difficulties associated with phase-change memory (PCM) devices in this context. One of the main challenges is the wide variation in conductance values during the SET/RESET processes, which can be harmful when it comes to applications like in-memory logic. (Christensen et al., 2022)

This binary storage capability in PCM-based systems is analogous to the synapses in biological neural networks, where the strength of connections between neurons can be modulated. It enables PCM-based neuromorphic systems to emulate synaptic plasticity, a crucial aspect of biological learning and memory processes. Consequently, these systems can perform cognitive tasks with remarkable efficiency and speed, making them a promising avenue for next-generation AI applications.

Additionally, their non-volatile nature ensures data retention even when power is turned off, a feature critical for edge computing and IoT devices, where energy efficiency and reliability are paramount. This characteristic forms the basis for conducting in-memory logical operations by manipulating voltage and resistance state variables (Sebastian et al., 2020). Notable applications of in-memory logic include tasks such as database queries (Giannopoulos et al., 2020) and hyper-dimensional computing (Karunaratne et al., 2020).

Another crucial property of PCM enabling neuromorphic computing is its capacity to achieve not only two discrete levels but a continuous range of resistance values (analog storage capability). This is typically accomplished by creating intermediate phase configurations through the application of partial RESET pulses. The analog storage capability plays a pivotal role in achieving matrix-vector multiplication (MVM) operations with a time complexity of O(1) by exploiting Kirchhoff's circuit laws. This capability finds significant use in deep neural network (DNN) inference (Joshi et al., 2020), where each synaptic layer of a deep neural network (DNN) can be mapped to a crossbar array of PCM devices. The industry is increasingly interested in this application due to its potential to significantly reduce latency and energy consumption compared to existing solutions. Additionally, in-memory MVM operations open doors to non-neuromorphic applications like linear solvers and compressed sensing recovery.

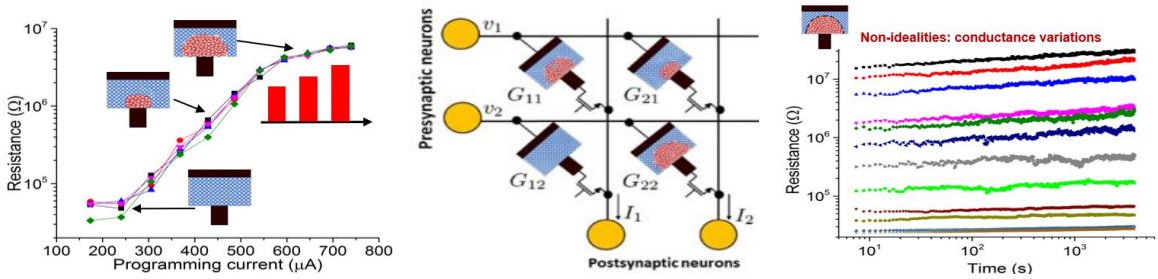


Figure 5. 2 : Analog storage is crucial for performing efficient matrix-vector multiply (MVM) operations, which play a central role in applications like deep neural network (DNN) inference. However, the presence of drift and noise in analog conductance values leads to less precise matrix-vector multiply operations. (Christensen et al., 2022)

The third key property enabling neuromorphic computing is the cumulative behavior, which arises from the crystallization kinetics of PCM. This property can be harnessed for implementing DNN training (Tsai et al., 2018) and plays a central role in realizing local learning rules like spike-timing-dependent plasticity in spiking neural networks (SNN) (Ambrogio et al., 2016). In both cases, the cumulative property is exploited to efficiently update synaptic weights and even to emulate neuronal dynamics (Tuma et al., 2016).

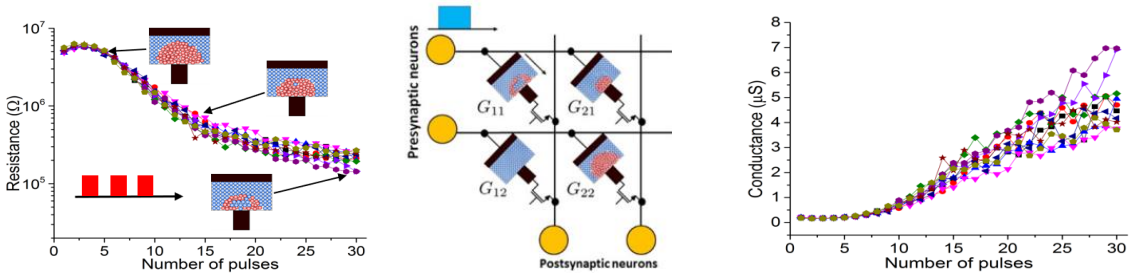


Figure 5. 3: The cumulative behavior is advantageous for applications like deep neural network (DNN) training and simulating neuronal and synaptic dynamics in Spiking Neural Networks (SNN). However, the non-linear and stochastic cumulative behavior can lead to less precise updates of synaptic weights.(Christensen et al., 2022)

5.2 The concept of multi-memristive synapse

The concept of the multi-memristive synapse is visually depicted in Figure 5.4. In this type of synapse, the synaptic weight is represented by the combined conductance of N individual memristive devices. To achieve synaptic efficacy, an input voltage corresponding to the activation level of a neuron is applied to all the constituent devices. The sum of the currents flowing through these individual devices forms the net synaptic output.

The implementation of synaptic plasticity in neuromorphic computing can involve a clock-based arbitration scheme that selects and programs only one device at a time. This scheme is used to address challenges related to multi-memristive synapses and synaptic efficacy. The selection of a device is typically done using a counter-based arbitration scheme, where one of the devices is chosen according to the value of a counter. This approach helps in reducing the effective number of programming operations of a synapse, further improving endurance-related issues. The selection is performed based on the arbitration module alone, without knowledge of the conductance values of the individual devices, which can lead to a non-zero probability that a potentiation or depression pulse will not result in an actual potentiation or depression of the synapse (Boybat et al., 2018). Additionally, a global clock-based arbitration scheme has been proposed for multi-memristive synaptic architecture to address the challenge of synaptic plasticity. This approach aims to provide an efficient way to implement synaptic plasticity in neuromorphic computing systems

In addition to the global selection clock, additional independent clocks, such as a potentiation frequency clock or a depression frequency clock, can be incorporated to further control the frequency of potentiation and depression events, as shown in Figure 5.4. This multi-memristive synapse concept offers a flexible mechanism for adjusting synaptic weights and can be applied in neural network systems with efficient management of updates and plasticity.

The multi-memristive synapse can be set up in two ways: a non-differential or a differential architecture.

In the non-differential architecture, each synapse is made up of N individual memristive devices. During synaptic plasticity, only one device among these N is chosen and modified to achieve the required change in synaptic strength. This design focuses on updating a single device to achieve potentiation or depression.

In contrast, the differential architecture involves two distinct sets of devices: one set represents the potentiation (G+) of the synapse, and the other set represents the depression (G-). Each of these sets contains N/2 devices. The synaptic conductance, G_{syn}, is determined as the difference between the total conductance of the G+ set and the total conductance of the G- set $G_{syn} = G+ - G-$. When there is a need for potentiation, one device from the G+ set is selected and enhanced, and when depression is required, one device from the G- set is chosen and modified.

The non-differential architecture, characterized by its simplicity and flexibility, comprises N individual memristive devices, with only one device chosen and modified during synaptic plasticity to achieve the necessary change in synaptic strength. This simplicity facilitates ease of understanding and

implementation. The architectures flexibility is evident in its capacity for independent modification of each device, offering a versatile approach to synaptic plasticity. However, notable drawbacks include the potential for mismatch or variability among the independent devices, which could compromise the precision of synaptic weight updates. Additionally, the non-differential architecture lacks inherent support for differential behavior, a limitation that may be a consideration for certain types of neural network computations where such behavior is desirable. Despite these cons, the architectures simplicity and flexibility contribute to its appeal, with careful consideration required to address potential challenges associated with device variability and the absence of inherent support for differential behavior.

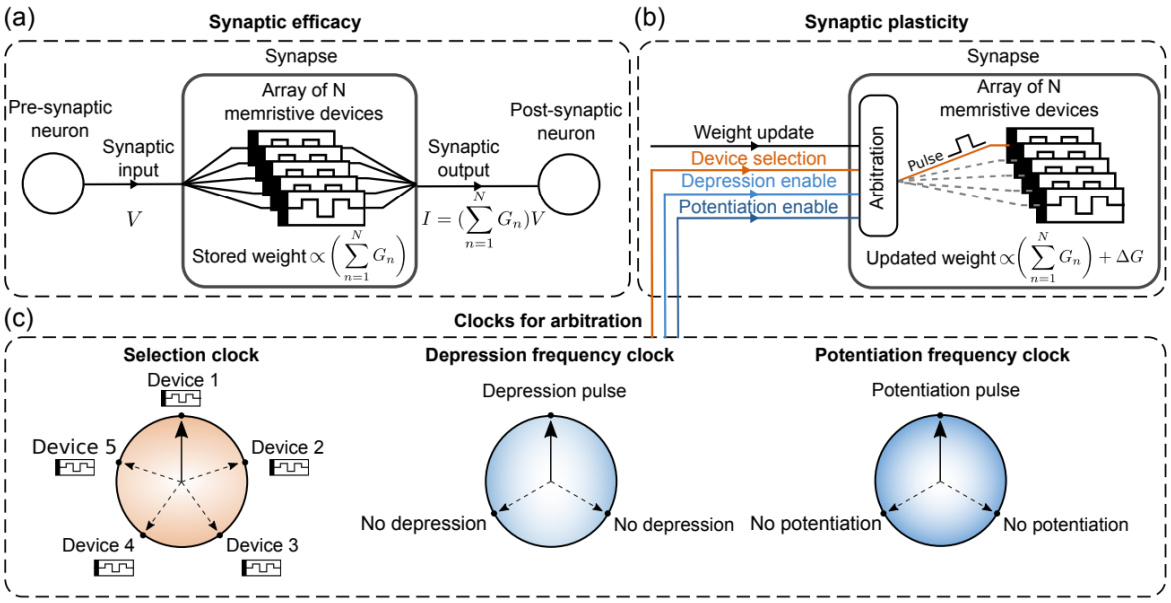


Figure 5. 4: The concept of a multi-memristive synapse involves the following key elements: (a) The net synaptic weight in a multi-memristive synapse is determined by the combined conductance (∑Gn) of multiple memristive devices. To achieve synaptic efficacy, a voltage signal, V, is applied to all these devices simultaneously. The resulting current passing through each device is added together to generate the synaptic output. (b) To implement synaptic plasticity, only one of the memristive devices is chosen during each synaptic update. This device's conductance is modified according to a learning algorithm by applying a suitable programming pulse. (c) A clock-based arbitration scheme is employed to select which device is programmed for synaptic plasticity at any given update. A global selection clock, with a duration equivalent to the number of devices representing a synapse, is used. During each synaptic update, the device indicated by the selection clock is programmed. Afterward, the selection clock is incremented by a fixed amount. Additionally, independent clocks for potentiation and depression frequencies can control the rate of potentiation and depression events. This multi-memristive synapse concept allows for flexible and dynamic adjustments of synaptic weights, making it suitable for various applications, including neural network training and emulating synaptic dynamics in spiking neural networks.(Boybat et al., 2018)

The differential architecture, of multi-memristive synapses, presents several advantages and disadvantages. On the positive side, the architecture inherently supports differential behavior, allowing for the representation of both potentiation and depression within the synapse. This feature proves

advantageous in specific neural network computations. Additionally, the relative difference between the two sets of devices, G+ and G-, offers potential mitigation of device mismatch or variability, enhancing the precision of synaptic weight updates. However, there are notable drawbacks to consider. The implementation of the differential architecture introduces increased complexity compared to its non-differential counterpart. This complexity arises from the involvement of two distinct sets of devices (G+ and G-) and necessitates additional circuitry to support the differential operation. The incorporation of this complexity may pose challenges in terms of understanding, implementation, and maintenance. Furthermore, the adoption of the differential architecture may require increased circuitry and overhead, impacting the overall design and implementation of the neuromorphic system. Despite these challenges, the architecture ability to inherently support differential behavior and mitigate device mismatch underscores its potential advantages for specific applications in neural network computations.

5.3 Programing synaptic weights

In this section, we will elucidate the conventional method for programming synaptic weights utilizing PCM (Phase Change Material) devices. These PCM devices as explained in previous chapters, consist of a slight layer of phase change material positioned between two metal electrodes, as depicted in Figure 5.5(a). This remarkable material can exist in either a state of high conductivity characterized by a crystalline structure or a state of low conductivity characterized by an amorphous structure. Typically, when these devices are first manufactured, the phase change material is in its crystalline state.

To change the conductance of the device, we apply specific current pulses. First, a high-amplitude current pulse, known as a “RESET pulse," is applied. This pulse generates enough heat (Joule heating) to melt a significant portion of the phase change material. If this pulse is abruptly stopped, the molten material quickly solidifies into the amorphous phase due to a process known as the glass transition.

To increase the conductance of the device, we use another type of current pulse called a "SET pulse." This pulse is designed to raise the temperature through Joule heating to a point above the crystallization temperature but below the melting point. This results in the recrystallization of some of the amorphous material.

The extent of crystallization depends on factors like the amplitude and duration of the SET pulse, as well as the number of such pulses applied. By gradually increasing the degree of crystallization in the amorphous region through the application of SET pulses, we can achieve a continuous range of conductance levels.

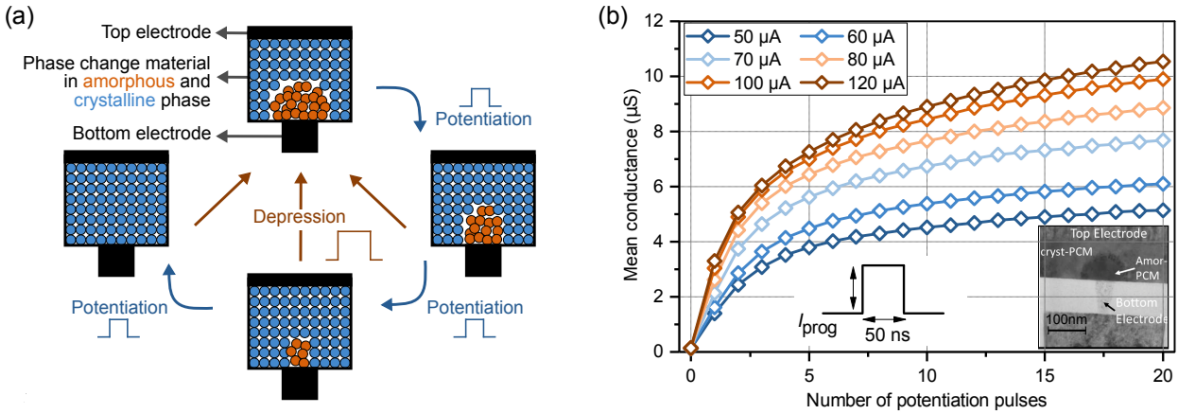


Figure 5. 5: a) A phase change memory device comprises three essential components: a top electrode, a phase change material, and a bottom electrode. When initially manufactured, the phase change material is in a crystalline state. By applying a RESET pulse, it can create an amorphous region within the material, causing a sudden decrease in its conductance. This drop in conductance occurs regardless of the initial state of the device. b)To gradually restore the crystalline state, SET pulses are applied. The evolution of the mean conductance is observed by modulating the programming current amplitude (Iprog). Each curve in the graph represents the average conductance measurements from multiple devices. The inset features a transmission electron microscopy (TEM) image displaying a phase change memory device. (Boybat et al., 2018)

In Figure 5.5(b), a noteworthy observation is presented, where the manipulation of final conductance values, akin to synaptic weights, is achieved through the adjustment of the programming current amplitude (referred to as Iprog), while keeping the pulse duration (width/fall) and the number of SET pulses constant. This intriguing finding underscores the remarkable degree of control and fine-tuning that can be exerted over the resulting synaptic weights by simply varying the intensity of the programming current, without altering other parameters such as pulse duration or the quantity of SET pulses. Nonetheless, it is important to note that achieving such modulation of synaptic weights is not always straightforward. In some instances, the need to simultaneously adjust multiple parameters, including the programming current (Iprog), pulse width, and pulse fall time, can make the programming of synaptic weights in silicon (PCM) devices a challenging and intricate task.

Furthermore, it is crucial to emphasize that beyond the programming complexity, the very process of devising and fabricating the PCM devices with the capability to effectively transfer and modulate synaptic weights presents a formidable challenge in terms of device design and fabrication. Particularly, the intricacies arise when engineering the pulse generator mechanisms, as they play a pivotal role in the precise manipulation of parameters like pulse width, pulse fall time, and programming current (Iprog). In these multifaceted challenges, there arises a pressing need to explore novel approaches aimed at achieving greater control over the final conductance levels within PCM devices.

In (Nandakumar et al., 2018) the authors present a comprehensive model of Phase-Change Memory (PCM) devices. This model effectively captures the accumulative behavior, conductance drift of PCM devices. To develop this model, an extensive experimental characterization was conducted, involving multiple PCM devices. Furthermore, the paper demonstrates the practical effectiveness of this model by using it to match experimentally observed array-level characteristics. Additionally, it highlights the model's utility in training both spiking and non-spiking artificial neural networks.

In their research, the authors began by conducting a detailed characterization of accumulative behavior resulting from the successive application of partial SET pulses in Phase-Change Memory (PCM) devices. To distinguish this accumulative behavior from conductance drift, the authors specifically examined the 50th read measurement. They graphically depicted the distribution of conductance values over pulse numbers in Fig. 5.6.

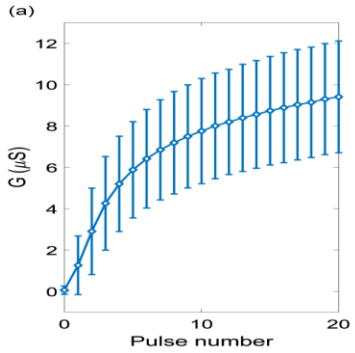


Figure 5. 6: Analyze the statistical aspects of cumulative conductance changes in relation to the number of partial SET pulses, with error bars representing one standard deviation.(Nandakumar et al., 2018)

The authors uncovered a notable trend in their observations: the average conductance change exhibited a significant magnitude at lower conductance values, gradually tapering off as conductance values increased. Additionally, a considerable degree of randomness was evident in the conductance values. This randomness primarily stemmed from the inherent unpredictability associated with the crystallization process, a phenomenon well documented in previous studies. Importantly, the authors observed that both the intra-device and inter-device variability within the array were of comparable magnitude, emphasizing the significance of this inherent randomness in PCM devices.

The authors integrate the various components of their model, encompassing accumulative behavior, conductance drift, and read noise, to create a comprehensive statistical model. This model is validated using experimental data with the aim of capturing the evolution of conductance values in a large set of devices after a specified time, T_0, following programming with a variable number of partial SET pulses.

Specifically, the objective is to determine the device's conductance, G(t), at any given time, t, after it has been initialized to approximately 0.1 μS and exposed to a sequence of 90 μA, 50 ns programming pulses with arbitrary time intervals between them. To simulate this behavior, three key quantities are recorded for each device:

- $G_i(T_0)$: The conductance after T_0 time following the *i*th programming pulse, where i ranges from 0 onwards.
- P_{mem}: A parameter that encapsulates the programming history of the device.
- t_p: The time of the last programming event, initially set to zero.

Based on the chosen initial conductance value, $G_0(T_0)$, Pmem is initialized to $P_{mem,0} = e^{-p0/\alpha}$, where p_0 represents the effective number of pulses required to achieve the initial conductance $G_0(T_0)$. Notably,

p_0 is zero for initialization around 0.1 µS, and for higher conductance values, it is determined empirically from the average conductance evolution curve presented in (Nandakumar et al., 2018).

After the initial initialization, for each subsequent programming event (Nth programming event), the P_{mem} parameter is updated. Specifically, $P_{mem,N}$ is calculated as follows:

$P_{mem,N} = P_{mem,N-1} * e^{-1/\alpha}$, where N varies incrementally starting from 1, 2, and so on.

Now, for a device with G(t) conductance that has undergone a total of N programming pulses, the conductance value G(t) at any given time t can be determined using the following method:

$$G(t) = G_N(T_0)\left(\frac{t - t_p}{T_p}\right)^{-v} + n_G$$

To validate the model, the authors initially employed it to assess the same set of experimental data that was used to establish the model parameters. Specifically, the model was applied to generate the distribution of conductance values as a function of the number of programming pulses. As depicted in Figure 5.7, the model predictions for both the mean and variance closely align with the corresponding experimental data. Moreover, it is evident that the distributions themselves exhibit remarkable similarity.

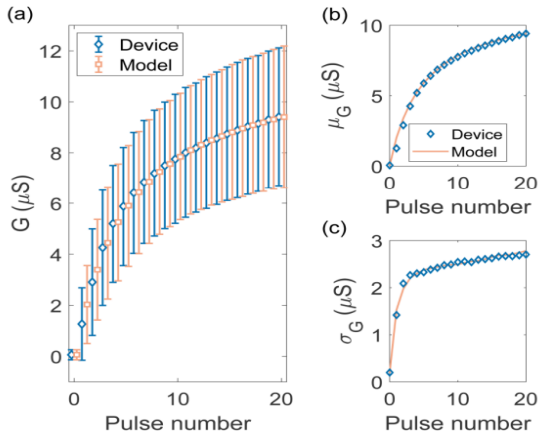


Figure 5. 7 : (a) Analyze how the conductance values acquired through the model change concerning the number of partial SET pulses and compare them to real experimental device data. (b) Examine how the mean conductance evolves in relation to the pulse count. (c) Investigate how the standard deviation of conductance varies with the number of pulses.(Nandakumar et al., 2018)

Interestingly, it is worth noting that, up to this point, there has been a notable absence of any frequency modulation technique proposed for the purpose of modulating the ultimate conductance values within phase change memory (PCM) devices. While various methods and strategies have been explored to control synaptic weights, the concept of leveraging frequency modulation as a means to achieve this goal remains largely uncharted territory. This represents a significant gap in current research and innovation within the field of PCM devices, as it highlights an untapped avenue with the potential to revolutionize the way we manipulate and fine-tune conductance levels, offering new possibilities for enhancing the capabilities of these devices.

An advantageous aspect of this technique lies in its straightforward implementation on silicon. This simplicity is attributed to the consistent application of the same pulse, eliminating the need for extensive parameter modulation. Notably, the modulation process is focused solely on frequency adjustments. This approach enhances the feasibility of integration into silicon-based systems. The minimized complexity in parameter modulation simplifies the implementation process, making the technique accessible for applications in neuromorphic computing.

This chapter presents a development in Phase Change Memory (PCM) cells by introducing a novel technique involving progressive set pulses with frequency modulation. This marks the first successful application of such modulation in PCM technology. The study demonstrates that applying a series of SET pulses results in a controlled increase in cell conductance until saturation (Gsat) is achieved. Adjusting the frequency allows precise control over Gsat, offering tailored performance for specific applications. A physics-based model, validated through simulations, attributes the conductance increase to the re-amorphization of a hollow region within the PCM material. The model mathematically describes the frequency modulation effect, aligning closely with experimental data. The research potential application in neuromorphic circuits, particularly as a synaptic device, signifies a promising strategy for advancing neuromorphic computing technologies. Overall, this study represents a significant new way to send synaptic weights later on using PCM technology, providing precise control over conductance levels with potential implications for artificial intelligence and computing.

5.4 Frequency modulation

5.4.1 Experiment

To study the conductance modulation in PCM In our research, we utilized a 16-kilobit (16kbit) 1T1R NOR array (Figure 4.14), which was constructed on a 300mm Fully Depleted Silicon-On-Insulator (FDSOI) wafer. Figure 4.3 in the study presents a Transmission Electron Microscopy (TEM) image along with a detailed description of a single Phase Change Memory (PCM) bit for reference.

This array is structured with Bitlines (BL) and Wordlines (WL), which are decoded using straightforward pass-gate Multiplexers (MUXs). These MUXs establish connections between the Input/Output (IO) ring, selector gates, and the top and bottom electrodes (TE and BE). A conventional 28nm FDSOI thin oxide NMOS (N-type Metal-Oxide-Semiconductor) serves as both the selector and current limiter within the well-established 1T1R (1 Transistor, 1 Resistor) configuration.

The heart of each PCM cell consists of a material known as $Ge_2Se_2Te_5$, we chose this material based on the results presented in Chapter 4, which demonstrate the lower drift coefficient of $Ge_2Se_2Te_5$ compared to the Ge-rich GST. This phase change material ($Ge_2Se_2Te_5$) which is sandwiched between Pt/Ti electrodes. This PCM cell is integrated between metal layers 5 and 6 within the wafer structure. Additionally, there is a heater element comprised of a narrow 5nm-wide TiN pillar that connects the BE to an underlying W (tungsten) plug.

To operate and synchronize the experimentation, an external microcontroller, specifically an Arduino Mega, has been programmed. This microcontroller collaborates with an Agilent B1500 instrument equipped with four Precision Measurement Units (PMUs). The parameter analyzer from the Agilent B1500 delivers the SET pulses in a burst sequence and communicates with the microcontroller through a trigger signal (Figure 4.20).

Upon receiving the trigger signal, the microcontroller is responsible for providing the necessary IO inputs to decode the array. It subsequently connects the required bit cell to the external IO interface. By employing this approach, the entire array can be programmed or read bit by bit, facilitating precise control and data acquisition in our experiments. (All the experimental setup is showing in Figure 4.18)

Our experimental approach, as illustrated in Figure 5.8, commenced with the crucial step of initializing the device conductances. To achieve this, we utilized a RESET pulse characterized by a specific current amplitude, denoted as I_{RESET}, which was set at 557µA. This RESET pulse was carefully designed with a pulse duration of 20ns for the rise time, a width of 500ns, and a subsequent fall time of 20ns. The purpose of this RESET pulse was to effectively reset the device to a full amorphous state, providing a standardized starting point for our subsequent investigations.

Following the initialization phase, we entered the core of our experiment. In this phase, we applied a series of partial SET pulses, each characterized by a consistent amplitude of 156 µA. These SET pulses were meticulously crafted with a pulse duration of 20ns for the rise time, a width spanning 10µs, and a fall time of 20ns. This part of the experiment was pivotal, as it induced controlled changes in the conductance of the devices, allowing us to investigate their behavior under precise modulation conditions.

Importantly, after the application of each of these SET pulses, we conducted multiple readings of the devices. Specifically, we performed a series of five readings for each pulse application. This approach to multiple readings was made in order to measure the drift after each SET.

Furthermore, we introduced an immediate conductance measurement step, which took place approximately 150 µs after the programming pulse was applied. This immediate measurement served

as an early snapshot of the conductance changes induced by the SET pulses, allowing us to capture initial dynamics and responses.

However, it is worth noting that the subsequent conductance measurements were taken at significantly longer time intervals. These measurements occurred on the order of seconds, rather than microseconds, as in the immediate measurement. This time delay between measurements was purposefully designed to observe the long-term behavior and stability of the conductance changes initiated by the SET pulses. It allowed us to track the evolution of conductance over time, uncovering any gradual saturation or further modulation effects.

Our experimental methodology was characterized by a systematic approach, from device initialization to controlled pulse application, multiple readings, and timed measurements. This comprehensive strategy was essential in capturing the full spectrum of conductance modulation dynamics in our experimental setup, providing valuable insights into the behavior of the PCM cells under various conditions.

In the course of our experimentation, we systematically collected multiple readings between the application of each set pulse, as depicted in the illustrative representation shown in Figure 5.8. This approach allowed us to observe a distinct and repeatable pattern in the behavior of the PCM cells. Upon the application of each set pulse, a noticeable increase in conductivity was consistently observed. This increment in conductivity was a direct result of the controlled programming induced by the set pulse. It is important to emphasize that this conductivity enhancement was a fundamental aspect of our study, and it highlighted the dynamic nature of the PCM cells in response to these pulses.

However, it is equally important to recognize that conductivity in the PCM cells was not solely subject to upward modulation. Concurrently, we observed a phenomenon known as "drift." Drift refers to the gradual and continuous decrease in conductivity over time. For a detailed explanation of drift, please refer to the discussion provided in Chapter 4.

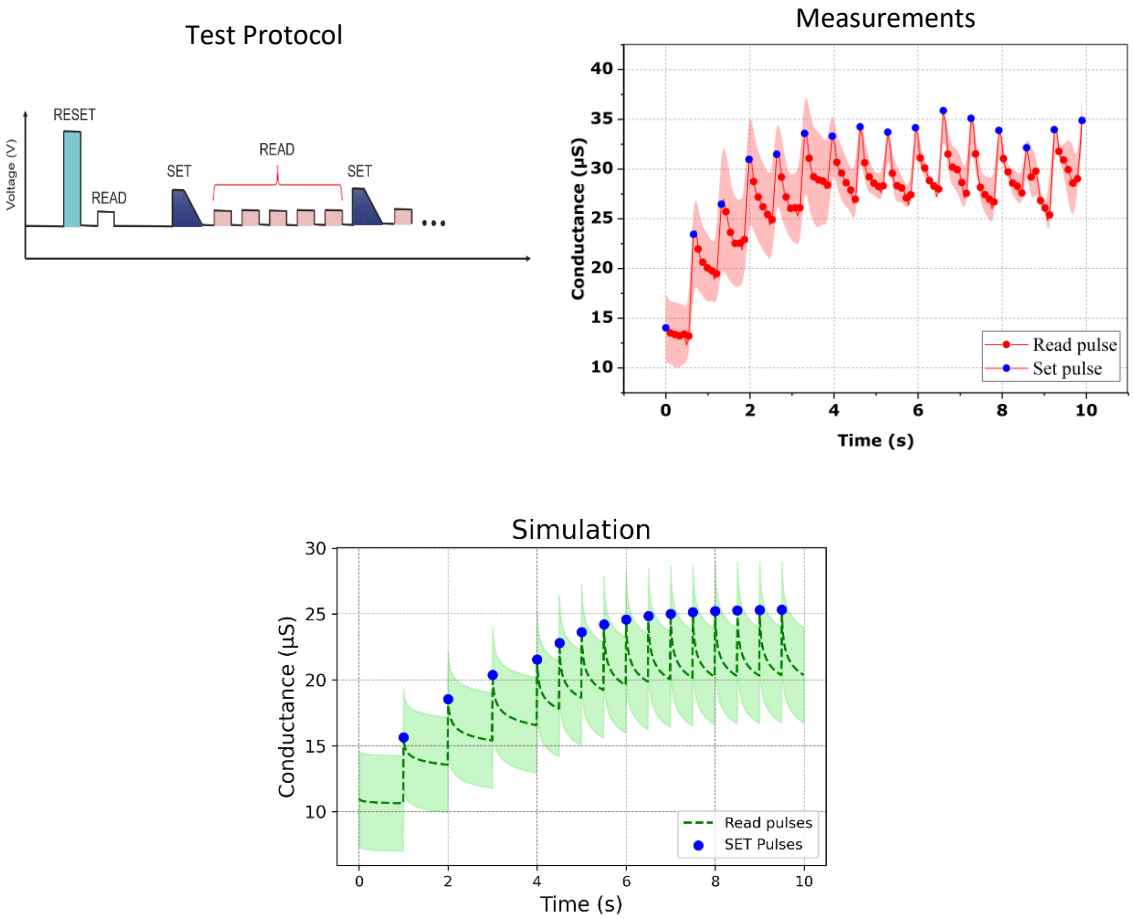


Figure 5. 8: (Right) the description of the measurements. (Left) Evolution of the average conductance and standard deviation for multiple devices. The following is the process: an initial RESET pulse is applied, and then a progressive SET pulses train with five read pulses to measure the drift in between the SET pulses (Trabelsi et al., 2022).(Bottom) simulation.

The interplay between the increase in conductivity resulting from the set pulses and the concurrent drift presented an intriguing dynamic. These two forces were consistently at odds with each other, creating a tug-of-war scenario within the PCM cells. This intricate balance between conductivity increase and drift was central to our findings.

Remarkably, this dynamic equilibrium eventually led to a noteworthy outcome the saturation of conductivity. In essence, as conductivity continued to increase with each set pulse, it was counteracted by the ongoing drift effect. This delicate balance ultimately reached a point of equilibrium, resulting in the observed conductivity saturation.

The simulation models the behavior of a PCM device shows in figure 5.8 present a spans a 10-second timeframe, with a time step of 0.01 seconds. The device is characterized by a conductance variable (conductance history) that evolves over time. The timing of external pulses, represented by an array of predetermined pulse times, triggers updates to the device's conductance. The simulation aims to capture the dynamic response of the device to these external stimuli. When a pulse occurs, the conductance is modified based on a mathematical expression that incorporates the timing constant (response rate) and a predetermined increase in conductance (G). The process accounts for the history of pulse application, ensuring a realistic response to the changing external conditions. The visualization of the simulation output includes a plot of the conductance over time, with dashed lines indicating the instances of pulse application. Additionally, the standard deviation of the conductance is calculated and depicted as a shaded region around the conductance curve. Peaks in the conductance curve, corresponding to significant changes in the device state, are identified and marked.

It is noteworthy to mention that this phenomenon of conductivity saturation aligns with previous findings reported by other researchers (Papandreou et al., 2011; M. Suri, Garbin, et al., 2013; W. Zhang & Ma, 2020). Our study not only corroborates these earlier observations but also provides a deeper understanding of the underlying dynamics and mechanisms governing conductivity modulation in PCM cells.

Our research continued with a series of measurements, following the same methodology as previously described, with one significant modification: this time, we conducted only a single reading after each SET pulse. To achieve comprehensive insights, we embarked on an extensive and systematic study. This study was designed to identify the optimal parameters for the SET pulses, specifically focusing on their rise, width, and fall times, in conjunction with variations in the programming current. This exploration of parameter combinations was instrumental in our quest to find the most effective programming strategies for these PCM cells.

By carefully evaluating the effects of different pulse parameters and programming currents, we aimed to uncover the conditions that would yield the most desirable outcomes in terms of conductivity modulation. This comprehensive investigation was crucial for gaining a deeper understanding of how these memory devices respond to various programming stimuli.

Concurrently, our study delved into the fine-tuning of pulse parameters and programming currents and we have used the following SET pulse (I_{set}=156µA 20ns/10µs/20ns, rise, width and fall times).

In our experimental protocol, we applied a series of 350 SET pulses, varying in frequency across a range from 1.5Hz to 22Hz. Each curve depicted in Figure 5.9 is derived from the average behavior observed across multiple distinct cells. This comprehensive approach allowed us to gain a comprehensive view of the responses of multiple PCM cells under differing frequency conditions.

Notably, the key observation from our findings is that the conductivity of the PCM cells displayed distinctive behavior. It reached a point of saturation (referred to as Gsat) after the application of these pulses. Exploring this phenomenon in detail, it is noteworthy that the specific Gsat value, indicative of saturation, exhibits a dependence on the applied frequency. Essentially, the frequency in this context

denotes the time interval between two successive SET pulses, and how influence on saturation values G_{sat}.

Consider the measurements present in Figure 5.9 at a relatively high frequency of 22Hz: Here, the saturation value nears an approximate 80 µS. This implies that with the rapid succession of SET pulses, the phase change material within the device experiences less time for relaxation between these pulses. This contributes to the higher saturation value observed at higher frequencies.

Conversely, when the pulse frequency is reduced to a mere 1.5 Hz, the saturation value noticeably diminishes to approximately 40 µS. In this case, the extended time intervals between successive SET pulses allow the material to partially cool and crystallize between pulses. As a result, we will have more time to the device to drift and this leads to a lower saturation value.

In essence, the frequency of applied pulses effectively dictates the temporal dynamics of the phase change process within the device, with higher frequencies favoring a more pronounced cumulative conductive effect due to limited time for relaxation and crystallization between pulses. This intricate relationship between pulse frequency and saturation thresholds opens up an attractive discover of this exploration within phase change memory devices.

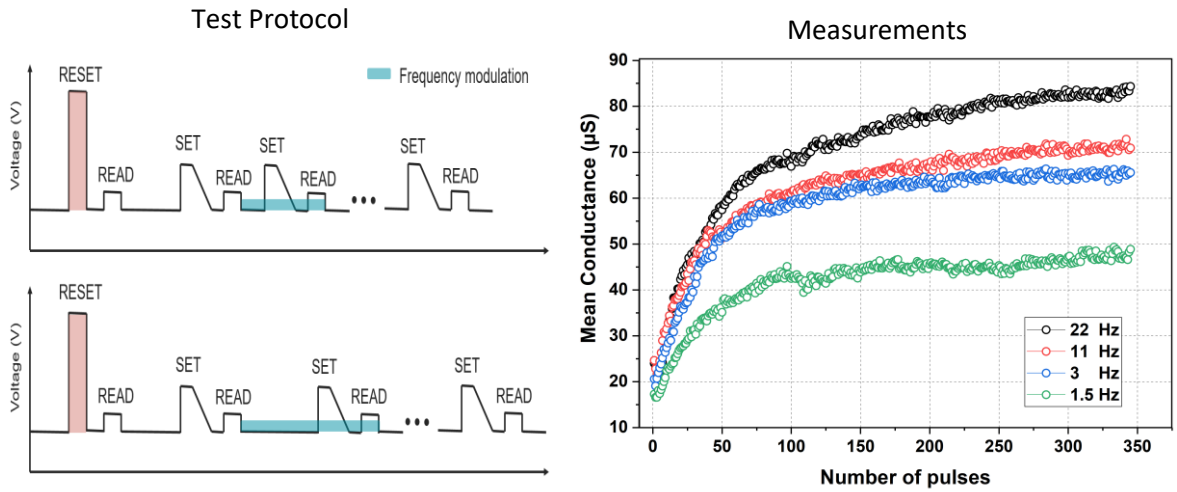


Figure 5. 9 : A schematic representation of the programming pulse train, incorporating frequency modulation, is depicted. The graph illustrates the average conductance increase observed across multiple devices throughout the experiment.

This finding unveiled a frequency-dependent effect on the conductivity saturation, a phenomenon that has significant implications for the controllability and adaptability of PCM technology. These findings offer valuable insights into the nuanced behavior of PCM cells and underscore the importance of considering frequency modulation as a means to fine-tune their conductance levels for diverse applications.

To gain a deeper understanding of the observed effect, we conducted a detailed investigation, emphasizing the critical role played by the initial pulses in our measurements. Figure 5.10 illustrates two frequencies, namely 22 Hz and 1.5 Hz, which were derived from the data presented in Figure 5.9. A focused examination of the first 75 pulses was undertaken to highlight the significance of this early stage in determining the final Gsat.

Our analysis revealed that the initial segment of these measurements significantly influences the ultimate Gsat outcome. To elucidate this, we computed the DeltaG, representing the difference between consecutive SET pulses. As depicted in Figure 5.10 on the right, the DeltaG values during the first 75 pulses for the 22 Hz frequency were notably larger than those observed for the 1.5 Hz frequency. This discrepancy in DeltaG values suggests that the 22 Hz frequency experiences less drifting time compared to the 1.5 Hz frequency, allowing for a more extended period conducive to cell crystallization.

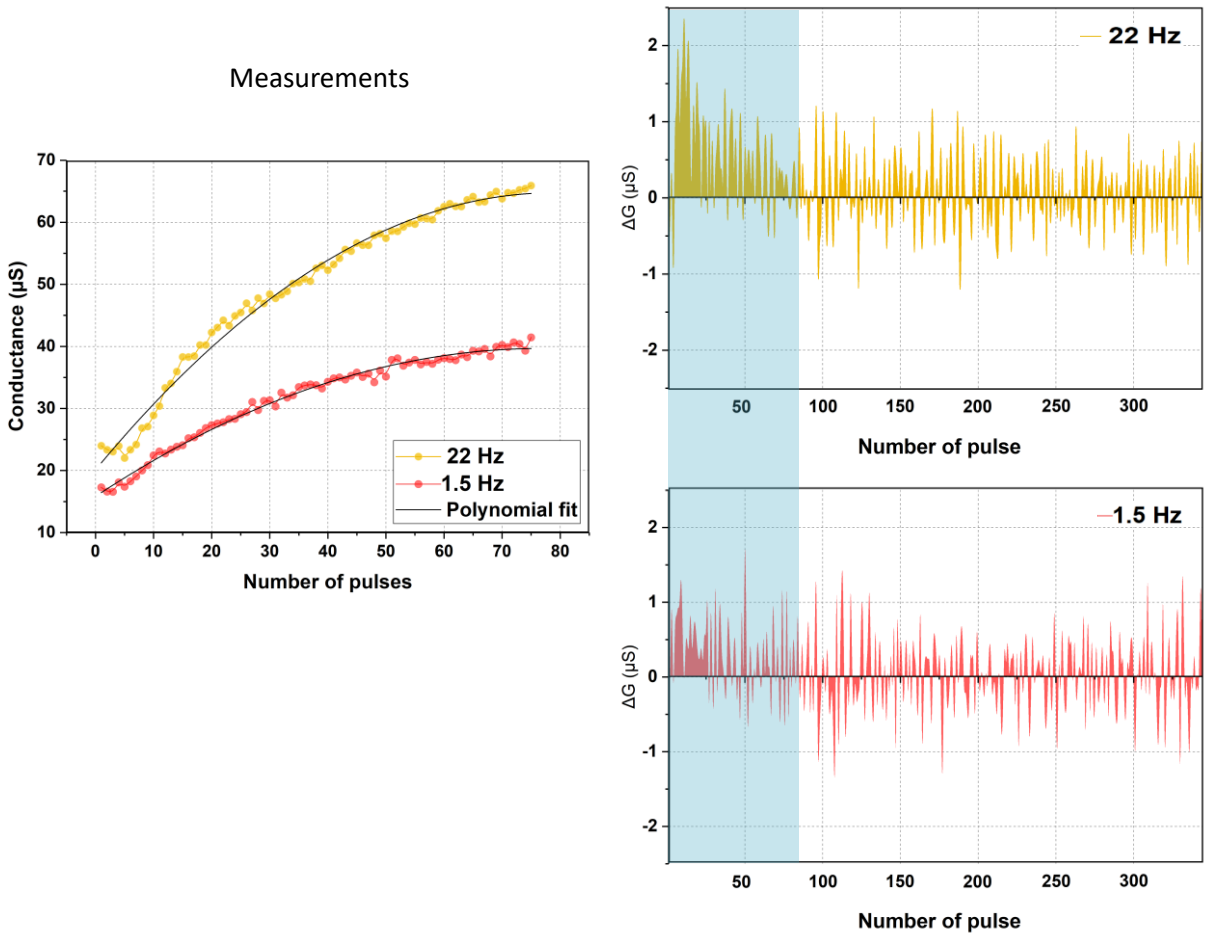


Figure 5. 10 : (Left) Snapshot of the first 75 pulse extracted from measurements made on 22 and 1.5Hz. (Right) calculated Delta G for both frequencies

Subsequent to this initial phase, a convergence in the trends of both DeltaG values becomes apparent. This convergence explains the observed saturation effect or plateau in our measurements. Notably, the consistent trends in DeltaG values beyond the initial 75 pulses indicate a shared behavior, providing insight into the saturation phenomenon.

Nevertheless, prior to drawing definitive conclusions, a prudent course of action entails conducting further in-depth investigations. To facilitate this, we initiated a comprehensive series of multiphysics simulations employing Comsol. These simulations were instrumental in obtaining valuable insights into the nuanced processes at play, particularly concerning the progressive crystallization phenomena. By leveraging the power of multiphysics simulations, we were able to delve deeper into the intricacies of the system, ensuring a more robust and informed foundation upon which to base our conclusions.

5.4.2 Simulations and measurements

In our recent publication (Cueto et al., 2023), we introduced an innovative approach for simulating Phase Change Memory (PCM) devices to investigate progressive crystallization. Our method integrates two fundamental components: the phase field model and the electro-thermal solver.

The phase field model serves as a foundational framework for simulating the behavior of phase change materials (PCMs), crucial for understanding the reversible phase transition between amorphous and crystalline states in PCM devices. At the atomic or molecular level, the phase field model mathematically describes the processes involved in this transition, including nucleation, growth, and the diffusion of atoms or molecules during the transformation.

A crucial aspect addressed by the phase field model is the transition from amorphous to crystalline states during data writing in a PCM cell. This is captured by the Allen-Cahn equation:

$$\frac{\partial \eta}{\partial t} = -L\eta \left(\frac{\partial f(\eta)}{\partial \eta} - \kappa \nabla^2 \eta \right)$$

Here, η represents the local crystallinity of the PCM material (1 in the crystalline phase, 0 in the amorphous phase), L_η is a parameter related to the interface mobility, $f(\eta)$ is the local free-energy density, and κ represents the nucleation term.

Importantly, the phase field model facilitates the representation of intermediate resistance states within PCM devices, capturing the graded synaptic strengths analogous to biological neurons. This is achieved by incorporating a nucleation model based on Classical Nucleation Theory (CNT):

$$\frac{\partial \eta}{\partial t} = \text{Nucleation Rate}$$

Complementing the phase field model is the electro-thermal solver, an essential component for simulating PCM devices that considers both electrical and thermal aspects. This solver accounts for electrical currents generating heat due to the material's resistance during data read and write operations. The solver simulates temperature dynamics within the PCM cell, considering the influence of heat on the phase transition process. The heat equation is given by:

$$\rho C_p \frac{\partial T}{\partial t} + \nabla \cdot (-k_{th} \nabla T) = \sigma (\nabla V)^2 + L \frac{dh}{d\eta} \frac{\partial \eta}{\partial t}$$

Here, σ, ρ, C_p, k_{th} and L represent density, heat capacity, thermal conductivity, electrical conductivity, and latent heat of melting, respectively.

The simulation focuses on the interplay between electrical currents, temperature fluctuations, and resulting changes in resistance within the PCM cell. By coupling the phase field model with the electro-thermal solver, this approach provides a comprehensive understanding of how these factors collectively influence resistance levels during data operations in PCM devices. This perspective is crucial for realistically replicating resistance variations, particularly in intermediate states, essential for neuromorphic computing where accurate emulation of synaptic plasticity is critical for learning and memory processes.

To model the correlation between the increase in saturation conductance G_{sat} and pulse frequency, we initiated a simulation procedure outlined in Figure 5.11. This simulation model incorporates trap-assisted tunneling, the heat equation, and nucleation/growth physics, as described comprehensively in (Cueto et al., 2023).As depicted in Figure 5.11, when the SET pulse is applied, a crucial sequence of events unfolds. The inner core of the amorphous dome undergoes a phase change, transitioning into a molten state due to the elevated temperature surpassing the material melting point. However, owing to the pulse rapid and steep falling edge (occurring over 20 nanoseconds), the inner core reverts

to its amorphous state. In contrast, a roughly spherical hollow region surrounding the inner core maintains a temperature below the melting point. It is within this region that nucleation processes occur, leading to the formation of GST (Germanium-Antimony-Tellurium) crystals, thereby enhancing conductivity.

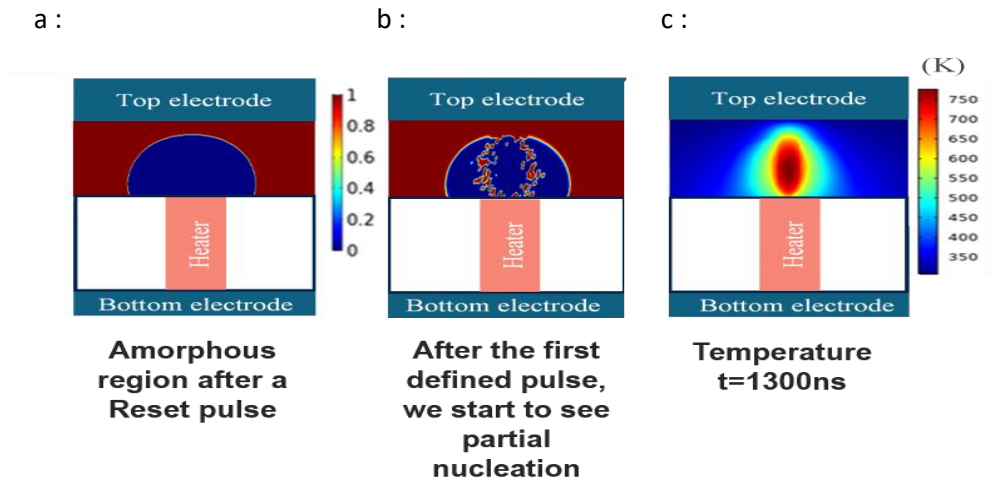


Figure 5. 11: Simulating the progressive crystallization phenomenon involves several stages. a) We start with an initial state that is completely reset. b) The inner core of the mushroom structure is subjected to a melt-quench process during each SET pulse. c) We analyze the temperature profile during an electrical SET pulse with a current of 150μA. (Trabelsi et al., 2022)

With each successive pulse, a cyclical pattern unfolds. The inner core is re-amorphized, and the volume of crystalline material around it steadily increases Figure 5.11.

In the recrystallization scenario, the increased current leads to the melting of the central part of the active domain, causing maximum nucleation at the periphery of the melted region (Fig. 5.11c). The simulation (Fig. 5.11b) shows that the high-current regime results in the re-amorphization of the GST mushroom core. Simultaneously, the surrounding GST slowly nucleates, leading to an observed increase in conductance.

In this scenario, the observed conductivity saturation can be explained by the counter effect of resistance drift in the amorphous region, which tends to decrease the conductance after each pulse. This observation provides a practical explanation for the changes in conductivity during recrystallization under high-current conditions.

In simpler terms, we can reasonably conclude that the conductivity behavior following the n+1th pulse adheres to a predictable trend, which is influenced by the established patterns observed in previous measurements. The equation that encapsulates this phenomenon can be expressed as follows:

$$G_{n+1}(t) = G_n \left(\frac{t}{t_{0T}}\right)^{-\nu} \quad (5.1)$$

- Here, G represents the conductivity of the material at a particular time point.
- The subscript "n" represents the discrete time steps or pulses, while "n+1" represents the next time step.
- The parameter " ν " and the drift pre-factor t_{0T} are essential factors that govern the behavior of G in response to these pulses.
- The drift pre-factor t_{0T} depends on the pulse period T, which is the time interval between consecutive pulses.

After each pulse, the conductivity G experiences changes as described by Equation 5.2:

$$G_1 = (1-\gamma)\, G_0 + \Delta G$$
$$G_2 = (1-\gamma)\, G_1 + \Delta G$$
$$\dots$$
$$G_n = (1-\gamma)\, G_{n-1} + \Delta G \quad (5.2)$$

- These equations illustrate how G evolves over time in response to the periodic pulses.
- The term "(1-γ)" represents the fraction of G retained from the previous time step after accounting for the drift, while ΔG represents the increase in conductivity due to additional crystallization.

Equation 5.3 defines γ:

$$(1-\gamma) := \left(\frac{t_{0T}}{T}\right)^{\nu} \quad (5.3).$$

- Here, γ quantifies the fraction of G retained after each pulse.

- It depends on the ratio of the drift pre-factor t_{0T} to the pulse period T raised to the power ν.

At the point of saturation, the increase in conductance is in equilibrium with the drift that occurred over a period, leading to $G_n=G_{n+1}=G_{sat}$. Equation 5.4 represents this equilibrium:

$$G_{sat} = (1-\gamma)\, G_{sat} + \Delta G$$

- This equation shows that at saturation, the retained conductance (G_{sat}) is in balance with the conductivity increase (ΔG) after each pulse.

Solving for γ, we obtain:

$$\gamma G_{sat} = \Delta G \quad (5.4).$$

In the practical implementation of our research, we conducted a series of drift experiments under carefully controlled conditions at room temperature. These experiments involved subjecting the material to a predefined set of pulses, as illustrated in Figure 5.12.

To determine the drift coefficient, denoted as ν and representing the slope of the drift curve, we employed a well-established analytical approach. This method entailed a meticulous analysis of the experimental data for over then 4000 devices, comparing the observed changes in resistivity after each pulse with the expected behavior as predicted by the drift equation.

Remarkably, our data analysis revealed a consistent pattern. The fraction of resistivity retained after each pulse, extracting the slope of each curve which corresponds to the drift coefficient $\boldsymbol{\nu}$, exhibited remarkable uniformity across a wide range of resistances showed in figure 5.12b. This observation is particularly noteworthy because it indicates that ν remains relatively insensitive to variations in experimental conditions and the specific characteristics of the material under examination.

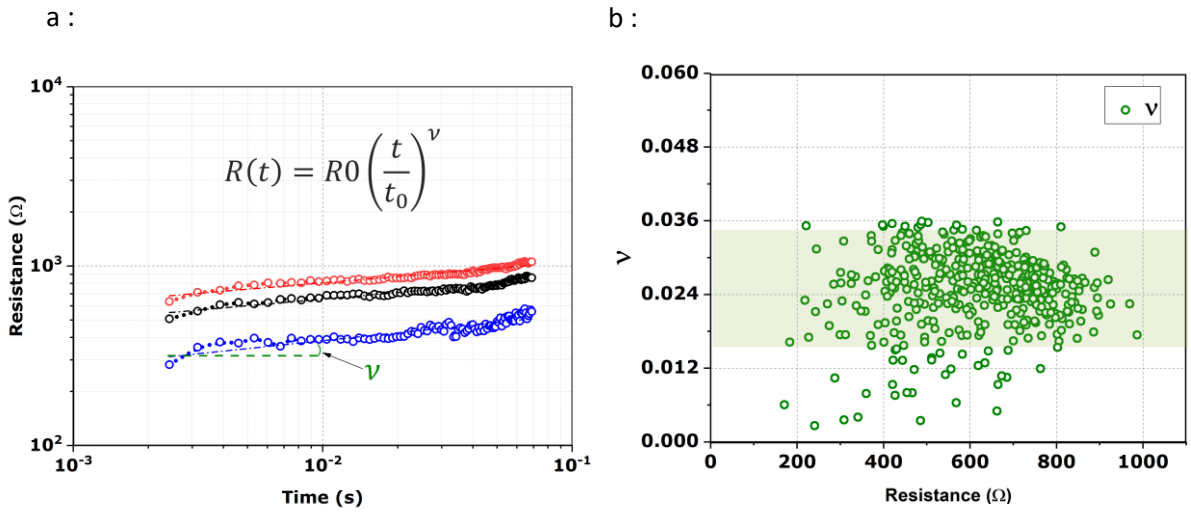


Figure 5. 12: The drift coefficient, which pertains to the rate of change in a certain parameter under room temperature conditions, underwent extensive measurements involving 4000 different devices. These measurements were conducted meticulously and resulted in an average drift coefficient value of 0.025. (Trabelsi et al., 2022)

In simpler terms, we have observed that the material response to the applied pulses displays a remarkable level of robustness and predictability. This consistency is best exemplified by the mean value of ν, which we painstakingly determined to be 0.025. Notably, when compared with other materials discussed in Chapter 4,ν, representing the slope of the drift curve.

Resistance drift is a common issue in various materials and devices, and it is a main focus of our research. Despite this ongoing challenge, our findings show a positive perspective. The material consistently handles resistance drift well, as seen in the steady value of ν.

This resilience, revealed through our detailed exploration, holds promise and goes beyond our specific study. It helps us understand the complex principles governing these systems and emphasizes the important role of resistance drift in our model.

As, we have uncovered a valuable parameter, ν, that helps us understand how a material's conductivity changes over time, especially when subjected to pulses. This parameter serves as a cornerstone in our analysis.

We utilize ν in Equation 5.2 to fit our experimental data showing in filled circles symbol, as demonstrated in Figure 5.13. The fit parameter we extract from this analysis, ΔG = 1.1μS, aligns remarkably well with the observed increase in conductance, as seen in Figure 5.13. This alignment is a crucial validation of our model's effectiveness in capturing real-world behavior.

Additionally, we extend the utility of our model to derive values for t_{0T}. Each value of t_{0T} corresponds to a specific period T, offering a customized insight into how the material responds to different pulse frequencies. This empirical approach has proven highly successful and is consistent with our experimental data, highlighting the robustness of our model.

Now, what makes this particularly intriguing is that the model's predictions closely match the experimental outcomes. This suggests that Equation 5.4 serves as a powerful analytical tool for connecting pulse frequency to the saturation conductance level (G_{sat}).

In Figure 5.13.b, we showcase our model expectations for G_{sat}. Here, we creatively represent t_{0T} as a power law function of the period T. This ingenious approach enables us to programmatically control the cell's conductance by adjusting the frequency of the applied pulses. Importantly, this modulation approach does not require us to vary the pulse amplitudes or engage in complex program/verify routines.

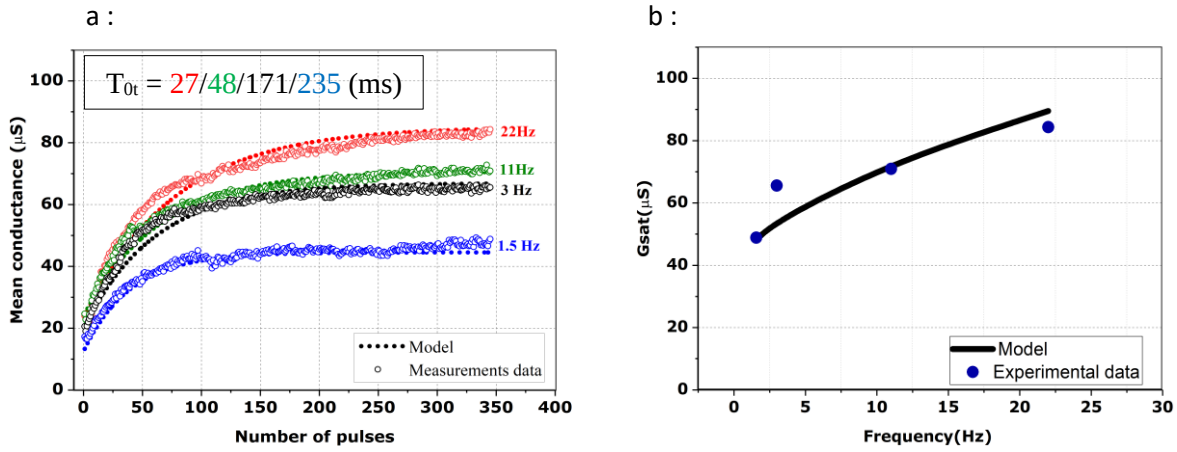


Figure 5. 13: (a) illustrates the average increase in conductance across 20 devices during the conducted experiment. This data is fitted with a ΔG value of 1.1µS, and the corresponding t0T values are provided in the accompanying box. b), there is a representation of the relationship between G_{sat} (saturation conductance) and the frequency of applied pulses. (Trabelsi et al., 2022)

In essence, Equation 5.4 of our model provides us with efficient means to map and control the conductance levels of the material in response to different pulse frequencies. This finding holds significant promise for various applications and contributes to a deeper understanding of the fundamental principles leading these intricate systems.

It might initially seem perplexing that the saturation conductance (G_{sat}) does not exhibit any dependence on the initial conductance value, G_0. However, this apparent paradox can be rationalized by considering the fundamental nature of G_{sat}. G_{sat} is the point at which the opposing influences of drift (the tendency for the conductance to change over time) and the increase in conductance reach an equilibrium. Therefore, by its very definition, G_{sat} should not be intricately linked to G_0, the starting point of the conductance.

To experimentally validate this concept, we designed an additional experiment. In this experiment, we meticulously prepared 20 cells, each with a distinct initial conductance value, spanning a range from 18 to 43µS (denoted as G_0). We subjected all these cells to the same pulse train, maintaining a consistent pulse frequency of 22 Hz.

The compelling results, as visually represented in Figure 5.14, provide empirical evidence that reinforces our understanding. These curves, which represent the average behavior of multiple cells for each initial G_0 value, intriguingly converge and saturate at approximately $G_{sat} \approx 80\mu S$. This outcome clearly illustrates that, regardless of the diverse starting points (G_0), all cells ultimately reach the same saturation conductance level G_{sat}.

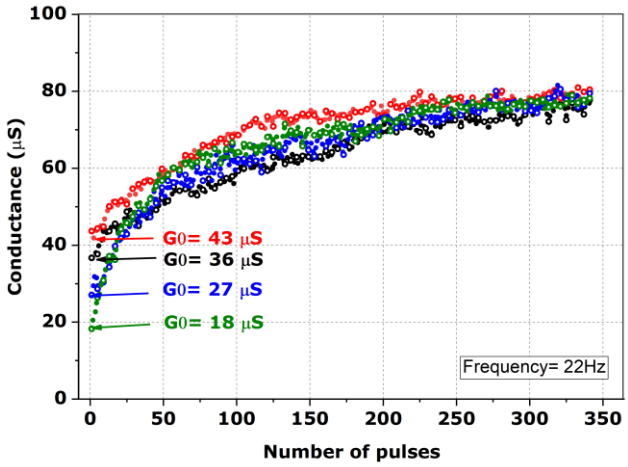


Figure 5. 14 : Starting with distinct initial conductance values (G0), when subjected to SET pulses of identical frequency, the data reveals that an equivalent Gsat (saturation conductance) is attained. The reported data represents the average results gathered from 20 devices. (Trabelsi et al., 2022)

In essence, this experimental confirmation underscores the intrinsic independence of G_{sat} from G_0. It demonstrates that G_{sat} is a characteristic property determined by the interplay of various factors within the material and the pulse train, and it remains constant regardless of the specific initial conductance values. This insight not only deepens our understanding of the underlying principles governing these systems but also has practical implications for the control and modulation of conductance levels in such devices.

Conclusion

In conclusion, the exploration of frequency modulation of conductance levels in phase-change-based systems represents a significant stride in advancing the understanding and application of multi-memristive concepts. The key attribute of phase change, as discussed in this chapter, underscores its pivotal role in shaping the programmability of synaptic weights, offering promising avenues for neural network applications.

The concept of multi-memristive systems introduces a nuanced perspective on memory and computation, demonstrating the potential for enhanced information processing capabilities. The chapter delves into the intricacies of programming synaptic weights, showcasing the versatility of phase-change materials in facilitating this crucial aspect of neural network functionality.

The focal point of this chapter, frequency modulation, emerges as a powerful technique for manipulating conductance levels in phase-change systems. Through both experimental and simulated approaches, the effectiveness of frequency modulation in achieving desired conductance changes has been demonstrated and measured. These findings contribute valuable insights to the practical implementation and optimization of frequency-based modulation techniques in the context of phase-change systems. The amalgamation of experimental results and simulations not only validates the feasibility of frequency modulation but also provides a comprehensive understanding of its potential applications. This chapter bridges the gap between theoretical concepts and empirical evidence, paving the way for future research and innovations in the realm of frequency-modulated conductance in phase-change systems. In essence, the exploration of frequency modulation in this chapter adds a crucial layer to the evolving landscape of memristive systems. By elucidating its effectiveness in programming synaptic weights, this work contributes to the broader field of neuromorphic computing, shows new possibilities for efficient and adaptive to program PCM cells.

Résumé

- Introduced and conceptualized multi-memristive systems.
- Demonstrated efficacy in manipulating conductance levels.
- Explored frequency modulation in phase-change systems.
- Delved into programming synaptic weights with phase-change materials.
- Validated frequency modulation through simulations, experiments and measurements.
- Impactful research guiding future studies and innovations.

Chapter 6

Convolution neural network inference using frequency modulation in computational phase-change memory

"In the digital ballet of computational phase-change memory, convolutional neural networks take center stage, pirouetting through layers of data to choreograph an exquisite performance of intelligent inference—a seamless blend of memory, computation, and cognitive artistry."

Dr. Maya Rodriguez, AI Systems Architect, NVIDIA Research.

6. Convolution neural network inference using frequency modulation in computational phase-change memory

In this chapter, we delve into the advancements in frequency modulation research within real-world scenarios. Our focus shifts from theoretical frameworks to practical applications, specifically exploring the manipulation of conductance in Phase Change Memory (PCM) devices. Through the use of identical SET pulses with varying frequencies, we uncover a method for precisely controlling conductance levels. This chapter highlights the simplicity and practicality of this method, demonstrating its transformative impact on PCM technology in tangible, everyday situations.

At the heart of this phenomenon lies a delicate equilibrium between two fundamental processes: nucleation and growth. Through extensive multiphysics simulations, we have examined these processes to gain a deep understanding of their interplay. With each applied pulse, we observe the re-amorphization of the central core composed of the phase-change material GST. It starts with nucleation in different locations and then the growth of nuclei in the surrounding materials. This leads to core re-amorphization.

This unique behavior offers a remarkable advantage the ability to reset the drift equation after every pulse, ultimately leading to a point of conductance saturation. At this equilibrium point, the increase in conductance harmoniously counterbalances the drift-induced decrease. What makes this system truly remarkable is its adaptability; by simply adjusting the frequency of the applied pulses, we can shift this equilibrium point, granting us precise control over the final conductance levels.

A remarkable feature of this technique is its complete independence from the initial conductance value. Instead, it relies solely on the carefully chosen pulse frequency, providing a level of control that was previously unattainable. This technique goes beyond our experimental setup and holds great promise for neuromorphic-inspired synaptic circuits.

In the context of neuromorphic computing systems, where PCM conductance values can directly represent synaptic weights, this approach eliminates the need for intricate program/verify routines and tailored programming pulses. As a result, it simplifies circuit design and operation significantly. This groundbreaking discovery marks a significant stride toward the development of more efficient and streamlined neuromorphic computing systems, revolutionizing the way we approach synaptic weight manipulation and paving the way for a new era of computational possibilities. By applying this innovative method to practical use cases, we uncover its potential for delivering substantial advantages and revolutionizing various fields and industries.

6.1 Convolution neural network

To evaluate the frequency modulation technique, we developed a Convolutional Neural Network (CNN) customized for MNIST dataset (Deng, 2012) for Handwritten Digit Classification. This deep learning model is made to handle complex task of automatically recognizing and categorizing handwritten digits within the MNIST dataset. The MNIST dataset, a fundamental benchmark in computer vision research, comprises a collection of 28x28 pixel grayscale images. Each image represents a handwritten digit ranging from 0 to 9.

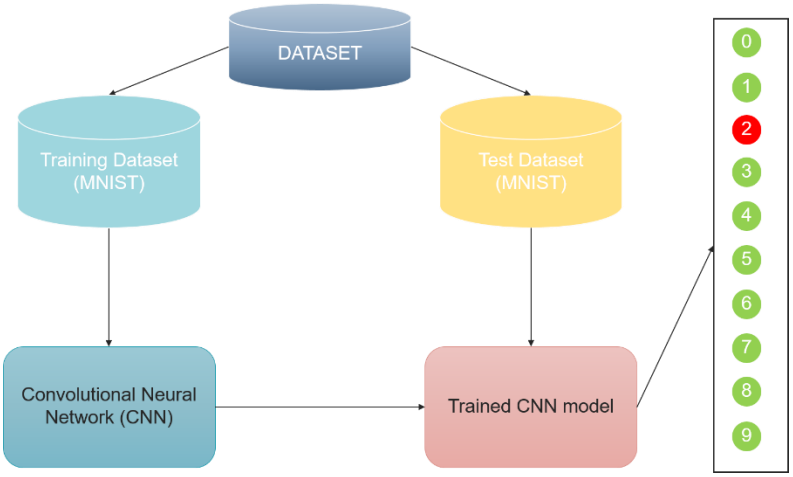


Figure 6. 1: Complete sequence of Convolutional Neural Network (CNN) workflow.

CNNs are the ideal choice for this task due to their intrinsic ability to extract and comprehend intricate local patterns and hierarchical features present in images. This architecture is tailored to exploit the spatial relationships and pixel correlations within these images, making it highly effective for digit recognition.

The typical architecture of such a CNN consists of multiple layers, each with a specific role. Convolutional layers employ small filters to slide over the input image, effectively detecting features like edges, corners, and textures. These layers enable the network to build a hierarchical representation of the digit's features. Subsequently, pooling layers reduce the spatial dimensions, focusing on the most salient information while reducing computational complexity.

The final part of the CNN is composed of fully connected layers, which act as a classifier. These layers process the extracted features and make a decision about which digit the input image represents. During the training process, the network learns to adjust its internal parameters through a process called backpropagation, optimizing them to minimize the error in classification. This training is accomplished by iteratively presenting the network with labeled examples from the MNIST dataset and adjusting its parameters until it achieves a high level of accuracy.

CNNs are good at recognizing digits in MNIST. This shows how deep learning used for understanding images. It is not just about digits; CNNs are used in many fields like recognizing images, finding objects, and analyzing medical images. Basically, CNNs are a big deal in the world of computer vision and machine learning, helping to make sense of visual information in all sorts of scientific areas.

The CNN architecture employed for this task is depicted in Figure 6.2.

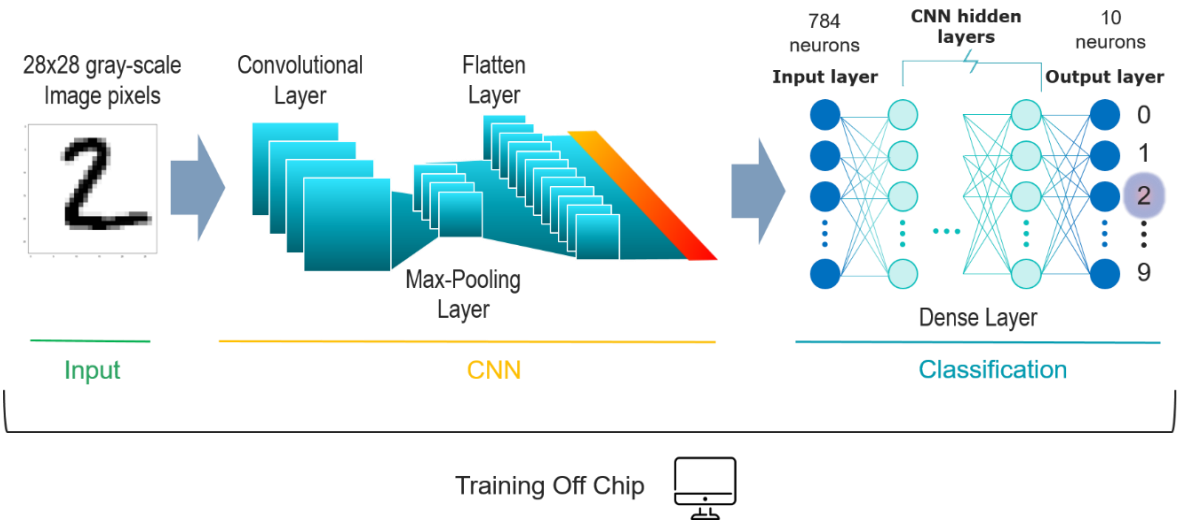


Figure 6. 2 : An illustrative representation of the intricate Convolutional Neural Network (CNN) architecture employed for the task of Handwritten Digit Classification. The network's structure includes convolutional layers responsible for detecting essential features, followed by activation functions like Rectified Linear Unit (ReLU) (Romanuke, 2017) it can be expressed mathematically as follows: f(x) = max(0,x) for non-linearity. Subsequently, max-pooling layers are employed to down sample and focus on salient information. The architecture concludes with fully connected layers for precise digit classification. This diagram encapsulates the essence of how deep learning and neural networks are harnessed to decipher handwritten digits with remarkable accuracy.

For the architecture, we used:

Convolutional layer:

- The convolutional layer functions as an ensemble of compact filters, specifically 3x3 weight matrices, systematically traversing the input image. These filters conduct localized assessments, analyzing discrete regions.
- Within each scrutinized region, a convolution operation occurs, involving the multiplication of filter values with corresponding pixel values and subsequent summation. The numerical output contributes to the output feature map.
- Convolutional filters aim to identify patterns in the image. In early layers, they detect rudimentary patterns like edges or corners, progressing to discern more intricate features in higher layers.

- The integration of multiple filters enables the convolutional layer to concurrently discern diverse features, enhancing its capability to capture nuanced information.

Max-Pooling layer:

- After the convolutional layer, the max-pooling layer is applied. Max-pooling serves two main functions: down sampling and feature selection.
- It divides each feature map into non-overlapping regions (e.g., 2x2 grids) and selects the maximum value from each region. This down sampling reduces the spatial dimensions of the feature maps, making them smaller.
- Down sampling helps reduce computational complexity and memory usage, making the model more efficient.
- Additionally, by selecting the maximum value, the max-pooling layer retains the most important information from each region while discarding less relevant details.

Flatten layer:

- After max-pooling, the data is still in a 2D grid format. The flatten layer's role is to reshape this data into a 1D vector.
- It essentially takes all the values from the feature maps and lines them up sequentially, preparing the data for input into fully connected layers.
- This transformation retains the hierarchical information learned by the convolutional layers but presents it in a form suitable for traditional neural network layers.

Dense layer (Output Layer):

- The dense layer is a traditional fully connected neural network layer. Each neuron (or unit) in this layer is connected to every neuron in the previous layer.
- The number of neurons in this layer corresponds to the number of classes in your classification problem (10 in our case). Each neuron represents a different class.
- The output of each neuron is a raw score, and these raw scores are then transformed using the softmax activation function.
- Softmax converts the raw scores into a probability distribution, where each value represents the probability that the input image belongs to the corresponding class.
- During inference, we can choose the class with the highest probability as the predicted class.

The model training process involves adjusting the internal parameters (weights and biases) using an optimization algorithm (Adam in our case) to minimize the loss function (categorical cross-entropy in

our case). This optimization process occurs over multiple iterations or epochs, gradually improving the model ability to make accurate predictions on the training data.

6.1.1 Understanding the Backpropagation Process in CNN Training

In the realm of machine learning and neural networks, the backpropagation algorithm stands as the bedrock of training. Specifically, in the context of Convolutional Neural Networks (CNNs), backpropagation is the driving force that enables these networks to learn from data and make accurate predictions, particularly in image classification tasks.

In this comprehensive explanation, the intricate steps involved in the backpropagation process within the developed code are dissected. This exploration delves into how CNNs progress through forward passes, compute loss, execute backward passes, adjust parameters, and iterate across epochs and batches. Through this detailed breakdown, light is shed on the mechanisms enabling a CNN to recognize complex patterns and features within images, ultimately leading to its ability to make informed classifications.

The backpropagation process in the context of training a Convolutional Neural Network (CNN) for handwritten digit recognition:

- **Forward pass:**

- At the start of each training iteration (epoch), a batch of training images (`train_images`) is passed through the neural network. These images are processed layer by layer.

Convolutional Layers: The convolution operation is a mathematical way of combining two functions to produce a third. In the context of CNNs, it involves sliding a filter or kernel (g) over the input image (f). At each position, the elements of the filter are multiplied element-wise with the corresponding elements of the input image, and the results are summed up. This process is repeated across the entire image, producing a feature map that highlights certain patterns. The equation $(f * g)(x, y)$ represents this operation at a specific location (x, y) in the feature map.

$$(f*g)(x,y)=\sum_{i=1}^{m}\sum_{j=1}^{n} f(i,j)\cdot g(x-i, y-j)$$

- Max-Pooling Layers: Max-pooling is a downsampling operation that reduces the spatial dimensions of the feature maps. The max-pooling operation $max\text{-}pooling(x,y)$ involves taking the maximum value within a defined neighborhood. If the pooling window is of size $(k\times l)$ the operation finds the maximum value among the elements in the window at each position (x,y) in the feature map. This helps retain the most important information from the feature maps while discarding less relevant details.

$$max\text{-}pooling(x,y)=max_{i=1}^{k} = max_{j=1}^{l} = f(x\times k+i, y\times l+j)$$

$f:$ represents the feature map.

x, k **:** are the spatial coordinates in the output feature map.

k, l**:** are the dimensions of the pooling window.

-Flatten Layer: The flatten layer is responsible for converting the 2D feature maps into a 1D vector. This is achieved by rearranging the elements of the matrix into a single line. The flattened vector maintains the hierarchical information learned by the convolutional layers but represents it in a format suitable for feeding into fully connected layers. The specific reshaping operation depends on the dimensions of the feature maps.

$$v = [\, f(1,1)\, f(1,2) \quad \dots \quad f(1,q)\, f(2,1) \quad \dots \quad f(p,q)\,]^{T}$$

Loss calculation:

- After the forward pass for each image in the batch, the model produces a set of class probabilities using the softmax activation function in the output layer. Given an output vector Z with elements Z_i for I =1,2,...,C (where C is the number of classes), the softmax activation function is defined as:

$$Softmax\,(z)_i = \frac{e^{Zi}}{\sum_{j=1}^{C} e^{Zj}}$$

e^{Zi} : computes the exponential of the raw score (model's initial output) Z_i

$\sum_{j=1}^{C} e^{Zj}$: calculates the sum of the exponentials of all the raw scores in the vector. This step normalizes the values, ensuring that the probabilities sum up to 1

- These probabilities are compared to the true labels for each image (one-hot encoded vectors) to compute the categorical cross-entropy loss. The loss quantifies how well the models predictions match the actual labels. Given the true one-hot encoded label vector y and the predicted probability vector $\hat{y}$ (output of the softmax), the categorical cross-entropy loss is calculated as:

$$Categorical\ Cross\text{-}Entropy\ Loss = -\sum_{j=1}^{C} y_i \,.\, log(\hat{y}_i)$$

(Log) is applied element-wise to the predicted probability vector $\hat{y}$. This operation penalizes the model more when it is confident about an incorrect prediction. If the predicted probability for the correct class is high, $log(\hat{y}_i)$ approaches 0, if it is low, the value diverges to negative infinity.

The true label vector y (which is one-hot encoded) is multiplied element-wise with the logarithmic-transformed predicted probabilities. This operation ensures that only the element corresponding to the correct class contributes to the overall loss.

➢ **Backward pass (Backpropagation):**

Computation of Gradients: For a parameter θ, the gradient $\frac{\partial Loss}{\partial \theta}$ is computed.

- $\frac{\partial Loss}{\partial \theta}$: This represents the partial derivative of the loss with respect to the parameter.

- Loss: The overall loss of the model, typically calculated using a loss function like categorical cross-entropy.

- θ : A parameter in the model, such as a weight or bias.

Layer-wise Backward Computation: For each layer, the local gradient is calculated using $\frac{\partial Loss}{\partial z}$

- $\frac{\partial Loss}{\partial z}$: The local gradient, indicating how the loss changes concerning the weighted sum (z) of inputs to the layer.

- z : The weighted sum of inputs to the layer, often calculated as $z = w \times a_{prev} + b$, where w is the weight matrix, a_{prev} is the output from the previous layer, and b is the bias.

Chain Rule for Parameter Gradients: $\frac{\partial Loss}{\partial \theta} = \frac{\partial Loss}{\partial z} \times \frac{\partial z}{\partial \theta}$

- $\frac{\partial Loss}{\partial \theta}$: The gradient of the loss with respect to the parameter θ.

- $\frac{\partial Loss}{\partial z}$: The local gradient calculated in the layer-wise backward computation.

- $\frac{\partial z}{\partial \theta}$: The sensitivity of the weighted sum z with respect to the parameter θ.

Adjustment of Parameters (Adam Optimizer Update):

$$\theta new = \theta old - \frac{learning_{rate} \times \hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$$

Where :

- θnew : The updated value of the parameter θ.

- θold : The current value of the parameter θ.

- $learning_{rate}$: A hyperparameter controlling the step size of the update. It adjusts parameters in the opposite direction of their gradients.

- $\hat{m}_t$: The biased first moment estimate (mean) of the gradient, computed as a moving average.

- $\hat{v}_t$: The biased second raw moment estimate (uncentered variance) of the gradient, also computed as a moving average.

- t: The time step.

- $\in$: A small constant to prevent division by zero.

Iteration over epochs and batches:

- The training process is organized into epochs, and during each epoch, the entire training dataset is processed in smaller batches.

- For each batch, the forward pass, loss calculation, backward pass, and parameter updates are performed.

- This process is repeated for multiple epochs, allowing the model to iteratively adjust its parameters and learn the underlying patterns in the data.

Convergence and stopping criteria:

- Training continues until certain stopping criteria are met. These criteria can include a maximum number of epochs ($Epochs_{max}$), achieving a satisfactory level of accuracy ($accuracy_{satisfactory}$), or a decrease in the loss reaching a plateau.

- The goal is for the model to converge, meaning that its parameters reach a stable state where further training does not significantly improve performance.

In essence, backpropagation is a complex but highly effective algorithm for training neural networks. It learns by iteratively adjusting model parameters based on gradients, making predictions, and comparing them to the true labels. This iterative process continues until the model performance converges to an acceptable level. The combination of convolutional layers, pooling, and dense layers in our CNN architecture enables the model to learn hierarchical features from images, which is especially powerful for tasks like image classification.

The following description elucidates the Backpropagation algorithm, which can be succinctly stated as:

Algorithm 1: Backpropagation

```
1. Initialize initial weights for all network inputs and outputs with small random values, typically in the range of -1 to 1.
2. Repeat the following steps until the maximum number of iterations is less than the specified value, and the Error
                Function is greater than the specified threshold:
3.   For each pattern in the training set, present the pattern to the network.
4.     //Forward propagate the input through the network:
5.       For each layer in the network:
6.         For each node in the layer:
7.           1. Compute the weighted sum of the nodes inputs.
8.           2. Add the threshold to the sum.
9.           3. Compute the activation for the node.
10         End
11.      End
12.
13.    //Backward propagate the errors through the network:
14.
15.      For each node in the output layer,
16.         calculate the error signal.
17.      End
18.
19.      For all hidden layers:
20.        For each node in the layer:
21.          1. Compute the signal error for the node.
22.          2. Update the weights of each node in the network.
23.        End
24.      End
25.
26.    //Calculate the global error using the Error Function.
27.
28.  End
29.End the loop when the maximum number of iterations is less than the specified value, and the Error Function is
                greater than the specified threshold.
```

In the field of machine learning and computer vision, Handwritten Digit Recognition is a specialized area focused on automatically discerning and categorizing handwritten numerals. This technology is crucial for various industries, playing a key role in tasks such as postal code recognition, bank check processing, and detailed analysis of digitized documents.

The core of this technology lies in its ability to make sense of the diverse array of handwritten digits encountered in real-world scenarios. To achieve this, the Handwritten Digit Recognition algorithm is subjected to a rigorous training regimen, an iterative process comprising numerous training epochs. During this journey, the algorithm is exposed to an extensive dataset brimming with handwritten digits, ranging from the elegant loops of number "0" to the sharp angles of number "7."

Algorithm: Cnn_handwritting_digit

```
1.  Import {…}
2.  # Set the number of filters in the convolutional layer
3.  num_filters =   # Number of filters
4.  # Set the size of the filters in the convolutional layer
5.  filter_size =   # Filter size
6.  # Set the size of the pooling operation
7.  pool_size =   # Pooling size
8.  # Create a Sequential model
9.  model = Sequential([  # Define a Sequential model
10.     Conv2D(num_filters, filter_size, input_shape=(28, 28, 1)),  # Convolutional layer
11.     MaxPooling2D(pool_size=pool_size),  # Max pooling layer
12.     Flatten(),  # Flatten layer
13.     Dense(10, activation='softmax'),  # Dense layer for classification
14. )
15. # Compile the model using the Adam optimizer, categorical crossentropy loss, and
    accuracy metric
16. model.compile('adam', loss='categorical_crossentropy', metrics=['accuracy'])
17. # Train the model on the training data with 11 epochs and use validation data
18. history = model.fit(
19.     train_images,  # Training images
20.     to_categorical(train_labels),  # One-hot encoded training labels
21.     epochs=11,  # Number of training epochs
22.     validation_data=(test_images, to_categorical(test_labels)),  # Validation data
23. )
```

As the algorithm progresses through epochs, it improves in discerning intricate patterns, subtleties, and distinctive features that differentiate each digit. During each epoch, the model refines its internal parameters a complex set of mathematical functions governing how it interprets and distinguishes handwritten digits. This iterative process involves refining the model over time, thereby enhancing its capacity to generalize learning and accurately recognize handwritten digits.

The result of these training epochs is a Handwritten Digit Recognition model powered by machine learning. With refined knowledge, the model swiftly and accurately predicts the identity of digits in new handwritten input. This capability significantly enhances efficiency in tasks like data entry and document processing, where both speed and accuracy are crucial.

Moreover, the impact of Handwritten Digit Recognition extends beyond its immediate applications. It serves as a fundamental building block for more advanced technologies, such as Optical Character Recognition (OCR), which broadens its scope to the interpretation of entire words and sentences. Additionally, it serves as a cornerstone in digit-based machine learning tasks, paving the way for

advancements in fields ranging from natural language processing to autonomous robotics. In essence, Handwritten Digit Recognition is a testament to the capabilities of machine learning and its potential to transform how we interact with handwritten information in the digital age.

6.1.2 Model training

Upon executing our algorithm, a critical step in the evaluation process involves comparing the performance of our model on two distinct datasets: the training set and the validation set. It is noteworthy that both of these datasets originate from the MNIST dataset, with the training set being utilized for training the model and the validation set serving as an independent benchmark for evaluating its generalization capabilities ideally, we anticipate our model to exhibit consistent and comparable performance on both sets. Such uniformity serves as an indicator of a well-generalized model.

However, if we find that the model performance on the validation set falls significantly short of its performance on the training set, it raises a red flag. This discrepancy suggests a potential problem: overfitting (Broadhurst & Kell, 2006; Walsh et al., 2016), characterized by a "high variance" issue. In essence, overfitting signifies that our model has become too specialized in learning from the training data, capturing noise or minor fluctuations rather than the underlying patterns. This overemphasis on the training data at the expense of generalization can lead to poor performance when faced with new, unseen data a situation we aim to avoid in building robust machine learning models.

After our thorough training of the algorithm (Cnn_handwritting_digit), the model reached an accuracy of 97% (Figure 6.3a), proving its reliability. Now, fully trained, we are ready to integrate it into different operations.

In this evaluation, we not only assess the model accuracy (Figure 6.3 a) but also closely scrutinize how the loss function evolves during training (Figure 6.3 b). The loss function is a fundamental metric that quantifies how well the model is learning from the data. We expect the loss to decrease progressively with each subsequent epoch, as the model refines its understanding of the underlying patterns in the data. Training Loss (Blue Line): represents the value of the loss function (specifically, categorical cross-entropy loss (Ho & Wookey, 2020)) on the training data at the end of each training epoch. An epoch is one complete pass through the entire training dataset.

Validation Loss (Orange Line): represents the value of the loss function on a separate validation dataset at the end of each epoch. The validation dataset is not used for training; it is used to evaluate how well the model generalizes to data it has not seen during training.

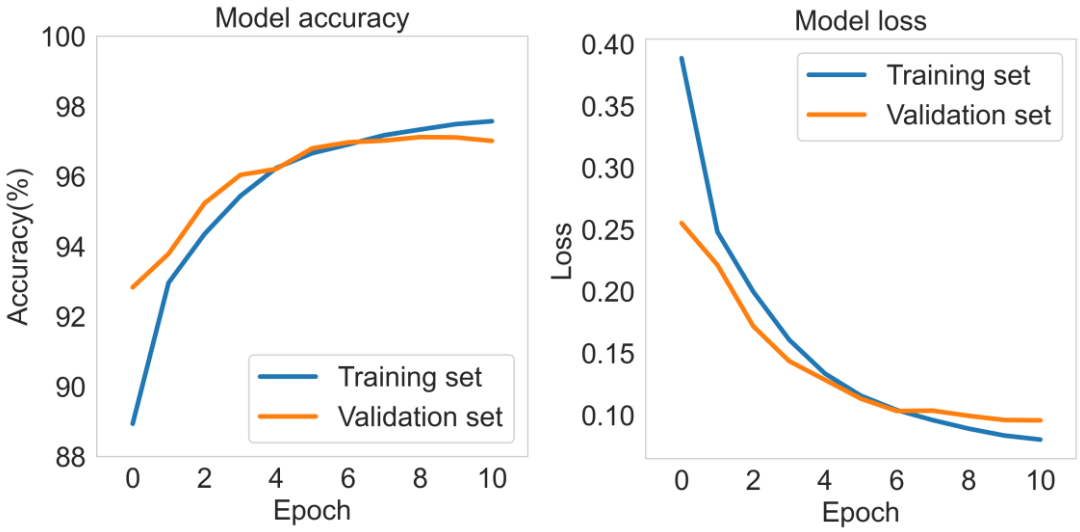


Figure 6. 3 : a) compare the model accuracy on both the training and validation sets, expecting similar performance. b) The loss function is crucial, quantifying how effectively the model learns. Ideally, this metric decrease over epochs, signifying improved pattern recognition. Discrepancies in performance or erratic loss trends can signal overfitting, emphasizing the need for model refinement to ensure robust generalization.

A successful convergence of the two loss curves, where both training and validation losses stabilize at a low value, signifies a well-trained model that demonstrates effective learning and generalization. Conversely, a significant gap between the two curves may indicate overfitting, emphasizing the importance of fine-tuning or adjusting the model architecture.

The central objective of this figure 6.4 is to effectively illustrate the performance of the model through the process of making predictions on a selected portion of the test data and subsequently visualizing the outcomes. The figure 6.4 adeptly guides users through the steps of loading a pre-trained Convolutional Neural Network (CNN) model, generating predictions, and visually pinpointing inaccuracies by highlighting them in a striking red hue. This graph is a useful tool that helps users quickly see where the model predictions differ from the actual labels.

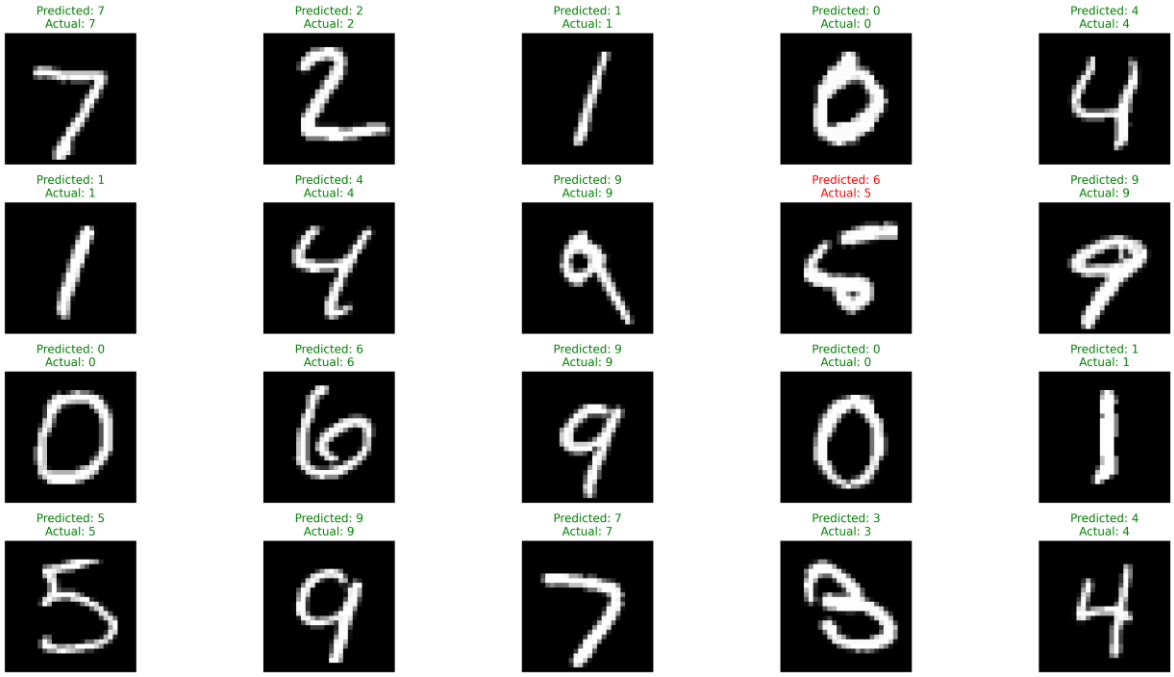


Figure 6. 4: Visualizing CNN Model Performance: Distinguishing Correct Predictions in Green and Incorrect Predictions in Red.

To visualize how well our model is performing for each class, we create a confusion matrix, as shown in Figure 6.5. The confusion matrix is a valuable tool for assessing the performance of a multi-class classification model like ours, especially when dealing with datasets like the MNIST dataset, which comprises 10 different classes, each representing a digit from 0 to 9.

In Figure 6.5, the confusion matrix breaks down how many samples from each class were correctly classified (true prediction) and how many were misclassified (wrong prediction) . Each row stands for the actual class, and each column stands for the predicted class. The diagonal shows the number of correct classifications (true prediction) for each class. Off-diagonal elements show misclassifications, with the row being the true class and the column being the predicted class. By looking at the accuracy rates for each class, we can make decisions about possible improvements to the model (Focus on classes with higher misclassification rates, considering strategies such as collecting more diverse data for those classes, adjusting hyper-parameters, applying data augmentation techniques, and evaluating the suitability of the model architecture. Additionally, addressing class imbalances or fine-tuning parameters may contribute to overall performance improvements)

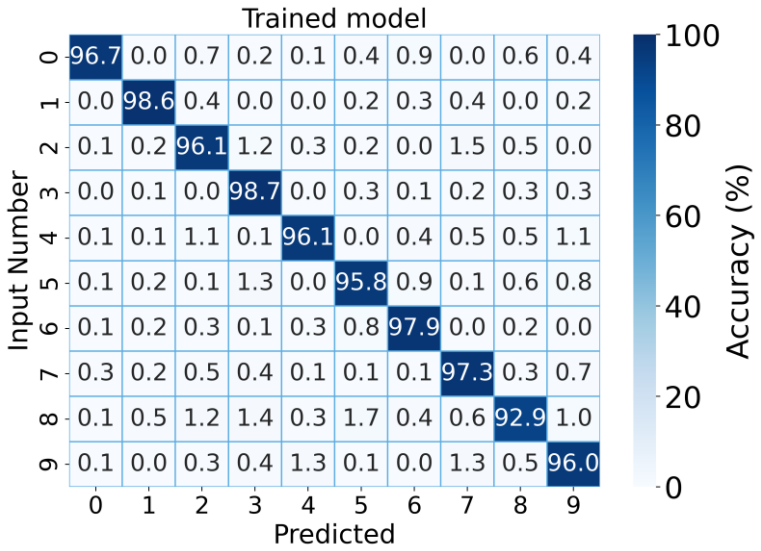


Figure 6. 5 : Confusion Matrix and Class Accuracy for MNIST Classification.

To emphasize a key point within this chapter, we embark on a comprehensive exploration of a broader frequency range, a crucial step towards fine-tuning our frequency modulation model introduced in Chapter 5. Our objective is to effectively calibrate this technique as we apply it to synaptic weight computations within a test Convolutional Neural Network (CNN) designed for the task of handwritten digit recognition.

Our approach involves a multi-stage transformation of these synaptic weights. Initially, we translate them into conductance values, a crucial intermediary step in our process. Subsequently, we convert these conductance values into their corresponding frequency representations. These frequencies are then transferred and encoded within a 16kbit phase change memory (PCM) array, which serves as a storage for the synaptic weights.

The aim of our investigation lies in how these encoded values within the PCM array can be harnessed to evaluate the performance of our test CNN. We take a thorough look at various redundancy schemes, each aimed at enhancing the resulting accuracy by coupling multiple PCM bits per weight. To predict the impact of redundancy on CNN accuracy, we have developed a concise yet effective model.

Additionally, we delve into the well-documented issue of PCM resistance drift and its repercussions on inference accuracy. Our exploration includes a detailed analysis of reliability considerations, with a specific focus on how PCM resistance drift evolves over time and its consequential effects on our inference accuracy tests.

6.2 Mapping synaptic weights

To improve our programming efforts, we used a frequency modulation technique. Before, we only explored four programming levels, however four levels was not enough for a real application. Therefore, we created a completely new testing program that allows us to test many different programming levels. This new program is different from our old ones, which had some issues, especially with handling high-frequency operations.

Algorithm: Main_routine

```
import {...}

# Define pulse configurations
Read_p = {...}
Set_1 = {...}
Res_p = {...}

# Create test environment
env = TTK.NewSession()
env.structure_type = "mad_16kbit_logic"
env.move = "None"
env.lot = {...}
env.waferNumber = 6
env.mask = {...}

# Set global parameters
env.scheduler.waferList = 6
env.scheduler.referenceScribe = {...}
env.scheduler.dieList = {...}
env.scheduler.pageList = 1
env.scheduler.scribeList = {...}

# Create and add test routine
Prog_1 = Progressive_Set({'A': Res_p, 'B': Set_1, 'R': Read_p, 'Repetitions_B': 50})
env.scheduler.New()
env.scheduler.Add('@PATTERN_FILE_onlyoneone0', Prog_1)

# Run the test
env.run()

# Cleanup
del env
# Measure and print the total execution time

end_time = time.monotonic()
print("Total execution time: " + str(timedelta(seconds=end_time - start_time)) + "
    seconds.")
```

The below Python code defines a class, `Progressive_Set`, which appears to managing and executing routines within a larger experimental or measurement system. The class is designed to configure a waveform generator (WFG) and initialize a digital input/output device (DIO) as part of its initialization process. The waveform generation process involves loading meta waveforms, adding waveforms for different patterns, and concatenating two waveforms. While the `execute` method is currently a placeholder, the code focus on controlling instruments, specifically waveform generators and digital input/output devices, and managing data through a reformatting process.

Algorithm: Progressive_set

```
1.  import {...}
2.  class Progressive_Set(ArrayPatternManager,RoutineTemplate):
3.          def __init__(self, params):
4.                  RoutineTemplate.__init__(self)
5.                  self.params = params
6.                  self.dataset = pd.DataFrame()
7.                  self.temp_data = pd.DataFrame()
8.          @RoutineTemplate.trackcall
9.          def control(self):
10.                 pass
11.         # Perform some initialization of the routine
12.         def initialize(self):
13.                 self.dataset = pd.DataFrame()
14.                 WFG = Resources.rack.instr["B1530"]
15.                 WFG.clear()
16.                 DIO = Resources.rack.instr['Arduino']
17.                 DIO.initialize()
18.         WFG.load_metawfms(list_of_pulses=['SET','B','READ'],params_of_pulses=self.p
    arams)
19.                 WFG.set_trigger({...})
20.                 WFG.add_waveforms(pattern_name='init', list_of_wfms=['SET', 'READ'],
    doMeasure=True)
21.                 WFG.add_waveforms(pattern_name='prog_set', list_of_wfms=['RESET',
    'READ'], doMeasure=True)
22.                 WFG.concatenate_two_waveforms({...})
23.         # Executing routine flow
24.         def execute(self):
25.         {...}
26.
27.         #Reformat the dataframe of results for saving
28.         def reformat_data(self):
29.          {...}
```

Using the right tools and a programming algorithm (importantly, we built this software from scratch, investing a big amount of time in development and debugging), we have achieved the capability to utilize nine distinct programming levels. These levels cover a wide frequency range, starting from 2Hz and going up to 1.6 kHz. It is crucial to emphasize that these programming levels also operate within ultra-precise temporal windows, spanning from a mere ≈ 45µs to a slightly more then 80µs. This temporal precision further bolsters the accuracy and fidelity of our experiments, allowing us to navigate the profound complexities of programming challenges with newfound confidence and sophistication. Consequently, our research endeavors have reached unprecedented heights, enriched by the expanded dimensions of programming exploration and investigation. Figure 6.6 vividly illustrates the diverse and intricate conductance distributions resulting from the application of nine distinct frequencies through our meticulously detailed frequency modulation technique, as expounded upon in comprehensive depth in Chapter 5. This graphical representation offers a rich and multifaceted glimpse into the intricate interplay of various frequencies, highlighting the nuanced and intricate patterns that emerge as a consequence of our methodical approach, as elucidated in the preceding chapter.

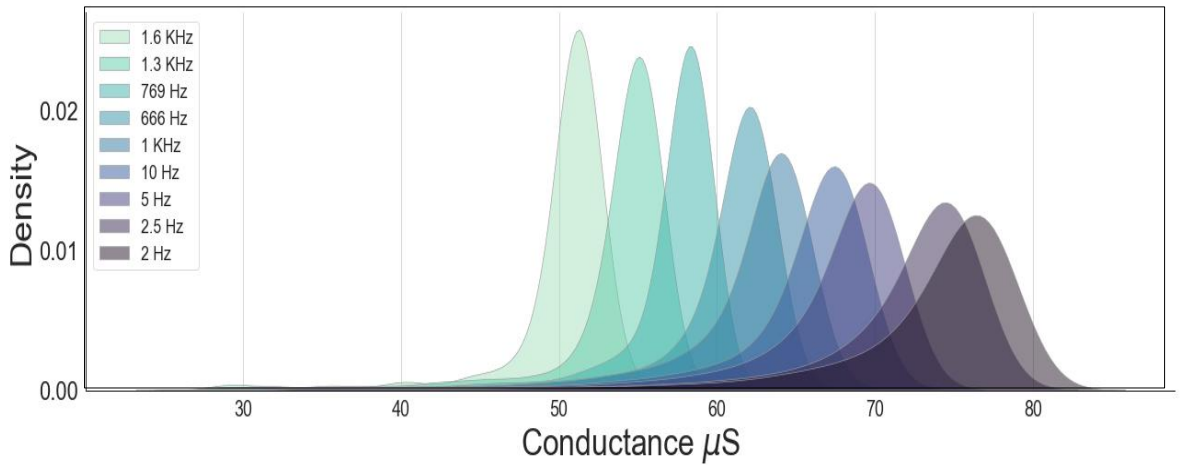


Figure 6. 6 : The conductance distribution among the nine programming levels.

Our immediate next step entails the fitting of this data with our sophisticated frequency modulation model, as evidenced in Figure 6.7.a. Notably, the striking alignment between our model and the measured data points is readily apparent, affirming the robustness of our approach in accurately capturing the underlying conductance distribution.

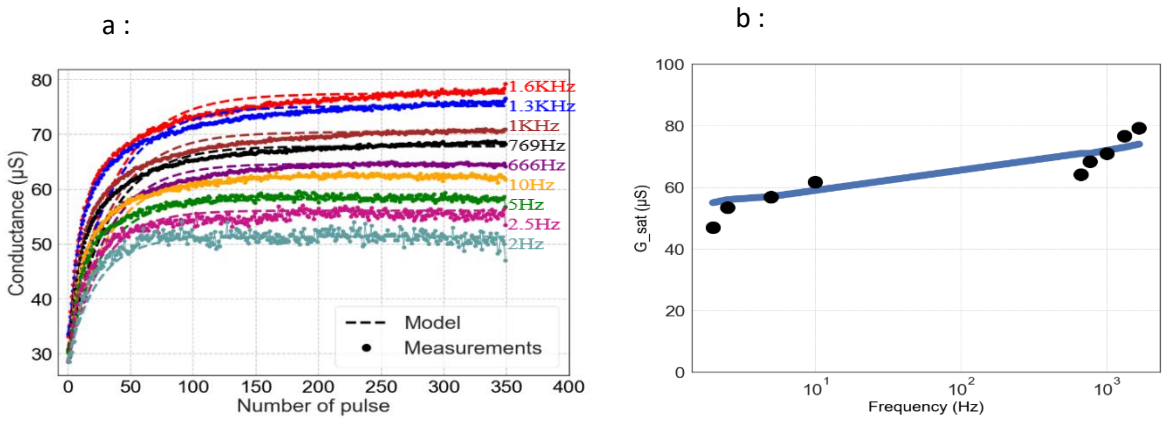


Figure 6. 7 : a) Extensive measurements characterizing the progressive SET behavior were meticulously conducted across multiple devices, each subjected to varying frequencies. Subsequently, meticulous calculations yielded averages for each device at the respective frequencies, with the application of our model fitting technique to enhance the precision of these outcomes. b) The resultant saturation values, represented as Gsat, were methodically examined in relation to the applied frequency. (Trabelsi et al., 2023)

In Figure 6.7.b, we present a graphical representation illustrating the median saturation conductance, denoted as Gsat, in relation to the applied frequency, denoted as F. Across a frequency range spanning from 2 Hz to 1.6 kHz, Gsat exhibits a variation, ranging from 45 to 80µS.

Equation ($G_{sat} = (1-\gamma)\, G_{sat} + \Delta G$) explained in chapter 5 establishes a crucial connection between the conductance at the current pulse, denoted as Gn (corresponding to the n-th pulse), and the preceding conductance. This equation, characterized by the parameter γ encompassing both frequency and drift coefficients as elucidated in (Trabelsi et al., 2022), and ΔG as a fit parameter representing the conductance increase per SET pulse without any drift, remarkably aligns with the programming curves shown in Figure 5.8, depicted by dashed lines. When saturation is reached, i.e., when Gn equals Gn-1, Equation 5.4 furnishes a valuable expression for Gsat as a function of the applied frequency. The ensuing figure, Figure 6.7.b, effectively juxtaposes both empirical data and the model predictions, offering a comprehensive view of the correspondence between the two.

The main challenge in conductance modulation approach using frequency is handling both intrinsic and extrinsic variabilities that consistently affect resistive devices. These variabilities can significantly hinder measurement precision in real-world situations. To thoroughly evaluate our programming algorithm's performance on a practical Convolutional Neural Network (CNN), we conducted an experimental investigation illustrated in Figure 6.8.

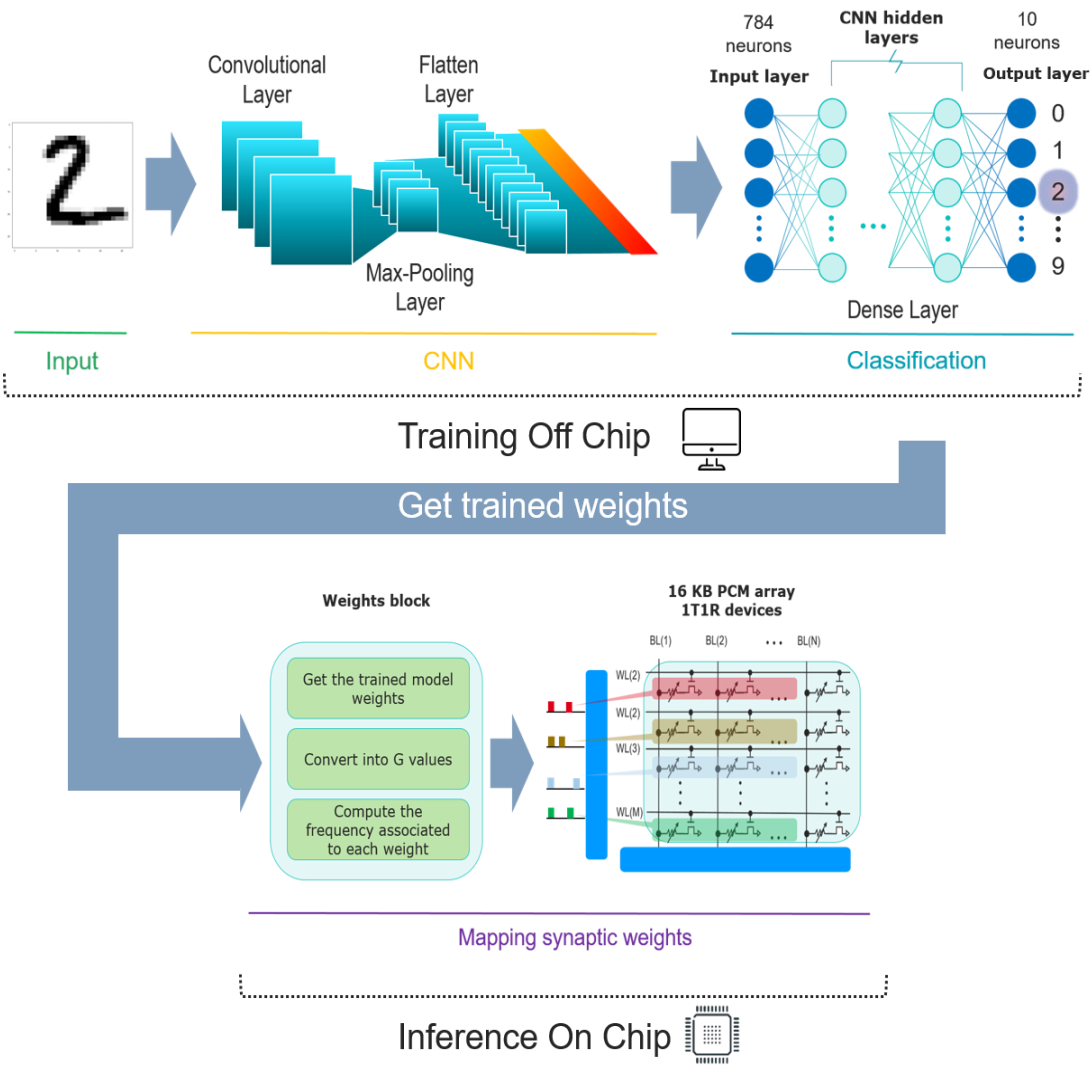


Figure 6. 8 : Our test measurements workflow is outlined in the schematic, offering a comprehensive view of our experimental setup. In this setup, a fully connected network features an input layer with 784 terminals, corresponding to the intricate details of a 28 × 28 pixel image. The output layer consists of 10 neurons, each aligned with one of the 10 distinct digits (0...9). This network architecture forms the basis for our testing and validation procedures. On the bottom right side of the schematic, we present the PCM 16-kilobit array, used in the experimental setup. This array acts as the substrate for programmable conductance modulation, offering a real interface for translating and encoding synaptic weights into conductance values. (Trabelsi et al., 2023).

For context, we had previously computed a set of weights tailored for a test CNN comprising 784 input neurons and 10 output neurons. These weights were obtained through the training of the CNN to solve the handwritten MNIST problem. Subsequently, we undertook the intricate task of converting these synaptic weights into target conductance values, thereby establishing a direct relationship with

frequencies. Each target conductance was meticulously programmed into a Phase Change Memory (PCM) bit, ensuring the faithful transfer of all weights.

To convert synaptic weights into conductance values, we utilized a precise mapping strategy. This strategy involved aligning the achievable range of conductance values, as shown in Figure 6.7.b, with corresponding weight values, using Equation 6.1 as our mathematical framework. In this equation, W represents the weight to be converted into the conductance value G, where G_{Min} is the minimum attainable conductance value observed in Figure 6.7.b, and W_{Min} is the lowest threshold weight requiring conversion.

$$G = \left(\frac{\Delta W}{\Delta G}\right)^{-1} |W + W_{Min}| + G_{Min} \qquad \text{Eq. 6.1}$$

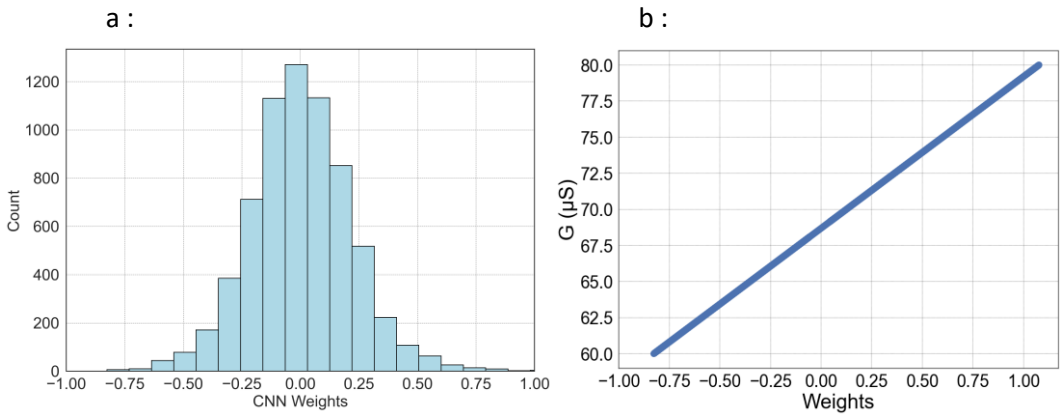


Figure 6. 9 : a) The distribution of synaptic weights within the Convolutional Neural Network (CNN) following the completion of the training process. b) By employing Equation 6.1, we have the capability to seamlessly convert the weights of our CNN model into corresponding conductance values. (Trabelsi et al., 2023)

The factor ΔW/ΔG plays a pivotal role in determining the granularity of programmable weights and is depending on the PCM device's capacity to span a comprehensive range of G values. This parameter exerts a direct and substantial impact on the precision and fidelity of the translation process.

Figure 6.9.a shows the initial distribution of CNN weights. For reference, Figure 6.9.b visually maps G values to their corresponding W values. Conversely, by inversely applying Equation 6.1, we can effortlessly convert G values back into their original W values. This step is important because after sending these weights to the PCM devices, we read them back and update the initial weights in the CNN algorithm with the PCM weights to assess accuracy.

The choice of using only one weight or both positive and negative weights in programming synaptic weights depends on the specific application and the underlying hardware. While most authors use both positive and negative weights, recent research has explored optimized weight programming strategies that involve complex weight programming optimization, including the use of both positive and negative conductances to achieve more accurate weight updates (Mackin et al., 2022). Additionally, the geometry of synaptic changes and plasticity can also influence the distribution of synaptic weights, with non-Euclidean distances potentially impacting weight distributions(Pogodin et al., 2023). Furthermore,

the characteristics of synaptic weight updates, such as linearity and symmetry, are crucial for energy-efficient and accurate learning operations in artificial synapse devices (Sahu et al., 2023).

The algorithm (Mapping synaptic weights) using one PCM per weight shows in details the steps taken to transmit the initial weights into silicon.

Algorithm 2: Mapping synaptic weights

```
g_target_values = [...all the G target values...]
frequency_values = [...corresponding frequency values...]
pcm_synaptic_weight = []  # Initialize an empty list to store results
for i in range(len(g_target_values)):
    target = g_target_values[i]
    frequency = frequency_values[i]  # Get the corresponding frequency value
    batch_size = 50  # Start with a smaller batch size
    max_batch_size = 400  # Maximum allowed batch size
    tolerance = 0.1  # Tolerance for deviation
    accuracy_achieved = False  # Flag to track whether desired accuracy is achieved
    while batch_size <= max_batch_size:
        send_pulses(batch_size)
        read_value = read_G()
        deviation = (read_value - target) / target
        if -tolerance <= deviation <= tolerance:
            accuracy_achieved = True  # Set the flag to True when the desired accuracy is achieved
            break
        batch_size += 50
    # Save the values regardless of whether the desired accuracy was achieved within the maximum   batch size
    pcm weight = {
        "Target": target,
        "Actual": read_value,
        "Batch Size": batch_size,
        "Accuracy Achieved": accuracy_achieved,
        "Frequency": frequency  # Include the corresponding frequency value in the result
    }
 pcm_synaptic_weight.append(pcm weight )
# At this point, 'pcm weight ' will contain the saved values for each iteration, along with the corresponding frequency
value
```

After completing the experiment, we obtained a comprehensive list of programmed synaptic weights from the PCM devices. The number of PCM devices used was dependent on the quantity of CNN weights (6760 weights). Subsequently, we retrieved and updated these PCM weights in the CNN model. Upon completing these updates, we proceeded to rerun our handwritten digit recognition program, specifically designed for this purpose. Executing this program with the updated weights allowed us to generate results for comparison with those obtained using the original CNN weights. This

comparison served as a validation step, enabling an evaluation of how well PCM-based synaptic weights performed compared to conventional CNN weights. To enhance the reliability and precision of our algorithm, we explored a strategy involving the collective use of PCM devices to encode a single weight.

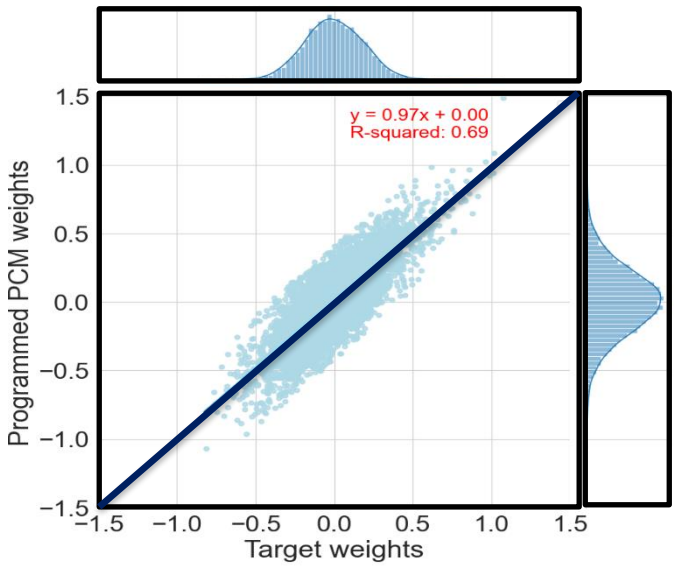


Figure 6. 10 : The graphical representation above illustrates the programmed weights within the Phase Change Memory (PCM) devices, with each weight being the result of an averaging process involving six individual PCM devices (each point corresponds to the average of six devices, totaling 6760 weights). These programmed weights are compared to their corresponding target weights. (Trabelsi et al., 2023).

This approach amalgamated individual PCM G values to calculate a weighted average accurately representing the desired weight. The comprehensive findings from this experimentation have been encapsulated in the graphical representation presented in Figure 6.10. In this figure, we present a scatter plot that contrasts the target weights against the corresponding PCM programmed weights. The x-axis represents the target weights, while the y-axis represents the PCM programmed weights. Ideally, all points should fall precisely on the diagonal line, indicating perfect alignment between target and programmed values. Upon visual inspection, the majority of points closely follow the diagonal line, suggesting a consistent and accurate encoding of weights by the PCM devices. This alignment signifies the effectiveness of the programming methodology in replicating the intended weights. However, it is important to note a slight dispersion observed around the diagonal, indicative of variability in the devices and potential drift. The scatter in the points suggests that, despite the overall alignment, there are instances where the programmed weights deviate slightly from the target values. This variability can be attributed to inherent device differences and factors contributing to drift over time.

Both the traditional techniques, such as amplitude with the pulse modulation, and the frequency modulation technique have been instrumental in controlling conductance response and optimizing weight programming strategies. These techniques have been crucial for achieving linear and symmetric synaptic weight updates, which are essential for energy-efficient and accurate learning

operations in artificial synapse devices. However the frequency modulation technique has the advantage for the easy silicon implementation.

6.3 Synaptic Behavior Simulation:

In the study of Phase Change Materials (PCMs), simulation serves as a tool for comprehending complex behaviors and predicting performance. This computational approach enables a detailed exploration of the underlying dynamics of PCMs, offering precise insights. The intricate nature of PCM systems, influenced by factors environmental variables, requires systematic examination through simulations.

Simulations contribute by creating virtual environments that replicate real-world scenarios, effectively capturing the complexity of PCM systems. To improve the reliability of predictions, experimentation with various configurations, such as 2, 4, 6, and 8 PCM devices, is essential. These diverse scenarios provide a comprehensive understanding of PCM behavior, ensuring robust and representative inferences for potential real-world applications.

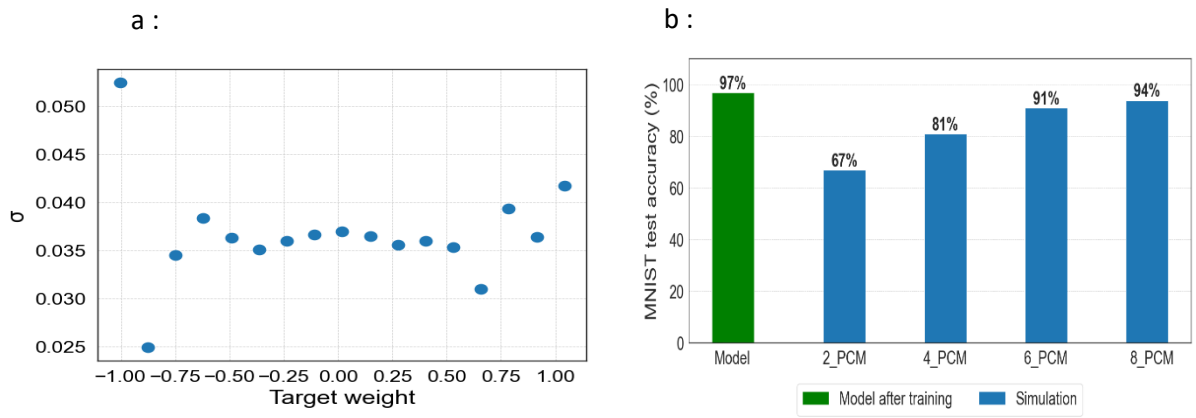


Figure 6. 11 : a) Comparing target weights to standard deviations. b) Following the training of our software model, in conjunction with Monte Carlo simulations, we achieved a remarkable test accuracy on the MNIST dataset. (Trabelsi et al., 2023)

Figure 6.11.a is an analysis based on weights, aiming to elucidate the distribution of these weights and how their standard deviation ('σ') (D. K. Lee et al., 2015) varies within distinct weight intervals. The methodology involves dividing the target weights into 20 predefined bins, facilitating a comprehensive examination of the weight distribution. The histogram is then computed, revealing the frequency distribution of target weights across these bins. The center of each bin is calculated to provide a representative value for that particular weight interval.

The mathematical insight lies in the calculation of standard deviations for each bin. Iterating through each bin, the boundaries are determined, and data points falling within each interval are collected. Subsequently, the standard deviation of these data points is calculated, offering a quantitative measure

of the dispersion or spread within that specific weight range. This process is iteratively repeated for all bins, yielding a series of standard deviation values corresponding to their respective bin centers. Figure 6.11.a provides a visual representation that aids in comprehending how the variability in target weights changes across different weight intervals. This analysis sheds light on which weight ranges exhibit higher or lower variability, setting the stage for our subsequent simulations showcased in Figure 6.11.b.

Figure 6.11.b. this figure elucidates the various schemas employed in the simulation, with the initial schema serving as our reference point for comparison the reference schema corresponds to a standard CNN model. The bar chart displays a progressive exploration, starting with the baseline CNN model. The second bar introduces the use of 2 PCM models, resulting in a test accuracy of 67%. Moving forward, the third bar highlights the implementation of 4 PCM models, leading to a substantial improvement with an accuracy of 81%. The trend continues with the fourth bar, where the utilization of 6 PCM models significantly elevates the accuracy to an impressive 91%. Finally, the fifth bar, harnessing the power of 8 PCM models, achieves a remarkable test accuracy of 94%.

The gradual increase in accuracy as we integrate more PCM models underscores the efficacy of this approach in enhancing model performance. This observation reveals a promising strategy for achieving superior results in tasks such as image classification.

In the specific context of simulating MNIST digit classification, Figure 6.12 presents four distinct heat maps generated with different configurations of PCM devices specifically, 2, 4, 6, and 8 PCM units. These heat maps serve as graphical representations of the PCM-based system's performance and behavior throughout the digit classification task and each row in the heat map corresponds to the class digit (0-9), while each column represents the predicted digit. Notably, the diagonal elements within these heat maps denote accurate predictions, while non-diagonal elements correspond to erroneous predictions. This straightforward visualization is instrumental in assessing the classification accuracy of the system. The simulation outcomes reveal a conspicuous trend: as the number of PCM devices incrementally increased from 2 to 8, there was a consistent and remarkable improvement in the accuracy of the classification system. One particularly result emerged with the deployment of 8 PCM devices, where the accuracy of the MNIST digit classification system surged impressively to 94%.

The achievement of 94% accuracy in the MNIST digit classification system using 8 phase-change memory (PCM) devices for inference testing is indeed impressive. However, the scalability of achieving high accuracy with 8 PCM devices to real-world applications involving large-scale neural networks is questionable. Realistic neural network models often require a significantly larger number of synaptic weights, and it is uncertain whether the performance observed with 8 PCM devices can be extrapolated to larger networks. The use of multiple PCM devices in parallel for synaptic weight programming introduces variability and reliability concerns. Variations in device characteristics, such as conductance levels and response times, could impact the overall accuracy and consistency of the neural network's performance. PCM devices are susceptible to manufacturing variability, which can lead to differences in conductance levels and switching behavior among individual devices. This variability could pose challenges in achieving consistent and reliable performance across a large number of PCM devices.

The sources of variability in PCM devices, including differences in crystallization behavior, nucleation kinetics, and thermal characteristics, can lead to variations in conductance levels and response times. Additionally, PCM devices may exhibit sensitivity to environmental factors such as temperature variations and electrical noise, which can introduce variability in their conductance behavior. Addressing the sources of variability and ensuring consistent performance across a larger number of PCM devices will be crucial for the practical implementation of PCM-based synaptic weight programming in neuromorphic computing systems. Thorough evaluation and further research are necessary to address these challenges and determine the suitability of PCM devices for large-scale neural network applications.

Simulations

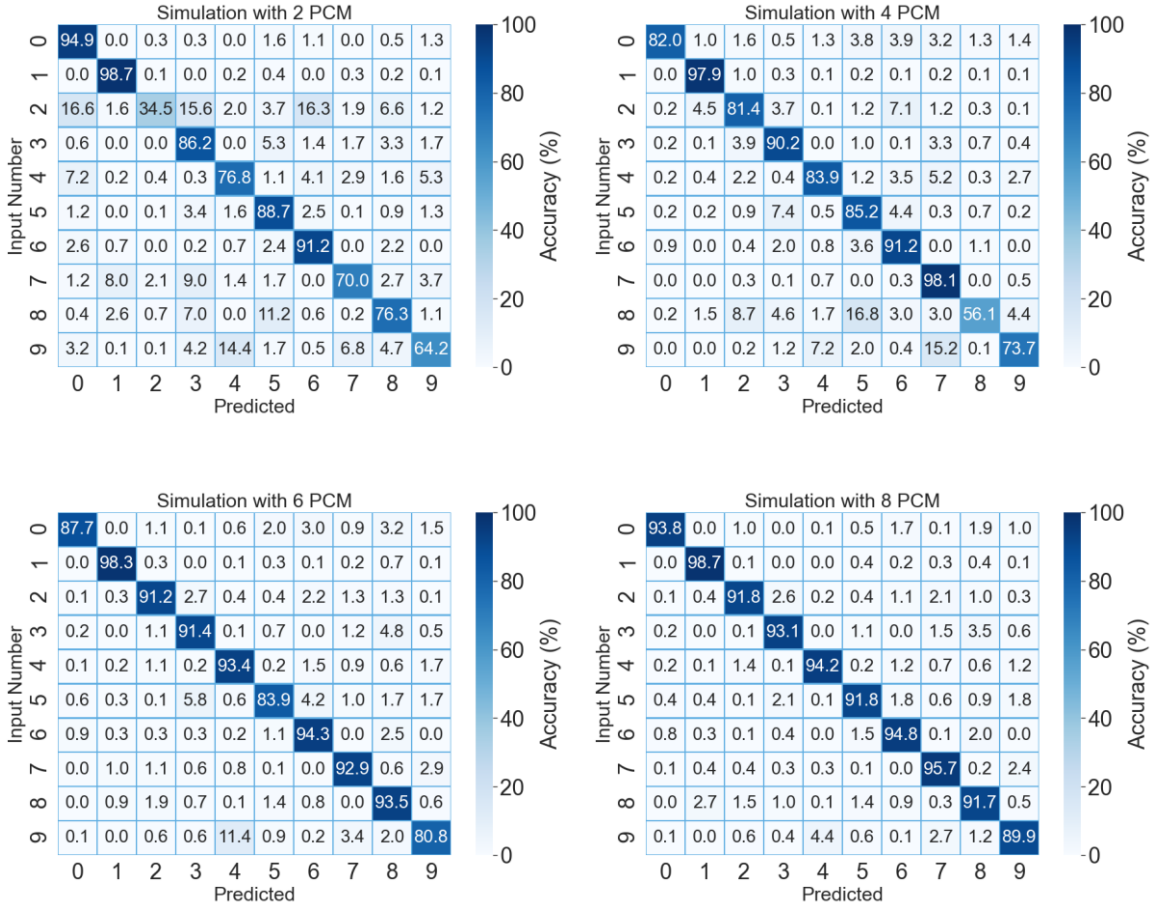


Figure 6. 12 : Analyzing the Influence of Phase Change Memory (PCM) Configurations on MNIST Digit Classification Accuracy through Inference Tests: A Study of PCM Units from 2 to 8

6.4 Performance Analysis

Following the simulation phase, we proceeded to real-world inference testing utilizing a 16Kbit array. The objective was to replicate the simulation methodology, incorporating multiple PCM devices per programmed weight. The procedural details are graphically illustrated in Figure 6.8, offering a succinct overview of the approach. Algorithm (Mapping synaptic weights) provides a comprehensive outline of the practical algorithm, systematically detailing each step.

The primary objective of the algorithm is to calibrate the 'pcm_synaptic_weight' parameter to align with specific target values denoted as 'g_target_values' and their corresponding frequencies. Each 'g_target_value' is associated with a calculated frequency, enhancing precision. The algorithm iteratively adjusts the batch size (iteration number) within predefined limits while continuously monitoring the deviation between observed and target values. This process ensures the calibration of the 'pcm_synaptic_weight' parameter to achieve alignment with the desired target values and their corresponding frequencies.

Start with smaller batch sizes and incrementally increasing them, the algorithm ensures accuracy within a predetermined tolerance range of $\pm$ 10%. It iteratively adjusts the batch size until the desired accuracy level is achieved or the batch size reaches its prescribed maximum limit. At each iteration, the algorithm records essential data points, including the target value, actual read value, batch size utilized, an accuracy indicator, and the corresponding frequency value.

These recorded results are systematically stored within the "pcm_synaptic_weight" list. The algorithm, responsive to real-time feedback, dynamically optimizes the "pcm_synaptic_weight" parameter. This approach serves as a pivotal tool in our experimental endeavors, ensuring effective calibration and performance optimization.

Prior to our final analysis, we conducted a detailed examination of both the CNN weights and the PCM weights. This examination is substantiated by the comprehensive plot presented below. Our analysis encompassed a multifaceted exploration of the characteristics and distributions of both PCM and CNN weights.

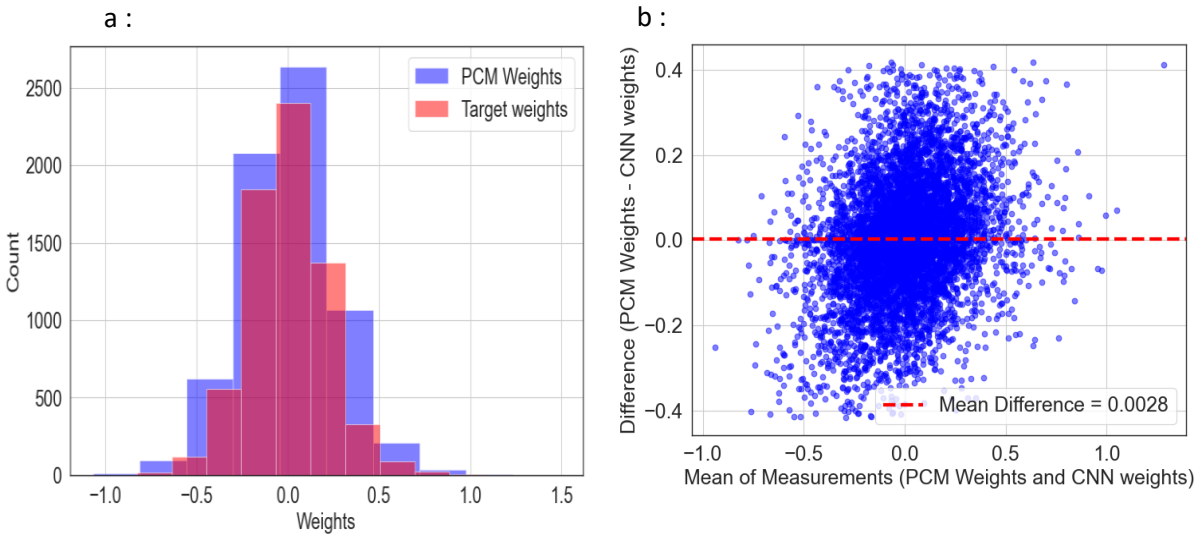


Figure 6. 13 : a) Exploratory Analysis of Weight Distributions: PCM Weights vs. Target Weight. b) Concordance Evaluation: Bland-Altman Plot of PCM Weights vs. CNN Weights

Figure 6.13.a. provides insights into the distribution patterns of two distinct datasets: 'PCM Weights' and 'Target Weights.' The histograms, represented by blue and red bars, respectively, offer a visual representation of how weight values are distributed within each dataset. The x-axis delineates different weight values, allowing for a comprehensive exploration of the data distribution. Notably, a discernible shift is observed between the two distributions, suggesting a systematic variation in weight values. This subtle displacement indicates a noteworthy distinction in the weight characteristics between the PCM-programmed weights and the target weights

In our analysis, the Bland-Altman plot (Figure 6.13.b) emerges as a pivotal tool for assessing the concordance between 'PCM Weights' and 'CNN Weights.' This graphical representation strategically positions the differences between corresponding measurements on the y-axis, with the mean of measurements plotted on the x-axis. In the context of the Bland-Altman plot, each blue data point represents the difference between a pair of measurements (one from 'PCM Weights' and one from 'CNN Weights'). The mean difference (0.0028) is the average of all these individual differences and is represented by the red dashed line on the plot. The standard deviation of differences (0.142) provides information about how much each individual difference deviates from the mean difference. A smaller standard deviation suggests that the individual differences tend to be close to the mean difference, indicating less variability or dispersion among the data points. On the other hand, a larger standard deviation suggests more variability, indicating that the individual differences are more spread out from the mean difference.

In Figures 6.14, shows two plots to gain insights into the relationship between 'PCM programmed weights' and 'CNN Weights.' On the left, a scatter plot vividly portrays this connection, introducing dimensions of size and color. The size of each point reflects its frequency, aiding in the identification of significant data points. Color variations provide additional context, highlighting different frequencies. This plot aims to unveil patterns and outliers, enhancing our understanding of the concordance or

divergence between 'PCM programmed weights' and 'CNN Weights.' The right side features a line plot, mapping the cumulative frequency against 'PCM programmed weights.' This visualization tracks how data points accumulate with increasing 'PCM programmed weights,' revealing distribution patterns and potential groupings. Together, these plots offer a comprehensive view, allowing us to identify relationships, outliers, and the overall distribution of weights within the dataset.

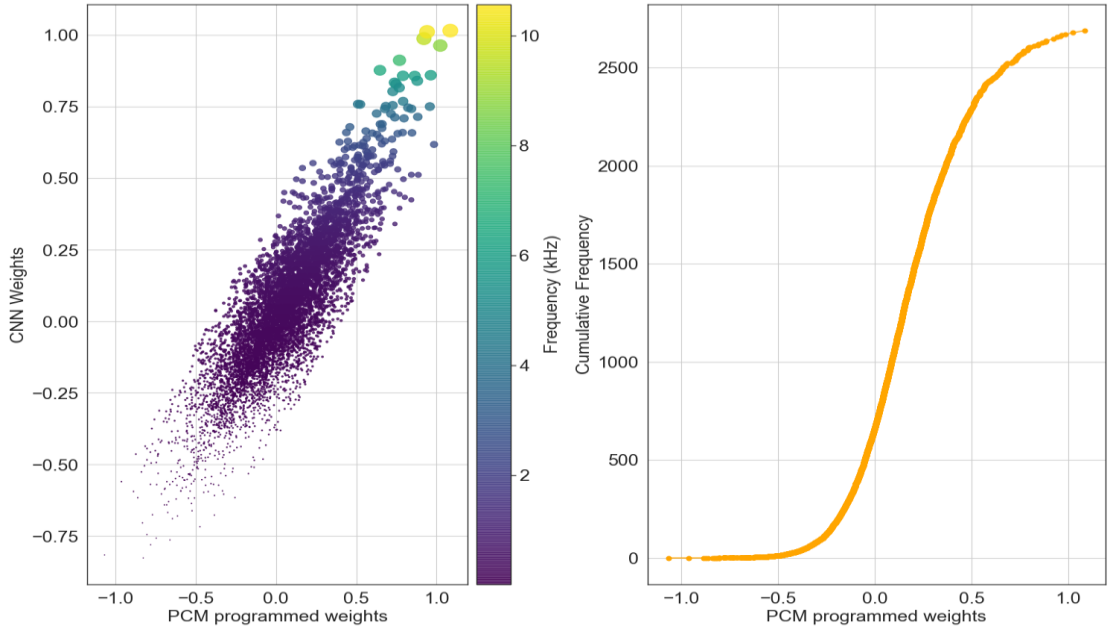


Figure 6. 14: Scatter plot and a line plot, each providing distinct insights into the relationship between 'PCM programmed weights' and 'CNN Weights.' The scatter plot employs size and color to indicate frequency, revealing patterns and outliers. The line plot depicts cumulative frequency in relation to 'PCM programmed weights,' offering a view of weight distribution patterns.

In Figure 6.15, we conduct a thorough analysis of PCM redundancy scenarios, examining levels of redundancy at 2, 4, 6, and 8. This investigation spans model performance, simulations, and PCM programmed weights, providing a comprehensive understanding. In an ideal scenario without weight transfer errors, our CNN model would achieve a remarkable 97% accuracy. However, real-world challenges introduce discrepancies between target and actual PCM cell values, impacting accuracy. To address this challenge, we explore the potential of PCM redundancy. Incorporating 4 PCM cells shows a significant recovery, boosting accuracy to nearly 80%. Undeterred, we continue to explore. The synergy of 6 PCM devices further elevates model accuracy to a commendable 90%. This crucial insight highlights the effectiveness of PCM redundancy in mitigating the impact of weight discrepancies, emphasizing its role as a powerful tool for enhancing model accuracy and resilience.

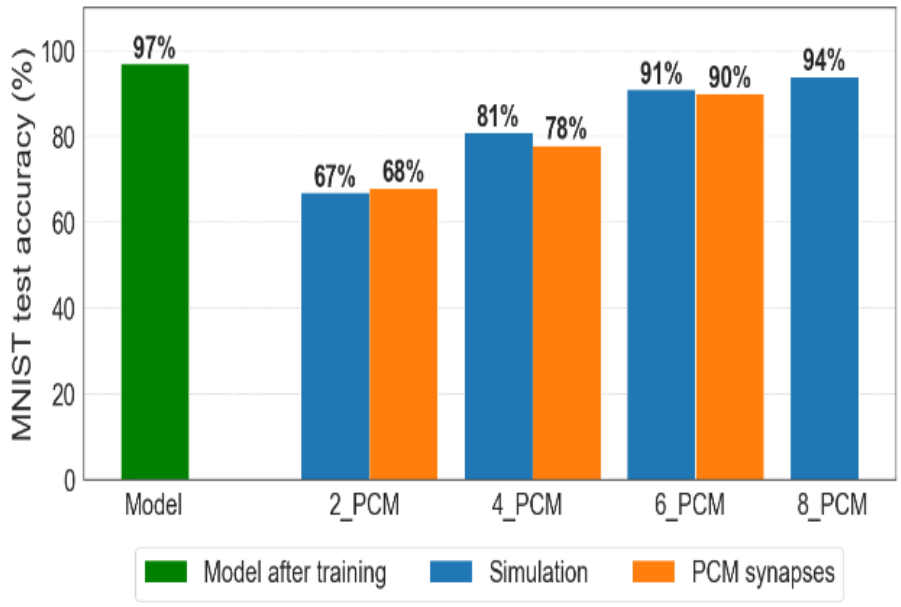


Figure 6. 15: Explore MNIST test accuracy through stages: initial CNN model training, weight transfer to PCM synapses (redundancy levels 2 to 8), and Monte Carlo simulations. (Trabelsi et al., 2023).

Our study reveals an alignment between simulations and real-world measurements, highlighting the robustness of our models. This synchronicity underscores the precision and accuracy of our approach, affirming the validity of our findings. The agreement between simulation and measurement confirms the reliability of our scientific efforts, enhancing confidence in our research. This alignment elucidates the dynamics of our studied phenomena and emphasizes the value of simulation as a potent tool for understanding and predicting real-world outcomes.

This also confirmed though Figure 6.16, we conduct a comprehensive exploration, into the PCM configurations and their impact on the accuracy of MNIST digit classification through inference tests. Our research specifically focuses on PCM units, ranging from 2 to 6, as we plot PCM weights. This investigation serves as a dynamic window into the symbiotic relationship between PCM configurations and the precision of MNIST digit classification for its different classes.

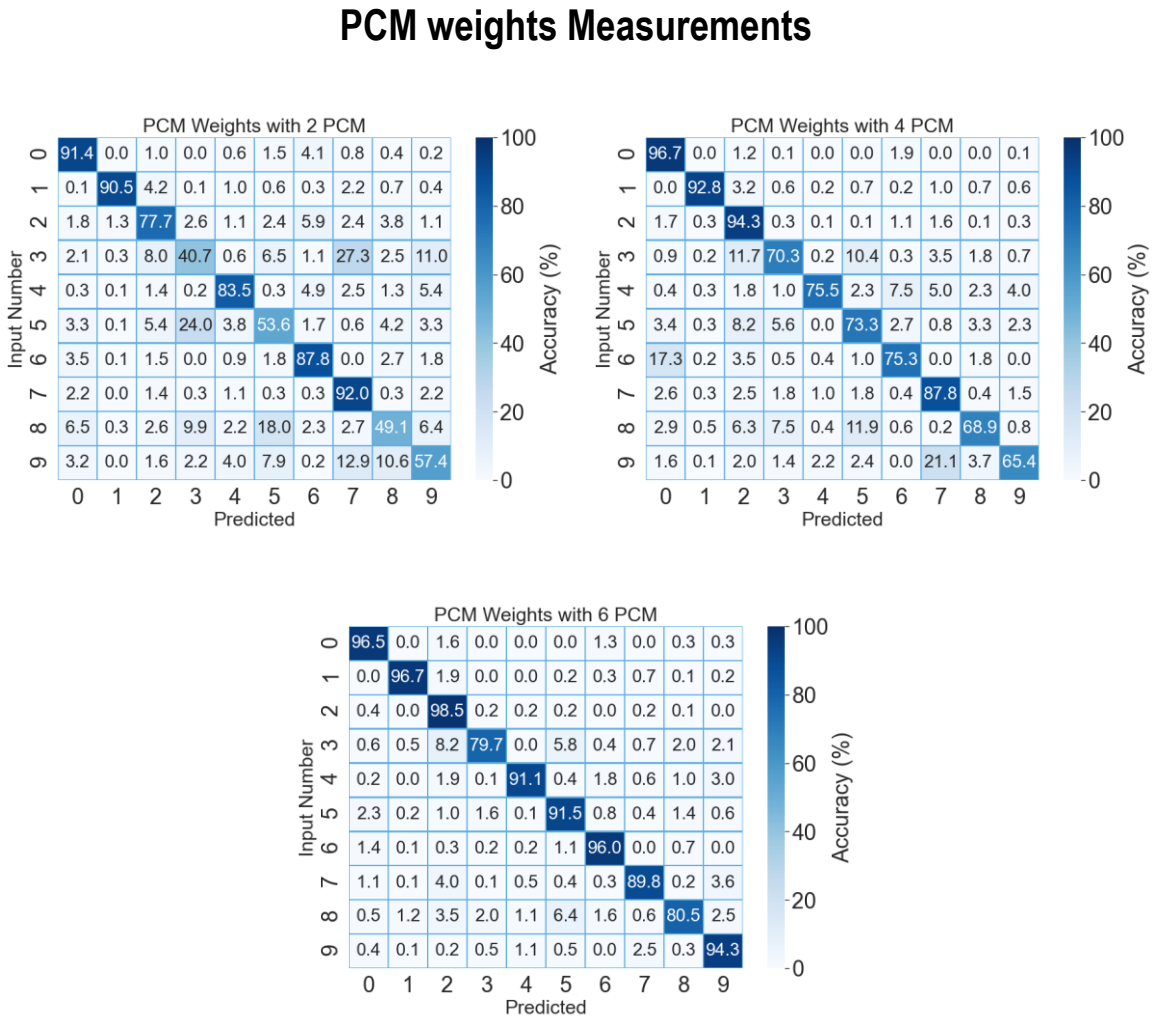


Figure 6. 16 : Heat maps of PCM devices (2 to 6) and their influence on MNIST digit classification accuracy through inference tests.

In the same context but using the traditional techniques and in other research introduced a mixed-precision architecture that combined a computational memory unit with PCM and a digital processing unit, achieving 97.73% test accuracy on the task of classifying handwritten digits (based on the MNIST dataset), within 0.6% of the software baseline (Nandakumar et al., 2020). An experimental demonstration and tolerancing of a large-scale neural network (165,000 synapses) using phase-change memory as the synaptic weight was conducted, providing insights into the practical implementation of PCM in neuromorphic systems(Ivanov, 2023). A methodology to train ResNet-type convolutional neural networks was introduced, resulting in no appreciable accuracy loss when using computational phase-change memory for inference. The study achieved a classification accuracy of 93.7% on the CIFAR-10 dataset and a top-1 accuracy of 71.6% on the ImageNet benchmark after mapping the trained weights to PCM (Joshi et al., 2020). The collective studies underscore the diverse applications and potential of Phase Change Memory (PCM) in implementing synaptic weights, particularly in the context of the MNIST dataset. Our implementation with an accuracy of 90%, positioning us favorably when compared to the current state-of-the-art. This achievement not only

highlights the promising capabilities of PCM but also reinforces our commitment to pushing the boundaries of neural network systems.

Next we conducted a thorough 24-hour investigation at room temperature to assess the potential impact of drift on our programmed weights. Our goal was to study the weight distribution and compare it with the initial values at t0. The compelling results, presented in Fig. 6.17, reveal a noteworthy revelation the distribution remains remarkably stable over our evaluation period.

To achieve a comprehensive understanding of weight stability, further tests, particularly those conducted at elevated temperatures, are deemed crucial. The relatively robust median value of G (approximately 65µS) offers reassurance, hinting at a low drift factor, as outlined in (Trabelsi et al., 2022).

In summary, our evaluation and the observed stability in the weight distribution establish a solid foundation. This sets the stage for continued research aimed at unraveling the intricacies of drift in our programmed weights, ensuring the reliability and longevity of our systems in real-world applications.

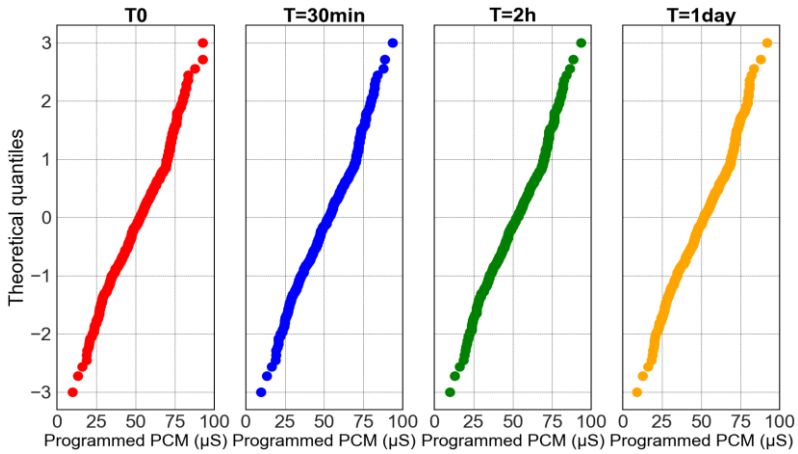


Figure 6. 17: Temporal Evolution of Drift Measurements: Understanding Stability and Variability Over Time (Trabelsi et al., 2023)

Our approach to frequency modulation brings several advantages over traditional methods, as depicted in Figure 6.18. We employ a straightforward Finite State Machine (FSM) for silicon implementation, providing an organized and simplified control flow. Unlike conventional program/verify approaches, our method eliminates the need to check conductance after each pulse, streamlining the process. The use of a lookup table for duty cycle determination based on target conductance simplifies programming. Additionally, our design of the pulse generator is made more direct by utilizing a consistent set pulse instead of increasing its potential or current.

The execution of array programming is efficiently managed by coupling the pulse generator with a burst counter. While Figure 6.18 does not illustrate it, the sensing circuit closely resembles that of a typical program/verify approach for compatibility. In summary, frequency modulation approach enhances efficiency by simplifying pulse generation and conductance checking in the programming process.

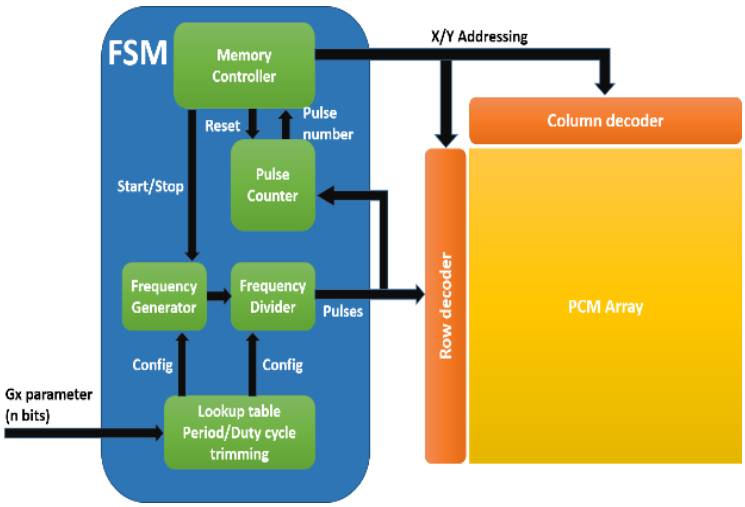


Figure 6. 18 : Efficient Weight Programming with Frequency Modulation: A Silicon Implementation Using a Finite State Machine (FSM)

In Figure 6.19, we introduce a simple user-friendly application that lets you draw digits and guess them using PCM weights. The app is made for fun and to show how PCM weights can work in recognizing digits. You draw a digit, and the app tries to guess it using PCM weights.

This app is like combining art and science. It is easy for users to draw and see how PCM weights can be useful in recognizing digits. It is not just a cool way to engage users; it also shows in real-time how PCM can be flexible in solving real-world problems. The app is a simple but effective way to see the power of PCM in action.

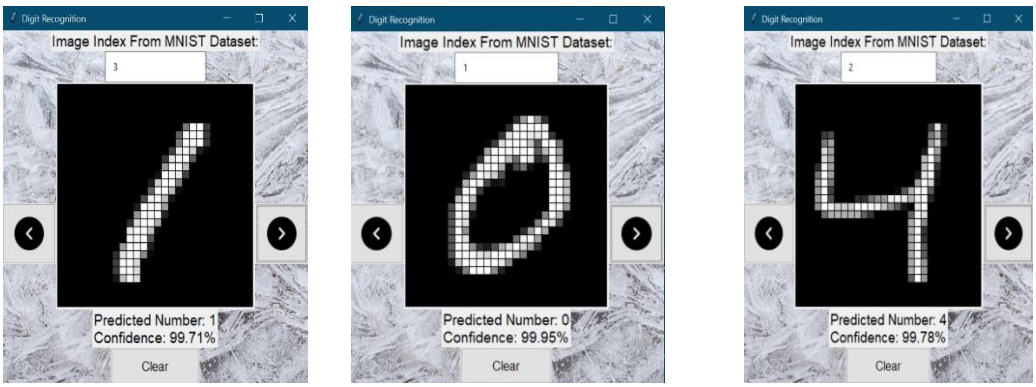


Figure 6. 19: Interactive MNIST Digit Recognition: Exploring PCM Programmed Weights through User-Drawn Input

The use of frequency modulation in sending synaptic weights into phase-change memory (PCM) offers significant benefits, particularly in the context of non-von Neumann computing such as in-memory computing and neuromorphic computing. However, a drawback is the considerable time required to run the experiment (If we focused on low frequencies) as chowing in Figure 6.20.a. To refine the estimation process for subsequent experiments beyond the initial three recorded runs ("1 day 5:49:31,"

"1 day 7:36:26," "1 day 23:20:42"), instead of relying solely on average times, the percentage increase in time between consecutive runs is calculated. Precisely, the percentage increase from the first to the second run and from the second to the third run is determined. Subsequently, these percentage increases can be applied to the duration of the last recorded run to extrapolate and estimate the times for the other experiments (Facing challenges during the initial setup led to system crashes when attempting to run all the experiments at once). To ensure successful execution, we find it necessary to restart the experiments (Figure 6.20.b). Notably, interrupting the machine during experimentation could potentially reduce the overall duration. Our observations reveal that the experiments typically range from approximately 1 day, 5 hours, to 1 day, 7 hours.

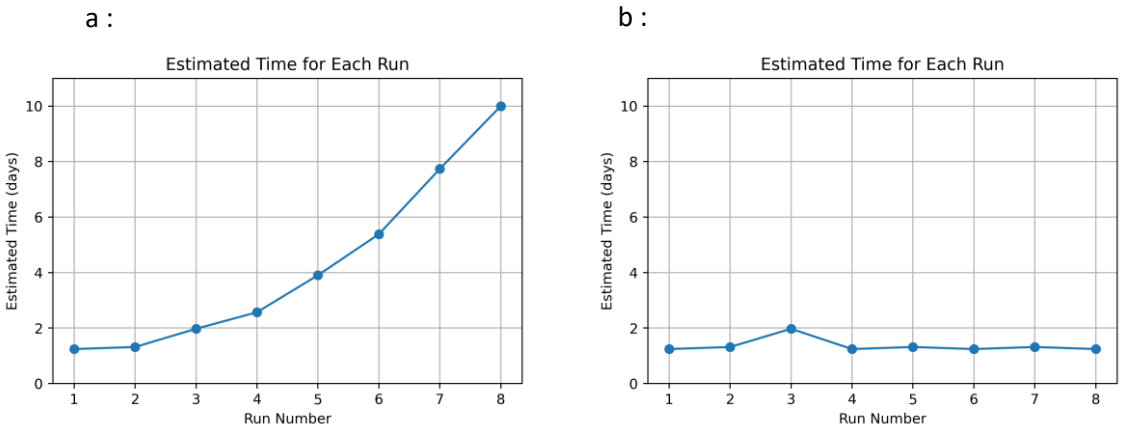


Figure 6. 20: a) Running time duration without restarting the experiment. b) Running time duration with restarting the experiment.

To address this timing issue further research and development are needed to optimize the efficiency of the process and reduce the time required for experimentation. Also other solution is parallel programming that offers a compelling solution to the challenges of executing problem-solving computations in parallel, thereby reducing the time required for complex problem solving. By breaking down the number of weights enables the simultaneous programing for the synaptic weights. Therefore, parallel programming presents a promising avenue for addressing the time-consuming nature of problem-solving computations.

6.5 PCM Reliability

Obtaining these results was not easy due to dealing with large number of reliability issues in PCM devices as conductance drift is which is a significant reliability issue for PCM (Figure 6.21.a). The impact of conductance drift on PCM synaptic devices has been studied, and it has been found to cause minor accuracy degradation of around 1% for both multilayer perceptron (MLP) and convolutional neural network (CNN) models (Oh et al., 2019).

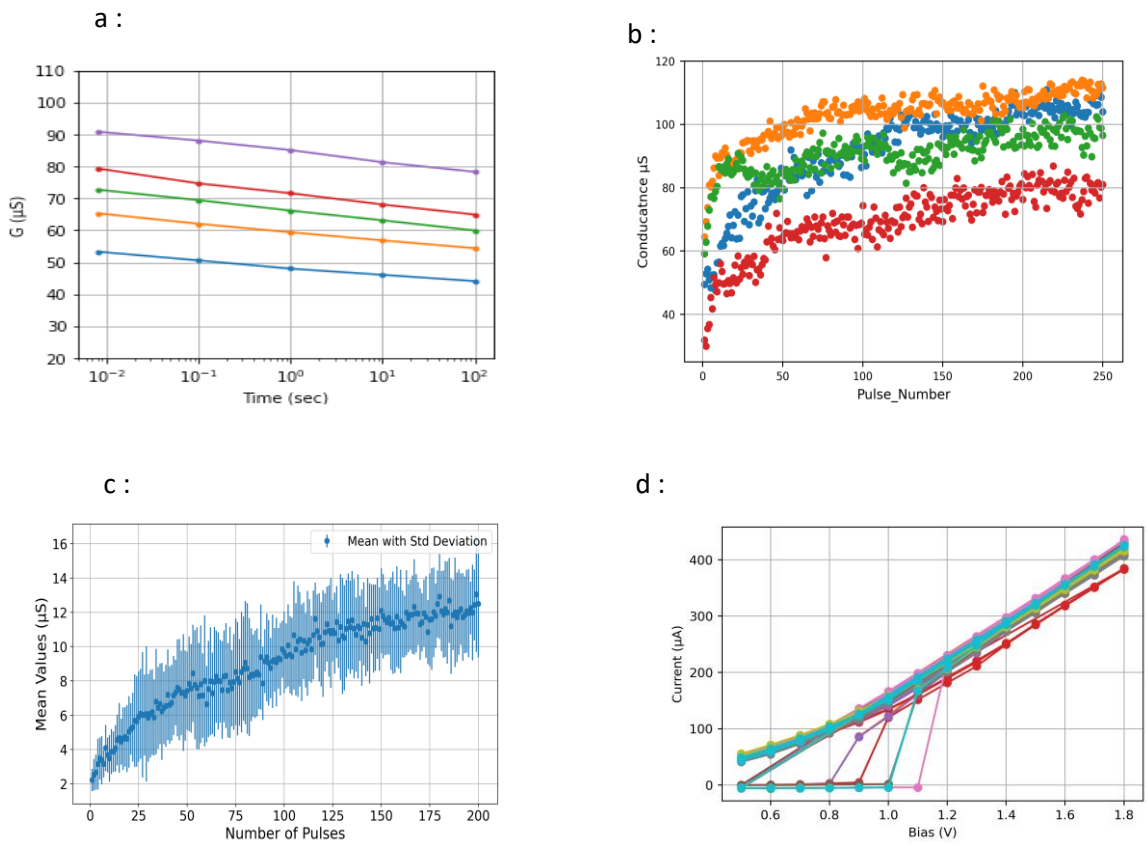


Figure 6. 21 : a) Drift measurements at room temperature starting from different G_{init} over 100 seconds. b) Cycle-to-Cycle analysis at the same frequency. c) Progressive crystallization performed using multiple devices with the same frequency. d) Threshold switching test performed on multiple devices.

Stochasticity and variability cycle-to-cycle and device-to-device of PCM behavior and retention (Figure 6.21.b) are also important factors. In this experiment, we ran the same test on the same device and at the same frequency. The results depicted in the figure reveal that the final saturation values differ, even in the initial stages. This discrepancy is evident in the significant standard deviation shown in Figure 6.21.c.

PCM inherently exhibits stochastic behavior in its operating principle, and data retention is influenced by factors such as resistance drift, noise, and temperature dependence. The observed stochasticity is further compounded by the fact that devices (we utilized 10 devices in this example) switch at different threshold switching voltages and with varying numbers of pulses (Figure 6.21.d). Understanding these factors is crucial for optimizing PCM technology for emerging non-von Neumann computing applications (Gallo & Sebastian, 2020).

Reading and programming disturbs in PCM arrays and endurance of PCM are additional important considerations. Reading and programming disturbs can affect the performance of PCM arrays, and the endurance of PCM refers to the number of read and write cycles that the device can reliably withstand. Mitigation strategies for resistance drift in neuromorphic systems have been proposed to address the impact of resistance drift on PCM devices (M. Suri, Garbin, et al., 2013).

Conclusion:

In conclusion, our work introduces an implementation of a frequency modulation scheme utilizing 28nm FDSOI technology on a 300mm silicon for synaptic weight transfer. This approach, converts pre-calculated weights into conductance values and further into frequency, showcasing the novel potential of this research. The subsequent step involves evaluating the accuracy of a CNN trained and programmed with PCM-based weights, affirming the practical application of our algorithm and validating the CNN model's precision.

Addressing inherent variability in PCM parameters through redundancy schemes is paramount to ensuring system robustness and reliability amid real-world variations. Our extensive experimentation underscores the effectiveness of the frequency modulation technique, achieving an accuracy rate of up to 90%. This outcome not only emphasizes the viability of our algorithm and PCM-based weight transfer scheme but also positions our work as a strong proof of concept for a CNN model based on frequency modulation compare with the state of the art.

In summary, our research represents a new technique, offering a fresh perspective on CNN inference through frequency modulation. By combining insights from weight mapping, simulations, and performance analysis, we present a new way for future advancements. The integration of frequency modulation and PCM technology holds promise for enhancing computational efficiency and reliability, influencing the trajectory of neural networks and making contributions to fields such as artificial intelligence and hardware development.

Résumé

- Explored CNN.
- Foundational understanding of CNN backpropagation for adaptability.
- Key focus on mapping synaptic weights for optimized CNN performance.
- Conducted extensive synaptic behavior simulations.
- Validated approach thorough performance analysis.
- Addressed PCM issues, ensuring robustness in real-world scenarios.
- Comprehensive contribution to CNN inference understanding.
- Demonstrated PCM reliability as a critical component.
- Pivotal advancement with broad implications for neural network applications.
- Positioned frequency modulation as a key player in the future of computing.

Conclusions and future work

"A thinker sees his own actions as experiments and questions--as attempts to find out something. Success and failure are for him answers above all."

Friedrich Nietzsche.

Phase-Change Memory (PCM) technology has shown great potential for neuromorphic applications. PCM is based on the reversible and rapid phase transition between amorphous and crystalline phases of certain materials, making it well-suited for synaptic realizations in neuromorphic computing (Nandakumar et al., 2018). The ability to alter the conductance levels in a controllable way and the progressive crystallization of the phase-change material make PCM devices particularly suitable for this application (Nandakumar et al., 2018).

Several studies have demonstrated the use of PCM as a synapse in ultra-dense large-scale neuromorphic systems, highlighting its energy-efficient methodology (M. Suri et al., 2011). Despite the inherent weaknesses of PCM-based neuromorphic devices, such as resistance drift and performance metrics, efforts are being made to address these challenges and advance the technology of artificial neural networks (L. Wang et al., 2017). Additionally, neuromorphic photonics devices based on phase-change materials have emerged as promising solutions for addressing neuromorphic computing challenges (Li et al., 2023). Therefore, PCM technology holds significant promise for neuromorphic applications, and ongoing research aims to overcome its limitations and further optimize its suitability for the embedded market requirements.

In the first chapter of our work, we explored Neuromorphic Computing. We started by understanding how our brains work with interconnected neurons. Scientists are trying to replicate this in computers to boost their processing power. We also learned about synapses, where learning happens in artificial

neural networks. We then covered different network designs, like feed-forward and convolutional neural networks, which help machines process data like humans. Finally, we discussed the challenges Neuromorphic Computing faces, including technical issues and ethical concerns and how solving these problems will lead to a more advanced and responsible era for Neuromorphic Computing.

In the second chapter, we explored semiconductor memory technologies and their applications in neuromorphic computing, covering CBRAM, OXRAM, FeRAM, STT-RAM, and PCRAM. We conducted a thorough comparison of these non-volatile memory (NVM) technologies to understand their strengths and weaknesses, aiding in informed decision-making for specific neuromorphic applications. Our investigation also extended to emerging memory technologies, emphasizing the ongoing importance of research and development. This knowledge equips us to connect memory technologies for brain-inspired computational systems. The chapter provides a solid foundation for further exploration into neuromorphic computing and its potential in cognitive computing and machine learning.

In the third chapter, we explored the combination of neuromorphic computing and non-volatile memory (NVM) technologies, highlighting essential concepts. We started by examining Spike-timing-dependent-plasticity (STDP), a key mechanism that mimics learning in biological neural networks. We then focused on the core operation of neuromorphic computing: vector-matrix multiplication, essential for tasks like pattern recognition. We discussed the phase transition of Phase-Change Memory (PCM), stressing its potential as both a synapse and a neuron. The chapter also covered various neuromorphic computing applications, from cognitive computing to artificial intelligence and robotics, showcasing its transformative potential.

In the fourth chapter, we investigate the PCM technology beyond the conventional von Neumann architecture, providing insights into the evolution of computing. We present a historical overview that traces PCM's progression from concept to practical use. The examination of the PCM cell and its fundamental functionalities is undertaken, including the presentation of SET/RESET operations. Understanding the phase-change mechanism and associated switching processes unveils PCM's memory and multilevel operations. Challenges such as resistance drift are highlighted, emphasizing the imperative for ongoing research to enhance stability and reliability. The chapter further explores phase-change materials, with a specific focus on $Ge_2Sb_2Te_5$ and Ge-rich GST, highlighting the role of material selection in optimizing PCM performance. A comparison between GST(225) and Ge-rich GST underscores the role of the phase-change material in the effectiveness of PCM technology, providing insights for future advancements.

In Chapter 5, we introduce a new method to modulate the conductance levels in PCM devices. This technique represents an improving how we use multi-memristive ideas. This chapter highlights the role of phase change in modelling the programmability of synaptic weights, promising potential for neural network applications. The concept of multi-memristive systems offers a nuanced perspective on memory and computation. The chapter explores the complexities of programming synaptic weights, emphasizing the versatility of phase-change materials in supporting this essential aspect of neural network functionality. In this chapter, frequency modulation emerges as a new technique for manipulating conductance levels in phase-change memory devices. Both experimental and simulated approaches demonstrate the effectiveness of frequency modulation in achieving desired conductance

changes. The combination of experimental results and simulations not only confirms the feasibility of frequency modulation but also enhances our understanding of its potential applications.

The final chapter of the thesis introduces the implementation of a frequency modulation scheme using 28nm FDSOI technology on a 300mm silicon for synaptic weight transfer. This approach converts pre-calculated weights into conductance values and then into frequency, showcasing the potential of this research. The next step involves assessing the accuracy of a CNN trained and programmed with PCM-based weights, confirming the practical application of our algorithm and validating the CNN model precision. Managing inherent variability in PCM parameters through redundancy schemes is crucial for ensuring system robustness amid real-world variations. Our experimentation demonstrates the effectiveness of the frequency modulation technique, achieving an accuracy rate of up to 90%. This result highlights not only the viability of our algorithm and PCM-based weight transfer scheme but also positions our work as a strong proof of concept for a CNN model based on frequency modulation compared to the state of the art.

To introduce the potential of the future works, there are several possibilities on the device side could be the development of advanced computational models to simulate the behavior of phase change materials on scaled architectures. These models could help in understanding the properties of the materials, thus contributing to the reduction of costs and the improvement of stability in phase change material devices. Additionally, further research on novel phase change materials with enhanced properties and sustainability could be another potential avenue for future works in this field.

On the algorithm development part the potential for improving various aspects of machine learning models. One notable benchmark for evaluating these improvements is to compare it with our MNIST dataset results. Some alternative datasets commonly used for comparative analysis include CIFAR-10 and CIFAR-100, which consist of color images across multiple classes, and ImageNet, a large-scale dataset encompassing a diverse range of object categories. By comparing outcomes across these datasets, we can gain deeper insights into algorithmic strengths and weaknesses.

In the context of frequency modulation, it is essential to refine our techniques, recognizing the challenge of extended experimental durations. To overcome this limitation, we can concentrate on optimizing the experimental setup, leveraging high-quality components, and integrating advanced modulation equipment. Exploring parallelization strategies for simultaneous experiments could substantially reduce the overall time investment, while real-time monitoring and analysis will further enhance the efficiency of our methodology.

References:

1. Ad, K., & Dc, W. (2015). A new spin on magnetic memories. Nature Nanotechnology, 10(3). https://doi.org/10.1038/nnano.2015.24

2. Alibart, F., Gao, L., Hoskins, B. D., & Strukov, D. B. (2012). High precision tuning of state for memristive devices by adaptable variation-tolerant algorithm. Nanotechnology, 23(7), 075201. https://doi.org/10.1088/0957-4484/23/7/075201

3. Alibart, F., Zamanidoost, E., & Strukov, D. B. (2013). Pattern classification by memristive crossbar circuits using ex situ and in situ training. Nature Communications, 4, 2072. https://doi.org/10.1038/ncomms3072

4. Alom, Md. Z., Taha, T., Yakopcic, C., Westberg, S., Hasan, M., Esesn, B., Awwal, A., & Asari, V. (2018). The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches.

5. Al-Qizwini, M., Barjasteh, I., Al-Qassab, H., & Radha, H. (2017). Deep learning algorithm for autonomous driving using GoogLeNet. 2017 IEEE Intelligent Vehicles Symposium (IV), 89–96. https://doi.org/10.1109/IVS.2017.7995703

6. Ambrogio, S., Ciocchini, N., Laudato, M., Milo, V., Pirovano, A., Fantini, P., & Ielmini, D. (2016). Unsupervised Learning by Spike Timing Dependent Plasticity in Phase Change Memory (PCM) Synapses. Frontiers in Neuroscience, 10. https://www.frontiersin.org/article/10.3389/fnins.2016.00056

7. Andrew, A. M. (2003). Spiking Neuron Models: Single Neurons, Populations, Plasticity. Kybernetes, 32(7/8). https://doi.org/10.1108/k.2003.06732gae.003

8. Antolini, A., Paolino, C., Zavalloni, F., Lico, A., Scarselli, E. F., Mangia, M., Pareschi, F., Setti, G., Rovatti, R., Torres, M. L., Carissimi, M., & Pasotti, M. (2023). Combined HW/SW Drift and Variability Mitigation for PCM-Based Analog In-Memory Computing for Neural Network Applications. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 13(1), 395–407. https://doi.org/10.1109/JETCAS.2023.3241750

9. Appeltant, L., Soriano, M. C., Van der Sande, G., Danckaert, J., Massar, S., Dambre, J., Schrauwen, B., Mirasso, C. R., & Fischer, I. (2011). Information processing using a single dynamical node as complex system. Nature Communications, 2, 468. https://doi.org/10.1038/ncomms1476

10. Arnaud, F., Zuliani, P., Reynard, J. P., Gandolfo, A., Disegni, F., Mattavelli, P., Gomiero, E., Samanni, G., Jahan, C., Berthelon, R., Weber, O., Richard, E., Barral, V., Villaret, A., Kohler, S., Grenier, J. C., Ranica, R., Gallon, C., Souhaite, A., … Cappelletti, P. (2018). Truly Innovative 28nm FDSOI Technology for Automotive Micro-Controller Applications embedding 16MB Phase Change

Memory. 2018 IEEE International Electron Devices Meeting (IEDM), 18.4.1-18.4.4. https://doi.org/10.1109/IEDM.2018.8614595

11. Barci, M., Guy, J., Molas, G., Vianello, E., Toffoli, A., Cluzel, J., Roule, A., Bernard, M., Sabbione, C., Perniola, L., & De Salvo, B. (2014). Impact of SET and RESET conditions on CBRAM high temperature data retention. 2014 IEEE International Reliability Physics Symposium, 5E.3.1-5E.3.4. https://doi.org/10.1109/IRPS.2014.6860677

12. Bayat, F. M., Hoskins, B., & Strukov, D. B. (2015). Phenomenological modeling of memristive devices. APPLIED PHYSICS A-MATERIALS SCIENCE & PROCESSING, 118(3), 779–786. https://doi.org/10.1007/s00339-015-8993-7

13. Benjamin, B. V., Gao, P., McQuinn, E., Choudhary, S., Chandrasekaran, A. R., Bussat, J.-M., Alvarez-Icaza, R., Arthur, J. V., Merolla, P. A., & Boahen, K. (2014). Neurogrid: A Mixed-Analog-Digital Multichip System for Large-Scale Neural Simulations. Proceedings of the IEEE, 102(5), 699–716. https://doi.org/10.1109/JPROC.2014.2313565

14. Berger, L. (1996). Emission of spin waves by a magnetic multilayer traversed by a current. Physical Review. B, Condensed Matter, 54(13), 9353–9358. https://doi.org/10.1103/physrevb.54.9353

15. Beyond von Neumann. (2020). Nature Nanotechnology, 15(7), 7. https://doi.org/10.1038/s41565-020-0738-x

16. Bichler, O., Suri, M., Querlioz, D., Vuillaume, D., DeSalvo, B., & Gamrat, C. (2012). Visual Pattern Extraction Using Energy-Efficient "2-PCM Synapse" Neuromorphic Architecture. IEEE Transactions on Electron Devices, 59(8), 2206–2214. https://doi.org/10.1109/TED.2012.2197951

17. Bogoslovskiy, N., & Tsendin, K. (2012). Physics of switching and memory effects in chalcogenide glassy semiconductors. Semiconductors, 46. https://doi.org/10.1134/S1063782612050065

18. Boniardi, M., Redaelli, A., Cupeta, C., Pellizzer, F., Crespi, L., D'Arrigo, G., Lacaita, A. L., & Servalli, G. (2014). Optimization metrics for Phase Change Memory (PCM) cell architectures. 2014 IEEE International Electron Devices Meeting, 29.1.1-29.1.4. https://doi.org/10.1109/IEDM.2014.7047131

19. Bordas, S., Clavaguer-Mora, M. T., Legendre, B., & Hancheng, C. (1986). Phase diagram of the ternary system Ge-Sb-Te: II. The subternary Ge-GeTe-Sb2Te3-Sb. Thermochimica Acta, 107, 239–265. https://doi.org/10.1016/0040-6031(86)85051-1

20. Böscke, T. S., Müller, J., Bräuhaus, D., Schröder, U., & Böttger, U. (2011). Ferroelectricity in hafnium oxide thin films. Applied Physics Letters, 99(10), 102903. https://doi.org/10.1063/1.3634052

21. Boybat, I., Le Gallo, M., Nandakumar, S. R., Moraitis, T., Parnell, T., Tuma, T., Rajendran, B., Leblebici, Y., Sebastian, A., & Eleftheriou, E. (2018). Neuromorphic computing with multi-memristive synapses. Nature Communications, 9(1), 2514. https://doi.org/10.1038/s41467-018-04933-y

22. Broadhurst, D. I., & Kell, D. B. (2006). Statistical strategies for avoiding false discoveries in metabolomics and related experiments. Metabolomics, 2(4), 171–196. https://doi.org/10.1007/s11306-006-0037-z

23. Burr, G. W., Brightsky, M. J., Sebastian, A., Cheng, H.-Y., Wu, J.-Y., Kim, S., Sosa, N. E., Papandreou, N., Lung, H.-L., Pozidis, H., Eleftheriou, E., & Lam, C. H. (2016). Recent Progress in Phase-Change Memory Technology. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 6(2), 146–162. https://doi.org/10.1109/JETCAS.2016.2547718

24. Burr, G. W., Narayanan, P., Shelby, R. M., Sidler, S., Boybat, I., di Nolfo, C., & Leblebici, Y. (2015). Large-scale neural networks implemented with non-volatile memory as the synaptic weight element: Comparative performance analysis (accuracy, speed, and power). 2015 IEEE International Electron Devices Meeting (IEDM), 4.4.1-4.4.4. https://doi.org/10.1109/IEDM.2015.7409625

25. Burr, G. W., Shelby, R. M., Sebastian, A., Kim, S., Kim, S., Sidler, S., Virwani, K., Ishii, M., Narayanan, P., Fumarola, A., Sanches, L. L., Boybat, I., Gallo, M. L., Moon, K., Woo, J., Hwang, H., & Leblebici, Y. (2017). Neuromorphic computing using non-volatile memory. Advances in Physics: X, 2(1), 89–124. https://doi.org/10.1080/23746149.2016.1259585

26. Burr, G. W., Shelby, R. M., Sidler, S., di Nolfo, C., Jang, J., Boybat, I., Shenoy, R. S., Narayanan, P., Virwani, K., Giacometti, E. U., Kurdi, B. N., & Hwang, H. (2015). Experimental Demonstration and Tolerancing of a Large-Scale Neural Network (165 000 Synapses) Using Phase-Change Memory as the Synaptic Weight Element. IEEE Transactions on Electron Devices, 62(11), 3498–3507. https://doi.org/10.1109/TED.2015.2439635

27. Bush, P., & Sejnowski, T. (1996). Inhibition synchronizes sparsely connected cortical neurons within and between columns in realistic network models. Journal of Computational Neuroscience, 3(2), 91–110. https://doi.org/10.1007/BF00160806

28. Camuñas-Mesa, L. A., Linares-Barranco, B., & Serrano-Gotarredona, T. (2019). Neuromorphic Spiking Neural Networks and Their Memristor-CMOS Hardware Implementations. Materials, 12(17), 2745. https://doi.org/10.3390/ma12172745

29. Chabi, D., Querlioz, D., Zhao, W., & Klein, J.-O. (2014). Robust Learning Approach for Neuro-Inspired Nanoscale Crossbar Architecture. ACM JOURNAL ON EMERGING TECHNOLOGIES IN COMPUTING SYSTEMS, 10(1), 5. https://doi.org/10.1145/2539123

30. Chen, A. (2016). A review of emerging non-volatile memory (NVM) technologies and applications. Solid-State Electronics, 125, 25–38. https://doi.org/10.1016/j.sse.2016.07.006

31. Chen, P.-Y., Lin, B., Wang, I.-T., Hou, T.-H., Ye, J., Vrudhula, S., Seo, J., Cao, Y., & Yu, S. (2015). Mitigating effects of non-ideal synaptic device characteristics for on-chip learning. 2015 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), 194–199. https://doi.org/10.1109/ICCAD.2015.7372570

32. Cheng, H. Y., Hsu, T. H., Raoux, S., Wu, J. Y., Du, P. Y., Breitwisch, M., Zhu, Y., Lai, E. K., Joseph, E., Mittal, S., Cheek, R., Schrott, A., Lai, S. C., Lung, H. L., & Lam, C. (2011). A high performance phase change memory with fast switching speed and high temperature retention by engineering the GexSbyTez phase change material. 2011 International Electron Devices Meeting, 3.4.1-3.4.4. https://doi.org/10.1109/IEDM.2011.6131481

33. Chicca, E., & Indiveri, G. (2020). A recipe for creating ideal hybrid memristive-CMOS neuromorphic processing systems. Applied Physics Letters, 116(12), 120501. https://doi.org/10.1063/1.5142089

34. Chicca, E., Stefanini, F., Bartolozzi, C., & Indiveri, G. (2014). Neuromorphic Electronic Circuits for Building Autonomous Cognitive Systems. Proceedings of the IEEE, 102(9), 1367–1388. https://doi.org/10.1109/JPROC.2014.2313954

35. Choi, S., Sheridan, P., & Lu, W. D. (2015). Data Clustering using Memristor Networks. Scientific Reports, 5, 10492. https://doi.org/10.1038/srep10492

36. Choi, W., Duraisamy, K., Kim, R. G., Doppa, J. R., Pande, P. P., Marculescu, D., & Marculescu, R. (2018). On-Chip Communication Network for Efficient Training of Deep Convolutional Networks on Heterogeneous Manycore Systems. IEEE Transactions on Computers, 67(5), 672–686. https://doi.org/10.1109/TC.2017.2777863

37. Choi, Y., Song, I., Park, M.-H., Chung, H., Chang, S., Cho, B., Kim, J., Oh, Y., Kwon, D., Sunwoo, J., Shin, J., Rho, Y., Lee, C., Kang, M. G., Lee, J., Kwon, Y., Kim, S., Kim, J., Lee, Y.-J., … Jeong, G. (2012). A 20nm 1.8V 8Gb PRAM with 40MB/s program bandwidth. 2012 IEEE International Solid-State Circuits Conference, 46–48. https://doi.org/10.1109/ISSCC.2012.6176872

38. Christensen, D. V., Dittmann, R., Linares-Barranco, B., Sebastian, A., Le Gallo, M., Redaelli, A., Slesazeck, S., Mikolajick, T., Spiga, S., Menzel, S., Valov, I., Milano, G., Ricciardi, C., Liang, S.-J., Miao, F., Lanza, M., Quill, T. J., Keene, S. T., Salleo, A., … Pryds, N. (2022). 2022 roadmap on neuromorphic computing and engineering. Neuromorphic Computing and Engineering, 2(2), 022501. https://doi.org/10.1088/2634-4386/ac4a83

39. Ciregan, D., Meier, U., & Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. 2012 IEEE Conference on Computer Vision and Pattern Recognition, 3642–3649. https://doi.org/10.1109/CVPR.2012.6248110

40. Dayarathna, M., Wen, Y., & Fan, R. (2016). Data Center Energy Consumption Modeling: A Survey. IEEE Communications Surveys & Tutorials, 18(1), 732–794. https://doi.org/10.1109/COMST.2015.2481183

41. Deng, L. (2012). The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]. IEEE Signal Processing Magazine, 29(6), 141–142. https://doi.org/10.1109/MSP.2012.2211477

42. Eaton, D. L. (1964). Electrical Conduction Anomaly of Semiconducting Glasses in the System As—Te—I. Journal of the American Ceramic Society, 47(11), 554–558. https://doi.org/10.1111/j.1151-2916.1964.tb13816.x

43. Eryilmaz, S. B., Kuzum, D., Jeyasingh, R. G. D., Kim, S., BrightSky, M., Lam, C., & Wong, H.-S. P. (2013). Experimental demonstration of array-level learning with phase change synaptic devices. 2013 IEEE International Electron Devices Meeting, 25.5.1-25.5.4. https://doi.org/10.1109/IEDM.2013.6724691

44. Eryilmaz, S. B., Kuzum, D., Yu, S., & Wong, H. S. P. (2015). Device and system level design considerations for analog-non-volatile-memory based neuromorphic architectures: 61st IEEE International Electron Devices Meeting, IEDM 2015. 2015 IEEE International Electron Devices Meeting, IEDM 2015, 4.1.1-4.1.4. https://doi.org/10.1109/IEDM.2015.7409622

45. Esser, S. K., Merolla, P. A., Arthur, J. V., Cassidy, A. S., Appuswamy, R., Andreopoulos, A., Berg, D. J., McKinstry, J. L., Melano, T., Barch, D. R., di Nolfo, C., Datta, P., Amir, A., Taba, B., Flickner, M. D., & Modha, D. S. (2016). Convolutional Networks for Fast, Energy-Efficient Neuromorphic

Computing. Proceedings of the National Academy of Sciences, 113(41), 11441–11446. https://doi.org/10.1073/pnas.1604850113

46. Feldman, D. E. (2012). The spike timing dependence of plasticity. Neuron, 75(4), 556–571. https://doi.org/10.1016/j.neuron.2012.08.001

47. Fujimoto, M., Koyama, H., Konagai, M., Hosoi, Y., Ishihara, K., Ohnishi, S., & Awaya, N. (2006). TiO2 Anatase Nanolayer on TiN Thin Film Exhibiting High-Speed Bipolar Resistive Switching. Applied Physics Letters, 89, 223509–223509. https://doi.org/10.1063/1.2397006

48. Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, 36(4), 193–202. https://doi.org/10.1007/BF00344251

49. Furber, S. (2016). Large-scale neuromorphic computing systems. Journal of Neural Engineering, 13(5), 051001. https://doi.org/10.1088/1741-2560/13/5/051001

50. Furber, S. B., Galluppi, F., Temple, S., & Plana, L. A. (2014). The SpiNNaker Project. Proceedings of the IEEE, 102(5), 652–665. https://doi.org/10.1109/JPROC.2014.2304638

51. Gallo, M. L., Kaes, M., Sebastian, A., & Krebs, D. (2015). Subthreshold electrical transport in amorphous phase-change materials. New Journal of Physics, 17(9), 093035. https://doi.org/10.1088/1367-2630/17/9/093035

52. Gallo, M. L., & Sebastian, A. (2020). An overview of phase-change memory device physics. Journal of Physics D: Applied Physics, 53(21), 213002. https://doi.org/10.1088/1361-6463/ab7794

53. Gamrat, C., Bichler, O., & Roclin, D. (2015). Memristive based device arrays combined with Spike based coding can enable efficient implementations of embedded neuromorphic circuits. 2015 IEEE International Electron Devices Meeting (IEDM), 4.5.1-4.5.7. https://doi.org/10.1109/IEDM.2015.7409626

54. Gao, L., Wang, I.-T., Chen, P.-Y., Vrudhula, S., Seo, J., Cao, Y., Hou, T.-H., & Yu, S. (2015). Fully parallel write/read in resistive synaptic array for accelerating on-chip learning. Nanotechnology, 26(45), 455204. https://doi.org/10.1088/0957-4484/26/45/455204

55. Garbin, D., Vianello, E., Bichler, O., Rafhay, Q., Gamrat, C., Ghibaudo, G., DeSalvo, B., & Perniola, L. (2015). HfO2-Based OxRAM Devices as Synapses for Convolutional Neural Networks. IEEE TRANSACTIONS ON ELECTRON DEVICES, 62(8), 2494–2501. https://doi.org/10.1109/TED.2015.2440102

56. Giannopoulos, I., Singh, A., Le Gallo, M., Jonnalagadda, V. P., Hamdioui, S., & Sebastian, A. (2020). In-Memory Database Query. Advanced Intelligent Systems, 2(12), 2000141. https://doi.org/10.1002/aisy.202000141

57. Godfrey, M. (1993). First Draft Report on the EDVAC by John von Neumann. IEEE Annals of the History of Computing, 15, 27–43.

58. Gokmen, T., & Vlasov, Y. (2016). Acceleration of Deep Neural Network Training with Resistive Cross-Point Devices: Design Considerations. Frontiers in Neuroscience, 10. https://www.frontiersin.org/articles/10.3389/fnins.2016.00333

59. Greengard, S. (2020). Neuromorphic chips take shape. Communications of the ACM, 63(8), 9–11. https://doi.org/10.1145/3403960

60. Gruning, A., & Bohte, S. M. (2014). Spiking Neural Networks: Principles and Challenges. Computational Intelligence.

61. Guo, P., Sarangan, A. M., & Agha, I. (2019). A Review of Germanium-Antimony-Telluride Phase Change Materials for Non-Volatile Memories and Optical Modulators. Applied Sciences, 9(3), 3. https://doi.org/10.3390/app9030530

62. Guy, J., Molas, G., Vianello, E., Longnos, F., Blanc, S., Carabasse, C., Bernard, M., Nodin, J. F., Toffoli, A., Cluzel, J., Blaise, P., Dorion, P., Cueto, O., Grampeix, H., Souchier, E., Cabout, T., Brianceau, P., Balan, V., Roule, A., … De Salvo, B. (2013). Investigation of the physical mechanisms governing data-retention in down to 10nm nano-trench Al2O3/CuTeGe conductive bridge RAM (CBRAM). 2013 IEEE International Electron Devices Meeting, 30.2.1-30.2.4. https://doi.org/10.1109/IEDM.2013.6724722

63. Herculano-Houzel, S. (2009). The human brain in numbers: A linearly scaled-up primate brain. Frontiers in Human Neuroscience, 3. https://www.frontiersin.org/articles/10.3389/neuro.09.031.2009

64. Ho, Y., & Wookey, S. (2020). The Real-World-Weight Cross-Entropy Loss Function: Modeling the Costs of Mislabeling. IEEE Access, 8, 4806–4813. https://doi.org/10.1109/ACCESS.2019.2962617

65. Hu, E. H., & Bloomfield, S. A. (2003). Gap junctional coupling underlies the short-latency spike synchrony of retinal alpha ganglion cells. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 23(17), 6768–6777. https://doi.org/10.1523/JNEUROSCI.23-17-06768.2003

66. Ielmini, D., & Lacaita, A. (2011). Phase change materials in non-volatile storage. Materials Today, 14, 600–607. https://doi.org/10.1016/S1369-7021(11)70301-7

67. Indiveri, G. (2021). Introducing 'Neuromorphic Computing and Engineering.' Neuromorphic Computing and Engineering, 1(1), 010401. https://doi.org/10.1088/2634-4386/ac0a5b

68. Indiveri, G., Linares-Barranco, B., Legenstein, R., Deligeorgis, G., & Prodromakis, T. (2013). Integration of nanoscale memristor synapses in neuromorphic computing architectures. Nanotechnology, 24(38), 384010. https://doi.org/10.1088/0957-4484/24/38/384010

69. Ivanov, A. (2023). Quantization of Deep Neural Networks to facilitate self-correction of weights on Phase Change Memory-based analog hardware (arXiv:2310.00337). arXiv. http://arxiv.org/abs/2310.00337

70. Izhikevich, E. M. (2006). Polychronization: Computation with Spikes. Neural Computation, 18(2), 245–282. https://doi.org/10.1162/089976606775093882

71. Jackson, B. L., Rajendran, B., Corrado, G. S., Breitwisch, M., Burr, G. W., Cheek, R., Gopalakrishnan, K., Raoux, S., Rettner, C. T., Padilla, A., Schrott, A. G., Shenoy, R. S., Kurdi, B. N., Lam, C. H., & Modha, D. S. (2013). Nanoscale Electronic Synapses Using Phase Change Devices. ACM JOURNAL ON EMERGING TECHNOLOGIES IN COMPUTING SYSTEMS, 9(2), 12. https://doi.org/10.1145/2463585.2463588

72. Jaguemont, J., Omar, N., Van den Bossche, P., & Mierlo, J. (2018). Phase-change materials (PCM) for automotive applications: A review. Applied Thermal Engineering, 132, 308–320. https://doi.org/10.1016/j.applthermaleng.2017.12.097

73. Jang, J.-W., Park, S., Burr, G. W., Hwang, H., & Jeong, Y.-H. (2015). Optimization of Conductance Change in Pr1–xCaxMnO3-Based Synaptic Devices for Neuromorphic Systems. IEEE Electron Device Letters, 36(5), 457–459. https://doi.org/10.1109/LED.2015.2418342

74. Jeyasingh, R., Fong, S. W., Lee, J., Li, Z., Chang, K.-W., Mantegazza, D., Asheghi, M., Goodson, K. E., & Wong, H.-S. P. (2014). Ultrafast Characterization of Phase-Change Material Crystallization Properties in the Melt-Quenched Amorphous Phase. Nano Letters, 14(6), 3419–3426. https://doi.org/10.1021/nl500940z

75. Jo, S. H., Chang, T., Ebong, I., Bhadviya, B. B., Mazumder, P., & Lu, W. (2010). Nanoscale memristor device as synapse in neuromorphic systems. Nano Letters, 10(4), 1297–1301. https://doi.org/10.1021/nl904092h

76. Joshi, V., Le Gallo, M., Haefeli, S., Boybat, I., Nandakumar, S. R., Piveteau, C., Dazzi, M., Rajendran, B., Sebastian, A., & Eleftheriou, E. (2020). Accurate deep neural network inference using computational phase-change memory. Nature Communications, 11(1), 2473. https://doi.org/10.1038/s41467-020-16108-9

77. Julliere, M. (1975). Tunneling between ferromagnetic films. Physics Letters A, 54(3), 225–226. https://doi.org/10.1016/0375-9601(75)90174-7

78. Kalb, J. A., Spaepen, F., & Wuttig, M. (2005). Kinetics of crystal nucleation in undercooled droplets of Sb- and Te-based alloys used for phase change recording. Journal of Applied Physics, 98(5), 054910. https://doi.org/10.1063/1.2037870

79. Karpov, V. G., Kryukov, Y. A., Savransky, S. D., & Karpov, I. V. (2007). Nucleation switching in phase change memory. Applied Physics Letters, 90(12), 123504. https://doi.org/10.1063/1.2715024

80. Karunaratne, G., Le Gallo, M., Cherubini, G., Benini, L., Rahimi, A., & Sebastian, A. (2020). In-memory hyperdimensional computing. Nature Electronics, 3(6), 6. https://doi.org/10.1038/s41928-020-0410-3

81. Kataeva, I., Merrikh-Bayat, F., Zamanidoost, E., & Strukov, D. (2015). Efficient training algorithms for neural networks based on memristive crossbar circuits. 2015 International Joint Conference on Neural Networks (IJCNN), 1–8. https://doi.org/10.1109/IJCNN.2015.7280785

82. Keysers, C., & Gazzola, V. (2014). Hebbian learning and predictive mirror neurons for actions, sensations and emotions. Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1644), 20130175. https://doi.org/10.1098/rstb.2013.0175

83. Khaled Ahmed, S., Mohammed Ali, R., Maha Lashin, M., & Fayroz Sherif, F. (2023). Designing a new fast solution to control isolation rooms in hospitals depending on artificial intelligence decision. Biomedical Signal Processing and Control, 79, 104100. https://doi.org/10.1016/j.bspc.2022.104100

84. Khvalkovskiy, A., Apalkov, D., Watts, S., Chepulskii, R., Beach, R., Ong, A., Tang, X., Driskill-Smith, A., Butler, W., Visscher, P., Lottis, D., Chen, E., Nikitin, V., & Krounbi, M. (2013). Basic principles of STT-MRAM cell operation in memory arrays. Journal of Physics D: Applied Physics, 46, 074001. https://doi.org/10.1088/0022-3727/46/7/074001

85. Kim, R. G., Doppa, J. R., Pande, P. P., Marculescu, D., & Marculescu, R. (2018). Machine Learning and Manycore Systems Design: A Serendipitous Symbiosis. Computer, 51(7), 66–77. https://doi.org/10.1109/MC.2018.3011040

86. Kim, W., Bruce, R. L., Masuda, T., Fraczak, G. W., Gong, N., Adusumilli, P., Ambrogio, S., Tsai, H., Bruley, J., Han, J.-P., Longstreet, M., Carta, F., Suu, K., & BrightSky, M. (2019). Confined PCM-based Analog Synaptic Devices offering Low Resistance-drift and 1000 Programmable States for Deep Learning. 2019 Symposium on VLSI Technology, T66–T67. https://doi.org/10.23919/VLSIT.2019.8776551

87. Koelmans, W. W., Sebastian, A., Jonnalagadda, V. P., Krebs, D., Dellmann, L., & Eleftheriou, E. (2015). Projected phase-change memory devices. Nature Communications, 6(1), 1. https://doi.org/10.1038/ncomms9181

88. Kolb, B., & Gibb, R. (2011). Brain Plasticity and Behaviour in the Developing Brain. Journal of the Canadian Academy of Child and Adolescent Psychiatry, 20(4), 265–276.

89. Krebs, D., Raoux, S., Rettner, C. T., Burr, G. W., Salinga, M., & Wuttig, M. (2009). Threshold field of phase change memory materials measured using phase change bridge devices. Applied Physics Letters, 95(8), 082101. https://doi.org/10.1063/1.3210792

90. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems, 25. https://proceedings.neurips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

91. Kumar, M., & Suri, M. (2023). Hybrid CMOS-PCM Ternary Logic for Digital Circuit Applications. IEEE Transactions on Nanotechnology, 22, 228–237. https://doi.org/10.1109/TNANO.2023.3272831

92. Kuzum, D., Jeyasingh, R. G. D., Lee, B., & Wong, H.-S. P. (2012). Nanoelectronic Programmable Synapses Based on Phase Change Materials for Brain-Inspired Computing. Nano Letters, 12(5), 2179–2186. https://doi.org/10.1021/nl201040y

93. Kuzum, D., Yu, S., & Wong, H.-S. P. (2013). Synaptic electronics: Materials, devices and applications. Nanotechnology, 24(38), 382001. https://doi.org/10.1088/0957-4484/24/38/382001

94. Le Gallo, M., Athmanathan, A., Krebs, D., & Sebastian, A. (2016). Evidence for thermally assisted threshold switching behavior in nanoscale phase-change memory cells. Journal of Applied Physics, 119(2), 025704. https://doi.org/10.1063/1.4938532

95. Le Gallo, M., & Sebastian, A. (2020). Chapter 3—Phase-change memory. In S. Spiga, A. Sebastian, D. Querlioz, & B. Rajendran (Eds.), Memristive Devices for Brain-Inspired Computing (pp. 63–96). Woodhead Publishing. https://doi.org/10.1016/B978-0-08-102782-0.00003-4

96. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 7553. https://doi.org/10.1038/nature14539

97. Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324. https://doi.org/10.1109/5.726791

98. Lee, B.-S., Shelby, R. M., Raoux, S., Retter, C. T., Burr, G. W., Bogle, S. N., Darmawikarta, K., Bishop, S. G., & Abelson, J. R. (2014). Nanoscale nuclei in phase change materials: Origin of

different crystallization mechanisms of Ge2Sb2Te5 and AgInSbTe. Journal of Applied Physics, 115(6). Scopus. https://doi.org/10.1063/1.4865295

99. Lee, D. K., In, J., & Lee, S. (2015). Standard deviation and standard error of the mean. Korean Journal of Anesthesiology, 68(3), 220–223. https://doi.org/10.4097/kjae.2015.68.3.220

100. Lee, H. D., Magyari-Köpe, B., & Nishi, Y. (2010). Model of metallic filament formation and rupture in NiO for unipolar switching. Physical Review B, 81(19), 193202. https://doi.org/10.1103/PhysRevB.81.193202

101. Lencer, D., Salinga, M., Grabowski, B., Hickel, T., Neugebauer, J., & Wuttig, M. (2008). A map for phase-change materials. Nature Materials, 7(12), 12. https://doi.org/10.1038/nmat2330

102. Li, T., Li, Y., Wang, Y., Liu, Y., Liu, Y., Wang, Z., Miao, R., Han, D., Hui, Z., & Li, W. (2023). Neuromorphic Photonics Based on Phase Change Materials. Nanomaterials, 13(11), 11. https://doi.org/10.3390/nano13111756

103. Lin, J.-W. (2017). Artificial neural network related to biological neuron network: A review. Advanced Studies in Medical Sciences, 5, 55–62. https://doi.org/10.12988/asms.2017.753

104. Liu, X., & Zeng, Z. (2022). Memristor crossbar architectures for implementing deep neural networks. Complex & Intelligent Systems, 8(2), 787–802. https://doi.org/10.1007/s40747-021-00282-4

105. Maass, W. (1997). Networks of spiking neurons: The third generation of neural network models. Neural Networks, 10(9), 1659–1671. https://doi.org/10.1016/S0893-6080(97)00011-7

106. Maass, W., & Markram, H. (2004). On the computational power of circuits of spiking neurons. JOURNAL OF COMPUTER AND SYSTEM SCIENCES, 69(4), 593–616. https://doi.org/10.1016/j.jcss.2004.04.001

107. Mackin, C., Rasch, M. J., Chen, A., Timcheck, J., Bruce, R. L., Li, N., Narayanan, P., Ambrogio, S., Le Gallo, M., Nandakumar, S. R., Fasoli, A., Luquin, J., Friz, A., Sebastian, A., Tsai, H., & Burr, G. W. (2022). Optimised weight programming for analogue memory-based deep neural networks. Nature Communications, 13(1), 1. https://doi.org/10.1038/s41467-022-31405-1

108. Maex, R., & Schutter, E. D. (2003). Resonant Synchronization in Heterogeneous Networks of Inhibitory Neurons. Journal of Neuroscience, 23(33), 10503–10514. https://doi.org/10.1523/JNEUROSCI.23-33-10503.2003

109. McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics, 5(4), 115–133. https://doi.org/10.1007/BF02478259

110. Mead, C. (1990). Neuromorphic electronic systems. Proceedings of the IEEE, 78(10), 1629–1636. https://doi.org/10.1109/5.58356

111. Meena, J. S., Sze, S. M., Chand, U., & Tseng, T.-Y. (2014). Overview of emerging nonvolatile memory technologies. Nanoscale Research Letters, 9(1), 526. https://doi.org/10.1186/1556-276X-9-526

112. Menzel, S., Böttger, U., Wimmer, M., & Salinga, M. (2015). Physics of the Switching Kinetics in Resistive Memories. Advanced Functional Materials, 25(40), 6306–6325. https://doi.org/10.1002/adfm.201500825

113. Merolla, P. A., Arthur, J. V., Alvarez-Icaza, R., Cassidy, A. S., Sawada, J., Akopyan, F., Jackson, B. L., Imam, N., Guo, C., Nakamura, Y., Brezzo, B., Vo, I., Esser, S. K., Appuswamy, R., Taba, B., Amir, A., Flickner, M. D., Risk, W. P., Manohar, R., & Modha, D. S. (2014). Artificial brains. A million spiking-neuron integrated circuit with a scalable communication network and interface. Science (New York, N.Y.), 345(6197), 668–673. https://doi.org/10.1126/science.1254642

114. Mikolajick, T., Müller, S., Schenk, T., Yurchuk, E., Slesazeck, S., Schroeder, U., Dünkel, S., Van Bentum, R., Kolodinski, S., Polakowski, P., & Müller, J. (2014). Doped Hafnium Oxide – An Enabler for Ferroelectric Field Effect Transistors. Advances in Science and Technology, Volume 95, 136–145. https://doi.org/10.4028/www.scientific.net/AST.95.136

115. Molas, G., & Nowak, E. (2021). Advances in Emerging Memory Technologies: From Data Storage to Artificial Intelligence. Applied Sciences, 11(23), 23. https://doi.org/10.3390/app112311254

116. Morris, R. G. (1999). D.O. Hebb: The Organization of Behavior, Wiley: New York; 1949. Brain Research Bulletin, 50(5–6), 437. https://doi.org/10.1016/s0361-9230(99)00182-3

117. Mott, N. F. (1971). Conduction in non-crystalline systems. The Philosophical Magazine: A Journal of Theoretical Experimental and Applied Physics, 24(190), 911–934. https://doi.org/10.1080/14786437108217058

118. Muhammad, U., Wang, W., Chattha, S. P., & Ali, S. (2018). Pre-trained VGGNet Architecture for Remote-Sensing Image Scene Classification. 2018 24th International Conference on Pattern Recognition (ICPR), 1622–1627. https://doi.org/10.1109/ICPR.2018.8545591

119. Mukti, I. Z., & Biswas, D. (2019). Transfer Learning Based Plant Diseases Detection Using ResNet50. 2019 4th International Conference on Electrical Information and Communication Technology (EICT), 1–6. https://doi.org/10.1109/EICT48899.2019.9068805

120. Nair, M. V., & Dudek, P. (2015). Gradient-descent-based learning in memristive crossbar arrays. 2015 International Joint Conference on Neural Networks (IJCNN), 1–7. https://doi.org/10.1109/IJCNN.2015.7280658

121. Nair, R. (2015). Evolution of Memory Architecture. Proceedings of the IEEE, 103(8), 1331–1345. https://doi.org/10.1109/JPROC.2015.2435018

122. Nandakumar, S. R., Le Gallo, M., Boybat, I., Rajendran, B., Sebastian, A., & Eleftheriou, E. (2018). A phase-change memory model for neuromorphic computing. Journal of Applied Physics, 124(15), 152135. https://doi.org/10.1063/1.5042408

123. Nandakumar, S. R., Le Gallo, M., Piveteau, C., Joshi, V., Mariani, G., Boybat, I., Karunaratne, G., Khaddam-Aljameh, R., Egger, U., Petropoulos, A., Antonakopoulos, T., Rajendran, B., Sebastian, A., & Eleftheriou, E. (2020). Mixed-Precision Deep Learning Based on Computational Memory. Frontiers in Neuroscience, 14. https://www.frontiersin.org/articles/10.3389/fnins.2020.00406

124. Nandakumar, S. R., & Rajendran, B. (2017). (Invited) Synaptic Plasticity in a Memristive Device below 500mV. ECS Transactions, 77(2), 31. https://doi.org/10.1149/07702.0031ecst

125. Navarro, G. (n.d.). Reliability analysis of embedded Phase-Change Memories based on innovative materials.

126. Navarro, G., Coue, M., Kiouseloglou, A., Noe, P., Fillot, F., Delaye, V., Persico, A., Roule, A., Bernard, M., Sabbione, C., Blachier, D., Sousa, V., Perniola, L., Maitrejean, S., Cabrini, A., Torelli, G., Zuliani, P., Annunziata, R., Palumbo, E., & Salvo, B. (2013). Trade-off between SET and data retention performance thanks to innovative materials for phase-change memory. 21.5.1-21.5.4. https://doi.org/10.1109/IEDM.2013.6724678

127. Neale, R. G., & Aseltine, J. A. (1973). The application of amorphous materials to computer memories. IEEE Transactions on Electron Devices, 20(2), 195–205. https://doi.org/10.1109/T-ED.1973.17628

128. Nguyen, C. (n.d.). Caractérisation électrique et modélisation de la dynamique de commutation résistive dans des mémoires OxRAM à base de HfO2.

129. Noé, P., Vallée, C., Hippert, F., Fillot, F., & Raty, J.-Y. (2017). Phase-change materials for non-volatile memory devices: From technological challenges to materials science issues. Semiconductor Science and Technology, 33(1), 013002. https://doi.org/10.1088/1361-6641/aa7c25

130. Oh, S., Huang, Z., Shi, Y., & Kuzum, D. (2019). The Impact of Resistance Drift of Phase Change Memory (PCM) Synaptic Devices on Artificial Neural Network Performance. IEEE Electron Device Letters, 40(8), 1325–1328. https://doi.org/10.1109/LED.2019.2925832

131. Orava, J., & Greer, A. L. (2017). Classical-nucleation-theory analysis of priming in chalcogenide phase-change memory. Acta Materialia, 139, 226–235. https://doi.org/10.1016/j.actamat.2017.08.013

132. Orava, J., Greer, A. L., Gholipour, B., Hewak, D. W., & Smith, C. E. (2012). Characterization of supercooled liquid Ge2Sb2Te5 and its crystallization by ultrafast-heating calorimetry. Nature Materials, 11(4), 279–283. https://doi.org/10.1038/nmat3275

133. Orava, J., Hewak, D. W., & Greer, A. L. (2015). Fragile-to-Strong Crossover in Supercooled Liquid Ag-In-Sb-Te Studied by Ultrafast Calorimetry. Advanced Functional Materials, 25(30), 4851–4858. https://doi.org/10.1002/adfm.201501607

134. Ovhsinsky, S. R. (2006). Analog neurons and neurosynaptic networks (United States US6999953B2). https://patents.google.com/patent/US6999953B2/en

135. Ovshinsky, S. R. (1968). Reversible Electrical Switching Phenomena in Disordered Structures. Physical Review Letters, 21(20), 1450–1453. https://doi.org/10.1103/PhysRevLett.21.1450

136. Ovshinsky, S. R. (1970). An introduction to ovonic research. Journal of Non-Crystalline Solids, 2, 99–106. https://doi.org/10.1016/0022-3093(70)90125-0

137. Papandreou, N., Sebastian, A., Pantazi, A., Breitwisch, M., Lam, C., Pozidis, H., & Eleftheriou, E. (2011). Drift-resilient cell-state metric for multilevel phase-change memory. 2011 International Electron Devices Meeting, 3.5.1-3.5.4. https://doi.org/10.1109/IEDM.2011.6131482

138. Parkin, S. S. P. (2004). Spintronic materials and devices: Past, present and future! IEDM Technical Digest. IEEE International Electron Devices Meeting, 2004., 903–906. https://doi.org/10.1109/IEDM.2004.1419328

139. Pershin, Y. V., & Di Ventra, M. (2011). Memory effects in complex materials and nanoscale systems. Advances in Physics, 60(2), 145–227. https://doi.org/10.1080/00018732.2010.544961

140. Pfeil, T., Potjans, T., Schrader, S., Potjans, W., Schemmel, J., Diesmann, M., & Meier, K. (2012). Is a 4-Bit Synaptic Weight Resolution Enough? – Constraints on Enabling Spike-Timing Dependent Plasticity in Neuromorphic Hardware. Frontiers in Neuroscience, 6, 90. https://doi.org/10.3389/fnins.2012.00090

141. Pirovano, A., Lacaita, A. L., Benvenuti, A., Pellizzer, F., & Bez, R. (2004). Electronic Switching in Phase-Change Memories. Electron Devices, IEEE Transactions On, 51, 452–459. https://doi.org/10.1109/TED.2003.823243

142. Pogodin, R., Cornford, J., Ghosh, A., Gidel, G., Lajoie, G., & Richards, B. (2023). Synaptic Weight Distributions Depend on the Geometry of Plasticity (arXiv:2305.19394). arXiv. http://arxiv.org/abs/2305.19394

143. Prezioso, M., Merrikh-Bayat, F., Hoskins, B. D., Adam, G. C., Likharev, K. K., & Strukov, D. B. (2015). Training and operation of an integrated neuromorphic network based on metal-oxide memristors. Nature, 521(7550), 61–64. https://doi.org/10.1038/nature14441

144. Pries, J., Sehringer, J. C., Wei, S., Lucas, P., & Wuttig, M. (2021). Glass transition of the phase change material AIST and its impact on crystallization. Materials Science in Semiconductor Processing, 134, 105990. https://doi.org/10.1016/j.mssp.2021.105990

145. Purves, D., Augustine, G. J., Fitzpatrick, D., Katz, L. C., LaMantia, A.-S., McNamara, J. O., & Williams, S. M. (2001). Excitatory and Inhibitory Postsynaptic Potentials. In Neuroscience. 2nd edition. Sinauer Associates. https://www.ncbi.nlm.nih.gov/books/NBK11117/

146. Querlioz, D., Bichler, O., Vincent, A., & Gamrat, C. (2015). Bioinspired Programming of Memory Devices for Implementing an Inference Engine. Proceedings of the IEEE, 103, 1398–1416. https://doi.org/10.1109/JPROC.2015.2437616

147. Qureshi, M., Srinivasan, V., & Rivers, J. (2009). Scalable high performance main memory system using phase-change memory technology. 37, 24–33. https://doi.org/10.1145/1555754.1555760

148. Rajendran, B., & Alibart, F. (2016). Neuromorphic Computing Based on Emerging Memory Technologies. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 6(2), 198–211. https://doi.org/10.1109/JETCAS.2016.2533298

149. Raoux, S., Burr, G. W., Breitwisch, M. J., Rettner, C. T., Chen, Y.-C., Shelby, R. M., Salinga, M., Krebs, D., Chen, S.-H., Lung, H.-L., & Lam, C. H. (2008). Phase-change random access memory: A scalable technology. IBM Journal of Research and Development, 52(4.5), 465–479. https://doi.org/10.1147/rd.524.0465

150. Romanuke, V. (2017). Appropriate Number and Allocation of ReLUs in Convolutional Neural Networks. Research Bulletin of the National Technical University of Ukraine “Kyiv Politechnic Institute,” 1, 1. https://doi.org/10.20535/1810-0546.2017.1.88156

151. Sahu, D. P., Park, K., Chung, P. H., Han, J., & Yoon, T.-S. (2023). Linear and symmetric synaptic weight update characteristics by controlling filament geometry in oxide/suboxide HfOx bilayer memristive device for neuromorphic computing. Scientific Reports, 13(1), 1. https://doi.org/10.1038/s41598-023-36784-z

152. Saïghi, S., Mayr, C. G., Serrano-Gotarredona, T., Schmidt, H., Lecerf, G., Tomas, J., Grollier, J., Boyn, S., Vincent, A. F., Querlioz, D., La Barbera, S., Alibart, F., Vuillaume, D., Bichler, O., Gamrat,

C., & Linares-Barranco, B. (2015). Plasticity in memristive devices for spiking neural networks. Frontiers in Neuroscience, 9. https://www.frontiersin.org/articles/10.3389/fnins.2015.00051

153. Schemmel, J., Briiderle, D., Griibl, A., Hock, M., Meier, K., & Millner, S. (2010). A wafer-scale neuromorphic hardware system for large-scale neural modeling. Proceedings of 2010 IEEE International Symposium on Circuits and Systems, 1947–1950. https://doi.org/10.1109/ISCAS.2010.5536970

154. Schuman, C. D., Kulkarni, S. R., Parsa, M., Mitchell, J. P., Date, P., & Kay, B. (2022). Opportunities for neuromorphic computing algorithms and applications. Nature Computational Science, 2(1), 1. https://doi.org/10.1038/s43588-021-00184-y

155. Schuman, C. D., Potok, T. E., Patton, R. M., Birdwell, J. D., Dean, M. E., Rose, G. S., & Plank, J. S. (2017). A Survey of Neuromorphic Computing and Neural Networks in Hardware. ArXiv:1705.06963 [Cs]. http://arxiv.org/abs/1705.06963

156. Sebastian, A., Le Gallo, M., Khaddam-Aljameh, R., & Eleftheriou, E. (2020). Memory devices and applications for in-memory computing. Nature Nanotechnology, 15(7), 7. https://doi.org/10.1038/s41565-020-0655-z

157. Sebastian, A., Le Gallo, M., & Krebs, D. (2014). Crystal growth within a phase change memory cell. Nature Communications, 5(1), 1. https://doi.org/10.1038/ncomms5314

158. Sebastian, A., Papandreou, N., Pantazi, A., Pozidis, H., & Eleftheriou, E. (2011). Non-resistance-based cell-state metric for phase-change memory. Journal of Applied Physics, 110(8). Scopus. https://doi.org/10.1063/1.3653279

159. Senn, W., & Fusi, S. (2005). Convergence of stochastic learning in perceptrons with binary synapses. Physical Review E, 71(6), 061907. https://doi.org/10.1103/PhysRevE.71.061907

160. Serrano-Gotarredona, T., Masquelier, T., Prodromakis, T., Indiveri, G., & Linares-Barranco, B. (2013). STDP and STDP variations with memristors for spiking neuromorphic learning systems. Frontiers in Neuroscience, 7. https://www.frontiersin.org/articles/10.3389/fnins.2013.00002

161. Servalli, G. (2009). A 45nm generation Phase Change Memory technology. 2009 IEEE International Electron Devices Meeting (IEDM), 1–4. https://doi.org/10.1109/IEDM.2009.5424409

162. Sharifshazileh, M., Burelo, K., Sarnthein, J., & Indiveri, G. (2021). An electronic neuromorphic system for real-time detection of high frequency oscillations (HFO) in intracranial EEG. Nature Communications, 12(1), 1. https://doi.org/10.1038/s41467-021-23342-2

163. Sherif, F. F., & Ahmed, K. S. (2022). Unsupervised clustering of SARS-CoV-2 using deep convolutional autoencoder. Journal of Engineering and Applied Science, 69(1), 72. https://doi.org/10.1186/s44147-022-00125-0

164. Shibata, T., & Ohmi, T. (1997). Neural microelectronics. International Electron Devices Meeting. IEDM Technical Digest, 337–342. https://doi.org/10.1109/IEDM.1997.650395

165. Sidler, S., Boybat, I., Shelby, R. M., Narayanan, P., Jang, J., Fumarola, A., Moon, K., Leblebici, Y., Hwang, H., & Burr, G. W. (2016). Large-scale neural networks implemented with Non-Volatile Memory as the synaptic weight element: Impact of conductance response. 2016 46th European Solid-

State Device Research Conference (ESSDERC), 440–443. https://doi.org/10.1109/ESSDERC.2016.7599680

166. Siegelmann, H., & Sontag, E. (1995). On the Computational Power of Neural Nets. JOURNAL OF COMPUTER AND SYSTEM SCIENCES, 50(1), 132–150. https://doi.org/10.1006/jcss.1995.1013

167. Singh, G., Chelini, L., Corda, S., Javed Awan, A., Stuijk, S., Jordans, R., Corporaal, H., & Boonstra, A.-J. (2018). A Review of Near-Memory Computing Architectures: Opportunities and Challenges. 2018 21st Euromicro Conference on Digital System Design (DSD), 608–617. https://doi.org/10.1109/DSD.2018.00106

168. Skelton, J. M., Loke, D., Lee, T., & Elliott, S. R. (2015). Ab Initio Molecular-Dynamics Simulation of Neuromorphic Computing in Phase-Change Memory Materials. ACS Applied Materials & Interfaces, 7(26), 14223–14230. https://doi.org/10.1021/acsami.5b01825

169. Skotnicki, T. (2007). Materials and device structures for sub-32 nm CMOS nodes. Microelectronic Engineering, 84(9), 1845–1852. https://doi.org/10.1016/j.mee.2007.04.091

170. Slonczewski, J. C. (1989). Conductance and exchange coupling of two ferromagnets separated by a tunneling barrier. Physical Review. B, Condensed Matter, 39(10), 6995–7002. https://doi.org/10.1103/physrevb.39.6995

171. Solomon, P. M. (2019). Analog neuromorphic computing using programmable resistor arrays. Solid-State Electronics, 155, 82–92. https://doi.org/10.1016/j.sse.2019.03.023

172. Song, S., & Das, A. (2020). Design Methodologies for Reliable and Energy-efficient PCM Systems. 2020 11th International Green and Sustainable Computing Workshops (IGSC), 1–3. https://doi.org/10.1109/IGSC51522.2020.9291024

173. Sonoda, K., Sakai, A., Moniwa, M., Ishikawa, K., Tsuchiya, O., & Inoue, Y. (2008). A compact model of phase-change memory based on rate equations of crystallization and amorphization. IEEE Transactions on Electron Devices, 55(7), 1672–1681. Scopus. https://doi.org/10.1109/TED.2008.923740

174. Sosso, G. C., Behler, J., & Bernasconi, M. (2012). Breakdown of Stokes-Einstein relation in the supercooled liquid state of phase change materials. Physica Status Solidi (b), 249(10), 1880–1885. https://doi.org/10.1002/pssb.201200355

175. Sousa, V., Navarro, G., Castellani, N., Coue, M., Cueto, O., Sabbione, C., Noe, P., Perniola, L., Blonkowski, S., Zuliani, P., & Annunziata, R. (2015). Operation fundamentals in 12Mb Phase Change Memory based on innovative Ge-rich GST materials featuring high reliability performance. 2015 Symposium on VLSI Technology (VLSI Technology), T98–T99. https://doi.org/10.1109/VLSIT.2015.7223708

176. Stanisavljevic, M., Athmanathan, A., Papandreou, N., Pozidis, H., & Eleftheriou, E. (2015). Phase-change memory: Feasibility of reliable multilevel-cell storage and retention at elevated temperatures. 2015 IEEE International Reliability Physics Symposium, 5B.6.1-5B.6.6. https://doi.org/10.1109/IRPS.2015.7112747

177. Sung, S. H., Kim, T. J., Shin, H., Im, T. H., & Lee, K. J. (2022). Simultaneous emulation of synaptic and intrinsic plasticity using a memristive synapse. Nature Communications, 13(1), 1. https://doi.org/10.1038/s41467-022-30432-2

178. Suri, Dr. M., Querlioz, D., Bichler, O., Palma, G., Vianello, E., Vuillaume, D., Gamrat, C., & DeSalvo, B. (2013). Bio-Inspired Stochastic Computing Using Binary CBRAM Synapses. IEEE Transactions on Electron Devices, 60, 2402–2409. https://doi.org/10.1109/TED.2013.2263000

179. Suri, M., Bichler, O., Hubert, Q., Perniola, L., Sousa, V., Jahan, C., Vuillaume, D., Gamrat, C., & DeSalvo, B. (2012). Interface Engineering of PCM for Improved Synaptic Performance in Neuromorphic Systems. 2012 4th IEEE International Memory Workshop, 1–4. https://doi.org/10.1109/IMW.2012.6213674

180. Suri, M., Bichler, O., Hubert, Q., Perniola, L., Sousa, V., Jahan, C., Vuillaume, D., Gamrat, C., & DeSalvo, B. (2013). Addition of HfO2 interface layer for improved synaptic performance of phase change memory (PCM) devices. SOLID-STATE ELECTRONICS, 79, 227–232. https://doi.org/10.1016/j.sse.2012.09.006

181. Suri, M., Bichler, O., Querlioz, D., Cueto, O., Perniola, L., Sousa, V., Vuillaume, D., Gamrat, C., & DeSalvo, B. (2011). Phase change memory as synapse for ultra-dense neuromorphic systems: Application to complex visual pattern extraction. 2011 International Electron Devices Meeting, 4.4.1-4.4.4. https://doi.org/10.1109/IEDM.2011.6131488

182. Suri, M., Bichler, O., Querlioz, D., Traore, B., Cueto, O., Perniola, L., Sousa, V., Vuillaume, D., Gamrat, C., & de Salvo, B. (2012). Physical aspects of low power synapses based on phase change memory devices. Journal of Applied Physics, 112(5), 054904. https://doi.org/10.1063/1.4749411

183. Suri, M., Garbin, D., Bichler, O., Querlioz, D., Vuillaume, D., Gamrat, C., & DeSalvo, B. (2013). Impact of PCM resistance-drift in neuromorphic systems and drift-mitigation strategy. 2013 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH), 140–145. https://doi.org/10.1109/NanoArch.2013.6623059

184. Terao, M., Morikawa, T., & Ohta, T. (2009). Electrical Phase-Change Memory: Fundamentals and State of the Art. Japanese Journal of Applied Physics, 48(8), 080001. https://doi.org/10.1143/JJAP.48.080001

185. Trabelsi, A., Cagli, C., Hirtzlin, T., Cueto, O., Cyrille, M. C., Vianello, E., Meli, V., Sousa, V., Bourgeois, G., & Andrieu, F. (2022). Frequency modulation of conductance level in PCM device for neuromorphic applications. ESSCIRC 2022- IEEE 48th European Solid State Circuits Conference (ESSCIRC), 129–132. https://doi.org/10.1109/ESSCIRC55480.2022.9911461

186. Tsai, H., Ambrogio, S., Narayanan, P., Shelby, R. M., & Burr, G. W. (2018). Recent progress in analog memory-based accelerators for deep learning. Journal of Physics D: Applied Physics, 51(28), 283001. https://doi.org/10.1088/1361-6463/aac8a5

187. Tsai, H., Ambrogio, S., Narayanan, P., Shelby, R. M., Mackin, C., & Burr, G. W. (2019). Analog memory-based techniques for accelerating the training of fully-connected deep neural networks (Conference Presentation). Novel Patterning Technologies for Semiconductors, MEMS/NEMS, and MOEMS 2019, 10958, 109580U. https://doi.org/10.1117/12.2515630

188. Tuchman, Y., Mangoma, T. N., Gkoupidenis, P., van de Burgt, Y., John, R. A., Mathews, N., Shaheen, S. E., Daly, R., Malliaras, G. G., & Salleo, A. (2020). Organic neuromorphic devices: Past, present, and future challenges. MRS Bulletin, 45(8), 619–630. https://doi.org/10.1557/mrs.2020.196

189. Tuma, T., Pantazi, A., Le Gallo, M., Sebastian, A., & Eleftheriou, E. (2016). Stochastic phase-change neurons. Nature Nanotechnology, 11(8), 8. https://doi.org/10.1038/nnano.2016.70

190. Tymoshchuk, P. V., & Wunsch, D. C. (2019). Design of a K-Winners-Take-All Model With a Binary Spike Train. IEEE Transactions on Cybernetics, 49(8), 3131–3140. https://doi.org/10.1109/TCYB.2018.2839691

191. Vianello, E., Molas, G., Longnos, F., Blaise, P., Souchier, E., Cagli, C., Palma, G., Guy, J., Bernard, M., Reyboz, M., Rodriguez, G., Roule, A., Carabasse, C., Delaye, V., Jousseaume, V., Maitrejean, S., Reimbold, G., De Salvo, B., Dahmani, F., … Liebault, J. (2012). Sb-doped GeS2 as performance and reliability booster in Conductive Bridge RAM. 2012 International Electron Devices Meeting, 31.5.1-31.5.4. https://doi.org/10.1109/IEDM.2012.6479145

192. Walsh, I., Pollastri, G., & Tosatto, S. C. E. (2016). Correct machine learning on protein sequences: A peer-reviewing perspective. Briefings in Bioinformatics, 17(5), 831–840. https://doi.org/10.1093/bib/bbv082

193. Wan, X., He, N., Liang, D., Xu, W., Wang, L., Lian, X., Liu, X., Xu, F., & Tong, Y. (2022). Memristive crossbar circuit for neural network and its application in digit recognition. Japanese Journal of Applied Physics, 61(6), 060905. https://doi.org/10.35848/1347-4065/ac6b01

194. Wang, J., Dong, X., Sun, G., Niu, D., & Xie, Y. (2011). Energy-efficient multi-level cell phase-change memory system with data encoding. 2011 IEEE 29th International Conference on Computer Design (ICCD), 175–182. https://doi.org/10.1109/ICCD.2011.6081394

195. Wang, L., Lu, S.-R., & Wen, J. (2017). Recent Advances on Neuromorphic Systems Using Phase-Change Materials. Nanoscale Research Letters, 12(1), 347. https://doi.org/10.1186/s11671-017-2114-9

196. Welcome to Fritzing. (n.d.). Retrieved August 24, 2023, from https://fritzing.org/

197. Wong, H.-S. P., Lee, H.-Y., Yu, S., Chen, Y.-S., Wu, Y., Chen, P.-S., Lee, B., Chen, F. T., & Tsai, M.-J. (2012). Metal–Oxide RRAM. Proceedings of the IEEE, 100(6), 1951–1970. https://doi.org/10.1109/JPROC.2012.2190369

198. Wong, H.-S. P., & Salahuddin, S. (2015). Memory leads the way to better computing. Nature Nanotechnology, 10(3), 191–194. https://doi.org/10.1038/nnano.2015.29

199. Wuttig, M., & Yamada, N. (2007). Wuttig, M. & Yamada, N. Phase-change materials for rewriteable data storage. Nat. Mater. 6, 824-832. Nature Materials, 6, 824–832. https://doi.org/10.1038/nmat2009

200. Xuan, X. (2008). Joule heating in electrokinetic flow. ELECTROPHORESIS, 29(1), 33–43. https://doi.org/10.1002/elps.200700302

201. Yamada, N., Ohno, E., Akahira, N., Nishiuchi, K., Nagata, K., & Takao, M. (1987). High Speed Overwritable Phase Change Optical Disk Material. Japanese Journal of Applied Physics, 26(S4), 61. https://doi.org/10.7567/JJAPS.26S4.61

202. Yamada, N., Ohno, E., Nishiuchi, K., Akahira, N., & Takao, M. (1991). Rapid-phase transitions of GeTe-Sb2Te3 pseudobinary amorphous thin films for an optical disk memory. Journal of Applied Physics, 69, 2849–2856. https://doi.org/10.1063/1.348620

203. Yang, X., Hou, Y., & He, H. (2019). A Processing-in-Memory Architecture Programming Paradigm for Wireless Internet-of-Things Applications. Sensors, 19, 140. https://doi.org/10.3390/s19010140

204. Yu, S., Chen, P.-Y., Cao, Y., Xia, L., Wang, Y., & Wu, H. (2015). Scaling-up resistive synaptic arrays for neuro-inspired architecture: Challenges and prospect. 2015 IEEE International Electron Devices Meeting (IEDM), 17.3.1-17.3.4. https://doi.org/10.1109/IEDM.2015.7409718

205. Yu, S., Kuzum, D., & Wong, H.-P. (2014). Design considerations of synaptic device for neuromorphic computing. 2014 IEEE International Symposium on Circuits and Systems (ISCAS), 1062–1065. https://doi.org/10.1109/ISCAS.2014.6865322

206. Yuasa, S., Hono, K., Hu, G., & Worledge, D. C. (2018). Materials for spin-transfer-torque magnetoresistive random-access memory. MRS Bulletin, 43(5), 352–357. https://doi.org/10.1557/mrs.2018.93

207. Yusiong, J. P. (2012). Optimizing Artificial Neural Networks using Cat Swarm Optimization Algorithm. International Journal of Intelligent Systems and Applications, 5, 69–80. https://doi.org/10.5815/ijisa.2013.01.07

208. Zhang, W., & Ma, E. (2020). Unveiling the structural origin to control resistance drift in phase-change memory materials. Materials Today, 41, 156–176. https://doi.org/10.1016/j.mattod.2020.07.016

209. Zhang, W., Mazzarello, R., Wuttig, M., & Ma, E. (2019). Designing crystallization in phase-change materials for universal memory and neuro-inspired computing. Nature Reviews Materials, 4(3), 3. https://doi.org/10.1038/s41578-018-0076-x

210. Zhang, X., Huang, A., Hu, Q., Xiao, Z., & Chu, P. K. (2018). Neuromorphic Computing with Memristor Crossbar. Physica Status Solidi (a), 215(13), 1700875. https://doi.org/10.1002/pssa.201700875

211. Zhong, Y., Li, Y., Xu, L., & Miao, X. (2015). Simple square pulses for implementing spike-timing-dependent plasticity in phase-change memory. PHYSICA STATUS SOLIDI-RAPID RESEARCH LETTERS, 9(7), 414–419. https://doi.org/10.1002/pssr.201510150

212. Zhou, W., Feng, D., Hua, Y., Liu, J., Huang, F., & Chen, Y. (2016). An Efficient Parallel Scheduling Scheme on Multi-partition PCM Architecture. Proceedings of the 2016 International Symposium on Low Power Electronics and Design, 344–349. https://doi.org/10.1145/2934583.2934610

213. Zuliani, P., Varesi, E., Palumbo, E., Borghi, M., Tortorelli, I., Erbetta, D., Dalla Libera, G., Pessina, N., Gandolfo, A., Prelini, C., Ravazzi, L., & Annunziata, R. (2013). Overcoming Temperature Limitations in Phase Change Memories With Optimized GexSbyTez. IEEE Transactions on Electron Devices, 60, 4020–4026. https://doi.org/10.1109/TED.2013.2285403

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 800945 — NUMERICS — H2020-MSCA-COFUND-2017