



HAL
open science

Mémoires résistives et technologies 3D monolithiques pour processeurs neuromorphiques impulsionnels et reconfigurables

Denys Ly

► **To cite this version:**

Denys Ly. Mémoires résistives et technologies 3D monolithiques pour processeurs neuromorphiques impulsionnels et reconfigurables. Micro et nanotechnologies/Microélectronique. Université Grenoble Alpes [2020-..], 2020. Français. NNT : 2020GRALT016 . tel-03027430

HAL Id: tel-03027430

<https://theses.hal.science/tel-03027430v1>

Submitted on 27 Nov 2020

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

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

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : NANO ELECTRONIQUE ET NANO TECHNOLOGIES

Arrêté ministériel : 25 mai 2016

Présentée par

Denys LY

Thèse dirigée par **Claire FENOUILLET-BERANGER**
et codirigée par **Elisa VIANELLO**

préparée au sein du **Laboratoire CEA/LETI**
dans l'**École Doctorale Electronique, Electrotechnique,**
Automatique, Traitement du Signal (EEATS)

Mémoires résistives et technologies 3D monolithiques pour processeurs neuromorphiques impulsionnels et reconfigurables

Resistive memories and 3D monolithic technologies for reconfigurable spiking neuromorphic processors

Thèse soutenue publiquement le **19 juin 2020**,
devant le jury composé de :

Madame CLAIRE FENOUILLET-BERANGER

INGENIEUR CHERCHEUR HDR, CEA GRENOBLE, Directeur de thèse

Madame JULIE GROLLIER

DIRECTRICE DE RECHERCHE, CNRS DELEGATION ILE-DE-FRANCE
SUD, Rapporteur

Monsieur JEAN-MICHEL PORTAL

PROFESSEUR DES UNIVERSITES, UNIVERSITE AIX-MARSEILLE,
Rapporteur

Monsieur DAMIEN QUERLIOZ

CHARGE DE RECHERCHE HDR, CNRS DELEGATION ILE-DE-
FRANCE SUD, Examineur

Monsieur GERARD GHIBAUDO

DIRECTEUR DE RECHERCHE, CNRS DELEGATION ALPES, Président

Madame ELISA VIANELLO

DOCTEUR-INGENIEUR, CEA GRENOBLE, Examineur





Abstract

Title:

Resistive memories and three-dimensional monolithic technologies for reconfigurable spiking neuromorphic processors

THE human brain is a complex, energy-efficient computational system that excels at cognitive tasks thanks to its natural capability to perform inference. By contrast, conventional computing systems based on the classic Von Neumann architecture require large power budget to execute such assignments. Herein comes the idea to build brain-inspired electronic computing systems, the so-called neuromorphic approach. In this thesis, we explore the use of novel technologies, namely Resistive Memories (RRAMs) and three-dimensional (3D) monolithic technologies, to enable the hardware implementation of compact, low-power reconfigurable Spiking Neural Network (SNN) processors. We first provide a comprehensive study of the impact of RRAM electrical properties on SNNs with RRAM synapses and trained with unsupervised learning (Spike-Timing-Dependent Plasticity (STDP)). In particular, we clarify the role of synaptic variability originating from RRAM resistance variability. Second, we investigate the use of RRAM-based Ternary Content-Addressable Memory (TCAM) arrays as synaptic routing tables in SNN processors to enable on-the-fly reconfigurability of network topology. For this purpose, we present in-depth electrical characterisations of two RRAM-based TCAM circuits: (i) the most common two-transistors/two-RRAMs (2T2R) RRAM-based TCAM, and (ii) a novel one-transistor/two-RRAMs/one-transistor (1T2R1T) RRAM-based TCAM, both featuring the smallest silicon area up-to-date. We compare both structures in terms of performance, reliability, and endurance. Finally, we explore the potential of 3D monolithic technologies to improve area-efficiency. In addition to the conventional monolithic integration of RRAMs in the back-end-of-line of CMOS technology, we examine the vertical stacking of CMOS over CMOS transistors. To this end, we demonstrate the full 3D monolithic integration of two tiers of CMOS transistors with one tier of RRAM devices and present electrical characterisations performed on the fabricated devices.

Keywords: Spiking neuromorphic processor, Resistive memory, 3D monolithic technology, Artificial synapse, Content-addressable memory, Synaptic routing table.

Résumé en français

Titre:

Mémoires résistives et technologies 3D monolithiques pour processeurs neuromorphiques impulsionsnels et reconfigurables

LE cerveau humain est un système computationnel complexe mais énergétiquement efficace qui excelle aux applications cognitives grâce à sa capacité naturelle à faire de l'inférence. À l'inverse, les systèmes de calculs traditionnels reposant sur la classique architecture de Von Neumann exigent des consommations de puissance importantes pour exécuter de telles tâches. Ces considérations ont donné naissance à la fameuse approche neuromorphique, qui consiste à construire des systèmes de calculs inspirés du cerveau. Dans cette thèse, nous examinons l'utilisation de technologies novatrices, à savoir les mémoires résistives (RRAMs) et les technologies tridimensionnelles (3D) monolithiques, pour permettre l'implémentation matérielle compacte de processeurs neuromorphiques impulsionsnels (SNNs) et reconfigurables à faible puissance. Dans un premier temps, nous fournirons une étude détaillée sur l'impact des propriétés électriques des RRAMs dans les SNNs utilisant des synapses à base de RRAMs, et entraînés avec des méthodes d'apprentissage non-supervisées (plasticité fonction du temps d'occurrence des impulsions, STDP). Notamment, nous clarifierons le rôle de la variabilité synaptique provenant de la variabilité résistive des RRAMs. Dans un second temps, nous étudierons l'utilisation de matrices de mémoires ternaires adressables par contenu (TCAMs) à base de RRAMs en tant que tables de routage synaptique dans les processeurs SNNs, afin de permettre la reconfigurabilité de la topologie du réseau. Pour ce faire, nous présenterons des caractérisations électriques approfondies de deux circuits TCAMs à base de RRAMs: (i) la structure TCAM la plus courante avec deux-transistors/deux-RRAMs (2T2R), et (ii) une nouvelle structure TCAM avec un-transistor/deux-RRAMs/un-transistor (1T2R1T), toutes deux dotées de la plus petite surface silicium à l'heure actuelle. Nous comparerons les deux structures en termes de performances, fiabilité et endurance. Pour finir, nous explorerons le potentiel des technologies 3D monolithiques en vue d'améliorer l'efficacité en surface. En plus de la classique intégration monolithique des RRAMs dans le retour en fin de ligne (back-end-of-line) des technologies CMOS, nous analyserons l'empilement vertical de transistors CMOS les uns au-dessus des autres. Pour cela, nous démontrerons la possibilité d'intégrer monolithiquement deux niveaux de transistors CMOS avec un niveau de dispositifs RRAMs. Cette preuve de concept sera appuyée par des caractérisations électriques effectuées sur les dispositifs fabriqués.

Mots-clés: Processeur neuromorphique impulsionsnel, Mémoire résistive, Technologie 3D monolithique, Synapse artificielle, Mémoire adressable par contenu, Table de routage synaptique.

Acknowledgements

"Everything has one end, only the sausage has two."

I have many people that I would like to thank for helping me get through this PhD. First of all, I must express my gratitude to my supervisors Claire Fenouillet-Béranger and Elisa Vianello without whom nothing would have even started. Their patience, support, and expertise over the past four years allowed us to come up with this exotic and multidisciplinary project and obtained interesting results. It has been a real pleasure to team up with them. Then, I would like to thank the rapporteurs and juries for reading this dissertation and evaluating my PhD work. Finally, this work would not have obviously been possible without the knowledge and advices of other people. Thus, I am deeply thankful to Bastien Giraud and Jean-Philippe Noël who adopted me in the middle of my PhD, Niccolo Castellani for his precious help, Laurent Brunet for training me in CoolCube™, and Damien Querlioz for providing me sound advices whenever I needed during my PhD.

Certainly, this PhD journey has been made possible thanks to numerous other people that gave me moral and/or professional support. I would like to start with the people from the LCM lab with whom I had the opportunity to share break times and talk. In particular, I am glad I have been able to meet many interns, PhD students, and post-docs with whom I could spend time at and outside of work: Alessandro B., Alessandro G., Anna-Lisa, Anthonin, Camille, Diego, Eduardo, Filippo, Joel, Juliana, Gilbert, Giuseppe, Giusy, Léo, Marios, Nicolas G., Paola, Thilo, and my office mate Thomas D. Of course, I am truly thankful to my friends I probably spent most of my time with. I was lucky I had the opportunity to meet all these people here and there, whether it be at work, school, climbing gym, or randomly. I tried not to forget anyone, and I apologise in advance to whomever I may have forgotten in the following list: Alexandra, Alexander, Alexandre, Alexandre M., Angelo R., Arnaud, Bartosz, Boris, Brune, Charles, Chhayarith, Chloé A., Claire, Clément C., Clément P., Daniele, Djinthana, Daphné, David, Dominique, Elie, Eliot, Erika V., Erwan D., Erwan L., Eva, Eve, Fanny T., Félix (le chômeur), Fernando, Florence, François (big thank !), Grégoire, Guillaume D., Guillaume, Hoël, Jean R., Jean-François P., Jean-Fred, Jessica, Julie N., Julie R., Léa S., Léo, Lisa, Lucas F., Lucien, Marc, Maryam, Maxime M., Mathieu L., Matthieu P., Max, Mickaël, Nicolas P., Nitish, Pauline, Patrice, Pierre R., Qiwei, Raphaël, Rayane, Saad, Simon De., Simon Du., Sindou, Sota, Sylvain, Sylvia, Théo, Thomas B., Thomas G., Thomas J., Tsy-Yeung, Yihong, and Youna.

There are still people I did not acknowledge yet because I wanted to give them a special thank. First, I want to say thank you to Laurent Rastello for his support

since I arrived in Grenoble. Next, I need to show my gratitude to my friends from my Taekwondo club. I have to start with Master Richard Passalacqua whose wisdom and knowledge allowed me to develop myself physically, mentally, and spiritually, his assistant Antoine Inchaurtieta whose energy and vitality have never failed to motivate me, Armel Cadiou without whom the club would not be doing as well, and Antoinette Passalacqua who is always taking good care of all of us. I express my deepest appreciation to every member of the club, and in particular I want to thank the following people: Amandine, Amédée, Chloé L., Coline, Cyrilline, Fanny I., Heidi, Jean M., Laurent F., Laureline, Léa C., Lena P., Louis-Marie, Lucas C., Mélanie, Mohamed Z., Nathan, Quentin, Sarah R., Sonia, Stéphanie, and Zahra. Last but definitely not the least, I will forever be in debt to my three bros Matthieu M., Elliot N-M., and Julien D. with whom I spent many days and nights watching Kung-Fu movies, getting salty at Smash Bros, Rokli-ing, sharing (Asian) food time, and reading amazing mangas on Instagram.

I dedicate this dissertation to every people I just acknowledged (as well as people I may have forgotten by mistake and to whom I apologise once more) and thank them tremendously again.

Table of contents

	Page
1 Introduction	1
1.1 From Von Neumann to neuromorphic computing	2
1.1.1 The Von Neumann bottleneck	2
1.1.2 The end of Moore’s law	3
1.2 New technology enablers	4
1.2.1 Resistive memory technology	4
1.2.2 Three-dimensional technology	13
1.3 The third generation of neural networks: Spiking neural network	17
1.3.1 Overview of spiking neural networks	18
1.3.2 Information coding and network routing	20
1.3.3 Hardware spiking neuron: the leaky integrate-and-fire neuron model	21
1.3.4 Hardware synapse implementation	23
1.3.5 Overview of fabricated neuromorphic processors	30
1.4 Goal of this PhD thesis	31
References: Introduction	35
2 Role of synaptic variability in resistive memory-based spiking neural networks with unsupervised learning	55
2.1 Introduction	56
2.1.1 Variability in biological brains	56
2.1.2 Synaptic variability in artificial spiking neural networks .	56
2.1.3 Goal of this chapter	57
2.2 Binary devices	58
2.2.1 Experimental characterisation	58

TABLE OF CONTENTS

2.2.2	Implications for a learning system: impact of binary RRAM-based synapse characteristics on the network performance	64
2.2.3	Conclusion	77
2.3	Analog devices	78
2.3.1	Goal of the section	78
2.3.2	Analog conductance modulation with non-volatile resistance-based memories	79
2.3.3	Learning rule and synapse behavioural model	81
2.3.4	Impact of the conductance response on spiking neural network learning performance	82
2.3.5	Discussion	86
References: Chapter 2		89
3	Synaptic routing reconfigurability of spiking neural networks with resistive memory-based ternary content-addressable memory systems	99
3.1	Content-addressable memory systems	100
3.1.1	Basics on content-addressable memories	100
3.1.2	Motivations for the implementation of resistive memory-based ternary content-addressable memories	101
3.1.3	Examples of ternary content-addressable memory applications	103
3.1.4	Goal of this chapter	107
3.2	Characterisation of resistive memory-based ternary content-addressable memories	108
3.2.1	Fabricated resistive memory-based ternary content-addressable memory circuits	108
3.2.2	Search operation principle	109
3.2.3	Common 2T2R TCAM circuit characterisation	110
3.2.4	Novel 1T2R1T TCAM circuit characterisation	122
References: Chapter 3		137
4	Three-dimensional monolithic integration of two layers of high-performance CMOS transistors with one layer of resistive memory devices	143
4.1	Goal of this chapter	144

4.2	Three-dimensional monolithic co-integration of resistive memories and CMOS transistors	144
4.2.1	CoolCube™ technology	144
4.2.2	Resistive memory integration	146
4.3	Electrical characterisation of the three-dimensional monolithic integration of two tiers of NMOS transistors with a tier of resistive memory devices	148
4.3.1	Basic functionality of bottom and top transistors	148
4.3.2	Characterisation of 1T1R structures	149
4.4	Discussion and conclusion	153
References: Chapter 4		157
5	Conclusion and perspectives	161
References: Conclusion		167
Appendices		171
A	Impact of resistive memory-based synapses on spiking neural network performance: Network topology	173
A.1	Network topology with binary devices	173
A.1.1	Car tracking	173
A.1.2	Digit classification	175
A.2	Network topology with analog devices	177
B	Impact of leaky integrate-and-fire neuron threshold value on spiking neural network performance	181
B.1	Car tracking	182
B.2	Digit classification	183
B.3	Impact of firing threshold variability	184
C	Robustness of spiking neural networks trained with unsupervised learning to input noise	189
References: Appendice		191
Résumé en français		195
1	Introduction	195

1.1	De Von Neumann au calcul neuromorphique	195
1.2	Les nouvelles solutions technologiques	196
1.2.1	Les mémoires résistives	196
1.2.2	Les technologies 3D monolithiques	199
1.3	Les réseaux de neurones impulsionnels	201
1.4	Objectif de ce travail de thèse de doctorat	202
2	Rôle de la variabilité synaptique dans les réseaux de neurones impulsionnels à base de mémoires résistives avec apprentissage non supervisé	207
2.1	Objectif de ce chapitre	207
2.2	Caractérisations électriques des RRAM	208
2.3	Implémentation des éléments synaptiques et règle d'apprentissage avec les mémoires résistives	210
2.4	Implications pour un système d'apprentissage: impact des caractéristiques des synapses à base de RRAM sur les performances d'un réseau	212
2.4.1	Topologie des réseaux de neurones impulsionnels	212
2.4.2	Impact de la fenêtre mémoire et variabilité conductive des RRAM	212
2.4.3	Impact du vieillissement des RRAM	215
2.5	Conclusion	216
3	Reconfigurabilité du routage synaptique des réseaux de neurones impulsionnels avec des mémoires ternaires adressables par contenu à base de mémoires résistives	219
3.1	Objectif de ce chapitre	219
3.2	Principes de base des mémoires adressables par contenu	220
3.3	Circuits de mémoires ternaires adressables par contenu à base de mémoires résistives	221
3.3.1	La cellule TCAM la plus commune deux-transistors/deux-RRAM (2T2R)	221
3.3.2	La nouvelle cellule TCAM un-transistor/deux-RRAM/un-transistor (1T2R1T)	222
3.3.3	Comparaison des deux structures TCAM	223
3.3.4	Intégration et fabrication des deux circuits TCAM à base de RRAM	224
3.4	Caractérisations électriques des circuits TCAM à base de RRAM	224

3.4.1	Fonctionnalité de base des circuits : caractérisation du temps de décharge de la ligne de match	224
3.4.2	Marge de détection et capacité de recherche	226
3.4.3	Caractérisation de l'endurance en recherche	229
3.5	Conclusion	229
4	Intégration tri-dimensionnelle monolithique de deux niveaux de transistors CMOS hautes performances avec un niveau de dispositifs de mémoires résistives	231
4.1	Objectif de ce chapitre	231
4.2	Intégration tri-dimensionnelle monolithique de mémoires résistives et transistors CMOS	231
4.2.1	La technologie CoolCube™	231
4.2.2	Intégration des mémoires résistives	232
4.3	Caractérisations électriques de l'intégration tri-dimensionnelle monolithique de deux niveaux de transistors NMOS et un niveau de mémoires résistives	233
4.4	Conclusion	236
5	Conclusion et perspectives	237
	Références : Résumé en français	239
	List of Figures	265
	List of Tables	268

TABLE OF CONTENTS

Introduction

Contents

1.1	From Von Neumann to neuromorphic computing .	2
1.1.1	The Von Neumann bottleneck	2
1.1.2	The end of Moore's law	3
1.2	New technology enablers	4
1.2.1	Resistive memory technology	4
1.2.2	Three-dimensional technology	13
1.3	The third generation of neural networks: Spiking neural network	17
1.3.1	Overview of spiking neural networks	18
1.3.2	Information coding and network routing	20
1.3.3	Hardware spiking neuron: the leaky integrate-and-fire neuron model	21
1.3.4	Hardware synapse implementation	23
1.3.5	Overview of fabricated neuromorphic processors . .	30
1.4	Goal of this PhD thesis	31

1.1 From Von Neumann to neuromorphic computing

1.1.1 The Von Neumann bottleneck

BIOLOGICAL brains are natural computing systems. The idea of taking inspiration from biological brains for designing computers can be dated back at least from the first draft of an Electronic Discrete Variable Automatic Computer (EDVAC) by John Von Neumann in 1945 [1]. Following the theoretical framework on neural computation by MacCulloch and Pitts [2], the Von Neumann's EDVAC was centered around computing elements behaving in a neuron-like manner (*i.e.* all-or-none elements) and transmitting stimuli along excitatory and inhibitory synapses. Yet the implementation was eventually not bio-inspired due to technological constraints and can be translated into three main parts: a *Central Processing Unit (CPU)*, the *memory*, and a *connecting element* between the CPU and the memory [3, 4]. This architecture paradigm - often named after his co-inventor as the *Von Neumann model* or *Von Neumann computer* - mainly relies on the exchange of data between the CPU and the memory through the connecting element [4–6]. Since then, it has dominated the computing paradigm mainly owing to its ease of programming [4, 7].

However, Von Neumann computers have two inherent drawbacks. The first problem is the sequential nature of the system: Von Neumann computers can only manipulate one operation at a time since the connecting element can only transmit a single word between the CPU and the memory [4, 8]. The second problem is the physical separation of computation cores (CPU) and the memory. Nowadays, computation can be as short as nanoseconds and memory accesses as long as milliseconds [9, 10]. Although these two problems were not critical back then, they now lead to a bottleneck - commonly referred to as the *Von Neumann bottleneck* [4] or the *memory wall* - that heavily constraints efficiency of current computing systems [5, 10–13] as shown in FIGURE 1.1.1 (a). This is particularly apparent with the growing importance of data-abundant applications [14–16], such as big data analytics and machine learning tasks [9], wherein most of the computational power and time are now spent in transmitting data back and forth between the CPU and the memory [7, 12, 17] (FIGURE 1.1.1 (b)). Therefore, this has motivated to rethink computation. One idea is to shift from the traditional Von Neumann architecture to *non-Von Neumann architectures*, for instance by merging computation cores and memories as depicted in FIGURE 1.1.1 (c). The biological brains are the best example of such systems featuring massively parallel networks of co-localised computational units, *neurons*, and memories, *synapses* [5, 8]. Herein comes the *neuromorphic* approach coined by Carver Mead in 1989 [18]. The neuromorphic engineering, or *neuromorphic computing*, aims to develop novel computing architectures based on Very Large Scale Integration (VLSI) systems that implement bio-inspired models from the neural system. It has emerged as an approach to tackle the issues presented by the Von Neumann architecture and more recently the challenges posed by the end of Moore's law [6, 19–21] by mimicking biological neural systems more accurately than what Von Neumann attempted.

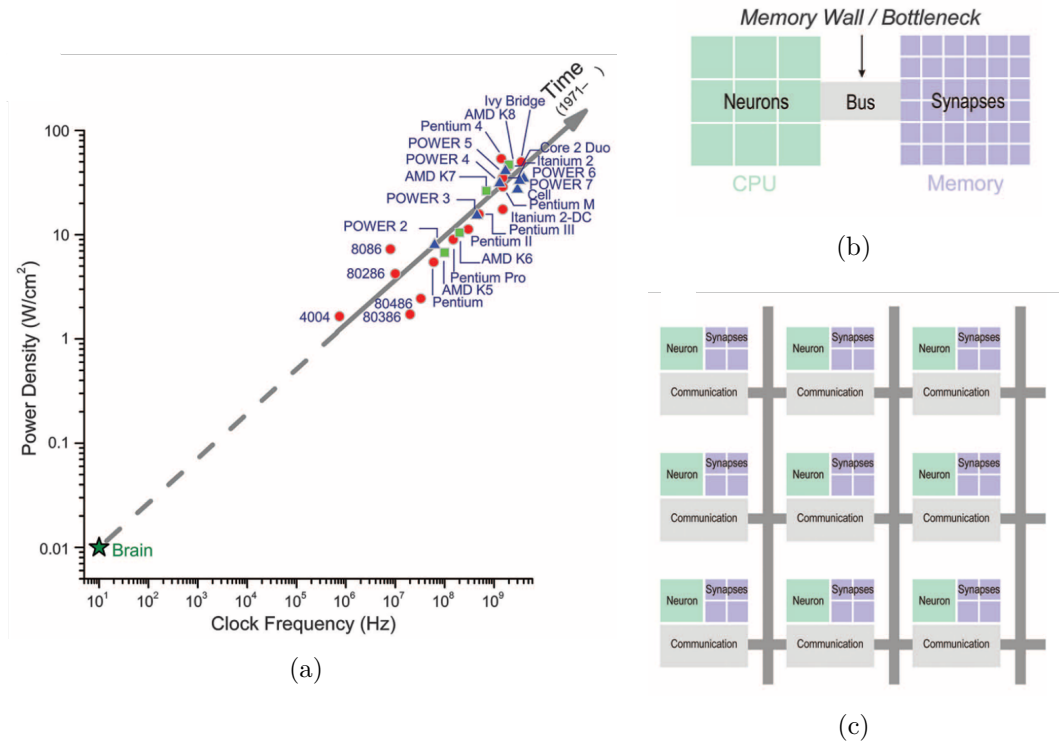


FIGURE 1.1.1: (a) Power density as a function of clock frequency. Current Von Neumann-based architectures are inefficient for representing massively interconnected neural networks. Brains differ from today's computers by their architecture: they feature a parallel, distributed architecture, whereas Von Neumann systems exhibit sequential, centralised architectures. (b) In Von Neumann-based architectures computation and memory units are physically separated by a bus leading to the so-called Von Neumann bottleneck. (c) Conceptual blueprint of a brain-like architecture wherein computation and memory are tightly co-localised. Reproduced from [22].

1.1.2 The end of Moore's law

Gordon E. Moore predicted in 1965 [23] that the number of components per integrated circuits would double every year at decreasing costs. This postulate remarkably held true ten years later when he extended it for the next decades [24]. Foreseeing a slowdown in its initial hypothesis, Moore anticipated an increase in the number of components per integrated circuits every two years rather than every year [25]. This has been known as the so-called *Moore's law* and has served as a goal for the semiconductor industry for more than fifty years [6, 21, 26]. The reasons for this improvement are several folds. They can be accounted for by an increase in die size thanks to a decreased density of defects at acceptable yields, new approaches in circuit and device design to benefit as much as possible from unused silicon areas, and scaling in device dimensions [25]. The latter has been the main motor of Moore's law trends, and the silicon area of Metal-Oxide-Semiconductor Field-Effect Transistor (MOSFET) halved every two years - *i.e.* the gate length of MOSFETs has been scaled down by roughly a factor 0.7x at every technology node. This has been possible by pure downscaling [27] until the 130-nm generation in the early 2000s [28] after which it was mandatory to innovate with new technics, such as the introduction of

strained silicon transistors (90-nm) [29], the use of other gate oxide materials (45-nm) [30], or the conversion from planar transistor to tri-dimensional structures with tri-gate transistors (FinFETs) (22-nm and below) [31, 32].

Nowadays, MOSFET downscaling continued into the sub-10 nm regime [33] with only a few companies - Intel [34], Samsung [35], and TSMC [36] - developing a 7-nm or even 5-nm technology node [37]. However, it becomes more and more challenging to scale down MOSFET transistors any further as we are now reaching fundamental physical limits. Transistor gates are currently as long as a few nanometers, that is the size of a few atoms, and it has been calculated that the minimum size of a computational switch cannot go below 1.5 nm [38]. Each new generation takes longer to be released - about 2.5 to 3 years instead of the normal 2-years rate [28] -, and the cost of lithographic equipment is exploding - it skyrocketed to several hundreds of millions of dollars for the latest technology nodes, whereas it costed only a few tens of thousands of dollars in 1968 [25]. This has motivated researchers to investigate *new devices* for logic and memory, *new integration processes*, such as three-dimensional monolithic integration, and *new computing architectures* much more energy-efficient than the Von Neumann architecture in order to perpetuate Moore's law trends [6, 19, 21].

The scope of this PhD thesis is to investigate the *hardware implementation of reconfigurable spiking neuromorphic processors exploiting new technologies*, namely *Resistive Memories (RRAMs)* and *three-dimensional (3D) monolithic technologies*. The following of this introduction provides the basics to grasp the challenges of this PhD work. Resistive memory technology and three-dimensional integration are first introduced. Then, an overview on spiking neural network systems is presented.

1.2 New technology enablers

This section will present an overview of the new technology enablers to implement neuromorphic systems based on non-Von Neumann architectures, namely resistive memory and three-dimensional integration.

1.2.1 Resistive memory technology

1.2.1.1 The memory hierarchy

Different memory technologies are available for data storage. They are usually classified into two broad categories: (i) *volatile memories* and (ii) *non-volatile memories*. FIGURE 1.2.1 (a) shows an overview of the most important memory technologies. While volatile memories lose the stored information shortly after the power supply is shut off, non-volatile memories permanently retain the stored data. The memory hierarchy of current Von Neumann computing systems employs different memory technologies to achieve a trade-off between cost and performance. The closer to the processor cores, the faster the memory needs to be as depicted in FIGURE 1.2.1 (b). The established memory technologies - namely Static Random Access Memory (SRAM), Dynamic Random Access

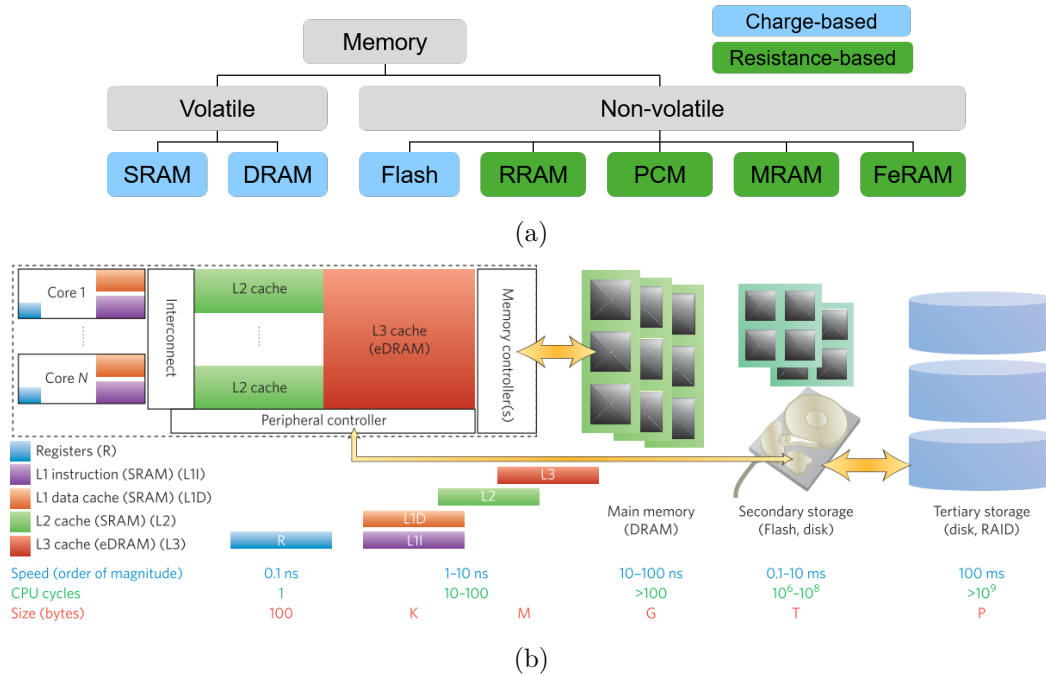


FIGURE 1.2.1: (a) Overview of established charge-based memories (blue) and new resistance-based non-volatile memories (green). (b) Memory hierarchy of today's computers. Speed, number of processors cycles (CPU cycles), and typical capacity (size) of the different memories are shown in the lower panel. The closer to processing cores (CPUs), the faster the memory (cache memory). Reproduced from [9].

Memory (DRAM), and Flash memory - are all based on charge storage, yet they exhibit very distinct characteristics. SRAM is implemented with six transistors. Therefore it is the most expensive memory because of its large silicon area consumption, but it is also the fastest. DRAM is cheaper than SRAM since it is implemented with one transistor and one capacitor but is also slower. In addition, it requires periodic refresh of the stored information to prevent data loss which increases the energy consumption. These two volatile memory technologies are used close to the processor (CPU) to enable fast operations (cache and main memory). On the other hand, Non-Volatile Memories (NVMs), such as Flash memory and hard drives, are used for non-volatile data storage. They are slower than SRAM and DRAM, however they do not consume stand-by power thanks to their non-volatility.

New NVM technologies have emerged [39] and have been intensively studied over the last decade. They fundamentally differ from charge-based memories as they do not store the information in a capacitor but deploy different physical mechanisms to change their electrical resistance state and often embody the concept of *memristors* [40]. More importantly, they can easily be integrated in the Back-End-Of-Line (BEOL) of advanced Complementary Metal-Oxide-Semiconductor (CMOS) process. The rest of this section will provide a quick overview of the major new non-volatile memories under research, namely Resistive Random Access Memory (RRAM), Phase-Change Memory (PCM), and Spin-Transfer-Torque Magnetic Random Access Memory (STT-MRAM), with a particular emphasis on RRAM. A focus on the challenges posed by RRAMs will

also be provided.

1.2.1.2 Overview of new non-volatile memory technology

Resistive random access memory

Resistive Random Access Memory (RRAM or ReRAM) is a type of memory consisting of a Metal-Insulator-Metal (MIM) structure wherein a thin metal oxide layer is sandwiched between two metal electrodes as depicted in FIGURE 1.2.2 (Bottom left). The basic principle of RRAMs relies on the formation and dissolution of a Conductive Filament (CF) in the oxide layer [41–43]. Initially, fresh RRAM samples are in a *pristine* state featuring a high resistance value. Upon the application of an initial forming voltage between the Top and Bottom Electrodes (TE and BE, respectively) during the so-called *forming* operation, the CF is created by soft dielectric breakdown. This process is reversible: the CF can be partially disrupted by applying a Reset voltage between the TE and BE during a *Reset* operation, and it can be formed again by applying a Set voltage during a *Set* operation. Forming the CF shunts the TE and BE and results in a drop of the RRAM electrical resistance - this leads to the Low Resistance State (LRS) -, whereas disrupting the CF disconnects the two electrodes and prevents current conduction in the oxide layer - this leads to the High Resistance State (HRS). Note that RRAMs in the HRS feature a lower resistance value than their pristine resistance value since the CF is only partially disrupted [41, 43–46]. For memory applications, the LRS and HRS are used to store one bit of information: the LRS is associated to a binary '1' and the HRS to a binary '0'. Switching back and forth between the LRS and HRS is called a *switching cycle* and can be repeated as many times as permitted by the RRAM technology [47–49]. The maximum number of switching cycles permitted by a technology defines its *programming endurance* - sometimes also termed *cycling endurance* or just *endurance*. Depending on their switching mode, RRAMs can be distinguished between *unipolar* devices wherein Set and Reset operations are performed with the same polarity - *i.e.* voltage biases are applied on the same electrode for both operations - or *bipolar* devices wherein Set and Reset polarities must be alternated. If the unipolar switching can symmetrically occur on both electrodes, it is also referred to as a *nonpolar* switching mode [44]. The switching mode depends on the choice of oxide layer and electrode materials [42, 50]. In some cases, both unipolar and bipolar switching modes can be observed in the same device [51]. Unipolar devices allow for reduced design complexity since both operations are performed with the same polarities. However, they typically require higher programming currents with respect to bipolar devices [46].

RRAM devices can be classified into (i) Oxide-based RAM (OxRAM) and (ii) Conductive-Bridge RAM (CBRAM). In OxRAM technology, the CF is composed of oxygen vacancies in the oxide layer [52]. In CBRAM technology, the CF is attributed to the migration of metallic cations, such as copper and silver [53]. OxRAM generally presents low resistance ratios between its HRS and LRS (≈ 10 -100) but good programming endurance ($> 10^{12}$ cycling operations), whereas CBRAM features higher resistance ratios (10^3 - 10^6) but lower endurance ($< 10^4$) [44, 46, 54, 55]. In this work, we will only focus on *bipolar OxRAM technology*. The most common oxide layer materials used today in RRAM

are HfO_x , AlO_x , TiO_x , and TaO_x [44]. FIGURE 1.2.2 illustrates the switching process in an OxRAM device. During the forming operation, oxygen atoms in the oxide layer drift towards the top electrode due to the application of a high electric field. This generates defects in the oxide layer and leads to the creation of a CF made of oxygen vacancies. The interface between the top electrode and the oxide layer acts like an oxygen reservoir [42, 44]. During Reset operations, the CF is disrupted by recombination of oxygen vacancies and oxygen atoms. Set operations reform the CF by pushing oxygen atoms back to the top electrode. Over the last decade, RRAMs have been seen as a promising candidate to replace Flash memories. Aside from their non-volatility property, RRAMs present numerous advantages, such as good programming endurance ($>10^{12}$ [56, 57]), non-destructive read operations, fast switching (below nanoseconds [58–60]), and low-current programming operations thanks to the filament nature of current conduction (tens of nanoamperes [61–65]). As the width of the CF can be smaller than 10 nm [52], RRAMs can potentially be scaled down below 10-nm dimensions [66]. Despite all its advantages, RRAM still faces two major roadblocks that have prevented it so far from being integrated in large arrays. First, the initial forming voltage (2-3V) is significantly higher than the operating voltage. Second, RRAM is strongly affected by extrinsic and intrinsic resistance variability arising from the fabrication process as well as the intrinsic stochastic nature of the CF formation. These challenges are discussed more in details in the next section.

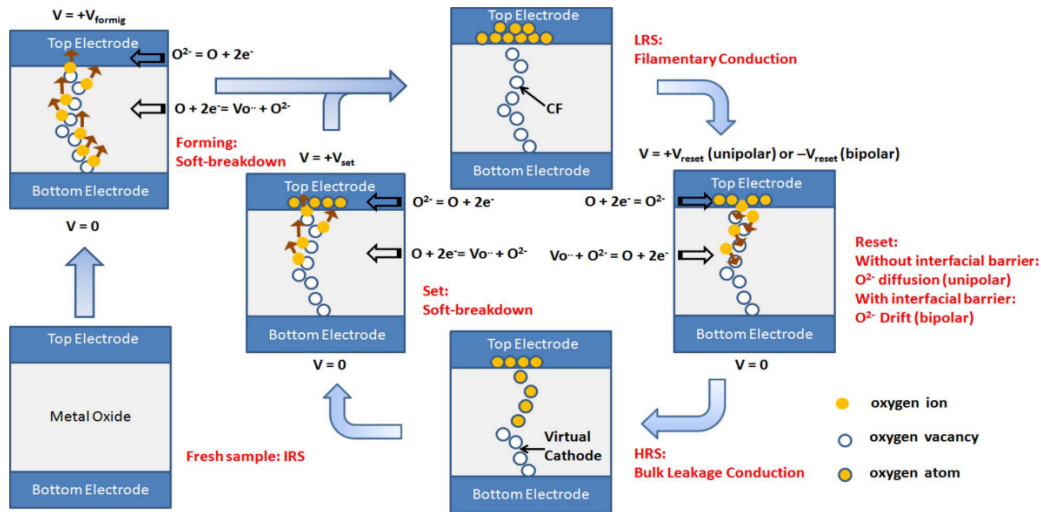


FIGURE 1.2.2: Schematic illustration of the switching process in Oxide-based Resistive Memories (OxRAMs). An initial forming process generates oxygen vacancies in the oxide layer by soft dielectric breakdown. Subsequent Set and Reset operations lead to the formation and dissolution of a Conductive Filament (CF) made of oxygen vacancies, respectively. The interface between the oxide layer and the top electrode acts like an oxygen reservoir. Reproduced from [44].

Phase-Change Random Access Memory

Phase-Change Random Access Memory (PCM or PCRAM) is composed of two electrodes sandwiching a chalcogenide glass that can change between a crystalline and an amorphous phase. The most used chalcogenide material in PCM is the ternary compound $\text{Ge}_2\text{Sb}_2\text{Te}_5$, also referred to as GST [67–70]. As for RRAM,

PCM stores one bit of information by modulating its electrical resistance. The resistance modulation relies on the transition between the crystalline and the amorphous phase of the chalcogenide material. The crystalline state features a low electrical resistance - corresponding to the Low Resistance State, LRS - while the amorphous state features a high electrical resistance - the High Resistance State, HRS. This transition occurs by passing a current through the material to heat it up by Joule heating. The advantage of PCM is that the ratio between the HRS and LRS resistance values is generally larger than that of RRAM technology, thus making it promising for multi-bits storage and facilitating its integration into large arrays. However, PCM technology suffers from *resistance drift* over time towards higher resistance values, in particular in the amorphous phase - *i.e.* mainly in the HRS [71]. This makes it difficult to distinguish the programmed states over time. Another drawback of PCM technology is its high current consumption during programming, especially during Reset operations. Since PCM is programmed by Joule heating, the programming current scales down with device area. Yet even for PCM devices scaled down to sizes smaller than 10 nm, programming current of the order of microamperes is still required [72, 73].

Spin-Transfer-Torque Magnetic Random Access Memory

Spin-Transfer-Torque Magnetic Random Access Memory (STT-MRAM) is a type of magnetic memory based on the most advanced currently available technology to achieve higher scalability. STT-MRAM stores the information (*i.e.* '0' or '1') in the magnetisation of ferromagnetic materials. It is composed of two ferromagnetic layers separated by a thin insulator layer. The basic principle of STT-MRAM relies on the switching of magnetisation of one ferromagnetic layer (the free layer), while the magnetisation of the other layer is fixed (the pinned layer) [9]. If the free layer has the same magnetisation as the pinned layer - the *parallel* configuration -, electrons have a higher probability to pass through the device. This corresponds to the LRS. Conversely, if the free layer has an opposite magnetisation - the *anti-parallel* configuration -, the device is in the HRS since the anti-parallel configuration prevents current conduction. STT-MRAM provides numerous advantages, such as low-energy programming, high speed, and almost unlimited programming endurance [12]. In addition, it shows high uniformity in its resistance states, unlike RRAM technology. However, one of the main drawbacks of STT-MRAM technology is its low resistance ratio between the LRS and the HRS. This requires the use of specific memory cell architecture to mitigate the low resistance ratio that limits STT-MRAM scalability [9, 12, 74].

Comparison of the main metrics and summary

Several prototypes of RRAM [78, 81], PCM [77], and STT-MRAM [76] have been demonstrated, up to several gigabits as reported in FIGURE 1.2.3 (a). Commercial products are already available by different companies, such as Intel and Micron with the 3D XPoint technology [82], Panasonic [83], Avalanche [84], or Everspin [85]. FIGURE 1.2.3 (b) compares reported programming energy as a function of cell area. Unlike PCM and STT-MRAM, programming energy of RRAM technology (encompassing OxRAM and CBRAM) does not scale

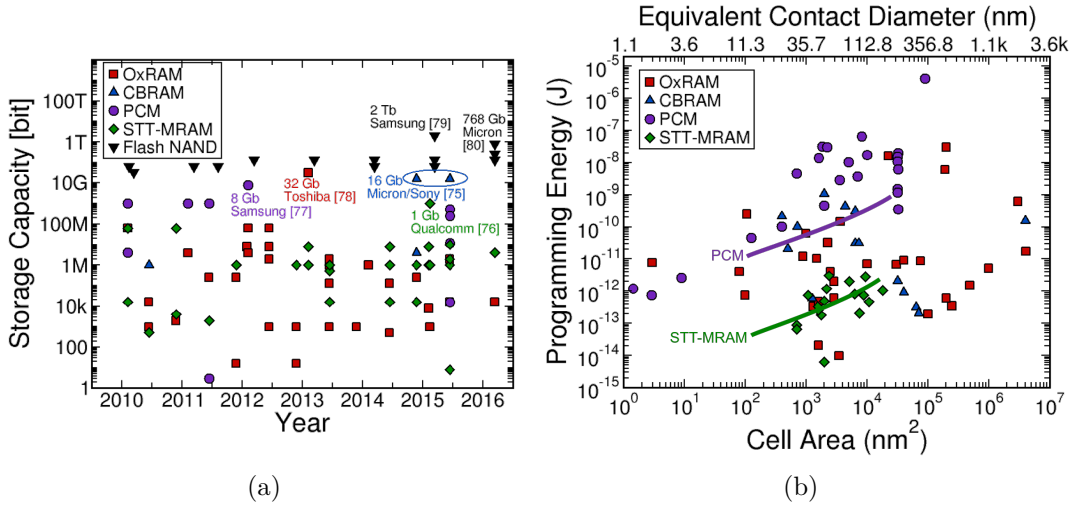


FIGURE 1.2.3: (a) Reported storage capacity over time of different non-volatile memory technologies. Multi-gigabits prototypes with the new non-volatile memories have been reported [75–78]. Data on Flash NAND technology are reported for comparison [79, 80]. (b) Programming energy as a function of cell area. Programming energy of Resistive Memories (RRAMs, encompassing OxRAM and CBRAM technologies) does not depend on cell area due to the filament conduction nature of RRAMs. Reproduced from [54].

with cell area due to the filament conduction nature of RRAMs. A significant advantage of these new non-volatile memories is that they can be monolithically integrated in three-dimension with CMOS logic circuits because their fabrication temperature is compatible with the back-end-of-line, and they are fabricated with materials commonly used in the semiconductor industry. This facilitates the implementation of in-memory computing architectures by physically bringing memory units close to processing cores and can improve computing efficiency by many orders of magnitude [12, 19]. In addition, each of these devices can be independently programmed bit by bit, whereas Flash memories require to erase whole blocks of kilobits whenever any individual bit in the array needs to be reprogrammed. Finally, they are two-terminals devices, unlike conventional three-terminals CMOS-based DRAMs or Flash memories, and can be integrated into so-called *crossbar arrays* in between densely packed word lines and bit lines. This allows for an extremely small bit area of only $4F^2/n$, where F is the minimal lithographic feature size and n the number of stacked layers [46].

1.2.1.3 Challenges of resistive memory technology

This part provides more details about the main metrics and challenges with RRAMs that will be discussed in this work, namely the memory window, the programming endurance, and the resistance variability.

In RRAM technology - encompassing OxRAM and CBRAM technologies - the resistance values of LRS and HRS depend on the programming conditions, *i.e.* the applied voltage, programming time, and programming current - related to the *compliance current*, I_{cc} . The compliance current I_{cc} is necessary to prevent cell

failures due to the abrupt increase of current during forming and Set operations. It can be defined by programming equipment or in practice by integrating in series a selector element, such as a diode or a transistor [43, 78, 86, 87]. It has been demonstrated that programming time exponentially depends on programming voltage [88–90]. Therefore, programming time is usually fixed and only the programming voltage varies. LRS resistance values are mostly defined by I_{cc} during the Set operation [46]. One of the universal characteristics of RRAM is that LRS resistance values have a power law dependency on I_{cc} as shown in FIGURE 1.2.4 (a): increasing I_{cc} results in lower LRS resistance values, R_{LRS} . On the other hand, HRS resistance values are mainly defined by Reset programming voltages [91], and using higher voltages during Reset operations leads to higher HRS resistance values, R_{HRS} (*cf* FIGURE 1.2.4 (b)). This provides important guidelines for programming RRAM devices. As explained previously, RRAM stores one bit of information in its LRS and HRS, thus it is fundamental to guarantee a sufficient ratio between both states, R_{HRS}/R_{LRS} , in order to discriminate them. Ideally, the ratio R_{HRS}/R_{LRS} , often called the *Memory Window* (MW), has to be maximised in order to facilitate the integration of RRAMs into large arrays. However, it has been demonstrated that a trade-off exists between the MW and programming endurance performance: higher MWs imply lower programming endurance [49, 54, 55, 90]. FIGURE 1.2.5 shows two typical RRAM endurance characterisations performed on a GeS_2/Ag (a) and a $\text{HfO}_2/\text{GeS}_2/\text{Ag}$ (b) RRAM stack, *i.e.* the evolution of LRS and HRS resistance values after different numbers of Set/Reset switching cycles [90]. During the cycling, HRS resistance values generally tend to decrease, and some cells can be permanently stuck in the LRS after a certain number of switching cycles [43, 44, 48, 92]. This results in a decrease of the MW. We define here the *programming endurance* as the maximum number of Set/Reset cycles we can perform with a stable MW. While it is possible to sustain a low constant MW of about 10 during 10^8 cycles (FIGURE 1.2.5 (a)), only 10^3 cycles can be performed with a large MW of 10^6 (FIGURE 1.2.5 (b)). FIGURE 1.2.5 (c) reports the memory window of different RRAM technologies associated to the corresponding endurance. For the sake of comparison, some data on PCM and STT-MRAM technologies are also reported. As it can be observed, endurance performance higher than 10^6 cycles for RRAM - to be comparable to Flash technology [12, 93] - is usually associated to low memory windows below 10-100. This low MW is critical for large memory arrays due to sneak paths issues [86]. Therefore, selector devices have to be integrated in series with each RRAM device to limit leakage currents - usually a CMOS transistor in the so-called one-transistor/one-RRAM (1T1R) structure. However, this limits storage density [94].

Another main drawback of RRAM technology is its high resistance variability - both across cycles and devices - inducing non-repeatable behaviours [43, 95, 96]. As it is illustrated in the endurance characterisations in FIGURE 1.2.5, LRS and HRS resistance values vary at every switching cycle - referred to as *cycle-to-cycle variability*. Cycle-to-cycle variability can be attributed to the stochastic nature of the conductive filament formation and dissolution. On the other hand, resistance variability also occurs across devices in a memory array - *device-to-device variability* - arising from external factors like fabrication process [97]. FIGURE 1.2.6 shows an example of resistance distribution measured on a 4-kbit

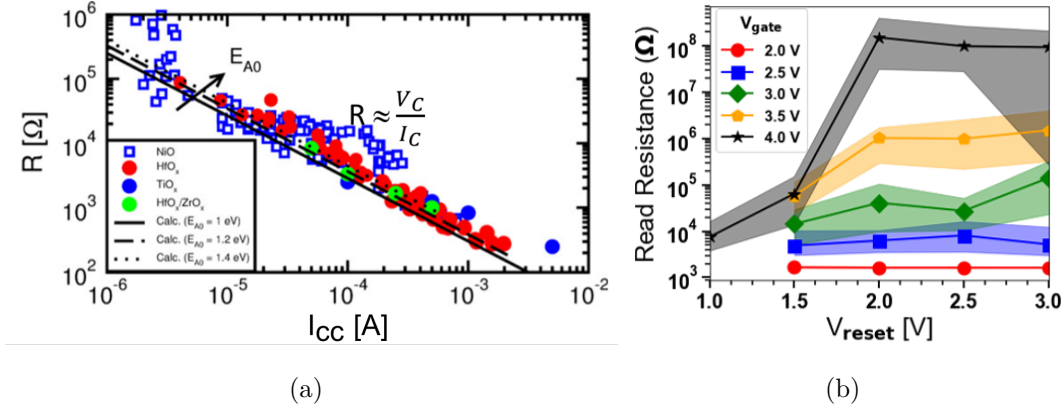


FIGURE 1.2.4: Low Resistance State (LRS) resistance value of different Resistive Memory (RRAM) technologies as a function of compliance current, I_{cc} , during a Set operation. A power law relationship exists between LRS resistance values and I_{cc} . Reproduced from [46]. (b) High Resistance State (HRS) resistance value as a function of the voltage applied during Reset operations, V_{reset} . Measurements have been performed on a TiN/HfO₂/Ti/TiN RRAM device. The mean HRS resistance value over 1000 Reset operations is shown (solid line) as well as the spread at two standard deviations (shaded area). Reproduced from [91].

TiN/HfO₂/Ti/TiN RRAM array [43]. While it is possible to reach a ratio of 2500 between the median HRS and LRS resistance values, the device-to-device resistance variability - mainly in the HRS [43, 44] - degrades this ratio down to 600 if one considers the ratio between HRS and LRS resistance values at -3σ and $+3\sigma$, respectively. Numerous works have tried to tackle and mitigate resistance variability, for instance by material and process engineering [98, 99]. Understanding better the physics of RRAM is still an active area.

A last issue that can be mentioned is the need of an initial *forming* operation [98]. In order to generate a sufficient amount of defects to initiate the switching [44], high forming voltages ($\approx 2-3$ V) associated with a high electric field (>10 MV/cm) are required [44, 54]. This is higher than the power supply voltage and any subsequent programming operations, and it is not desirable for practical applications. In addition, it constraints the transistor used as a selector in order to prevent any degradation at such a high voltage [13]. Therefore, there have been significant efforts in the literature to design *forming-free* RRAM devices [64, 100–102]. For instance, it has been found that the forming voltage is linearly dependent on the thickness of the oxide layer, and HfO_x-based RRAMs can be free of the forming operation below 3 nm [100]. However, it may severely decrease HRS resistance values and the memory window. It has also been reported that forming voltages can be reduced by engineering around the fabrication process [44].

To summarise, main challenges of RRAM technology are:

- the low memory window ($<10-100$) in order to ensure sufficient programming endurance ($>10^6$ cycles)
- the high cycle-to-cycle and device-to-device resistance variability that limits the memory window

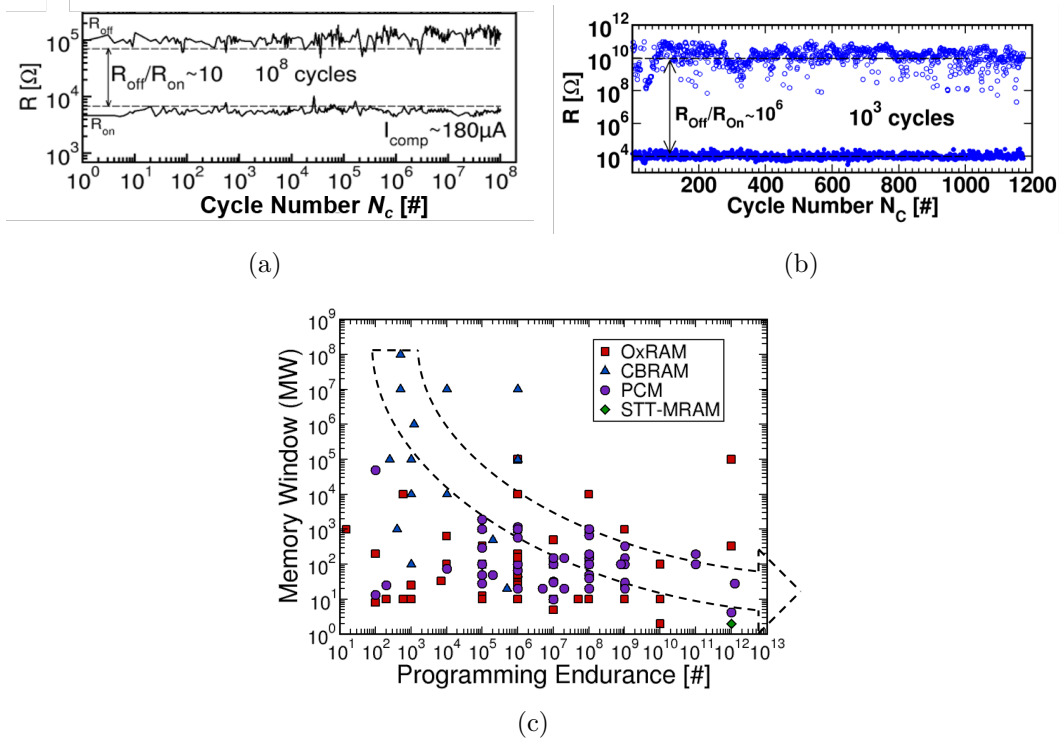


FIGURE 1.2.5: Typical endurance characterisations performed on (a) a GeS_2/Ag and (b) a $\text{HfO}_2/\text{GeS}_2/\text{Ag}$ Resistive Memory (RRAM) stack. While it is possible to sustain a low resistance ratio $R_{\text{off}}/R_{\text{on}}$ of 10 during 10^8 switching cycles, only 10^3 switching cycles can be performed with a large resistance ratio of 10^6 . Reproduced from [90]. (c) Reported Memory Window (MW) as a function of programming endurance for different RRAM technologies. Data on Phase-Change Memory (PCM) and Spin-Torque-Transfer Magnetic Memory (STT-MRAM) are reported for comparison. A general trend of lower MWs with higher endurance performance is observed. Reproduced from [54].

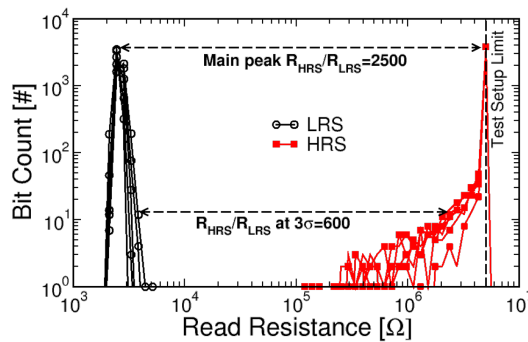


FIGURE 1.2.6: High Resistance State (HRS, red) and Low Resistance State (LRS, black) resistance distributions measured on a 4-kbit $\text{TiN}/\text{HfO}_2/\text{Ti}/\text{TiN}$ Resistive Memory (RRAM) array, after one Reset/Set cycle, respectively. While a resistance ratio of 2500 is measured between the median HRS and LRS resistance values, it is reduced to 600 at three standard deviations, 3σ , due to device-to-device resistance variability. Reproduced from [43].

- the high forming voltage

1.2.2 Three-dimensional technology

Another technological solution to continue Moore’s law trends is to benefit from the third dimension, *i.e.* the vertical axis. Three-dimensional (3D) integration allows to pack more components on a given silicon area. We first provide an overview of 3D integration with Resistive Memories (RRAMs, presented in SECTION 1.2.1), then 3D integration with Complementary Metal-Oxide-Semiconductor (CMOS) technology.

1.2.2.1 Three-dimensional integration of resistive memories

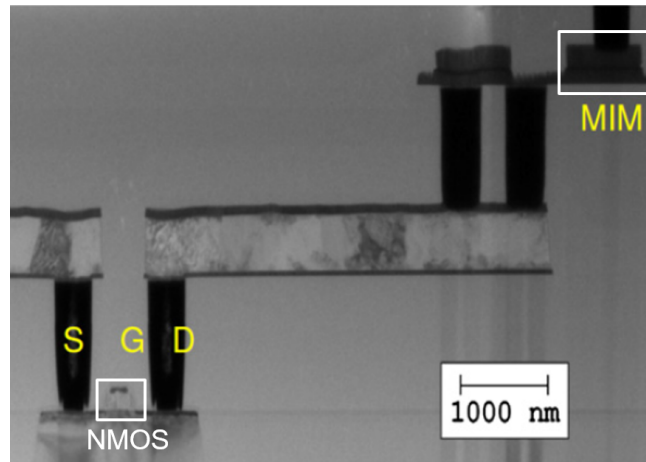


FIGURE 1.2.7: Transmission electron microscopy of a TiN/HfO₂/Ti/TiN RRAM fabricated on top of a NMOS transistor. RRAMs have been integrated in the back-end-of-line. Reproduced from [103].

Two main concepts were proposed with Resistive Memories (RRAMs) in order to benefit from the third dimension. They take advantage of RRAM Back-End-Of-Line (BEOL) CMOS process compatibility and the simple Metal-Insulator-Metal (MIM) structure of RRAM. The first concept consists in stacking one or several layers of RRAM devices directly on top of CMOS logic circuits in the BEOL [19, 104]. For instance, RRAMs can be fabricated on top of NMOS or PMOS transistor contacts in the so-called one-transistor/one-RRAM (1T1R) structure as shown in FIGURE 1.2.7. Another case in point are cross-point architectures with RRAMs [105, 106] wherein memory cells are located in between densely stacked word lines and bit lines (*cf* FIGURE 1.2.8(a)), such as the 3D XPoint technology of Intel and Micron [82]. This design allows to use every word- and bit-line for two consecutive layers of memory devices, thus halving the number of metal layers. The second possible 3D integration concept is called *Vertical RRAM* (VRRAM) wherein MIM layers are integrated vertically in a pillar (*cf* FIGURE 1.2.8 (b)) [61, 62, 107–110]. The pillar is a common vertical electrode (the metal (M), *e.g.* TiN/Ti) whose sidewall is covered by the resistive switching layer oxide (the insulator (I), *e.g.* HfO_x). Horizontal metal layers are stacked on top of each other and form the other electrode (the metal (M), *e.g.* TiN). Memory elements are located where the horizontal electrode surrounds the vertical one.

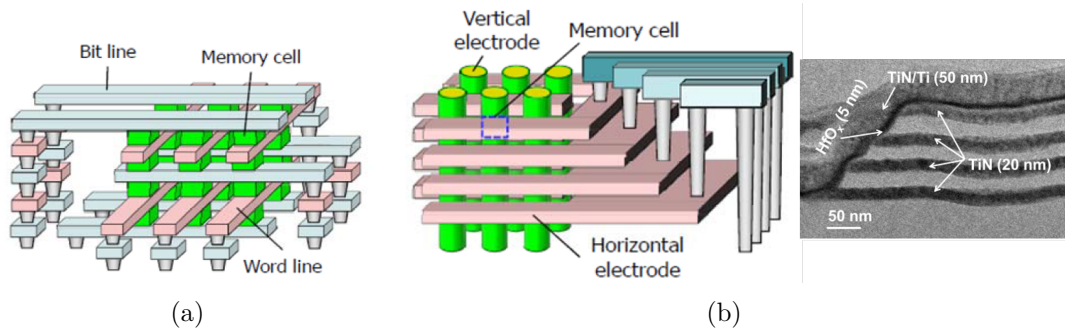
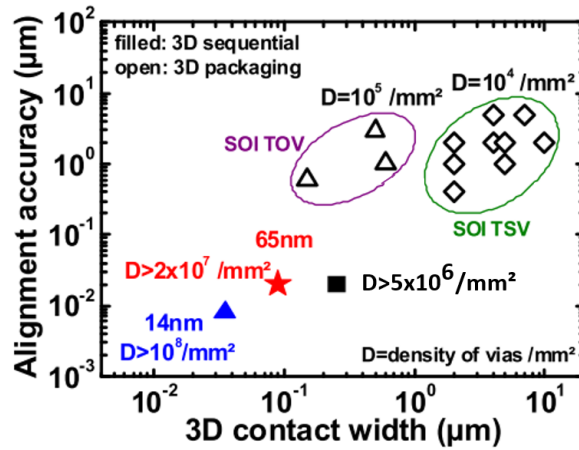
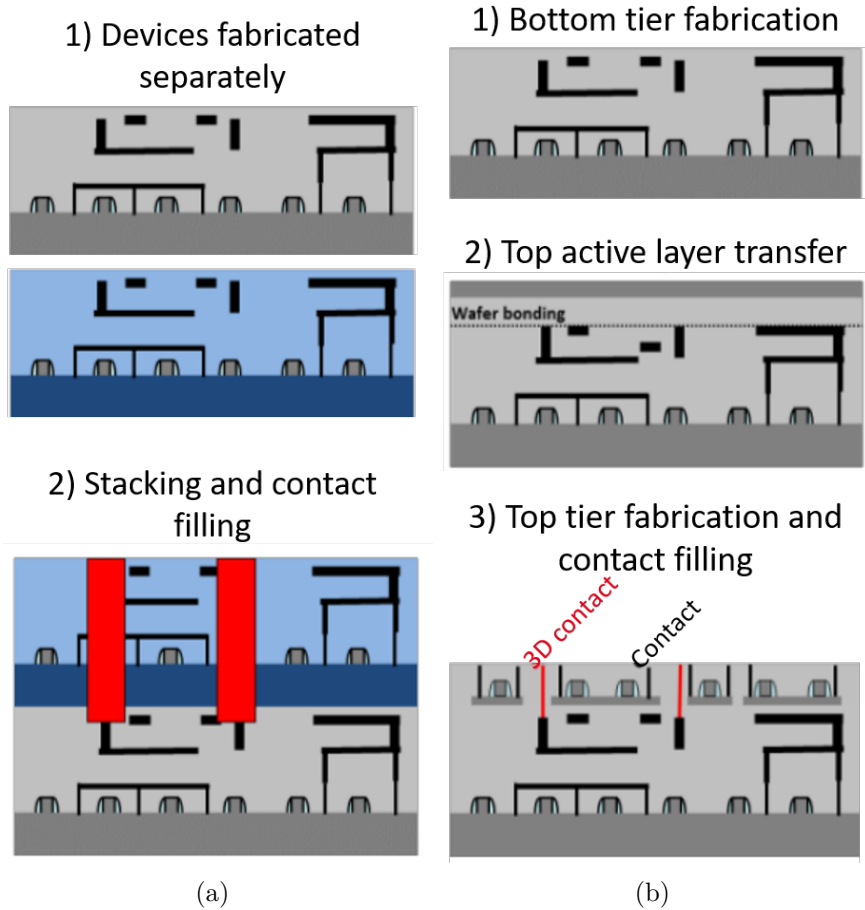


FIGURE 1.2.8: (a) Schematic drawing of a three-dimensional (3D) cross-point Resistive Memory (RRAM) structure. RRAM cells are located in between densely stacked word-lines and bit-lines. Reproduced from [109]. (b) (Left) Schematic drawing of Vertical RRAM (VRRAM) arrays (reproduced from [109]), and (Right) transmission electron microscopy of a four-layers TiN/Ti/HfO_x/TiN VRRAM (reproduced from [110]).

1.2.2.2 Three-dimensional integration of CMOS transistors

Although RRAMs can easily be fabricated on top of CMOS transistors thanks to their low-temperature process, stacking several layers of CMOS transistors on top of each other poses more challenges. Two 3D integration types can be distinguished: (i) the parallel integration, and (ii) the sequential integration as depicted in FIGURE 1.2.9. In the parallel integration - sometimes called *3D packaging* (FIGURE 1.2.9 (a)) - the different layers (*tiers*) of transistors are processed separately, then vertically stacked and connected afterward (for instance with Through-Silicon Via (TSV) or Through-Oxise Via (TOV)). In the sequential integration - also termed *monolithic integration* (FIGURE 1.2.9 (b)) - every tier of transistors is fabricated directly on top of the previous one. The parallel integration has the advantage of much simpler manufacturing process with respect to the sequential integration, yet it suffers from lower alignment accuracy since it requires to align two whole tiers together (see FIGURE 1.2.9 (c)). On the other hand, in the sequential integration, a tier is fabricated directly on top of the previous one. Therefore, it can reach alignment accuracy at the transistor scale as it only depends on lithographic alignment on the stepper. This is a significant advantage of monolithic integration over parallel integration as it allows to improve interconnection density by a factor 50x (10^{10} vias/cm² vs $2 \cdot 10^8$ vias/cm², respectively) [113]. Yet the major downside of monolithic integration is its fabrication process: the fabrication of a tier can degrade performance of lower tiers if process temperature is not kept low enough [111, 113, 114]. For instance, it has been demonstrated that transistors fabricated in a 28-nm Silicon On Insulator (SOI) process can sustain thermal budget up to 500°C for 5 hours without noticeable performance degradation (*cf* FIGURE 1.2.10 (a)) [115]. At higher thermal budget (*i.e.* temperature and process time), degradation comes from silicide deterioration of CMOS transistors as well as from a slight dopant deactivation [115]. However, standard thermal budget to manufacture a transistor is usually higher than 1000°C [112]. As a result, it is mandatory to either improve transistor thermal stability and/or adapt manufacturing process to enable low-temperature transistor fabrication, while



(c)

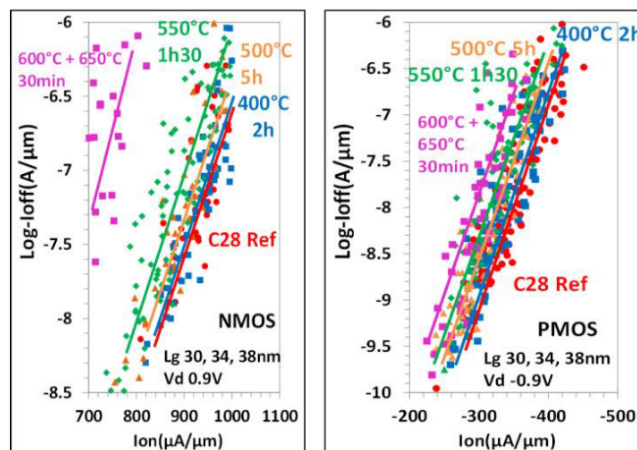
FIGURE 1.2.9: Schematic illustrations of three-dimensional (3D) (a) parallel integration, and (b) sequential integration. In the parallel integration both layers are fabricated separately, then vertically stacked and connected. In the sequential integration the top layer is fabricated directly on top of the bottom layer. Reproduced from [111]. (c) Alignment accuracy as a function of 3D contact width. 3D sequential integration allows for higher alignment accuracy than parallel integration since it only depends on lithographic alignment on the stepper. Reproduced from [112].

sustaining high transistor performance for every tier [116]. Different research groups [117–121] have proposed solutions to overcome these problems, however they generally require complex and expensive fabrication process [119, 120], for instance with the use of III-V materials, or result in lower electrical performance of top tiers [117, 118, 121]. CoolCube™ technology [112] developed by CEA-Leti is an example of 3D monolithic integration of SOI CMOS transistors. Currently, it allows for monolithic integration of two tiers of CMOS transistors in a 65-nm SOI process without noticeable degradation of electrical performance. In addition, it is fabricated with conventional foundry process. Therefore, it is compatible with industrial requirements, in particular with hard contamination constraints. A first layer of CMOS transistors is fabricated in a conventional SOI 65-nm design rules CMOS over CMOS process (the *bottom tier*). Then, the active area of the next layer (the *top tier*) is obtained by transferring a new SOI substrate on top of the bottom tier by oxide bonding. Finally, top tier transistors are fabricated directly on the new top active area with alignment accuracy at the transistor scale as shown in the transmission electron microscopy in FIGURE 1.2.10 (b). CHAPTER 4 will describe more in details the fabrication process of CoolCube™ technology. Preliminary studies have evaluated potential benefits of CoolCube™ integration, for instance on a 3D Field-Programmable Gate Array (FPGA) architecture in [122]. In the 3D FPGA architecture, memory components are placed on the bottom tier and logic circuits on the top tier in order to keep a good global performance. Compared to a planar FPGA architecture, the 3D FPGA architecture can reduce area consumption by 55% and the energy-delay product by 47%. In terms of cost benefits, a cost model developed in [123] has predicted benefits of the order of 50% with 300-mm² (12-inch) wafers.

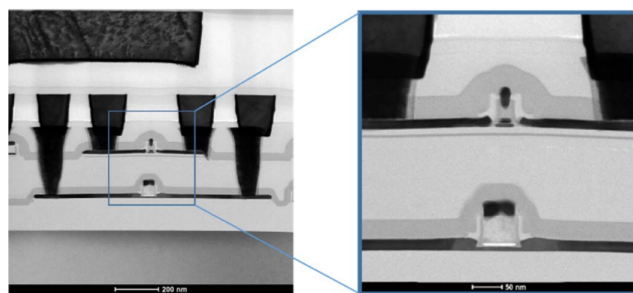
1.2.2.3 Motivations of this work

The first advantage of 3D integration is the possibility to pack more components on a given silicon area, thus increasing component density. In particular, 3D monolithic integration could virtually grant access to a new technology node by stacking two tiers fabricated at a previous CMOS technology node, while being potentially more economically advantageous than developing a new technology node [111, 123]. Another advantage of such integration is that interconnections are in average shorter than those of for a planar integration. This results in less parasitic capacitance as well as less routing congestion [111, 112, 122, 124]. Also, 3D integration can facilitate heterogenous integration [119–121]. N3XT [19] is an example of computing systems implementing many novel technologies, such as RRAM, STT-MRAM, and carbon nanotubes, integrated in a 3D monolithic technology. Such systems are expected to provide significant gains in performance - up to 1000x in energy-delay product. Another case in point is the gas sensor chip fabricated and tested in [104] wherein four layers have been monolithically fabricated - one layer of silicon FET logic circuits, two layers of carbon nanotubes, and one layer of RRAM.

In this thesis, we propose the *3D monolithic integration of several layers of high-performance CMOS transistors with Resistive Memories (RRAMs)*. To this end, we will demonstrate the monolithic co-integration of two tiers of CMOS



(a)



(b)

FIGURE 1.2.10: (a) (Left) NMOS and (Right) PMOS I_{off}/I_{on} performance for different thermal annealings. Transistor performance can be ensured for annealing up to 500°C for 5 hours. Reproduced from [115]. (b) Transmission electron microscopy of two tiers of NMOS transistors fabricated in a 3D monolithic 65-nm SOI process with CoolCube™ technology. Reproduced from [112].

transistors using CoolCube™ technology of CEA-Leti [112] with an additional tier of RRAMs fabricated directed on top of the two tiers of CMOS transistors in the BEOL. CHAPTER 4 will describe the fabrication process of RRAMs with CoolCube™ CMOS transistors and show the electrical functionality of the integration.

1.3 The third generation of neural networks: Spiking neural network

This section presents an overview of Spiking Neural Network (SNN) systems and the main building blocks of hardware SNNs.

1.3.1 Overview of spiking neural networks

1.3.1.1 Biological brains

It is now known that brains are much more efficient at computing than conventional computers based on the Von Neumann architecture. The human brain computes with a meagre power budget of 10-20 W ([125, Section 5.8.2]), improving on Von Neumann computers by millions of fold in terms of power efficiency [126–130]. It is an extremely complex computational engine consisting of 10^{11} computing elements, the *neurons*, densely interconnected by more than 10^{14} connections, the *synapses* [131]. Although most of neural computation is still to be understood, it is widely accepted that memory is stored in synapses, while computation takes place in neurons [132–135]. Neurons are cellular units specialised for the processing of cellular signals. They are mainly composed of a soma, several dendrites, and an axon as shown in FIGURE 1.3.1, and they communicate between each other via electrical signal events, the *Action Potentials* (APs), transmitted along synapses [136]. APs are sharp electrical pulses of about 100 mV and 1 ms. Input APs coming from other neurons are received by the dendrites and integrated inside the soma. At rest, the *membrane potential* of neurons - *i.e.* the difference in electric potential between the inside and outside of neurons - is typically in the range of -40 to -90 mV. Upon integrating APs, the membrane potential fluctuates. If it goes below the resting membrane potential, nothing really happens. If it goes above the resting potential and reaches the *threshold potential* level, the neuron fires an AP from its axon to other neurons it is connected to. Axon terminals connect to other neuron dendrites through terminal buttons forming synapses. Pre- and post-synaptic terminals are physically separated by a *synaptic cleft* whose length is in the order of 20 nm. Two types of synapses can be distinguished: (i) chemical synapses, and (ii) electrical synapses. In chemical synapses, the most abundant type of synapses, synaptic transmissions are carried out by release of neurotransmitters stored in *synaptic vesicles* that bind to receptors at the postsynaptic terminal. In electrical synapses, ions can directly diffuse between pre- and post-synaptic terminals.

Brain efficiency can be accounted for by several factors. First, brains are massively parallel computing systems: all synapses and neurons can transmit and process information in parallel. Second, they use short and low-voltage pulses at low operating frequencies (10-100 Hz) for communication [68, 130] resulting in high resource-efficiency [137, 138]. Finally, another factor of efficiency may lie in the vast heterogeneity and diversification of brain circuit elements [131]. Synapses and neurons continuously adapt over time via *learning* [135]. In particular, the strength of each synapse, the *synaptic weight*, can be tuned to facilitate or prevent the transmission of APs. During learning, synaptic weights are constantly adjusted to respond to specific cognitive tasks.

1.3.1.2 The different generations of neural networks

Three different generations of neural networks can be distinguished. The first generation is based on the seminal works of Rosenblatt [139] on the *perceptron*.

1.3. THE THIRD GENERATION OF NEURAL NETWORKS: SPIKING NEURAL NETWORK

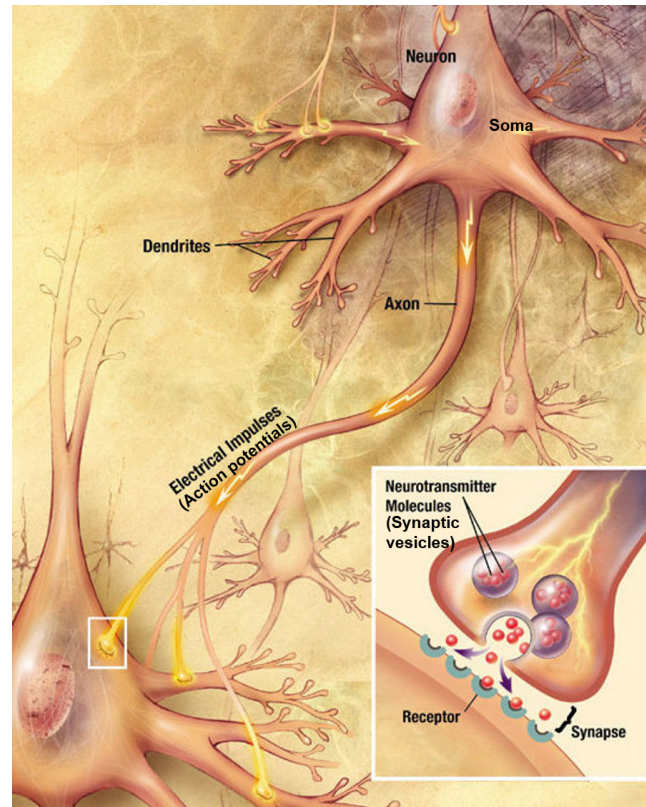


FIGURE 1.3.1: Drawing of two connected neurons. A neuron is mainly composed of a soma, several dendrites, and an axon. Neurons transmit electrical signal events (action potentials) along their axon connected to other neuron dendrites. Axon terminals connect to other neuron dendrites through terminal buttons forming synapses.

Perceptrons are based on McCulloch-Pitts neurons [2] and act as all-or-none elements. They output a boolean value depending on if their inputs reach a certain threshold value or not. The second generation is based on so-called *Artificial Neural Networks* (ANNs). ANNs are composed of a collection of computing units, the *artificial neurons*, interconnected by *weighted synapses*. Unlike perceptrons, artificial neurons can provide a continuous set of possible output values by applying an *activation function* to the weighted sum of their inputs. This second generation of neural networks have provided solutions to many artificial intelligence applications, such as pattern recognition, natural language processing, forecasting and prediction, or speech recognition [12, 15, 140, 141]. As illustrations of the success of ANNs we can cite the facial recognition system DeepFace by Facebook [142] and AlphaGo by Google DeepMind [126] that defeated one of the best human professional players in the full-size game of Go, something that was thought not to be possible before at least another decade. More recently, Google DeepMind introduced their program AlphaStar that reached the rank of Grandmaster at the real-time strategy game StarCraft II - the highest league above 99.8% of officially ranked human players [128]. This is highly promising for applications requiring real-time decisions, such as self-driving cars or robotics.

Although ANNs have been introduced to provide more bio-plausible neural networks with respect to perceptrons, they eventually deviated from biology to

focus on performance instead of power efficiency [143, 144]. For instance, training AlphaStar [128] required the use of 384 third-generation Tensor Processing Units for 44 days. By contrast, *Spiking Neural Networks* (SNNs) aim to reproduce biological brains in a closer manner with the promise of achieving high energy-efficiency systems [16, 145–150]. This gives rise to the third generation of neural networks [151]. As in biological brains, SNNs mainly rely on the exchange of spikes between neurons - the action potentials - that are transmitted along weighted synapses. Information is encoded in the timing and spiking rate of spikes [152–155]. For neuronal processing, SNNs employ spiking neurons, also called *integrate-and-fire neurons* (IF neurons) [148, 156]. IF neurons sum input spikes - *integrate* - whose amplitudes are modulated by synaptic weights, and they emit a spike - *fire* - when the summation goes above a threshold level. In the scope of this PhD thesis, we will focus on SNNs. The rest of this section will present the main building blocks of hardware SNN systems.

1.3.2 Information coding and network routing

1.3.2.1 The address event representation communication protocol

In Spiking Neural Networks (SNNs) neurons communicate using *spikes*. The information can be encoded in the time of occurrence of spikes, time difference between consecutive spikes, or spiking rates [152–155]. From a hardware implementation point of view, the Address Event Representation (AER) [157, 158] has been proposed as an efficient communication protocol for SNNs based on time-multiplexing. A "brute force" approach to transmit spikes between neurons would be to use one wire for each pair of neurons, *i.e.* N wires for N pairs of neurons [159]. In the AER protocol each neuron is assigned an address that is encoded as a digital word. When a neuron fires a spike, also called an *event*, its address is sent across a shared data bus to a receiver circuit using asynchronous digital circuits [22, 129, 130, 145, 147, 160–162]. The receiver decodes the address and transmits an event to every neuron paired with the spiking neuron. The AER allows to reduce the number of wires from N to $\approx \log_2(N)$ [159]. Note that a handshake protocol is required to ensure that only one address is transmitted in the shared digital bus at a time. Therefore, it is crucial that all events can be processed and transmitted quick enough to prevent routing congestion [163, 164].

1.3.2.2 Impact of network topology

In biological brains part of neural computation lies in the network topology, *i.e.* the connectivity scheme between neurons (the *connectome*) [165–167]. For instance, neurons have the ability to extend their neurites, *i.e.* dendrites and axons, to find appropriate synaptic partners [136]. The growth of neurites is not random, and neurites seek for particular targets on a basis of trials-and-errors [168]. Another example is the synaptic pruning process that mainly occurs in the early childhood and puberty [169]. Synaptic pruning consists in the elimination of synapses and may be an energy-saving process wherein redundant

synapses are eliminated. For neuromorphic systems, it has been shown that the choice of network topology affects network outcomes, performance, and energy [170–173]. Therefore, it is crucial that neuromorphic processors are not bound to a fixed topology and can adapt their synaptic connections to a specific task [163, 164]. The use of synaptic Lookup Tables (LUTs) associated with the AER protocol is an efficient method to enable network topology reconfigurability [129, 145, 159, 161, 163]. Synaptic LUTs store the addresses of pre-synaptic neurons, and they map them with the addresses of paired post-synaptic neurons. When a neuron spikes, its address is searched inside the synaptic LUTs. This allows to retrieve the addresses of post-synaptic neuron it is virtually connected to. As a result, the network topology can be modified by simply reprogramming synaptic LUTs. However, routing congestion can appear if the time to process an event, in particular the time to search an address in the LUTs, is longer than the time between two consecutive events.

FIGURE 1.3.2 shows the three main neural network architectures: (i) Fully-Connected Neural Network (FCNN), (ii) Convolutional Neural Network (CNN), and (iii) Recurrent Neural Network (RNN). FCNN is the simplest topology wherein every neuron of a layer is connected to every neuron of the next layer. It is mainly used for tasks such as classification or detection [174, 175]. CNNs are designed to process data that come in the form of multiple arrays, for example a colour image composed of three two-dimensional arrays containing pixel intensities in the three colour channels [15]. Neurons in a convolutional layer are organised in *feature maps* wherein each neuron is connected to a small subset of neurons, *receptive field*, of the previous layer [176]. Convolutional layers require fewer synapses than fully-connected layers, and they have the advantage to be insensitive to the spatial location of specific features in the inputs. For instance, if a CNN is sensitive to specific motifs from input images, it can detect them whatever their spatial location in the inputs. CNNs are mainly used for visual processing like face recognition and have achieved so far the highest performance in image classification [177]. The third main architecture is the RNN wherein feedback loops are included inside the network topology. This allows to store an input information, while processing new inputs. For this reason, RNNs are often used in tasks that involve sequential inputs, such as speech and language recognition.

CHAPTER 3 focuses on the implementation of synaptic LUTs for SNNs, and more details are provided in this chapter.

1.3.3 Hardware spiking neuron: the leaky integrate-and-fire neuron model

A canonical neuron model used in SNNs is the *Leaky Integrate-and-Fire* (LIF) neuron [156]. As biological neurons, LIF neurons rely on the integration of synaptic input currents and fire a spike when the integration value reaches a certain threshold [178, 179]. From a system point of view, LIF neurons receive input currents from *excitatory or inhibitory synapses*. If the synapse is excitatory, the current is positive. Otherwise, it is negative. The stronger a synapse, *i.e.* the higher its synaptic weight, the higher the current absolute value. These

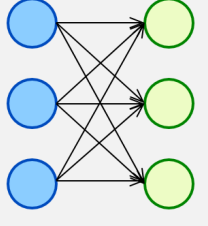
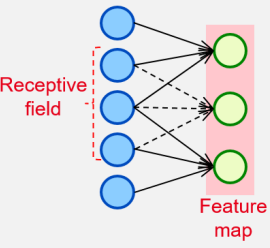
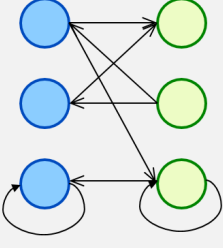
	Fully-Connected Neural Network (FCNN)	Convolutional Neural Network (CNN)	Recurrent Neural Network (RNN)
			
Major Applications	Simple classification Handwritten character recognition, ...	Visual processing Face recognition, image classification, ...	Sequential data Processing, transition, speech recognition, ...
Number of Layers	3-10 layers	5-100 layers	3-5 layers

FIGURE 1.3.2: The three main neural network architectures: Fully-Connected Neural Network (FCNN), Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN). Adapted from [12].

input currents are then integrated by LIF neurons. That is, LIF neurons can be modelled by an internal state variable, X , that evolves according to a first-order differential equation [16]:

$$\tau_{leak} \frac{dX}{dt} + X = I_{input} \quad (1.3.1)$$

wherein X represents the integrated current value, τ_{leak} is an integration time constant, and I_{input} is the input synaptic current. Upon receiving input synaptic currents, X increases (or decreases if the synapse is inhibitory). Between two integrations, X exponentially decreases with a time constant τ_{leak} . When X reaches a certain current threshold value, I_{th} , the neuron emits a spike, and X is reset to zero. After emitting a spike, LIF neurons are unable to integrate any input synaptic current for a *refractory period*. *Lateral inhibition* can also be implemented: when a LIF neuron spikes, it also inhibits neighbouring neurons from integrating input currents for a certain duration t_{inhib} . This allows to implement winner-take-all systems to prevents different neurons from being selective to similar features [174, 180].

FIGURE 1.3.3 (a) shows a simple LIF neuron circuit originally proposed in [181]. The capacitance C_{mem} , referred to as the membrane capacitance, models the membrane of a biological neuron [148]. FIGURE 1.3.3 (b) illustrates the evolution of the membrane capacitance potential, V_{mem} , during the generation of an action potential. Upon being fed by excitatory input currents, I_{in} , C_{mem} charges up (*integration*). Respectively, inhibitory currents (not shown) remove charges from C_{mem} . In the absence of input currents, C_{mem} discharges to its resting potential (ground in this case) through *leakage currents controlled by the gate voltage V_{lk} and a time constant τ_{leak} dependent on C_{mem} value*. V_{mem} is compared to a threshold voltage, V_{thr} , using a basic transconductance amplifier [148]. When V_{mem} exceeds V_{thr} , an action potential is generated in a similar

way as in biological neurons: an increase in sodium conductance (modelled by I_{Na}) creates the up-swing of the spike, and a delayed increase in potassium conductance (modelled by I_K) creates the down-swing. More precisely, when V_{mem} exceeds V_{thr} , the output of the first inverter goes low and activates the transistor M3. This charges up C_{mem} through I_{Na} and pulls up V_{mem} . At the same time, the second inverter charges up the capacitance C_K at a speed controlled by the current I_{Kup} . As soon as the voltage on C_K is high enough to activate the transistor M2, the potassium current, I_K , pulls down the voltage V_{mem} . *The current I_{Kup} controls the spike width.* Finally, C_K is discharged through the current I_{Kdn} . As long as the voltage on C_K is high enough, C_{mem} cannot integrate any incoming input current. That is, *I_{Kdn} controls the refractory period of the neuron.*

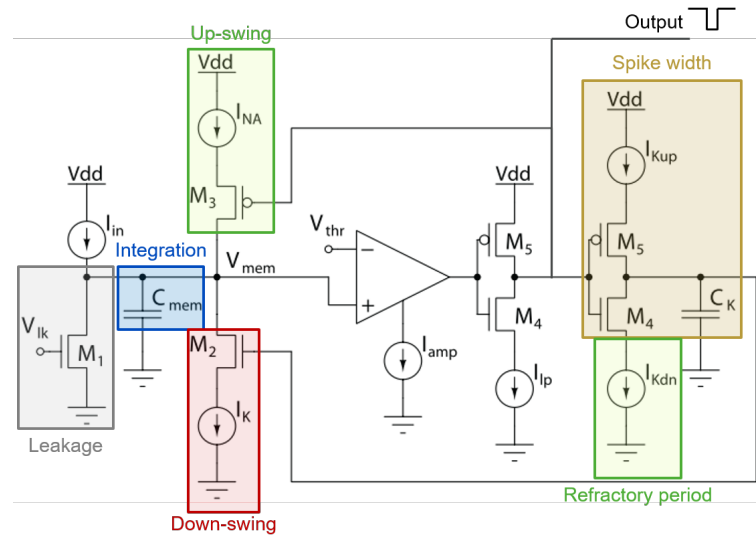
Many silicon hardware implementation of LIF neurons have been proposed based on standard VLSI CMOS technology [148, 182–184]. However, they usually require the use of big capacitors ($\approx pF$) to achieve biological time constants τ_{leak} (≈ 10 ms) which consumes significant silicon area. Recent works have sought to propose novel neuron implementations with better area-efficiency, in particular with the use of new non-volatile memories presented in SECTION 1.2.1 [146, 185–187] or new materials like Mott insulators [188, 189]. This is out of the scope of this work. As synapses outnumber neurons in the human brain by several orders of magnitude (10^{14} vs 10^{11} , respectively), most of the efforts in the literature have been focused on efficient designs of synaptic circuits. This is discussed in the next section.

1.3.4 Hardware synapse implementation

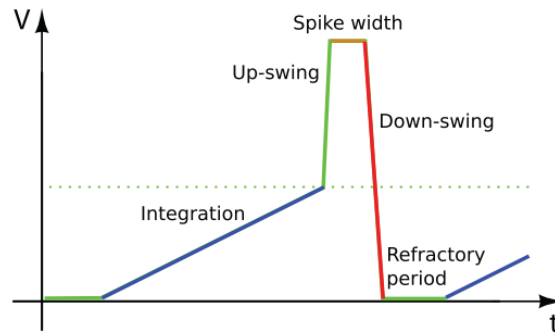
In the following, we will refer to the neuron transmitting spikes along a synapse as the input neuron or *pre-synaptic neuron*, and to the neuron receiving spikes as the output neuron or *post-synaptic neuron*.

1.3.4.1 Synaptic plasticity and learning rule

It is now known that all chemical synapses, the most abundant types of biological synapses, are plastic: their strength, or *synaptic weights*, undergo experience-dependent changes during brain development and over time as a result of learning [136]. These changes persist on time scales ranging from milliseconds to minutes for the *short-term plasticity*, and up to days or years for the *long-term plasticity*. Short-term plasticity is mainly driven by spiking activity of pre-synaptic neurons, whereas long-term plasticity can be a function of both pre- and post-synaptic neurons activities [136]. Short-term plasticity can be subdivided into Short-Term Facilitation (STF) and Short-Term Depression (STD). STF is a transient increase in synaptic weight that results in more neurotransmitters released at each successive action potential. On the other hand, STD is a transient decrease in synaptic weight due to the progressive depletion of the pool of synaptic vesicles available at each action potential. Long-term plasticity can be distinguished between (i) *Long-Term Potentiation* (LTP) wherein synaptic weights undergo a long-lasting increase, and (ii) *Long-Term Depression* (LTD) wherein synaptic



(a)



(b)

FIGURE 1.3.3: (a) Example of a bio-inspired silicon-based Leaky Integrate-and-Fire (LIF) neuron circuit from [181]. (b) Evolution of the membrane capacitance potential, V_{mem} , during the generation of an action potential. Reproduced from [148].

weights undergo a long-lasting decrease. LTP and LTD are broad terms that only define the direction of change in synaptic efficacy, and the different cellular and molecular mechanisms involved in these changes will not be reviewed here. Because of its long-lasting nature, long-term plasticity is widely believed to be responsible of learning and memory.

In neuromorphic systems, the network internal parameters - in particular synaptic weights - are modified during the learning phase, or training phase, following a *learning rule*. Learning rules can be classified into three main paradigms: (i) supervised learning, (ii) unsupervised learning, and (iii) reinforcement learning [15, 190]. In *supervised learning*, the network is trained using labelled input data (*e.g.* dog pictures are associated with the label "dog", cat pictures with the label "cat", ...) and/or using an external teacher. A common supervised learning rule used to train feed-forward artificial neural networks is the gradient-descent back-propagation algorithm [191], wherein synaptic weights are adjusted at each iteration in order to minimise the error rate between the actual and correct output. By contrast, *unsupervised learning* does not need training data to be

labelled; there is no notion of correct or incorrect output. The typical task of such machine learning algorithms consists in identifying similarities between the inputs and organising them based on these similarities [174, 192–195]. This is a considerable advantage over supervised algorithms since today’s big data applications often come with large volume of unlabelled and unstructured data [14]. In *reinforcement learning*, training data are not presented but collected by an agent in an environment so as to maximise some notion of reward [128, 196]. In this work, we will only focus on *unsupervised learning paradigms*.

An exemplary bio-inspired unsupervised learning paradigm suitable for training SNNs is the so-called *Spike-Timing-Dependent Plasticity* (STDP) [174, 192–195, 197]. The most known STDP rule is the long-term plasticity induced by *pairs of pre- and post-synaptic spikes* [132, 198] which was first experimentally observed in 1998 by Bi and Poo [133] (reported in FIGURE 1.3.4). The change in synaptic weight depends on the difference between the spike timing of a pre- and a post-synaptic neuron: if the post-synaptic neuron spikes shortly after the pre-synaptic neuron (within a time window of about 100 ms), the synaptic weight increases (*Long-Term Potentiation* (LTP) event, right-hand side of FIGURE 1.3.4). This facilitates the transmission of future spikes. Otherwise, the synaptic weight decreases (*Long-Term Depression* (LTD) event, left-hand side of FIGURE 1.3.4). The shorter the time difference, the higher the synaptic weight change in accordance with Hebb’s postulate [132]: ”Neurons that fire together, wire together”. This form of STDP is often referred to as the *Hebbian STDP*. Other forms of STDP have been observed in biology, such as the anti-Hebbian STDP wherein LTP events are induced by post-synaptic neurons firing before pre-synaptic neurons, and vice-versa, the symmetric Hebbian STDP, or symmetric anti-Hebbian STDP [135, 199]. However, these forms of *pair-based STDP* fail to replicate recent *triplet-based STDP* experiments (*i.e.* two pre-synaptic spikes and one post-synaptic spike, or two post-synaptic spike and one pre-synaptic spike to induce synaptic changes) [200]. Spike-Rate-Dependent Plasticity (SRDP) [201, 202] is another example of unsupervised learning paradigm wherein LTP is induced when the pre-synaptic neuron fires with a high frequency (20-100 Hz), while LTD is induced for low-frequency spiking (1-5 Hz). In this work, we will focus on long-term plasticity based on the simple *Hebbian STDP rule*.

1.3.4.2 Overview of synaptic implementations

A plethora of artificial synapses has been reported in the literature [12, 68, 203–207]. Traditional synaptic designs made use of conventional CMOS technology. A typical CMOS-based synapse is the Differential-Pair Integrator (DPI) synapse proposed in [203] and depicted in FIGURE 1.3.5. The DPI synapse features bio-inspired synaptic properties, such as tunable synaptic weights and realistic synaptic dynamics, and it can be integrated in VLSI spike-based neural systems as recently demonstrated on the Dynamic Neuromorphic Asynchronous Processors (DYNAPs) chip [145]. However, this implementation consumes valuable silicon area due to the use of several MOSFET transistors for one synapse - four n-FETs, two p-FETs, and one capacitor [145]. In addition, synaptic parameters are often stored in centralised volatile memories which increases power consumption [91]. This is detrimental for the implementation of neuromorphic processors with the

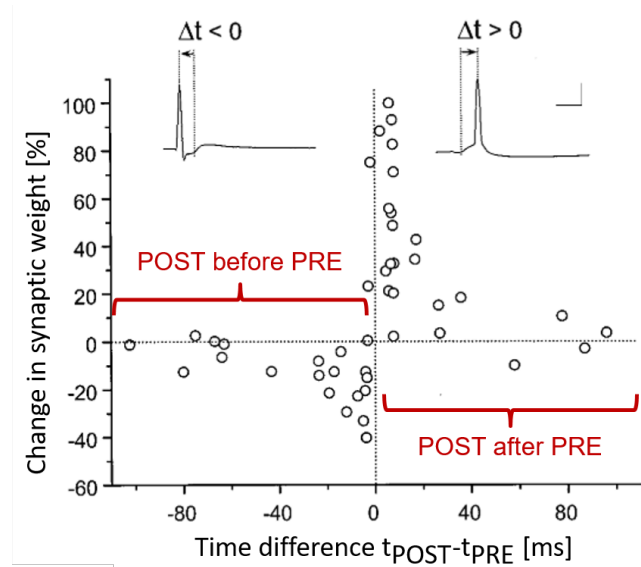


FIGURE 1.3.4: Experimental Spike-Timing Dependent Plasticity (STDP) observed by Bi and Poo [133]. If a post-synaptic neuron spikes shortly after a pre-synaptic neuron within a time window of about 100 ms (right-hand side), the synaptic weight increases. Otherwise, the synaptic weight decreases (left-hand side). Adapted from [133].

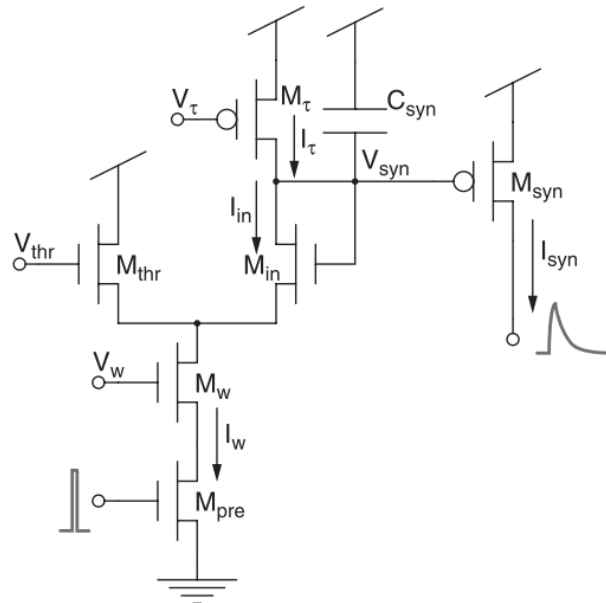


FIGURE 1.3.5: Differential-Pair Integrator (DPI) synapse. Reproduced from [203].

same granularity as biological human brains (more than 10^{14} synapses). New non-volatile resistance-based memories presented in SECTION 1.2.1 - encompassing the presented Resistive Memory (RRAM), Phase-Change Memory (PCM), and Spin-Torque-Transfer Magnetic Memory (STT-MRAM), and will be referred to as *memristors* in the following - have been foreseen as suitable candidates over the last decade to emulate artificial synapses thanks to their many common properties with biological synapses. They are two-terminals devices, scalable to sizes similar to biological synaptic clefts (≈ 20 nm), and can store synaptic weights

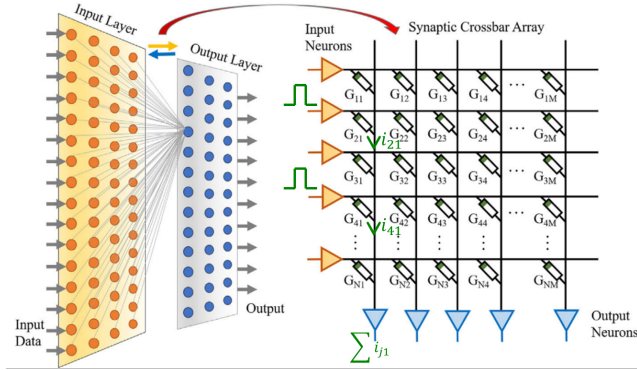


FIGURE 1.3.6: Schematic illustration of a vector-matrix multiplication performed by a memristor crossbar array in a single read cycle. Multiplication operations are performed on each memory element by Ohm’s law, while accumulate operations are performed on every column or row by Kirchhoff’s law. Reproduced from [207].

in their conductance value. In particular, they can adjust their conductance value over time with low-power programming pulses and retain them thanks to their non-volatility property. This facilitates the implementation of local, long-term synaptic plasticity algorithms, just like the STDP rule [16, 192–195, 197, 208, 209]. Finally, memristors can be integrated into crossbar arrays which is promising to implement hardware accelerators during both training and inference [87, 205, 210–214]. In both steps, hardware optimised for matrix multiplication, such as Graphical or Tensor Processing Units (GPUs and TPUs) [126, 128], or Application-Specific Integrated Circuits (ASICs) [215–218], are required due to the large number of Multiply-And-Accumulate (MAC) operations performed between the weights of the network and the input data. By contrast, memristor crossbar arrays can efficiently perform parallel MAC operations wherein multiplication operations are performed directly on memory devices at every cross-point by Ohm’s law, and the resulting currents are accumulated along rows or columns with Kirchhoff’s law (FIGURE 1.3.6) [87].

Many synaptic implementations based on single memristor devices have been reported. One of the earliest demonstrations of RRAM-based synapses was implemented with a nanoscale Ag/Si-based active layer (see FIGURE 1.3.7 (a)) [219]. As biological synapses, the resistance value could be gradually tuned in an analog fashion by controlling the motion of Ag ions in the active layer upon the application of voltage pulses (*cf* FIGURE 1.3.7 (b)). In addition, STDP-based learning capability was demonstrated by interconnecting two CMOS-based LIF neurons with the nanoscale RRAM. A Time-Division Multiplexing (TDM) approach was used to capture the spike timing difference between the two neurons and map it to the width of a pulse to be applied on the synaptic device. The results are reported in FIGURE 1.3.7 (c) and proved the capability of RRAM technology to implement electronic synapses capable of STDP learning. However, the use of the TDM approach increases circuit complexity. Another approach to implement STDP was demonstrated by Yu et al. [220] in a TiN/HfO_x/AlO_x/Pt RRAM stack based on a direct overlap scheme of pre- and post-synaptic spikes. For this purpose, the authors in [220] showed that the resistance value could be controlled by the amplitude of programming pulses. Then, pre- and post-synaptic

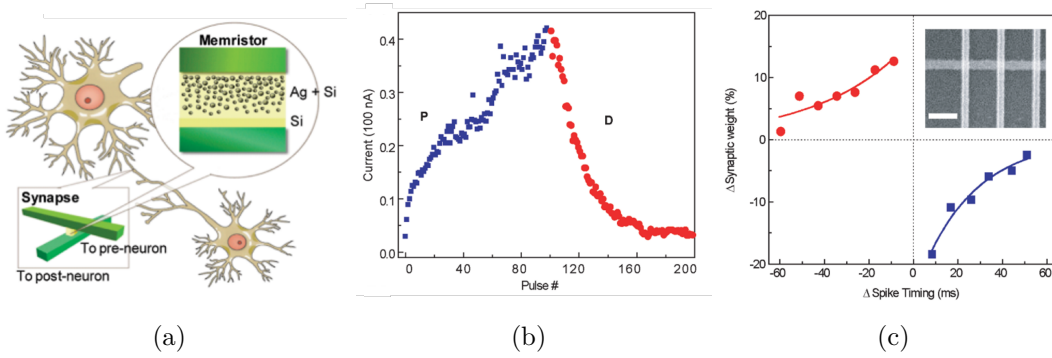


FIGURE 1.3.7: (a) Schematic illustration of a Resistive Memory (RRAM) used as a synapse between a pre- and post-synaptic neuron. (b) Conductance response (represented by the read current at 1 V) of a Ag/Si-based active layer RRAM after a series of 100 identical potentiation pulses (3.2 V for 300 μ s) followed by 100 identical depression pulses (-2.8 V for 300 μ s). (c) Experimental demonstration of Spike-Timing-Dependent Plasticity (STDP) measured on the Ag/Si-based RRAM. Timing difference between the pre- and post-synaptic neuron, Δ Spike Timing, was captured and mapped with a time-division multiplexing approach. Reproduced from [219].

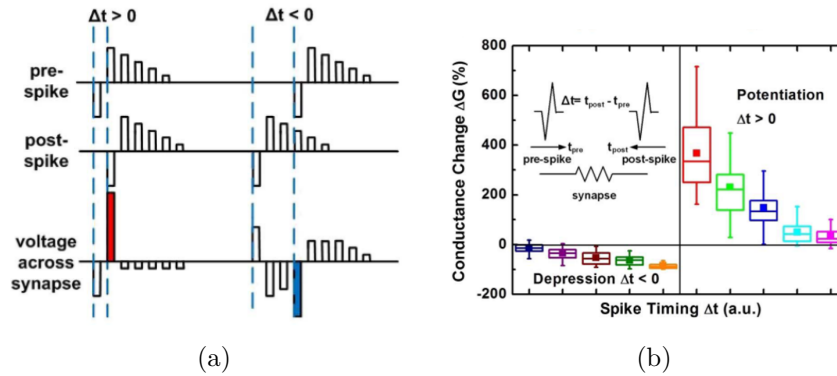


FIGURE 1.3.8: (a) Pre- and post-synaptic spike sequences to enable Spike-Timing-Dependent Plasticity (STDP) with pulse amplitude modulation. Pulse amplitudes are: -1.4 V, 1 V, 0.9 V, 0.8 V, 0.7 V, and 0.6 V (pre-synaptic spikes); -1 V, 1.4 V, 1.3 V, 1.2 V, 1.1 V, and 1 V (post-synaptic spikes). (b) Experimental demonstration of STDP measured on a TiN/HfO_x/AlO_x/Pt RRAM stack using the previous STDP scheme. Reproduced from [220].

spikes were carefully designed via sequences of single pulses with decreasing amplitude (*cf* FIGURE 1.3.8 (a)) such that only the direct overlap of pre- and post-spikes induces a conductance change in the RRAM device as demonstrated in FIGURE 1.3.8 (b). Similar approaches with PCM technology have also been reported [67].

The use of single memristor devices to implement artificial synapses provides good opportunities to build extremely dense neuromorphic circuits. However, it also results in serious issues, such as sneak-path currents in crossbar arrays and the risk to degrade memristor devices due to the lack of current limiters [43, 78, 86, 221]. Therefore, hybrid CMOS/memristor structures have been proposed in the so-

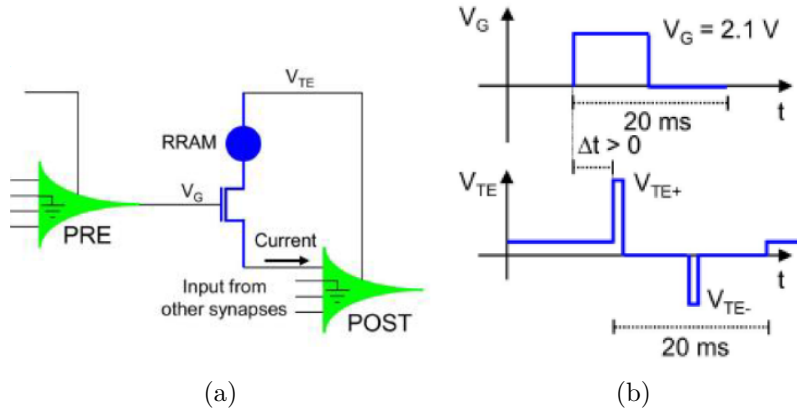


FIGURE 1.3.9: (a) Example of a one-transistor/one-RRAM (1T1R) synapse connecting a pre- and post-synaptic neuron. The 1T1R synapse is activated via the transistor gate only at pre-synaptic pulses. (b) Pre- (top) and post-synaptic (bottom) pulses to enable Spike-Timing-Dependent Plasticity (STDP) learning in the 1T1R synapse. Only the overlap between pre- and post-synaptic pulses induces an increase ($\Delta t > 0$) or decrease ($\Delta t < 0$) in conductance of the RRAM. Reproduced from [222].

called one-transistor/one-resistor (1T1R) structure [87, 192, 193, 208, 209, 222–227]. FIGURE 1.3.9 (a) shows an example of 1T1R synapse based on a MOSFET in series with a TiN/HFO_x/TiN RRAM cell connecting a pre- and post-synaptic neuron [222]. The pre-synaptic neuron biases the gate of the 1T1R synapse, while the post-synaptic neuron biases the Top Electrode (TE) voltage of the RRAM, V_{TE} . V_{TE} is set at a low constant voltage that induces a current proportional to the synaptic weight - *i.e.* RRAM conductance - across the 1T1R synapse whenever a pre-synaptic spike occurs. This current is integrated by the post-synaptic neuron. When the post-synaptic neuron fires a spike, it emits a short (1 ms) positive pulse followed after 10 ms by a short (1 ms) negative pulse applied on the RRAM TE as illustrated in FIGURE 1.3.9 (b). Pre- and post-synaptic spikes are designed so that only their overlap within ± 10 ms produces a conductance change: if the time difference is positive, the positive post-synaptic pulse is applied on the RRAM TE which sets the device to its Low Resistance State (LRS, potentiation event). Otherwise, the post-synaptic negative pulse is applied which resets the device to its High Resistance State (HRS, depression event). Numerous RRAM-based 1T1R synapse designs capable of STDP learning have been studied [87, 192, 193, 208, 209, 222–227]. Note that most of RRAM technologies are binary, *i.e.* they only switch between their two distinct states, LRS and HRS, after a potentiation or depression event, respectively. This can be detrimental for network performance as analog modulation is often necessary. This will be discussed more in details in CHAPTER 2. Many other hybrid synaptic structures have been reported [105, 228–230], for instance 2T1R synapses [231, 232] that provide more flexibility in the emulation of biological process thanks to the use of an additional transistor, and will not be reviewed.

The downside of the overlapping scheme to emulate STDP learning is that it increases circuit complexity and degrades data throughput since it relies on the overlapping of long pre- and post-synaptic pulses [201, 233, 234]. To overcome

this issue, other synaptic implementations that are not based on overlapping pulses have been proposed, such as the differential synapse proposed in [234] composed of twenty transistors and two memristors. Although the differential synapse is way less dense than classic 1T1R synapses, it can potentially scale better with large arrays of neurons and synapses since it allows to use all synapses in parallel and eliminates sneak-path issues. Another possibility is the use of second-order memristors [235–237] - another class of memristors. Unlike first-order memristors, their conductance change is also governed by second-order state variables like the internal temperature in $\text{Ta}_2\text{O}_{5-x}$ -based cells presented in [236]. Applications of pre- or post-synaptic spikes on $\text{Ta}_2\text{O}_{5-x}$ -based synapses result in a transient increase in the internal temperature that decays in time. This naturally provides an internal timing mechanism. Thus, post-synaptic spikes induce LTP events only if synaptic devices have first been heated by pre-synaptic pulses. Respectively, LTD events are induced at pre-synaptic spikes only if synaptic devices have first been heated by a post-synaptic pulse.

Many other three-terminals synaptic devices have also been reported [238–244] and will not be reviewed here. Other learning rules have been demonstrated with memristors, such as the widely studied supervised back-propagation algorithms [141, 175, 245, 246], triplet-based STDP [247], short-term plasticity rules [224, 248–251], and spike-rate-dependent plasticity [252], and will not be analysed in this section neither. In the scope of this work we will focus on *RRAM-based 1T1R synapses with unsupervised STDP learning*. As explained in SECTION 1.2.1.3, RRAMs pose major challenges for memory applications, namely their low memory window associated with rather low programming endurance, and their high resistance variability. However, the impact of these problems for neuromorphic applications is still obscure and will be analysed in CHAPTER 2. In particular, we will *clarify the role of synaptic variability arising from RRAM resistance variability*. Another issue is the intrinsic binary nature of RRAM technology that can degrade performance of neuromorphic systems [16, 175, 194, 195, 241, 245, 246, 253–256]. This will also be studied in CHAPTER 2.

1.3.5 Overview of fabricated neuromorphic processors

Many fabricated spiking neural network processors have been reported in the literature [22, 129, 130, 145, 161, 162]. TABLE 1.1 summarises the main features of each neuromorphic processor. While these works are excellent proofs-of-concept of neuromorphic system capabilities, there is still plenty of room for improvement from a technological, design, and circuit point of view. First, synapses are implemented fully in CMOS technology, such as the DPI synapse [145, 203], or with digital circuits associated with SRAM [22, 129, 161, 162]. Such implementations consume valuable silicon real-estate. Second, network parameters are often stored in centralised memories - generally in SRAM or DRAM [22, 129, 130, 161]. This does not truly eliminate the Von Neumann bottleneck as network parameters need to be read and transferred from a digital bus during computation. Third, this leads to static power consumption since SRAM and DRAM are volatile. Fourth, network topology needs to be

reconfigurable. The reported processors store the network topology inside Synaptic Routing Tables (SRTs) based either on SRAM or DRAM [22, 130, 161, 162], or Content-Addressable Memories (CAMs) [129, 145] to enable network reconfigurability. It has been demonstrated that the latter solution based on CAMs is more efficient to prevent routing congestion of spiking events and to relax constraints on the maximum number of neurons and synapses inside each neuromorphic core [163] thanks to the fast, parallel search capability of CAMs [257–260]. However, CAM-based SRTs are typically implemented with conventional SRAM-based CAM structures like in DYNAPs [145], and chip area can be saved with the use of RRAM-based CAM structures [261–266]. For instance, CAM circuits in DYNAPs consume more silicon area (31.7%) than both neuron and synapse circuits (22.8%). Finally, the processors do not always permit on-line training, and appropriate learning schemes and circuits are still to be developed.

	SpiNNaker [129]	TrueNorth [22]	Neurogrid [130]	HiAER [161]	Loihi [162]	DYNAPs [145]
Technology	130 nm	28 nm	180 nm	130 nm	14 nm	180 nm
Neurons/core	~100	256	64 kbit	16 kbit	1 kbit	256
Synapses/core	~10 ⁵	64 kbit	N/A	~1 kbit	N/A	16 kbit
Neuron Type	Digital CMOS	Digital CMOS	Mixed analog/ Digital CMOS	Digital CMOS	Digital CMOS	Mixed analog/ Digital CMOS
Synapse Type	Digital CMOS	Digital CMOS	Analog CMOS	Digital CMOS	Digital CMOS	Analog CMOS (DPI synapse)
Routing Scheme	SRT (CAM-based)	SRT (SRAM-based)	SRT (SRAM-based)	SRT (stored in off-chip DRAM)	SRT (SRAM-based)	SRT (CAM-based integrated within neurons)
Learning	On-line	Off-line	Off-line	Off-line	On-line	On-line

SRT=Synaptic Routing Table

TABLE 1.1: Summary of reported silicon-proven multi-core spiking neuromorphic processors [22, 129, 130, 145, 161, 162].

1.4 Goal of this PhD thesis

The main objective of this PhD thesis is *the use of resistive memories and three-dimensional monolithic technologies to enable the hardware implementation of bio-inspired reconfigurable Spiking Neural Networks (SNNs)*. Recently, multi-core SNN processors have been demonstrated [22, 129, 130, 145, 147, 161, 162], yet there is still plenty of room for improvement as presented in the previous section, especially in terms of *performance, energy-efficiency, and silicon area consumption*. To this aim, optimisations can be achieved on the major building blocks of SNN cores, namely:

- Synaptic arrays with adjustable synaptic weights
- Synaptic routing tables for on-the-fly network topology reconfigurability
- Neuron circuits with adjustable parameters

- On-line learning circuitry

Resistive Memories (RRAMs) have been seen as suitable candidates for the implementation of area- and energy-efficient SNN processors. However, there are still significant roadblocks that prevent their integration into large memory arrays for standard memory applications, namely their *low resistance ratio and high resistance variability*. Yet RRAM requirements for SNN processors are still unclear, and comprehensive studies of the impact of RRAM electrical properties and variability on SNN performance and reliability are still to be provided. In the framework of this study, we will focus on (i) *arrays of RRAMs to implement adjustable synaptic weights*, and (ii) *arrays of RRAM-based Ternary Content-Addressable Memories (TCAMs) to implement synaptic routing tables for on-the-fly network topology reconfigurability*. The very goal of this PhD thesis work is to *thoroughly evaluate the impact of RRAM electrical characteristics on these two building blocks and provide guidelines to optimise RRAM programming* by means of extensive electrical characterisations and simulations. In addition, we will open up perspectives to further improve SNN area-efficiency from a technological point of view by demonstrating the 3D monolithic co-integration of high-performance CMOS transistors with RRAM technology. It is also worth noting that all the results presented in this dissertation are not specifically bounded to RRAM technology, and the proposed guidelines to optimise SNN performance can be applied to any technology that can replace RRAM technology - such as phase-change memory, magnetic memory, ...

This dissertation is organised as follows:

Chapter 2: Role of synaptic variability in resistive memory-based spiking neural networks with unsupervised learning

In this chapter, we study the implementation of artificial synapses with RRAMs in SNNs trained with the unsupervised spike-timing-dependent plasticity learning paradigm. For this purpose, two canonical applications are simulated: (i) a detection application, and (ii) a character classification. We first present electrical characterisations measured on multi-kilobits RRAM arrays. Then, we evaluate the impact of RRAM electrical properties on SNN learning performance by means of system-level simulations calibrated on RRAM electrical characterisations. In particular, we clarify the role of synaptic variability - arising from RRAM cycle-to-cycle and device-to-device resistance variability.

Chapter 3: Synaptic routing reconfigurability of spiking neural network with resistive memory-based ternary content-addressable memory systems

In this chapter, we study the implementation of synaptic routing tables with RRAM-based Ternary Content-Addressable Memories (TCAMs). We first present extensive electrical characterisations performed on a RRAM-based TCAM circuit implementing the most common two-transistors/two-RRAMs (2T2R) bitcell structure. Then, we present a new RRAM-based TCAM bitcell in a one-transistor/two-RRAMs/one-transistor (1T2R1T) configuration featuring a similar silicon area to that of the previous 2T2R structure. The proposed 1T2R1T TCAM structure aims to overcome the main limitations of the most common 2T2R TCAM. Extensive electrical characterisations are performed

on a 1T2R1T TCAM circuit, and electrical results obtained on both TCAM circuits are compared. We finally propose design optimisation to improve TCAM performance and reliability.

Chapter 4: Three-dimensional monolithic integration of two layers of high-performance CMOS transistors with one layer of resistive memory devices

In this chapter, we demonstrate the full co-integration of two layers of CMOS transistors fabricated in a three-dimensional sequential (3D monolithic) technology with one layer of RRAM devices monolithically fabricated on top of the two layers of CMOS transistors. Devices have been fabricated in a conventional 65-nm Silicon On Insulator (SOI) CMOS over CMOS process. We first introduce the process flow of the integration. We then present electrical characterisations performed on the fabricated devices to demonstrate the functionality of the integration.

Chapter 5: Conclusion and perspectives

This chapter concludes the dissertation by summarising the main results presented in this PhD thesis work and by giving short perspectives for potential future works.

References: Introduction

- [1] John Von Neumann and Michael D. Godfrey. “First Draft of a Report on the EDVAC”. *IEEE Annals of the History of Computing*, 15(4):27–75, aug 1993. ISSN 10586180. doi: 10.1109/85.238389.
- [2] W. McCulloch and W. Pitts. “A logical calculus of the ideas immanent in nervous activity (reprinted from 1943)”. *Bulletin of Mathematical Biology*, 52(1/2):99–115, 1990.
- [3] M D Godfrey and D F Hendry. “The Computer as von Neumann Planned It”. *IEEE Annals of the History of Computing*, 15(1):11–21, 1993. ISSN 10586180. doi: 10.1109/85.194088.
- [4] J. Backus. “Can Programming Be Liberated from the von Neumann Style? A Functional Style and Its Algebra of Programs”. *Communications of the ACM*, 21(8):613–641, 1978. doi: 10.1145/359576.359579.
- [5] Carver Mead. “Neuromorphic Electronic Systems”. *Proceedings of the IEEE*, 78(10):1629–1636, 1990.
- [6] Thomas N Theis and H.-S Philip Wong. “The End of Moore’s Law: A New Beginning for Information Technology”. *Computing in Science Engineering*, 19(2):41–50, 2017. doi: 10.1109/MCSE.2017.29.
- [7] Igor L Markov. “Limits on fundamental limits to computation”. *Nature*, 512(7513):147–154, 2014. ISSN 14764687. doi: 10.1038/nature13570.
- [8] Massimiliano Di Ventra and Yuriy V. Pershin. “The parallel approach”. *Nature Physics*, 9(4):200–202, apr 2013. ISSN 17452481. doi: 10.1038/nphys2566.
- [9] H. S. Philip Wong and Sayeef Salahuddin. “Memory leads the way to better computing”. *Nature Nanotechnology*, 10(3):191–194, 2015. ISSN 17483395. doi: 10.1038/nnano.2015.29.
- [10] M. T. Bohr. “Interconnect scaling—the real limiter to high performance ULSI”. In *Proceedings of International Electron Devices Meeting*, pages 241–244. Institute of Electrical and Electronics Engineers (IEEE), 1995. doi: 10.1109/iedm.1995.499187.
- [11] Saber Moradi and Rajit Manohar. “The impact of on-chip communication on memory technologies for neuromorphic systems”. *Journal of Physics D: Applied Physics*, 52(1):1–25, 2019. ISSN 13616463. doi: 10.1088/1361-6463/aae641.
- [12] Hongsik Jeong and Luping Shi. “Memristor devices for neural networks”. *Journal of Physics D: Applied Physics*, 52(2):023003, 2019. ISSN 13616463. doi: 10.1088/1361-6463/aae223.
- [13] Yuan Taur, Douglas A. Buchanan, et al. “CMOS scaling into the nanometer regime”. *Proceedings of the IEEE*, 85(4):486–503, 1997. ISSN 00189219. doi: 10.1109/5.573737.

- [14] Yen Kuang Chen, Jatin Chhugani, et al. “Convergence of recognition, mining, and synthesis workloads and its implications”. *Proceedings of the IEEE*, 96(5):790–807, 2008. ISSN 00189219. doi: 10.1109/JPROC.2008.917729.
- [15] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. *Nature*, 521(7553):436–444, 2015. ISSN 14764687. doi: 10.1038/nature14539.
- [16] Damien Querlioz, Olivier Bichler, Adrien Francis Vincent, and Christian Gamrat. “Bioinspired Programming of Memory Devices for Implementing an Inference Engine”. *Proceedings of the IEEE*, 103(8):1398–1416, aug 2015. ISSN 00189219. doi: 10.1109/JPROC.2015.2437616.
- [17] J Wang, X Wang, et al. “14.2 A Compute SRAM with Bit-Serial Integer/Floating-Point Operations for Programmable In-Memory Vector Acceleration”. In *2019 IEEE International Solid - State Circuits Conference - (ISSCC)*, pages 224–226. IEEE, 2019. ISBN 9781538685310. doi: 10.1109/ISSCC.2019.8662419.
- [18] Carver Mead. *Analog VLSI and Neural Systems*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989. ISBN 0-201-05992-4.
- [19] M. M. Sabry Aly, M. Gao, et al. “Energy-Efficient Abundant-Data Computing: The N3XT 1,000x”. *Computer*, 48(12):24–33, 2015. doi: 10.1109/MC.2015.376.
- [20] T.J. Biggerstaff. “Moore’s law: Change or Die”. *IEEE Software*, 13(1):4, 1996. doi: 10.1109/MS.1996.476277.
- [21] Wolfgang Arden, Michel Brillouet, et al. “”More-than-Moore” (ITRS Whitepaper)”. Technical report, International Technology Roadmap for Semiconductors (ITRS). URL http://www.itrs2.net/uploads/4/9/7/7/49775221/irc-itrs-mtm-v2{_}3.pdf.
- [22] Paul A. Merolla, John V. Arthur, et al. “A million spiking-neuron integrated circuit with a scalable communication network and interface”. *Science*, 345(6197):668–673, 2014. ISSN 10959203. doi: 10.1126/science.1254642.
- [23] Gordon E. Moore. “Cramming more components onto integrated circuits”. *Electronics*, 38(8), 1965.
- [24] G. E. Moore. “Progress in Digital Integrated Electronics”. In *International Electron Devices Meeting*,, pages 11–13, 1975.
- [25] Gordon E. Moore. “Lithography and the future of Moore’s law”. In *Electron-Beam, X-Ray, EUV, and Ion-Beam Submicrometer Lithographies for Manufacturing V*, volume 2437, pages 2–17. SPIE, 1995. doi: 10.1117/12.209151.
- [26] Liming Xiu. “Time Moore: Exploiting Moore’s Law from the Perspective of Time”, dec 2019. ISSN 19430590.
- [27] Robert Dennard, Fritz Gaensslen, et al. “Design of ion-implanted MOSFET’s with very small physical dimensions”. *IEEE Journal of Solid State Circuits*, 9(5):259–268, 1974. ISSN 0018-9200. doi: 10.1109/JSSC.1974.1050511.
- [28] Mark T. Bohr and Ian A. Young. “CMOS Scaling Trends and beyond”. *IEEE Micro*, 37(6):20–29, 2017. ISSN 02721732. doi: 10.1109/MM.2017.4241347.
- [29] T Ghani, M Armstrong, et al. “A 90nm High Volume Manufacturing Logic Technology Featuring Novel 45nm Gate Length Strained Silicon CMOS Transistors”. In *Technical Digest - International Electron Devices Meeting*, pages 978–980, 2003. doi: 10.1109/iedm.2003.1269442.

- [30] K Mistry, C Allen, et al. “A 45nm Logic Technology with High-k+Metal Gate Transistors, Strained Silicon, 9 Cu Interconnect Layers, 193nm Dry Patterning, and 100% Pb-free Packaging”. In *2007 IEEE International Electron Devices Meeting*, pages 247–250, 2007. doi: 10.1109/IEDM.2007.4418914.
- [31] C Auth, C Allen, et al. “A 22nm high performance and low-power CMOS technology featuring fully-depleted tri-gate transistors, self-aligned contacts and high density MIM capacitors”. In *2012 Symposium on VLSI Technology (VLSIT)*, pages 131–132, 2012. ISBN 9781467308458. doi: 10.1109/VLSIT.2012.6242496.
- [32] S Natarajan, M Agostinelli, et al. “A 14nm logic technology featuring 2nd-generation FinFET, air-gapped interconnects, self-aligned double patterning and a 0.0588 um² SRAM cell size”. In *Technical Digest - International Electron Devices Meeting, IEDM*, volume 2015-Febru, pages 3.7.1–3.7.3, 2015. ISBN 9781479980017. doi: 10.1109/IEDM.2014.7046976.
- [33] S Narasimha, B Jagannathan, et al. “A 7nm CMOS technology platform for mobile and high performance compute application”. In *2017 IEEE International Electron Devices Meeting (IEDM)*, pages 29.5.1–29.5.4, 2017. doi: 10.1109/IEDM.2017.8268476.
- [34] Ian Cutress. “Intel’s Manufacturing Roadmap from 2019 to 2029 Back Porting, 7nm, 5nm, 3nm, 2nm, and 1”, 2019. URL <https://www.anandtech.com/show/15217/intels-manufacturing-roadmap-from-2019-to-2029>.
- [35] David Schor. “Samsung 5 nm and 4 nm Update WikiChip Fuse”, 2019. URL <https://fuse.wikichip.org/news/2823/samsung-5-nm-and-4-nm-update/>.
- [36] TSMC. “Future R&D Plans”, 2019. URL https://www.tsmc.com/english/dedicatedFoundry/technology/future{_}rd.htm.
- [37] WikiChip. “Technology Node”, 2019. URL https://en.wikichip.org/wiki/technology{_}node.
- [38] Victor V. Zhirnov, Ralph K. Cavin, James A. Hutchby, and George I. Bourianoff. “Limits to binary logic switch scaling - A gedanken model”. *Proceedings of the IEEE*, 91(11): 1934–1939, 2003. ISSN 00189219. doi: 10.1109/JPROC.2003.818324.
- [39] Dmitri B Strukov, Gregory S Snider, Duncan R Stewart, and R Stanley Williams. “The missing memristor found”. *Nature*, 453(7191):80–83, 2008. ISSN 1476-4687. doi: 10.1038/nature06932.
- [40] L Chua. “Memristor-The missing circuit element”. *IEEE Transactions on Circuit Theory*, 18(5):507–519, 1971. ISSN 2374-9555. doi: 10.1109/TCT.1971.1083337.
- [41] Ximeng Guan, Shimeng Yu, and H. S.Philip Wong. “On the switching parameter variation of metal-oxide RRAM - Part I: Physical modeling and simulation methodology”. *IEEE Transactions on Electron Devices*, 59(4):1172–1182, 2012. ISSN 00189383. doi: 10.1109/TED.2012.2184545.
- [42] P. Gonon, C. Vallee, C. Mannequin, M. Saadi, and .F Jomni. “Mechanisms of resistance switching in nanometric metal oxides and their dependence on electrodes”. In *Proceedings of the IEEE International Conference on Properties and Applications of Dielectric Materials*, volume 2015-Octob, pages 56–59, 2015. ISBN 9781479989034. doi: 10.1109/ICPADM.2015.7295207.
- [43] A. Grossi, E. Nowak, et al. “Fundamental variability limits of filament-based RRAM”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 4.7.1–4.7.4, 2016. ISBN 9781509039012. doi: 10.1109/IEDM.2016.7838348.
- [44] H. S.Philip Wong, Heng Yuan Lee, et al. “Metal-oxide RRAM”. *Proceedings of the IEEE*, 100(6):1951–1970, 2012. ISSN 00189219. doi: 10.1109/JPROC.2012.2190369.

- [45] Zizhen Jiang, Yi Wu, et al. “A Compact model for metal-oxide resistive random access memory with experiment verification”. *IEEE Transactions on Electron Devices*, 63(5): 1884–1892, 2016. ISSN 00189383. doi: 10.1109/TED.2016.2545412.
- [46] Daniele Ielmini. “Resistive switching memories based on metal oxides: Mechanisms, reliability and scaling”. *Semiconductor Science and Technology*, 31(6):063002, 2016. ISSN 13616641. doi: 10.1088/0268-1242/31/6/063002.
- [47] Y. Y. Chen, R. Roelofs, et al. “Tailoring switching and endurance / retention reliability characteristics of HfO₂ / Hf RRAM with Ti, Al, Si dopants”. In *Digest of Technical Papers - Symposium on VLSI Technology*, pages 1–2. IEEE, 2014. ISBN 9781479933310. doi: 10.1109/VLSIT.2014.6894403.
- [48] R. Degraeve, A. Fantini, et al. “Quantitative endurance failure model for filamentary RRAM”. In *Digest of Technical Papers - Symposium on VLSI Technology*, pages T188–T189. JSAP, 2015. ISBN 9784863485013. doi: 10.1109/VLSIT.2015.7223673.
- [49] C. Nail, G. Molas, et al. “Understanding RRAM endurance, retention and window margin trade-off using experimental results and simulations”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 4.5.1–4.5.4, 2016. ISBN 9781509039012. doi: 10.1109/IEDM.2016.7838346.
- [50] Caidie Cheng, Yiqing Li, et al. “Bipolar to unipolar mode transition and imitation of metaplasticity in oxide based memristors with enhanced ionic conductivity”. *Journal of Applied Physics*, 124(15), oct 2018. ISSN 10897550. doi: 10.1063/1.5037962.
- [51] L. Goux, Y. Y. Chen, et al. “On the gradual unipolar and bipolar resistive switching of TiN/HfO₂/Pt Memory Systems”. *Electrochemical and Solid-State Letters*, 13(6):11–14, 2010. ISSN 10990062. doi: 10.1149/1.3373529.
- [52] Deok Hwang Kwon, Kyung Min Kim, et al. “Atomic structure of conducting nanofilaments in TiO₂ resistive switching memory”. *Nature Nanotechnology*, 5(2):148–153, 2010. ISSN 17483395. doi: 10.1038/nnano.2009.456.
- [53] Yuchao Yang, Peng Gao, et al. “Observation of conducting filament growth in nanoscale resistive memories”. *Nature Communications*, 3(1):732, 2012. ISSN 20411723. doi: 10.1038/ncomms1737.
- [54] H.-S. P. Wong, C. Ahn, et al. “Stanford Memory Trends”, 2018. URL <https://nano.stanford.edu/stanford-memory-trends/>.
- [55] L Perniola, G Molas, et al. “Universal Signatures from Non-Universal Memories: Clues for the Future.”. In *2016 IEEE 8th International Memory Workshop, IMW 2016*, pages 1–3, 2016. ISBN 9781467388313. doi: 10.1109/IMW.2016.7495295.
- [56] Young Bae Kim, Seung Ryul Lee, et al. “Bi-layered RRAM with unlimited endurance and extremely uniform switching”. In *Digest of Technical Papers - Symposium on VLSI Technology*, pages 52–53. IEEE, 2011. ISBN 9784863481640.
- [57] Chung-wei Hsu, I-ting Wang, Chun-li Lo, Ming-chung Chiang, and Wen-yueh Jang. “Self-rectifying bipolar TaO_x/TiO₂ RRAM with superior endurance over 10¹² cycles for 3D high-density storage-class memory”. In *Digest of Technical Papers - Symposium on VLSI Technology*, pages T166–T167, 2013. ISBN 9784863483477.
- [58] K. Tsunoda, K. Kinoshita, et al. “Low power and high speed switching of Ti-doped NiO ReRAM under the unipolar voltage source of less than 3 V”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 767–770, 2007. doi: 10.1109/IEDM.2007.4419060.

- [59] H. Y. Lee, Y. S. Chen, et al. “Evidence and solution of over-RESET problem for HfOX based resistive memory with sub-ns switching speed and high endurance”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 19.7.1–19.7.4, 2010. ISBN 9781424474196. doi: 10.1109/IEDM.2010.5703395.
- [60] L. Goux, A. Fantini, et al. “Role of the Ta scavenger electrode in the excellent switching control and reliability of a scalable low-current operated TiN/Ta₂O₅/Ta RRAM device”. In *Digest of Technical Papers - Symposium on VLSI Technology*, pages 1–2. IEEE, 2014. ISBN 9781479933310. doi: 10.1109/VLSIT.2014.6894401.
- [61] Qing Luo, Xiaoxin Xu, et al. “Demonstration of 3D vertical RRAM with ultra low-leakage, high-selectivity and self-compliance memory cells”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 10.2.1–10.2.4, 2015. ISBN 9781467398930. doi: 10.1109/IEDM.2015.7409667.
- [62] Chung Wei Hsu, Chia Chen Wan, et al. “3D vertical TaOx/TiO₂ RRAM with over 103 self-rectifying ratio and sub-uA operating current”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 10.4.1–10.4.4, 2013. ISBN 9781479923076. doi: 10.1109/IEDM.2013.6724601.
- [63] X. P. Wang, Z. Fang, et al. “Highly compact 1T-1R architecture (4F²footprint) involving fully CMOS compatible vertical GAA nano-pillar transistors and oxide-based RRAM cells exhibiting excellent NVM properties and ultra-low power operation”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 20.6.1–20.6.4, 2012. ISBN 9781467348706. doi: 10.1109/IEDM.2012.6479082.
- [64] Wanki Kim, Sung Il Park, et al. “Forming-free nitrogen-doped AlOX RRAM with sub-uA programming current”. In *Digest of Technical Papers - Symposium on VLSI Technology*, pages 22–23. IEEE, 2011. ISBN 9784863481640.
- [65] L. Goux, A. Fantini, et al. “Ultralow sub-500nA operating current high-performance TiN/Al₂O₃/HfO₂/Hf/TiN bipolar RRAM achieved through understanding-based stack-engineering”. In *Digest of Technical Papers - Symposium on VLSI Technology*, pages 159–160, 2012. ISBN 9781467308458. doi: 10.1109/VLSIT.2012.6242510.
- [66] B. Govoreanu, G. S. Kar, et al. “10×10nm² Hf/HfOx crossbar resistive RAM with excellent performance, reliability and low-energy operation”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 31.6.1–31.6.4. IEEE, 2011. ISBN 9781457705052. doi: 10.1109/IEDM.2011.6131652.
- [67] Duygu Kuzum, Rakesh G.D. Jeyasingh, Byoungil Lee, and H. S. Philip Wong. “Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing”. *Nano Letters*, 12(5):2179–2186, 2012. ISSN 15306984. doi: 10.1021/nl201040y.
- [68] Duygu Kuzum, Shimeng Yu, and H. S. Philip Wong. “Synaptic electronics: Materials, devices and applications”. *Nanotechnology*, 24(38):382001, sep 2013. ISSN 09574484. doi: 10.1088/0957-4484/24/38/382001.
- [69] Matthias Wuttig and Noboru Yamada. “Phase-change materials for rewriteable data storage”. *Nature Materials*, 6(11):824–832, 2007. ISSN 14764660. doi: 10.1038/nmat2009.
- [70] H. S. Philip Wong, Simone Raoux, et al. “Phase change memory”. *Proceedings of the IEEE*, 98(12):2201–2227, 2010. ISSN 00189219. doi: 10.1109/JPROC.2010.2070050.
- [71] M. Suri, D. Garbin, et al. “Impact of PCM resistance-drift in neuromorphic systems and drift-mitigation strategy”. In *2013 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*, pages 140–145, 2013. doi: 10.1109/NanoArch.2013.6623059.
- [72] Feng Xiong, Myung-ho Bae, et al. “Self-Aligned Nanotube-Nanowire Phase Change Memory”. *Nano Letters*, 13(2):464–469, 2013. doi: 10.1021/nl3038097.

- [73] Chiyui Ahn, Zizhen Jiang, et al. “A 1TnR array architecture using a one-dimensional selection device”. In *Digest of Technical Papers - Symposium on VLSI Technology*, pages 1–2, 2014. ISBN 9781479933310. doi: 10.1109/VLSIT.2014.6894404.
- [74] L Tillie, E Nowak, et al. “Data retention extraction methodology for perpendicular STT-MRAM”. In *2016 IEEE International Electron Devices Meeting (IEDM)*, pages 27.3.1–27.3.4, 2016. doi: 10.1109/IEDM.2016.7838492.
- [75] J Zahurak, K Miyata, et al. “Process integration of a 27nm, 16Gb Cu ReRAM”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 6.2.1–6.2.4, 2014. doi: 10.1109/IEDM.2014.7046994.
- [76] C. Park, J. J. Kan, et al. “Systematic optimization of 1 Gbit perpendicular magnetic tunnel junction arrays for 28 nm embedded STT-MRAM and beyond”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 26.2.1–26.2.4, 2015. ISBN 9781467398930. doi: 10.1109/IEDM.2015.7409771.
- [77] Youngdon Choi, Ickhyun Song, et al. “A 20nm 1.8V 8Gb PRAM with 40MB/s program bandwidth”. In *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, pages 46–47, 2012. ISBN 9781467303736. doi: 10.1109/ISSCC.2012.6176872.
- [78] Tz Yi Liu, Tian Hong Yan, et al. “A 130.7-mm² 2-layer 32-gb reram memory device in 24-nm technology”. *IEEE Journal of Solid-State Circuits*, 49(1):140–153, jan 2013. ISSN 00189200. doi: 10.1109/JSSC.2013.2280296.
- [79] H Kim, J Lim, et al. “1GB/s 2Tb NAND flash multi-chip package with frequency-boosting interface chip”. In *2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers*, pages 1–3, 2015. doi: 10.1109/ISSCC.2015.7062964.
- [80] T Tanaka, M Helm, et al. “A 768Gb 3b/cell 3D-floating-gate NAND flash memory”. In *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 142–144, jan 2016. doi: 10.1109/ISSCC.2016.7417947.
- [81] Richard Fackenthal, Makoto Kitagawa, et al. “A 16Gb ReRAM with 200MB/s write and 1GB/s read in 27nm technology”. In *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, pages 338–339. IEEE, 2014. ISBN 9781479909186. doi: 10.1109/ISSCC.2014.6757460.
- [82] Intel. “Intel Optane Technology”. URL <https://www.intel.com/content/www/us/en/architecture-and-technology/intel-optane-technology.html>.
- [83] Panasonic. “Panasonic Starts World’s First Mass Production of ReRAM Mounted Microcomputers”, 2013. URL <https://news.panasonic.com/global/press/data/2013/07/en130730-2/en130730-2.html{\#}1>.
- [84] Avalanche-Technology. “Avalanche Technology Products”. URL <http://www.avalanche-technology.com/products/>.
- [85] Chinh Nguyen, Dona Burkard, Kelvin Dobbins, and Chuck Bohac. “MRAM Improvements to Automotive Non- Volatile Memory Storage”. Technical report, 2016. URL <https://www.everspin.com/technical-papers>.
- [86] A. Levisse, B. Giraud, J. P. Noel, M. Moreau, and J. M. Portal. “SneakPath compensation circuit for programming and read operations in RRAM-based CrossPoint architectures”. In *2015 15th Non-Volatile Memory Technology Symposium, NVMTS 2015*, 2015. ISBN 9781509021260. doi: 10.1109/NVMTS.2015.7457426.
- [87] Yue Zha, Etienne Nowak, and Jing Li. “Liquid Silicon: A Nonvolatile Fully Programmable Processing-In-Memory Processor with Monolithically Integrated ReRAM for Big Data/Machine Learning Applications”. In *IEEE Symposium on VLSI Circuits, Digest of Technical Papers*, pages C206–C207, 2019. ISBN 9784863487185. doi: 10.23919/VLSIC.2019.8778064.

-
- [88] Rainer Waser and Masakazu Aono. “Nanoionics-based resistive switching memories”. *Nature Materials*, 6(11):1833–840, 2007. ISSN 1476-4660. doi: 10.1038/nmat2023.
- [89] E. Vianello, O. Thomas, et al. “Back-end 3D integration of HfO₂-based RRAMs for low-voltage advanced IC digital design”. In *ICICDT 2013 - International Conference on IC Design and Technology, Proceedings*, pages 235–238. IEEE, 2013. ISBN 9781467347419. doi: 10.1109/ICICDT.2013.6563344.
- [90] E. Vianello, O. Thomas, et al. “Resistive Memories for Ultra-Low-Power embedded computing design”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 6.3.1–6.3.4, 2014. ISBN 9781479980017. doi: 10.1109/IEDM.2014.7046995.
- [91] Thomas Dalgaty, Melika Payvand, et al. “Hybrid neuromorphic circuits exploiting non-conventional properties of RRAM for massively parallel local plasticity mechanisms”. *APL Materials*, 7(8), 2019. ISSN 2166532X. doi: 10.1063/1.5108663.
- [92] Alessandro Grossi, Elisa Vianello, et al. “Resistive RAM endurance: Array-level characterization and correction techniques targeting deep learning applications”. *IEEE Transactions on Electron Devices*, 66(3):1281–1288, 2019. ISSN 00189383. doi: 10.1109/TED.2019.2894387.
- [93] A A Chien and V Karamcheti. “Moore’s Law: The First Ending and a New Beginning”. *Computer*, 46(12):48–53, 2013. ISSN 1558-0814. doi: 10.1109/MC.2013.431.
- [94] J. Sandrini, M. Barlas, et al. “OxRAM for embedded solutions on advanced node: scaling perspectives considering statistical reliability and design constraints”. In *IEEE International Electron Devices Meeting (IEDM)*, pages 30.5.1–30.5.4, 2019. ISBN 9781728140322. doi: 10.1109/iedm19573.2019.8993484.
- [95] Yoshifumi Nishi, Ulrich Bottger, Rainer Waser, and Stephan Menzel. “Crossover from Deterministic to Stochastic Nature of Resistive-Switching Statistics in a Tantalum Oxide Thin Film”. *IEEE Transactions on Electron Devices*, 65(10):4320–4325, oct 2018. ISSN 00189383. doi: 10.1109/TED.2018.2866127.
- [96] C. Cagli, G. Piccolboni, et al. “About the intrinsic resistance variability in HfO₂-based RRAM devices”. In *Joint International EUROSOL Workshop and International Conference on Ultimate Integration on Silicon-ULIS, EUROSOL-ULIS 2017 - Proceedings*, pages 31–34, 2017. ISBN 9781509053131. doi: 10.1109/ULIS.2017.7962593.
- [97] Christopher H Bennett, Diana Garland, Robin B Jacobs-Gedrim, Sapan Agarwal, and Matthew J Marinella. “Wafer-Scale TaOx Device Variability and Implications for Neuromorphic Computing Applications”. In *IEEE International Reliability Physics Symposium Proceedings*, pages 1–4, 2019. ISBN 9781538695043. doi: 10.1109/IRPS.2019.8720596.
- [98] M. Barlas, A. Grossi, et al. “Improvement of HfO₂ based RRAM array performances by local Si implantation”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 14.6.1–14.6.4, 2017. ISBN 9781538635599. doi: 10.1109/IEDM.2017.8268392.
- [99] A Bricalli, E Ambrosi, et al. “SiO_x-based resistive switching memory (RRAM) for crossbar storage/select elements with high on/off ratio”. In *2016 IEEE International Electron Devices Meeting (IEDM)*, pages 4.3.1–4.3.4, 2016. doi: 10.1109/IEDM.2016.7838344.
- [100] H. Y. Lee, P. S. Chen, et al. “Low power and high speed bipolar switching with a thin reactive ti buffer layer in robust HfO₂ based RRAM”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 1–4, 2008. ISBN 9781424423781. doi: 10.1109/IEDM.2008.4796677.

- [101] W. C. Chien, Y. R. Chen, et al. “A forming-free WO_xresistive memory using a novel self-aligned field enhancement feature with excellent reliability and scalability”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 19.2.1–19.2.4. IEEE, 2010. ISBN 9781424474196. doi: 10.1109/IEDM.2010.5703390.
- [102] B. Govoreanu, D. Crotti, et al. “A-VMCO: A novel forming-free, self-rectifying, analog memory cell with low-current operation, nonfilamentary switching and excellent variability”. In *Digest of Technical Papers - Symposium on VLSI Technology*, pages T132–T133. JSAP, 2015. ISBN 9784863485013. doi: 10.1109/VLSIT.2015.7223717.
- [103] D. Walczyk, Ch Walczyk, et al. “Resistive switching characteristics of CMOS embedded HfO₂-based 1T1R cells”. *Microelectronic Engineering*, 88(7):1133–1135, 2011. ISSN 01679317. doi: 10.1016/j.mee.2011.03.123.
- [104] Max M. Shulaker, Gage Hills, et al. “Three-dimensional integration of nanotechnologies for computing and data storage on a single chip”. *Nature*, 547(7661):74–78, 2017. ISSN 14764687. doi: 10.1038/nature22994.
- [105] Fabien Alibart, Elham Zamanidoost, and Dmitri B Strukov. “Pattern classification by memristive crossbar circuits using ex situ and in situ training”. *Nature Communications*, 4(1):2072, 2013. doi: 10.1038/ncomms3072.
- [106] Hyung Dong Lee, S. G. Kim, et al. “Integration of 4F2 selector-less crossbar array 2Mb ReRAM based on transition metal oxides for high density memory applications”. In *Digest of Technical Papers - Symposium on VLSI Technology*, pages 151–152. IEEE, 2012. ISBN 9781467308458. doi: 10.1109/VLSIT.2012.6242506.
- [107] G. Piccolboni, G. Molas, et al. “Investigation of HfO₂/Ti based vertical RRAM - Performances and variability”. In *2014 14th Annual Non-Volatile Memory Technology Symposium, NVMTS 2014*, 2014. ISBN 9781479942039. doi: 10.1109/NVMTS.2014.7060867.
- [108] Haitong Li, T F Wu, S Mitra, and H . P Wong. “Device-architecture co-design for hyperdimensional computing with 3d vertical resistive switching random access memory (3D VRRAM)”. In *2017 International Symposium on VLSI Technology, Systems and Application (VLSI-TSA)*, pages 1–2, 2017. doi: 10.1109/VLSI-TSA.2017.7942490.
- [109] S. Park, M. K. Yang, et al. “A non-linear ReRAM cell with sub-1uA ultralow operating current for high density vertical resistive memory (VRRAM)”. In *Technical Digest - International Electron Devices Meeting*, pages 20.8.1–20.8.4, 2012. ISBN 9781467348713. doi: 10.1109/IEDM.2012.6479084.
- [110] H Li, K Li, et al. “Four-layer 3D vertical RRAM integrated with FinFET as a versatile computing unit for brain-inspired cognitive information processing”. In *2016 IEEE Symposium on VLSI Technology*, pages 1–2, 2016. doi: 10.1109/VLSIT.2016.7573431.
- [111] Cao-Minh Lu. *Fabrication de CMOS à basse température pour l’intégration 3D séquentielle*. PhD thesis, Université Grenoble Alpes, 2017.
- [112] L. Brunet, P. Batude, et al. “First demonstration of a CMOS over CMOS 3D VLSI CoolCube™ integration on 300mm wafers”. In *Digest of Technical Papers - Symposium on VLSI Technology*, pages 1–2, 2016. ISBN 9781509006373. doi: 10.1109/VLSIT.2016.7573428.
- [113] Perrine Batude. *Intégration à trois dimensions séquentielle : Etude, fabrication et caractérisation*. PhD thesis, Institut National Polytechnique de Grenoble - INPG, 2009.
- [114] Jessy Micout. *Fabrication et caractérisation de transistor réalisée à basse température pour l’intégration 3D séquentielle*. PhD thesis, Université Grenoble Alpes, 2019.

-
- [115] C. Fenouillet-Beranger, B. Previtali, et al. “FDSOI bottom MOSFETs stability versus top transistor thermal budget featuring 3D monolithic integration”. In *European Solid-State Device Research Conference*, pages 110–113, 2014. ISBN 9781479943784. doi: 10.1109/ESSDERC.2014.6948770.
- [116] L. Pasini, P. Batude, et al. “High performance CMOS FDSOI devices activated at low temperature”. In *Digest of Technical Papers - Symposium on VLSI Technology*, pages 1–2, 2016. ISBN 9781509006373. doi: 10.1109/VLSIT.2016.7573407.
- [117] Soon Moon Jung, Hoon Lim, et al. “High speed and highly cost effective 72M bit density S3 SRAM technology with doubly stacked Si layers, peripheral only CoSix layers and Tungsten shunt W/L scheme for standalone and embedded memory”. In *Digest of Technical Papers - Symposium on VLSI Technology*, pages 82–83, 2007. ISBN 9784900784031. doi: 10.1109/VLSIT.2007.4339736.
- [118] Chang Hong Shen, Jia Min Shieh, et al. “Heterogeneously integrated sub-40nm low-power epi-like Ge/Si monolithic 3D-IC with stacked SiGeC ambient light harvester”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 3.6.1–3.6.4, 2014. ISBN 9781479980017. doi: 10.1109/IEDM.2014.7046975.
- [119] T. Irisawa, K. Ikeda, et al. “Demonstration of ultimate CMOS based on 3D stacked InGaAs-OI/SGOI wire channel MOSFETs with independent back gate”. In *Digest of Technical Papers - Symposium on VLSI Technology*, pages 1–2, 2014. ISBN 9781479933310. doi: 10.1109/VLSIT.2014.6894395.
- [120] V. Deshpande, V. Djara, et al. “Advanced 3D Monolithic hybrid CMOS with Sub-50 nm gate inverters featuring replacement metal gate (RMG)-InGaAs nFETs on SiGe-OI Fin pFETs”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 8.8.1–8.8.4, 2015. ISBN 9781467398930. doi: 10.1109/IEDM.2015.7409658.
- [121] M. M. Shulaker, K. Saraswat, H. . P. Wong, and S. Mitra. “Monolithic three-dimensional integration of carbon nanotube FETs with silicon CMOS”. In *Digest of Technical Papers - Symposium on VLSI Technology*, pages 1–2, 2014. ISBN 9780979806476. doi: 10.1017/S1431927607073709.
- [122] Ogun Turkyilmaz, Gerald Cibrario, Olivier Rozeau, Perrine Batude, and Fabien Clermidy. “3D FPGA using high-density interconnect Monolithic Integration”. In *Proceedings - Design, Automation and Test in Europe, DATE*, pages 1–4, 2014. ISBN 9783981537024. doi: 10.7873/DATE2014.351.
- [123] Daniel Gitlin, Maud Vinet, and Fabien Clermidy. “Cost model for monolithic 3D integrated circuits”. In *2016 SOI-3D-Subthreshold Microelectronics Technology Unified Conference, S3S 2016*, pages 1–2, 2016. ISBN 9781509043903. doi: 10.1109/S3S.2016.7804408.
- [124] W. Rhett Davis, Eun Chu Oh, Ambarish M. Sule, and Paul D. Franzon. “Application exploration for 3-D integrated circuits: TCAM, FIFO, and FFT case studies”. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 17(4):496–506, 2009. ISSN 10638210. doi: 10.1109/TVLSI.2008.2009352.
- [125] Rahul Sarpeshkar. *Efficient precise computation with noisy components : extrapolating from an electronic cochlea to the brain*. PhD thesis, California Institute of Technology, 1997.
- [126] David Silver, Aja Huang, et al. “Mastering the game of Go with deep neural networks and tree search”. *Nature*, 529(7587):484–489, 2016. ISSN 14764687. doi: 10.1038/nature16961.
- [127] D Silver, J Schrittwieser, K Simonyan, I Antonoglou Nature, and Undefined 2017. “Mastering the game of Go without human knowledge”. *Nature*, 550(7676):354, 2016.

- [128] Oriol Vinyals, Igor Babuschkin, et al. “Grandmaster level in StarCraft II using multi-agent reinforcement learning”. *Nature*, 575(7782):350–354, 2019. ISSN 14764687. doi: 10.1038/s41586-019-1724-z.
- [129] Steve B. Furber, Francesco Galluppi, Steve Temple, and Luis A. Plana. “The SpiNNaker project”. *Proceedings of the IEEE*, 102(5):652–665, 2014. ISSN 00189219. doi: 10.1109/JPROC.2014.2304638.
- [130] Ben Varkey Benjamin, Peiran Gao, et al. “Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations”. *Proceedings of the IEEE*, 102(5):699–716, 2014. ISSN 00189219. doi: 10.1109/JPROC.2014.2313565.
- [131] Vijay Balasubramanian. “Heterogeneity and Efficiency in the Brain”. *Proceedings of the IEEE*, 103(8):1346–1358, 2015. ISSN 00189219. doi: 10.1109/JPROC.2015.2447016.
- [132] D. O. Hebb. *The Organization of Behavior*. Wiley, New York, 1949. doi: 10.2307/1418888.
- [133] Guo Qiang Bi and Mu Ming Poo. “Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type”. *Journal of Neuroscience*, 18(24):10464–10472, 1998. ISSN 02706474. doi: 10.1523/jneurosci.18-24-10464.1998.
- [134] Guo-Qiang Bi and Mu-Ming Poo. “Synaptic Modification by Correlated Activity: Hebb’s Postulate Revisited”. *Annual Review of Neuroscience*, 24(1):139–166, 2001. ISSN 0147-006X. doi: 10.1146/annurev.neuro.24.1.139.
- [135] L. F. Abbott and Sacha B. Nelson. “Synaptic plasticity: Taming the beast”. *Nature Neuroscience*, 3(11s):1178–1183, 2000. ISSN 15461726. doi: 10.1038/81453.
- [136] Marc W Halterman. *Neuroscience, 3rd Edition*, volume 64. Wolters Kluwer Health, Inc. on behalf of the American Academy of Neurology, 2005. doi: 10.1212/01.WNL.0000154473.43364.47.
- [137] David Attwell and Simon B Laughlin. “An energy budget for signaling in the grey matter of the brain”. *Journal of Cerebral Blood Flow and Metabolism*, 21(10):1133–1145, 2001. ISSN 0271678X. doi: 10.1097/00004647-200110000-00001.
- [138] Fengyun Zhu, Rubin Wang, Xiaochuan Pan, and Zhenyu Zhu. “Energy expenditure computation of a single bursting neuron”. *Cognitive Neurodynamics*, 13(1):75–87, 2019. ISSN 18714099. doi: 10.1007/s11571-018-9503-3.
- [139] F. Rosenblatt. *The Perceptron, a Perceiving and Recognizing Automaton*. Cornell Aeronautical Laboratory, 1957.
- [140] Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. “Gradient-based learning applied to document recognition”. *Proceedings of the IEEE*, 86(11):2278–2323, 1998. ISSN 00189219. doi: 10.1109/5.726791.
- [141] Abu Sebastian, Tomas Tuma, et al. “Temporal correlation detection using computational phase-change memory”. *Nature Communications*, 8(1), 2017. ISSN 20411723. doi: 10.1038/s41467-017-01481-9.
- [142] Yaniv Taigman, Ming Yang, Marc’Aurelio Ranzato, and Lior Wolf. “DeepFace: Closing the gap to human-level performance in face verification”. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014. ISBN 9781479951178. doi: 10.1109/CVPR.2014.220.
- [143] Emma Strubell, Ananya Ganesh, and Andrew McCallum. “Energy and Policy Considerations for Deep Learning in NLP”. 2019. doi: 10.18653/v1/p19-1355.

-
- [144] John Naughton. “Can the planet really afford the exorbitant power demands of machine learning?”, 2019. URL <https://www.theguardian.com/commentisfree/2019/nov/16/can-planet-afford-exorbitant-power-demands-of-machine-learning>.
- [145] Saber Moradi, Ning Qiao, Fabio Stefanini, and Giacomo Indiveri. “A Scalable Multi-core Architecture with Heterogeneous Memory Structures for Dynamic Neuromorphic Asynchronous Processors (DYNAPs)”. *IEEE Transactions on Biomedical Circuits and Systems*, 12(1):106–122, feb 2018. ISSN 19324545. doi: 10.1109/TBCAS.2017.2759700.
- [146] Alice Mizrahi, Tifenn Hirtzlin, et al. “Neural-like computing with populations of superparamagnetic basis functions”. *Nature Communications*, 9(1), dec 2018. ISSN 20411723. doi: 10.1038/s41467-018-03963-w.
- [147] G. Indiveri, F. Corradi, and N. Qiao. “Neuromorphic architectures for spiking deep neural networks”. In *2015 IEEE International Electron Devices Meeting (IEDM)*, pages 4.2.1–4.2.4. IEEE, 2015. doi: 10.1109/IEDM.2015.7409623.
- [148] Giacomo Indiveri, Bernabe Linares-Barranco, et al. “Neuromorphic silicon neuron circuits”. *Frontiers in Neuroscience*, 5(73):1–23, 2011. ISSN 16624548. doi: 10.3389/fnins.2011.00073.
- [149] Sebastian Schmitt, Johann Klahn, et al. “Neuromorphic hardware in the loop: Training a deep spiking network on the BrainScaleS wafer-scale system”. *Proceedings of the International Joint Conference on Neural Networks*, 2017-May:2227–2234, 2017. doi: 10.1109/IJCNN.2017.7966125.
- [150] Kaushik Roy, Akhilesh Jaiswal, and Priyadarshini Panda. “Towards spike-based machine intelligence with neuromorphic computing”. *Nature*, 575(7784):607–617, 2019. ISSN 14764687. doi: 10.1038/s41586-019-1677-2. URL <http://dx.doi.org/10.1038/s41586-019-1677-2>.
- [151] Wolfgang Maass. “Networks of spiking neurons: The third generation of neural network models”. *Neural Networks*, 10(9):1659–1671, 1997. ISSN 08936080. doi: 10.1016/S0893-6080(97)00011-7.
- [152] Michael N. Shadlen and William T. Newsome. “Noise, neural codes and cortical organization”. *Current Opinion in Neurobiology*, 4(4):569–579, 1994. ISSN 09594388. doi: 10.1016/0959-4388(94)90059-0.
- [153] David Ferster and Nelson Spruston. “Cracking the neuronal code”. *Science*, 270(5237):756–757, 1995. ISSN 00368075. doi: 10.1126/science.270.5237.756.
- [154] J. J. Hopfield. “Pattern recognition computation using action potential timing for stimulus representation”. *Nature*, 376(6535):33–36, 1995. doi: 10.1038/376033a0.
- [155] Farzad Farkhooi, Eilif Muller, and Martin P Nawrot. “Adaptation reduces variability of the neuronal population code”. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 83(5):1–4, 2011. ISSN 15393755. doi: 10.1103/PhysRevE.83.050905.
- [156] Nicolas Brunel and Mark C.W. Van Rossum. “Lapicque’s 1907 paper: From frogs to integrate-and-fire”. *Biological Cybernetics*, 97(5-6):337–339, 2007. ISSN 03401200. doi: 10.1007/s00422-007-0190-0.
- [157] Steve Deiss, Rodney Douglas, Mike Fischer, Misha Mahowald, and Tony Matthews. “Address-Event Asynchronous Local Broadcast Protocol”, 1994. URL <https://www.ini.uzh.ch/~amw/scx/aeprotocol.html>.
- [158] Kwabena A. Boahen. “Point-to-point connectivity between neuromorphic chips using address events”. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 47(5):416–434, 2000. ISSN 10577130. doi: 10.1109/82.842110.

REFERENCES: INTRODUCTION

- [159] David H Goldberg, Gert Cauwenberghs, and Andreas G Andreou. “Analog VLSI spiking neural network with address domain probabilistic synapses”. In *ISCAS 2001. The 2001 IEEE International Symposium on Circuits and Systems (Cat. No.01CH37196)*, pages 241–244, 2001. ISBN 0780366859.
- [160] Wen Yu and Edgar N. Sanchez, editors. *Advances in Computational Intelligence*, volume 61. Springer-Verlag Berlin Heidelberg, 1 edition, 2009. doi: 10.1007/978-3-642-03156-4.
- [161] Jongkil Park, Theodore Yu, Siddharth Joshi, Christoph Maier, and Gert Cauwenberghs. “Hierarchical Address Event Routing for Reconfigurable Large-Scale Neuromorphic Systems”. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10):2408–2422, oct 2017. ISSN 21622388. doi: 10.1109/TNNLS.2016.2572164.
- [162] Mike Davies, Narayan Srinivasa, et al. “Loihi: A Neuromorphic Manycore Processor with On-Chip Learning”. *IEEE Micro*, 38(1):82–99, 2018. ISSN 02721732. doi: 10.1109/MM.2018.112130359.
- [163] Vladimir Kornijcuk, Jongkil Park, et al. “Reconfigurable Spike Routing Architectures for On-Chip Local Learning in Neuromorphic Systems”. *Advanced Materials Technologies*, 4(1):1800345, jan 2019. ISSN 2365709X. doi: 10.1002/admt.201800345.
- [164] Vladimir Kornijcuk and Doo Seok Jeong. “Recent Progress in Real-Time Adaptable Digital Neuromorphic Hardware”. *Advanced Intelligent Systems*, 1900030:1900030, 2019. ISSN 2640-4567. doi: 10.1002/aisy.201900030.
- [165] Marcus Kaiser. “A tutorial in connectome analysis : Topological and spatial features of brain networks”. *NeuroImage*, 57(3):892–907, 2011. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2011.05.025. URL <http://dx.doi.org/10.1016/j.neuroimage.2011.05.025>.
- [166] Juan Wang, Reza Khosrowabadi, et al. “Alterations in Brain Network Topology and Structural-Functional Connectome Coupling Relate to Cognitive Impairment”. *Frontiers in Aging Neuroscience*, 10(December):1–15, 2018. ISSN 1663-4365. doi: 10.3389/fnagi.2018.00404.
- [167] Fabrizio Damicelli, Claus C. Hilgetag, Marc-Thorsten Hutt, and Arnaud Messe. “Topological reinforcement as a principle of modularity emergence in brain networks”. *Network Neuroscience*, 3(2):589–605, 2019. doi: 10.1162/netn.a_00085.
- [168] Peter G.H. Clarke. “The limits of brain determinacy”. *Proceedings of the Royal Society B: Biological Sciences*, 279(1734):1665–1674, may 2012. ISSN 14712970. doi: 10.1098/rspb.2011.2629.
- [169] Gal Chechik, Isaac Meilijson, and Eytan Ruppin. “Synaptic Pruning in Development: A Computational Account”. *Neural Computation*, 10(7):1759–1777, 1998. ISSN 08997667. doi: 10.1162/089976698300017124.
- [170] Kenneth O. Stanley and Risto Miikkulainen. “Efficient evolution of neural network topologies”. In *Proceedings of the 2002 Congress on Evolutionary Computation, CEC 2002*, volume 2, pages 1757–1762, 2002. ISBN 0780372824. doi: 10.1109/CEC.2002.1004508.
- [171] Suojun Lu, Jian’an Fang, Aike Guo, and Yueqing Peng. “Impact of network topology on decision-making”. *Neural Networks*, 22(1):30–40, 2009. ISSN 08936080. doi: 10.1016/j.neunet.2008.09.012.
- [172] Longwen Huang, Yuwei Cui, Danke Zhang, and Si Wu. “Impact of noise structure and network topology on tracking speed of neural networks”. *Neural Networks*, 24(10):1110–1119, 2011. ISSN 08936080. doi: 10.1016/j.neunet.2011.05.018.

- [173] Mayra Z. Pimenta, Cesar Henrique Comin, Francisco A. Rodrigues, and Luciano Da F. Costa. “The impact of Interconnecting Topologies on SOM Neural Networks”. In *Proceedings of the International Joint Conference on Neural Networks*, pages 1–6. IEEE, 2018. ISBN 9781509060146. doi: 10.1109/IJCNN.2018.8489044.
- [174] Olivier Bichler, Damien Querlioz, Simon J. Thorpe, Jean Philippe Bourgoin, and Christian Gamrat. “Unsupervised features extraction from asynchronous silicon retina through spike-timing-dependent plasticity”. In *Proceedings of the International Joint Conference on Neural Networks*, pages 859–866. IEEE, 2011. ISBN 9781457710865. doi: 10.1109/IJCNN.2011.6033311.
- [175] Geoffrey W. Burr, Robert M. Shelby, et al. “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, jul 2015. ISSN 00189383. doi: 10.1109/TED.2015.2439635.
- [176] E Vianello, D Garbin, et al. “Multiple binary OxRAMs as synapses for convolutional neural networks”. In *Advances in Neuromorphic Hardware Exploiting Emerging Nanoscale Devices. Cognitive Systems Monographs*, volume 31, pages 109–127. Springer Verlag, 2017. doi: 10.1007/978-81-322-3703-7-6.
- [177] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. *Commun. ACM*, 60(6):84–90, 2017. doi: 10.1145/3065386.
- [178] A. N. Burkitt. “A review of the integrate-and-fire neuron model: I. Homogeneous synaptic input”. *Biological Cybernetics*, 95(1):1–19, 2006. ISSN 03401200. doi: 10.1007/s00422-006-0068-6.
- [179] A. N. Burkitt. “A review of the integrate-and-fire neuron model: II. Inhomogeneous synaptic input and network properties”. *Biological Cybernetics*, 95(2):97–112, 2006. ISSN 03401200. doi: 10.1007/s00422-006-0082-8.
- [180] S Yu. “Orientation classification by a winner-take-all network with oxide RRAM based synaptic devices”. In *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1058–1061, 2014. doi: 10.1109/ISCAS.2014.6865321.
- [181] A. Van Schaik. “Building blocks for electronic spiking neural networks”. *Neural Networks*, 14(6-7):617–628, 2001. ISSN 08936080. doi: 10.1016/S0893-6080(01)00067-3.
- [182] Ning Qiao and Giacomo Indiveri. “Scaling mixed-signal neuromorphic processors to 28 nm FD-SOI technologies”. *Proceedings - 2016 IEEE Biomedical Circuits and Systems Conference, BioCAS 2016*, pages 552–555, 2016. doi: 10.1109/BioCAS.2016.7833854.
- [183] Syed Ahmed Aamir, Paul Muller, Andreas Hartel, Johannes Schemmel, and Karlheinz Meier. “A highly tunable 65-nm CMOS LIF neuron for a large scale neuromorphic system”. In *European Solid-State Circuits Conference*, volume 2016-October, pages 71–74, 2016. ISBN 9781509029723. doi: 10.1109/ESSCIRC.2016.7598245.
- [184] Antoine Joubert, Bilel Belhadj, and Rodolphe Héliot. “A robust and compact 65 nm LIF analog neuron for computational purposes”. In *2011 IEEE 9th International New Circuits and Systems Conference, NEWCAS 2011*, pages 9–12. IEEE, 2011. ISBN 9781612841359. doi: 10.1109/NEWCAS.2011.5981206.
- [185] Tomas Tuma, Angeliki Pantazi, Manuel Le Gallo, Abu Sebastian, and Evangelos Eleftheriou. “Stochastic phase-change neurons”. *Nature Nanotechnology*, 11(8):693–699, 2016. ISSN 17483395. doi: 10.1038/nnano.2016.70.
- [186] Angeliki Pantazi, Stanisław Woźniak, Tomas Tuma, and Evangelos Eleftheriou. “All-memristive neuromorphic computing with level-tuned neurons”. *Nanotechnology*, 27(35):355205, 2016. doi: 10.1088/0957-4484/27/35/355205.

- [187] Zhongrui Wang, Saumil Joshi, et al. “Fully memristive neural networks for pattern classification with unsupervised learning”. *Nature Electronics*, 1(2):137–145, 2018. ISSN 25201131. doi: 10.1038/s41928-018-0023-2.
- [188] Matthew D Pickett, Gilberto Medeiros-Ribeiro, and R Stanley Williams. “A scalable neuristor built with Mott memristors”. *Nature Materials*, 12(2):114–117, 2013. ISSN 1476-4660. doi: 10.1038/nmat3510.
- [189] Coline Adda, Benoit Corraze, et al. “Mott insulators: A large class of materials for Leaky Integrate and Fire (LIF) artificial neuron”. *Journal of Applied Physics*, 124(15), 2018. ISSN 10897550. doi: 10.1063/1.5042756.
- [190] Johanni Brea and Wulfram Gerstner. “Does computational neuroscience need new synaptic learning paradigms?”. *Current Opinion in Behavioral Sciences*, 11:61–66, 2016. ISSN 23521546. doi: 10.1016/j.cobeha.2016.05.012.
- [191] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. “Learning representations by back-propagating errors”. *Nature*, 323(6088):533–536, 1986. ISSN 00280836. doi: 10.1038/323533a0.
- [192] Stefano Ambrogio, Nicola Ciocchini, et al. “Unsupervised learning by spike timing dependent plasticity in phase change memory (PCM) synapses”. *Frontiers in Neuroscience*, 10:1–12, 2016. ISSN 1662453X. doi: 10.3389/fnins.2016.00056.
- [193] Peter U. Diehl and Matthew Cook. “Unsupervised learning of digit recognition using spike-timing-dependent plasticity”. *Frontiers in Computational Neuroscience*, 9 (AUGUST), aug 2015. ISSN 16625188. doi: 10.3389/fncom.2015.00099.
- [194] Alexander Serb, Johannes Bill, et al. “Unsupervised learning in probabilistic neural networks with multi-state metal-oxide memristive synapses”. *Nature Communications*, 7, sep 2016. ISSN 20411723. doi: 10.1038/ncomms12611.
- [195] Erika Covi, Stefano Brivio, et al. “Analog memristive synapse in spiking networks implementing unsupervised learning”. *Frontiers in Neuroscience*, 10(OCT), oct 2016. ISSN 1662453X. doi: 10.3389/fnins.2016.00482.
- [196] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath. “Deep Reinforcement Learning: A Brief Survey”. *IEEE Signal Processing Magazine*, 34(6):26–38, 2017. doi: 10.1109/MSP.2017.2743240.
- [197] Konstantin Zarudnyi, Adnan Mehonic, et al. “Spike-timing dependent plasticity in unipolar silicon oxide RRAM devices”. *Frontiers in Neuroscience*, 12(FEB), feb 2018. ISSN 1662453X. doi: 10.3389/fnins.2018.00057.
- [198] Henry Markram, Joachim Lübke, Michael Frotscher, and Bert Sakmann. “Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs”. *Science*, 275(5297): 213–215, 1997. ISSN 00368075. doi: 10.1126/science.275.5297.213.
- [199] Daniel E. Feldman. “The Spike-Timing Dependence of Plasticity”, 2012. ISSN 08966273.
- [200] Jean Pascal Pfister and Wulfram Gerstner. “Triplets of spikes in a model of spike timing-dependent plasticity”. *Journal of Neuroscience*, 26(38):9673–9682, 2006. ISSN 02706474. doi: 10.1523/JNEUROSCI.1425-06.2006.
- [201] Joseph M. Brader, Walter Senn, and Stefano Fusi. “Learning real-world stimuli in a neural network with spike-driven synaptic dynamics”. *Neural Computation*, 19(11): 2881–2912, 2007. ISSN 08997667. doi: 10.1162/neco.2007.19.11.2881.
- [202] Guy Rachmuth, Harel Z Shouval, Mark F Bear, and Chi Sang Poon. “A biophysically-based neuromorphic model of spike rate- and timing-dependent plasticity”. *Proceedings of the National Academy of Sciences*, 108(49):E1266—E1274, 2011. ISSN 00278424. doi: 10.1073/pnas.1106161108.

-
- [203] Chiara Bartolozzi and Giacomo Indiveri. “Synaptic dynamics in analog VLSI”. *Neural Computation*, 19(10):2581–2603, 2007. ISSN 08997667. doi: 10.1162/neco.2007.19.10.2581.
- [204] Shimeng Yu. “Neuro-Inspired Computing with Emerging Nonvolatile Memory”. *Proceedings of the IEEE*, 106(2):260–285, feb 2018. ISSN 00189219. doi: 10.1109/JPROC.2018.2790840.
- [205] Changhyuck Sung, Hyunsang Hwang, and In Kyeong Yoo. “Perspective: A review on memristive hardware for neuromorphic computation”. *Journal of Applied Physics*, 124(15):151903, oct 2018. ISSN 10897550. doi: 10.1063/1.5037835.
- [206] Jingrui Wang and Fei Zhuge. “Memristive Synapses for Brain-Inspired Computing”. *Advanced Materials Technologies*, 4(3):1800544, mar 2019. ISSN 2365709X. doi: 10.1002/admt.201800544.
- [207] Teng Zhang, Ke Yang, et al. “Memristive Devices and Networks for Brain-Inspired Computing”. *Physica Status Solidi - Rapid Research Letters*, 13(8):1900029, 2019. ISSN 18626270. doi: 10.1002/pssr.201900029.
- [208] M. Suri, D. Querlioz, et al. “Bio-Inspired Stochastic Computing Using Binary CBRAM Synapses”. *IEEE Transactions on Electron Devices*, 60(7):2402–2409, 2013. doi: 10.1109/TED.2013.2263000.
- [209] T. Werner, D. Garbin, et al. “Real-time decoding of brain activity by embedded Spiking Neural Networks using OxRAM synapses”. In *IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 2318–2321, 2016. ISBN 9781479953417. doi: 10.1109/ISCAS.2016.7539048.
- [210] Geoffrey W Burr, Robert M Shelby, et al. “Neuromorphic computing using non-volatile memory”. *Advances in Physics: X*, 2(1):89–124, 2017. ISSN 23746149. doi: 10.1080/23746149.2016.1259585.
- [211] Hsinyu Tsai, Stefano Ambrogio, Pritish Narayanan, Robert M Shelby, and Geoffrey W Burr. “Recent progress in analog memory-based accelerators for deep learning”. *Journal of Physics D: Applied Physics*, 51(28):283001, 2018. doi: 10.1088/1361-6463/aac8a5.
- [212] A Fumarola, P Narayanan, et al. “Accelerating machine learning with Non-Volatile Memory: Exploring device and circuit tradeoffs”. In *2016 IEEE International Conference on Rebooting Computing (ICRC)*, pages 1–8, oct 2016. doi: 10.1109/ICRC.2016.7738684.
- [213] Robert Legenstein. “Nanoscale connections for brain-like circuits”. *Nature*, 521(7550):37–38, 2015. doi: 10.1038/521037a.
- [214] Reiji Mochida, Kazuyuki Kouno, et al. “A 4M synapses integrated analog ReRAM based 66.5 TOPS/W neural-network processor with cell current controlled writing and flexible network architecture”. *Digest of Technical Papers - Symposium on VLSI Technology*, 2018-June:175–176, 2018. ISSN 07431562. doi: 10.1109/VLSIT.2018.8510676.
- [215] Jaehyeong Sim, Jun Seok Park, et al. “A 1.42TOPS/W deep convolutional neural network recognition processor for intelligent IoE systems”. In *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, pages 264–265. IEEE, 2016. ISBN 9781467394666. doi: 10.1109/ISSCC.2016.7418008.
- [216] Bert Moons, Roel Uytterhoeven, Wim Dehaene, and Marian Verhelst. “Envision: A 0.26-to-10TOPS/W subword-parallel dynamic-voltage-accuracy-frequency-scalable Convolutional Neural Network processor in 28nm FDSOI”. In *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, volume 60, pages 246–247. IEEE, 2017. ISBN 9781509037575. doi: 10.1109/ISSCC.2017.7870353.

- [217] Giuseppe Desoli, Nitin Chawla, et al. “A 2.9TOPS/W deep convolutional neural network SoC in FD-SOI 28nm for intelligent embedded systems”. In *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, pages 238–239. IEEE, 2017. ISBN 9781509037575. doi: 10.1109/ISSCC.2017.7870349.
- [218] Yu Hsin Chen, Tushar Krishna, Joel S. Emer, and Vivienne Sze. “Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks”. *IEEE Journal of Solid-State Circuits*, 52(1):127–138, 2017. ISSN 00189200. doi: 10.1109/JSSC.2016.2616357.
- [219] Sung Hyun Jo, Ting Chang, et al. “Nanoscale memristor device as synapse in neuromorphic systems”. *Nano Letters*, 10(4):1297–1301, 2010. ISSN 15306984. doi: 10.1021/nl904092h.
- [220] Shimeng Yu, Yi Wu, Rakesh Jeyasingh, Duygu Kuzum, and H. S. Philip Wong. “An electronic synapse device based on metal oxide resistive switching memory for neuromorphic computation”. *IEEE Transactions on Electron Devices*, 58(8):2729–2737, 2011. ISSN 00189383. doi: 10.1109/TED.2011.2147791.
- [221] M Alayan, E Vianello, et al. “In-depth investigation of programming and reading operations in RRAM cells integrated with Ovonic Threshold Switching (OTS) selectors”. In *2017 IEEE International Electron Devices Meeting (IEDM)*, pages 2.3.1–2.3.4, 2017. doi: 10.1109/IEDM.2017.8268311.
- [222] S Ambrogio, S Balatti, et al. “Neuromorphic Learning and Recognition With One-Transistor-One-Resistor Synapses and Bistable Metal Oxide RRAM”. *IEEE Transactions on Electron Devices*, 63(4):1508–1515, 2016. ISSN 1557-9646. doi: 10.1109/TED.2016.2526647.
- [223] Daniele Garbin, Elisa Vianello, et al. “HfO₂-Based OxRAM Devices as Synapses for Convolutional Neural Networks”. *IEEE Transactions on Electron Devices*, 62(8):2494–2501, 2015. ISSN 00189383. doi: 10.1109/TED.2015.2440102.
- [224] T. Werner, E. Vianello, et al. “Experimental demonstration of short and long term synaptic plasticity using OxRAM multi k-bit arrays for reliable detection in highly noisy input data”. In *IEEE International Electron Devices Meeting (IEDM)*, pages 16.6.1–16.6.4, 2016. ISBN 9781509039029. doi: 10.1109/IEDM.2016.7838433.
- [225] E. Vianello, T. Werner, et al. “Resistive memories for spike-based neuromorphic circuits”. In *2017 IEEE 9th International Memory Workshop, IMW 2017*, pages 1–6, 2017. ISBN 9781509032723. doi: 10.1109/IMW.2017.7939100.
- [226] Giacomo Pedretti, Valerio Milo, et al. “Stochastic Learning in Neuromorphic Hardware via Spike Timing Dependent Plasticity with RRAM Synapses”. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 8(1):77–85, mar 2018. ISSN 21563357. doi: 10.1109/JETCAS.2017.2773124.
- [227] S Ambrogio, S Balatti, F Nardi, S Facchinetti, and D Ielmini. “Spike-timing dependent plasticity in a transistor-selected resistive switching memory”. *Nanotechnology*, 24(38):384012, sep 2013. doi: 10.1088/0957-4484/24/38/384012.
- [228] Chaoxing Wu, Tae Whan Kim, Hwan Young Choi, Dmitri B Strukov, and J Joshua Yang. “Flexible three-dimensional artificial synapse networks with correlated learning and trainable memory capability”. *Nature Communications*, 8(1), 2017. ISSN 20411723. doi: 10.1038/s41467-017-00803-1.
- [229] Valerio Milo, Giacomo Pedretti, et al. “A 4-Transistors/1-Resistor Hybrid Synapse Based on Resistive Switching Memory (RRAM) Capable of Spike-Rate-Dependent Plasticity (SRDP)”. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 26(12):2806–2815, 2018. ISSN 10638210. doi: 10.1109/TVLSI.2018.2818978.

-
- [230] Peng Yao, Huaqiang Wu, et al. “Face classification using electronic synapses”. *Nature Communications*, 8(May):1–8, 2017. ISSN 20411723. doi: 10.1038/ncomms15199.
- [231] Zhongqiang Wang, Stefano Ambrogio, Simone Balatti, and Daniele Ielmini. “A 2-transistor/1-resistor artificial synapse capable of communication and stochastic learning in neuromorphic systems”. *Frontiers in Neuroscience*, 9:1–11, 2015. ISSN 1662453X. doi: 10.3389/fnins.2014.00438.
- [232] S. Kim, M. Ishii, et al. “NVM neuromorphic core with 64k-cell (256-by-256) phase change memory synaptic array with on-chip neuron circuits for continuous in-situ learning”. In *2015 IEEE International Electron Devices Meeting (IEDM)*, pages 17.1.1–17.1.4. IEEE, 2015. ISBN 9781467398930. doi: 10.1109/IEDM.2015.7409716.
- [233] H Mostafa, C Mayr, and G Indiveri. “Beyond spike-timing dependent plasticity in memristor crossbar arrays”. In *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 926–929, 2016. doi: 10.1109/ISCAS.2016.7527393.
- [234] Manu V. Nair, Lorenz K. Muller, and Giacomo Indiveri. “A differential memristive synapse circuit for on-line learning in neuromorphic computing systems”. *Nano Futures*, 1(3), 2017. ISSN 23991984. doi: 10.1088/2399-1984/aa954a.
- [235] Yuriy V. Pershin and Massimiliano Di Ventra. “Neuromorphic, digital, and quantum computation with memory circuit elements”. *Proceedings of the IEEE*, 100(6):2071–2080, 2012. ISSN 00189219. doi: 10.1109/JPROC.2011.2166369.
- [236] Sungho Kim, Chao Du, et al. “Experimental demonstration of a second-order memristor and its ability to biorealistically implement synaptic plasticity”. *Nano Letters*, 15(3): 2203–2211, 2015. ISSN 15306992. doi: 10.1021/acs.nanolett.5b00697.
- [237] M. A. Zidan, Y. Jeong, and W. D. Lu. “Temporal Learning Using Second-Order Memristors”. *IEEE Transactions on Nanotechnology*, 16(4):721–723, 2017. doi: 10.1109/TNANO.2017.2710158.
- [238] S. Brink, S. Nease, and P. Hasler. “Computing with networks of spiking neurons on a biophysically motivated floating-gate based neuromorphic integrated circuit”. *Neural Networks*, 45:39–49, sep 2013. ISSN 08936080. doi: 10.1016/j.neunet.2013.02.011.
- [239] Sungho Kim, Jinsu Yoon, Hee Dong Kim, and Sung Jin Choi. “Carbon Nanotube Synaptic Transistor Network for Pattern Recognition”. *ACS Applied Materials and Interfaces*, 7(45):25479–25486, oct 2015. ISSN 19448252. doi: 10.1021/acsami.5b08541.
- [240] Sungmin Hwang, Hyungjin Kim, et al. “System-level simulation of hardware spiking neural network based on synaptic transistors and if neuron circuits”. *IEEE Electron Device Letters*, 39(9):1441–1444, sep 2018. ISSN 07413106. doi: 10.1109/LED.2018.2853635.
- [241] Suhwan Lim, Jong Ho Bae, et al. “Hardware-based Neural Networks using a Gated Schottky Diode as a Synapse Device”. In *Proceedings - IEEE International Symposium on Circuits and Systems*, volume 2018-May, 2018. ISBN 9781538648810. doi: 10.1109/ISCAS.2018.8351152.
- [242] Sung Yun Woo, Kyu Bong Choi, et al. “Synaptic device using a floating fin-body MOSFET with memory functionality for neural network”. *Solid-State Electronics*, 156: 23–27, jun 2019. ISSN 00381101. doi: 10.1016/j.sse.2019.02.011.
- [243] Matthew Jerry, Pai Yu Chen, et al. “Ferroelectric FET analog synapse for acceleration of deep neural network training”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 6.2.1–6.2.4, 2017. ISBN 9781538635599. doi: 10.1109/IEDM.2017.8268338.

- [244] Jianshi Tang, Douglas Bishop, et al. “ECRAM as Scalable Synaptic Cell for High-Speed, Low-Power Neuromorphic Computing”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 13.1.1–13.1.4. IEEE, 2018. ISBN 9781728119878. doi: 10.1109/IEDM.2018.8614551.
- [245] Alessandro Fumarola, Severin Sidler, et al. “Bidirectional Non-Filamentary RRAM as an Analog Neuromorphic Synapse, Part II: Impact of Al/Mo/Pr 0.7 Ca 0.3 MnO 3 Device Characteristics on Neural Network Training Accuracy”. *IEEE Journal of the Electron Devices Society*, 6(1):169–178, 2018. ISSN 21686734. doi: 10.1109/JEDS.2017.2782184.
- [246] Giorgio Cristiano, Massimo Giordano, et al. “Perspective on training fully connected networks with resistive memories: Device requirements for multiple conductances of varying significance”. *Journal of Applied Physics*, 124(15), oct 2018. ISSN 10897550. doi: 10.1063/1.5042462.
- [247] Mostafa Rahimi Azghadi, Bernabe Linares-Barranco, Derek Abbott, and Philip H.W. Leong. “A Hybrid CMOS-Memristor Neuromorphic Synapse”. *IEEE Transactions on Biomedical Circuits and Systems*, 11(2):434–445, 2017. ISSN 19324545. doi: 10.1109/TBCAS.2016.2618351.
- [248] Takeo Ohno, Tsuyoshi Hasegawa, et al. “Short-term plasticity and long-term potentiation mimicked in single inorganic synapses”. *Nature Materials*, 10(8):591–595, 2011. ISSN 14764660. doi: 10.1038/nmat3054.
- [249] Sizhao Li, Fei Zeng, et al. “Synaptic plasticity and learning behaviours mimicked through Ag interface movement in an Ag/conducting polymer/Ta memristive system”. *J. Mater. Chem. C*, 1(34):5292–5298, 2013. doi: 10.1039/C3TC30575A.
- [250] S. La Barbera, A. F. Vincent, D. Vuillaume, D. Querlioz, and F. Alibart. “Short-term to long-term plasticity transition in filamentary switching for memory applications”. In *2015 International Conference on Memristive Systems (MEMRISYS)*, pages 1–2, 2015. doi: 10.1109/MEMRISYS.2015.7378402.
- [251] C. H. Bennett, S. La Barbera, et al. “Exploiting the short-term to long-term plasticity transition in memristive nanodevice learning architectures”. In *2016 International Joint Conference on Neural Networks (IJCNN)*, volume 2016-October, pages 947–954, 2016. ISBN 9781509006199. doi: 10.1109/IJCNN.2016.7727300.
- [252] V. Milo, G. Pedretti, et al. “Demonstration of hybrid CMOS/RRAM neural networks with spike time/rate-dependent plasticity”. *Technical Digest - International Electron Devices Meeting, IEDM*, pages 16.8.1–16.8.4, 2017. ISSN 01631918. doi: 10.1109/IEDM.2016.7838435.
- [253] Johannes Bill and Robert Legenstein. “A compound memristive synapse model for statistical learning through STDP in spiking neural networks”. *Frontiers in Neuroscience*, 8(DEC), 2014. ISSN 1662453X. doi: 10.3389/fnins.2014.00412.
- [254] Robert M. Shelby, Geoffrey W. Burr, Irem Boybat, and Carmelo Di Nolfo. “Non-volatile memory as hardware synapse in neuromorphic computing: A first look at reliability issues”. In *IEEE International Reliability Physics Symposium Proceedings*, volume 2015-May, pages 6A11–6A16, 2015. ISBN 9781467373623. doi: 10.1109/IRPS.2015.7112755.
- [255] Sungho Kim, Meehyun Lim, Yeamin Kim, Hee Dong Kim, and Sung Jin Choi. “Impact of Synaptic Device Variations on Pattern Recognition Accuracy in a Hardware Neural Network”. *Scientific Reports*, 8(1), dec 2018. ISSN 20452322. doi: 10.1038/s41598-018-21057-x.
- [256] Irem Boybat, Cecilia Giovinazzo, et al. “Multi-ReRAM synapses for artificial neural network training”. In *Proceedings - IEEE International Symposium on Circuits and Systems*, volume 2019-May, pages 1–5, 2019. ISBN 9781728103976. doi: 10.1109/ISCAS.2019.8702714.

- [257] L. Chisvin and R. J. Duckworth. “Content-addressable and associative memory: alternatives to the ubiquitous RAM”. *Computer*, 22(7):51–64, 1989. doi: 10.1109/2.30732.
- [258] Bruce Gamache, Zachary Pfeffer, and Sunil P. Khatri. “A fast ternary CAM design for IP networking applications”. In *Proceedings - International Conference on Computer Communications and Networks, ICCCN*, pages 434–439. Institute of Electrical and Electronics Engineers Inc., 2003. ISBN 0780379454. doi: 10.1109/ICCCN.2003.1284205.
- [259] Anthony J. McAuley and Paul Francis. “Fast routing table lookup using CAMs”. In *Proceedings - IEEE INFOCOM*, volume 3, pages 1382–1891. Institute of Electrical and Electronics Engineers (IEEE), mar 1993. ISBN 0818635800. doi: 10.1109/infcom.1993.253403.
- [260] Kostas Pagiamtzis and Ali Sheikholeslami. “Content-addressable memory (CAM) circuits and architectures: A tutorial and survey”. *IEEE Journal of Solid-State Circuits*, 41(3): 712–727, mar 2006. ISSN 00189200. doi: 10.1109/JSSC.2005.864128.
- [261] Meng Fan Chang, Chien Chen Lin, et al. “A 3T1R Nonvolatile TCAM Using MLC ReRAM for Frequent-Off Instant-On Filters in IoT and Big-Data Processing”. *IEEE Journal of Solid-State Circuits*, 52(6):1664–1679, jun 2017. ISSN 00189200. doi: 10.1109/JSSC.2017.2681458.
- [262] Chien Chen Lin, Jui Yu Hung, et al. “A 256b-wordlength ReRAM-based TCAM with 1ns search-time and 14x improvement in wordlength-energyefficiency-density product using 2.5T1R cell”. In *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, pages 136–137. Institute of Electrical and Electronics Engineers Inc., feb 2016. ISBN 9781467394666. doi: 10.1109/ISSCC.2016.7417944.
- [263] Jing Li, Robert K. Montoye, Masatoshi Ishii, and Leland Chang. “1 Mb 0.41 μm^2 2T-2R cell nonvolatile TCAM with two-bit encoding and clocked self-referenced sensing”. *IEEE Journal of Solid-State Circuits*, 49(4):896–907, 2014. ISSN 00189200. doi: 10.1109/JSSC.2013.2292055.
- [264] Meng Fan Chang, Ching Hao Chuang, et al. “Designs of emerging memory based non-volatile TCAM for Internet-of-Things (IoT) and big-data processing: A 5T2R universal cell”. In *Proceedings - IEEE International Symposium on Circuits and Systems*, volume 2016-July, pages 1142–1145. Institute of Electrical and Electronics Engineers Inc., jul 2016. ISBN 9781479953400. doi: 10.1109/ISCAS.2016.7527447.
- [265] Meng Fan Chang, Lie Yue Huang, et al. “A ReRAM-Based 4T2R Nonvolatile TCAM Using RC-Filtered Stress-Decoupled Scheme for Frequent-OFF Instant-ON Search Engines Used in IoT and Big-Data Processing”. *IEEE Journal of Solid-State Circuits*, 51(11):2786–2789, nov 2016. ISSN 00189200. doi: 10.1109/JSSC.2016.2602218.
- [266] Cheol Kim, Sung Gi Ahn, Jisu Min, and Kee Won Kwon. “Power efficient and reliable nonvolatile TCAM with Hi-PFO and semi-complementary driver”. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 66(2):605–615, feb 2019. ISSN 15498328. doi: 10.1109/TCSI.2018.2867005.

Role of synaptic variability in resistive memory-based spiking neural networks with unsupervised learning

Contents

2.1	Introduction	56
2.1.1	Variability in biological brains	56
2.1.2	Synaptic variability in artificial spiking neural networks	56
2.1.3	Goal of this chapter	57
2.2	Binary devices	58
2.2.1	Experimental characterisation	58
2.2.2	Implications for a learning system: impact of binary RRAM-based synapse characteristics on the network performance	64
2.2.3	Conclusion	77
2.3	Analog devices	78
2.3.1	Goal of the section	78
2.3.2	Analog conductance modulation with non-volatile resistance-based memories	79
2.3.3	Learning rule and synapse behavioural model	81
2.3.4	Impact of the conductance response on spiking neural network learning performance	82
2.3.5	Discussion	86

2.1 Introduction

2.1.1 Variability in biological brains

VARIABILITY is ubiquitous in any computational system, and brains are no exception. Neurons and synapses - the fundamental computational units in brains - are noisy devices [1, 2] due to effects such as the stochastic nature of ion channels [2], time to replenish synaptic vesicle pools [3–5], background synaptic activity [1, 6, 7], jitter in spike timings of action potentials [1], or stochasticity in neurite (*i.e.* axons and dendrites) growth [8]. It has been shown that randomness in the brain leads to trial-to-trial variability in neurons: if a neuron is repeatedly driven with identical stimuli, its response varies from trial to trial. This can have an impact on the behaviour or decision-making [1, 7, 9, 10]. While it is clear that variability is present in the brain, its implications are still not entirely understood. Several studies have suggested that the brain may actually benefit from noise and variability [9–14]. As an example, up to ninety percent of presynaptic signals in the brain do not elicit postsynaptic signals (*synaptic failures* since synaptic vesicle pools are not readily releasable at any time and take as much as a few seconds to be fully replenished [3, 5, 15]). Yet the predominance of synaptic failures may have a functional role providing an energy-saving mechanism with less important spikes being filtered out and relevant spikes being transmitted [16]. Another interesting statement is that noise facilitates brains to explore more possible solutions to a specific problem [8, 13, 17, 18] which prevents them from being stuck in suboptimal solutions. This assumption is supported by the work of Rokni et al. [18] wherein the authors showed that different combinations of synaptic weights can produce the same output and that noise helps probe different synaptic configurations during learning.

2.1.2 Synaptic variability in artificial spiking neural networks

These biological considerations can have important implications in nanoelectronics as today multiple bio-inspired hardware architectures are being developed incorporating nanodevices. Many of these architectures encode neuron values as *spikes* [19, 20] - in so-called Spiking Neural Networks (SNNs) - which can lead to high energy-efficiencies. These architectures also incorporate the brain-inspired principle of learning, in the way the synaptic connections among neurons are created, modified, and preserved. Many works on neuromorphic architectures have already studied the implementation of electronic synapses [21–24] with technologies such as Complementary Metal-Oxide Semiconductor (CMOS) [25–28], carbon nanotubes [29, 30], magnetic memories [31–33], phase-change memories [34–38], or resistive memories [39–58]. Recently, efforts were rather focused on new non-volatile resistance-based memories such as Resistive Memories (RRAMs) to implement electronic synapses in artificial spiking neural networks thanks to their compatibility with advanced CMOS technology nodes [59–61]. In addition,

RRAMs resemble biological synapses as they are two-terminals devices, can encode synaptic weights in their conductance value, and have the ability to modulate their conductance state over time in the same way biological synapses modulate their synaptic weight during learning. They also benefit from many advantages such as non-volatility, fast switching speed, and high scalability [59, 62]. This enables their integration in dense arrays to connect many silicon-based neurons [26, 63–66]. Various approaches have been proposed to implement learning with RRAMs, such as the so-called supervised back-propagation algorithm [34, 47, 48, 56, 67–69] or the bio-inspired unsupervised Spike-Timing-Dependent Plasticity (STDP) learning rule [33, 37, 70–74].

Despite the numerous advantages listed above, RRAMs pose challenges as they suffer from many non-idealities that can have an impact on learning and inference performance. One drawback is the high conductance variability - both across cycles and devices - inducing *synaptic variability*, *i.e.* non-repeatable behaviours [60, 75–78]. It has been demonstrated that RRAM-based neural networks are intrinsically robust to synaptic variability - with supervised [25, 28, 29, 34, 35, 46, 47, 52, 79, 80] or unsupervised learning [33, 37, 41, 42, 45, 49, 53, 54, 81, 82] - but a clear study explaining the origin of this robustness is still to be provided. In particular, it is still to be understood if neural networks are simply robust to synaptic variability or if synaptic variability could actually be beneficial, in the same manner that noise might be beneficial to biological brains. For instance, Mahvash et al. [29] showed that synaptic variability plays a beneficial role in the reliability of spike generation, wherein synapses with high synaptic variability produce spike trains with reproducible timing. Another recent work by Pedretti et al. [41] demonstrated that adding noise in the input data presented to the network during learning is beneficial as it accelerates learning by a factor 3x.

2.1.3 Goal of this chapter

While several comprehensive studies of the impact of RRAM electrical properties with supervised algorithms have been reported [28, 34, 35, 48, 80, 83–85], little has been done with unsupervised learning [33, 81, 82, 86]. In this chapter, we provide a comprehensive insight of RRAM electrical requirements for artificial Spiking Neural Network (SNN) systems with unsupervised learning by Spike-Timing-Dependent Plasticity (STDP). A fully-connected feed-forward SNN topology with leaky integrate-and-fire neurons and RRAM-based synapses is adopted. We focus on two different applications: a detection task [87] and a classification task [69]. Electrical characterisations of RRAMs are provided. SNN simulations have been calibrated on the electrical characterisation results. In SECTION 2.2, we consider synaptic elements implemented with binary devices. In SECTION 2.3, we consider synaptic elements implemented with analog devices. The impact of RRAM characteristics on artificial SNN learning performance is investigated, namely:

Binary devices : the memory window, cycle-to-cycle and device-to-device conductance variability, aging.

Analog devices : the limited number of synaptic levels, the asymmetry between potentiation and depression, the non-linearity of the conductance response.

2.2 Binary devices

In this section, we focus on binary devices - in particular binary RRAMs-, *i.e.* devices switching between two distinct conductance states: a High Conductance State (HCS) and a Low Conductance State (LCS).

2.2.1 Experimental characterisation

2.2.1.1 Resistive memory device characteristics

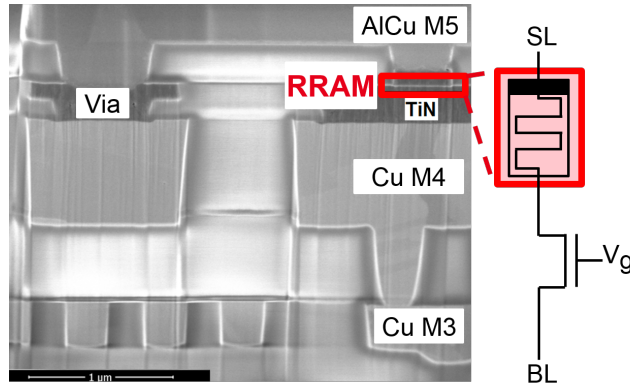


FIGURE 2.2.1: (Left) Scanning electron microscope cross-section of the TiN/HfO₂/Ti/TiN (100 nm/10 nm/10 nm/100 nm) RRAM cell integrated on top of the fourth Cu metal layer. (Right) Schematic view of the 1T1R cell configuration. The NMOS transistor is used as a selector device.

We focus on HfO₂-based oxide-based RRAM cells integrated in the back-end-of-line of a 130-nm CMOS process [60]. They consist of a capacitor-like metal-insulator-metal structure in series with a selector device. The memory element integration starts on top of the fourth metal layer (Cu). The scanning electron microscope cross-section of a 300-nm diameter HfO₂-based RRAM is shown in FIGURE 2.2.1 (Left). The RRAM devices are composed of a TiN/HfO₂/Ti/TiN stack where layers are 100 nm/10 nm/10 nm/100 nm thick. A NMOS transistor in series with the memory element is used as a selector device in a 1T1R configuration as depicted in FIGURE 2.2.1 (Right). This allows each memory device of the array to be read from and written to individually, and also regulates the compliance current - which usually defines the programming current, I_{prog} - during programming operations. Each 1T1R structure in the matrix is addressed using a Source Line (SL) and a Bit Line (BL) which connect to the top electrode of the device and the source of the transistor, respectively. The RRAMs require an initial forming process wherein a positive voltage of 4 V is applied on RRAM top electrodes ($V_{\text{SL}}=4.0$ V, $V_{\text{BL}}=\text{GND}$, $I_{\text{prog}}=65$ μA ,

$t_{\text{pulse}}=100$ ns) and during which they switch from a pristine state featuring a very low conductance value (<500 pS) to a higher conductance state. Upon the application of a positive voltage on RRAM top (SL) or bottom (BL) electrodes, the RRAM devices exhibit a reversible switching between a High Conductance State (HCS) (Set operation) and a Low Conductance State (LCS) (Reset operation), respectively. During a Set operation, the Set voltage, V_{Set} , is applied on SL while BL is grounded. During a Reset operation, the Reset voltage, V_{Reset} , is applied on BL while SL is grounded. The conductance value of RRAM devices after a programming operation depends on the programming conditions (applied voltage, programming current I_{prog} , and pulse duration). However, it varies across cycles (cycle-to-cycle variability) and devices (device-to-device variability).

All measurements presented in this section have been performed on a 4-kbit 1T1R HfO₂-based RRAM array. In order to study the impact of RRAM electrical properties on spiking neural network performance, RRAMs have been programmed with four different programming conditions. FIGURE 2.2.2 shows the cumulative distributions of HCS and LCS associated with each programming condition. The cumulative distributions are measured on all cells on the 4-kbit array after one cycle (one Set and one Reset operations on the 4 kbit devices). TABLE 2.1 summarises the different programming conditions. Programming conditions B1 and B2 consume less programming energy than condition A, whereas programming conditions C feature the highest programming energy consumption. In memory applications, RRAMs are used to store one bit of information: RRAMs in HCS are associated with a binary '1' value, and RRAMs in LCS are associated with a binary '0' value. Therefore, it is fundamental that HCS and LCS distributions do not overlap so that each state can be properly detected. This is the case for programming conditions A and C (FIGURE 2.2.2 (Top left and right)), whereas HCS and LCS distributions overlap for programming conditions B1 and B2 (FIGURE 2.2.2 (Bottom left and right)). The appropriate separation of HCS and LCS distributions is characterised by the memory window at 3σ , $MW_{3\sigma}$. The $MW_{3\sigma}$ is defined as the ratio between the HCS conductance value at -3σ , $\text{HCS}_{-3\sigma}$, and the LCS conductance value at $+3\sigma$, $\text{LCS}_{+3\sigma}$, of the conductance distributions:

$$MW_{3\sigma} = \frac{\text{HCS}_{-3\sigma}}{\text{LCS}_{+3\sigma}} \quad (2.2.1)$$

FIGURE 2.2.3 (a) shows the evolution of HCS and LCS during one million Set/Reset switching cycles with programming conditions A, measured on the 4-kbit array. Solid lines represent the median values of HCS and LCS distributions ($\text{HCS}_{0\sigma}$ and $\text{LCS}_{0\sigma}$) which remain constant for the 10^6 switching cycles. However, the conductance values at $\pm 3\sigma$, represented by the dotted lines, evidence an increase of conductance variability in both HCS and LCS due to RRAM aging. This causes a reduction of the memory window at 3σ . After 10^5 switching cycles, the HCS and LCS distributions start overlapping and it is no longer possible to use the RRAMs for memory applications. After 10^6 switching cycles, oxide breakdowns can be observed in some cells causing these *broken* cells to be stuck in the HCS. We define the *programming endurance* as the maximum number of programming operations before oxide breakdowns occur. FIGURE 2.2.3 (b)

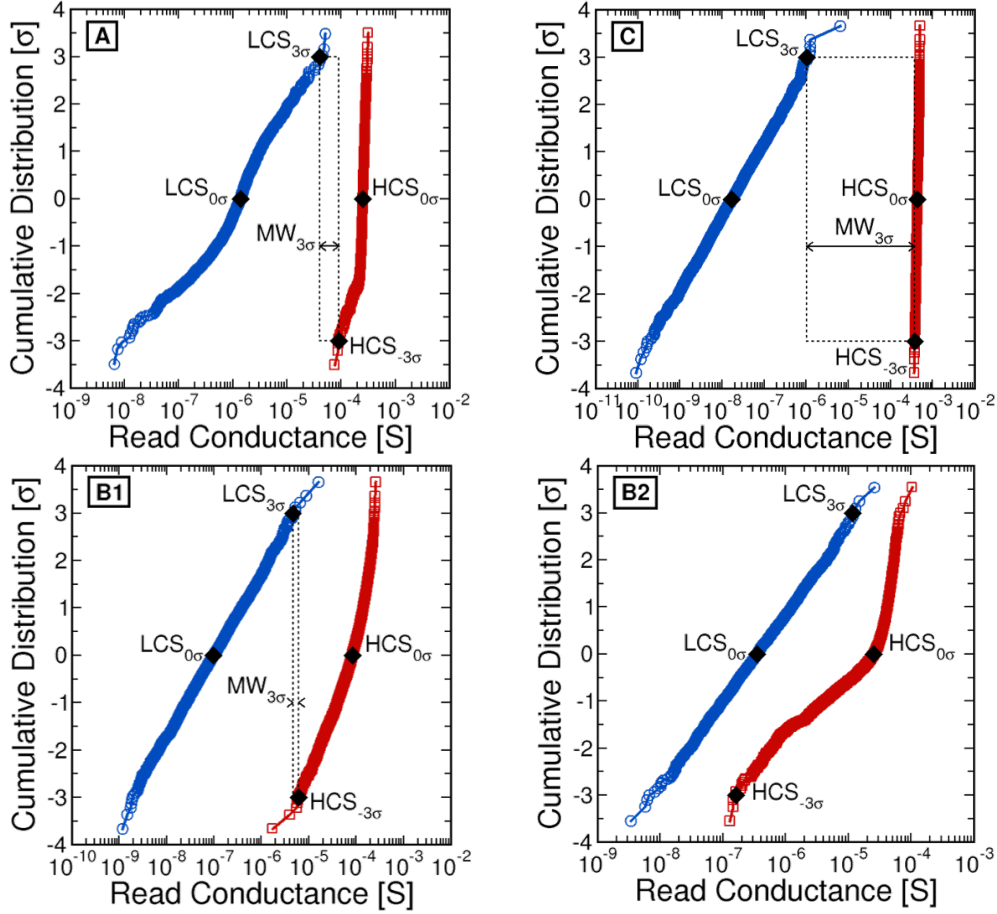


FIGURE 2.2.2: Cumulative distributions of the LCS and HCS distributions measured on the 4-kbit array (Top left) after 1000 switching cycles with condition A, (Top right) with condition C, (Bottom left) with condition B1, and (Bottom right) with condition B2. TABLE 2.1 summarises the parameters of each programming condition. These distributions represent the device-to-device variability.

Programming conditions		A	C	B1	B2
Voltage [V]	V_{Set}	2	2	2	2
	V_{Reset}	2.5	2.5	2.5	2.5
$I_{\text{prog,set}}$ [μA]		250	500	57	20
$V_{\text{g,reset}}$ [μA]		3	3.5	3.5	3.5
Energy [pJ/spike]	E_{Set}	50	100	11.4	4
	E_{Reset}	62.5	125	14.25	5
$\sigma_{\text{G,HCS}}$ [$\log_{10}(\text{S})$]		0.05	0.02	0.28	0.53
$\sigma_{\text{G,LCS}}$ [$\log_{10}(\text{S})$]		0.49	0.64	0.58	0.54
$\text{MW}_{3\sigma}$ [#]		3	370	1.3	0.014
Endurance [#]		10^6	$\approx 10^2$	$\approx 10^6$	$\approx 10^8$

TABLE 2.1: Programming conditions used in this work, with $t_{\text{pulse}}=100$ ns. The programming energy is defined in EQUATION 2.2.3.

shows the evolution of HCS (red circle) and LCS (blue square) conductance variability during the endurance cycling of FIGURE 2.2.3 (a). Here, we define the *conductance variability* as the standard deviation of the base-10 logarithm

of the conductance distributions:

$$\begin{aligned}\sigma_{G,HCS} &= \text{std}[\log_{10}(G_{HCS})] \\ \sigma_{G,LCS} &= \text{std}[\log_{10}(G_{LCS})]\end{aligned}\tag{2.2.2}$$

with G_{HCS} and G_{LCS} the conductance distributions of HCS and LCS, respectively. This definition of conductance variability allows to translate the absolute standard deviation of the conductance values into the standard deviation in terms of orders of magnitude of the distributions [49]. TABLE 2.1 presents the different metrics for each programming condition. The programming energy in Set and Reset has been calculated as:

$$\begin{aligned}E_{\text{Set}} &= V_{\text{Set}} * I_{\text{prog,set}} * t_{\text{pulse}} \\ E_{\text{Reset}} &= V_{\text{Reset}} * I_{\text{prog,set}} * t_{\text{pulse}}\end{aligned}\tag{2.2.3}$$

$MW_{3\sigma}$, programming endurance, and conductance variability of both HCS and LCS depend on the programming conditions (programming current, I_{prog} , and the amplitude of Set/Reset voltage pulses) [88–90]. FIGURE 2.2.4 shows the conductance variability, defined in EQUATION 2.2.2, as a function of the median conductance value, measured on the 4-kbit array for several programming conditions. The conductance variability is constant at roughly 0.5 for conductance values lower than $77.5 \mu\text{S}$ and then decreases with the median conductance value. In order to increase the memory window, it is necessary to apply stronger Reset programming conditions in order to decrease the LCS median conductance value, and/or apply stronger Set programming conditions to decrease HCS variability and increase HCS median conductance value. However, this implies an increase in programming power consumption. In addition, it has been demonstrated that a trade-off exists between the memory window and the programming endurance: higher memory windows imply lower programming endurance [88–91]. In this work, we focus on the four representative programming conditions which are reported in FIGURE 2.2.4 with the filled symbols. TABLE 2.1 summarises the parameters of each condition:

- **A:** compromise between programming endurance and $MW_{3\sigma}$ (suited conditions for standard memory applications)
- **B1 and B2:** low programming power consumption, high variability in both HCS and LCS, and low $MW_{3\sigma}$ (cannot be used for memory applications due to the low window margin)
- **C:** highest $MW_{3\sigma}$ amongst the four conditions, high programming power consumption, low HCS variability, and low programming endurance.

2.2.1.2 Implementation of synaptic elements and learning rule with resistive memories

Many implementations of RRAM-based synapses seek an analog conductance modulation under identical pulses in both programming directions: when consecutive Set (*potentiation*) or Reset (*depression*) pulses are applied,

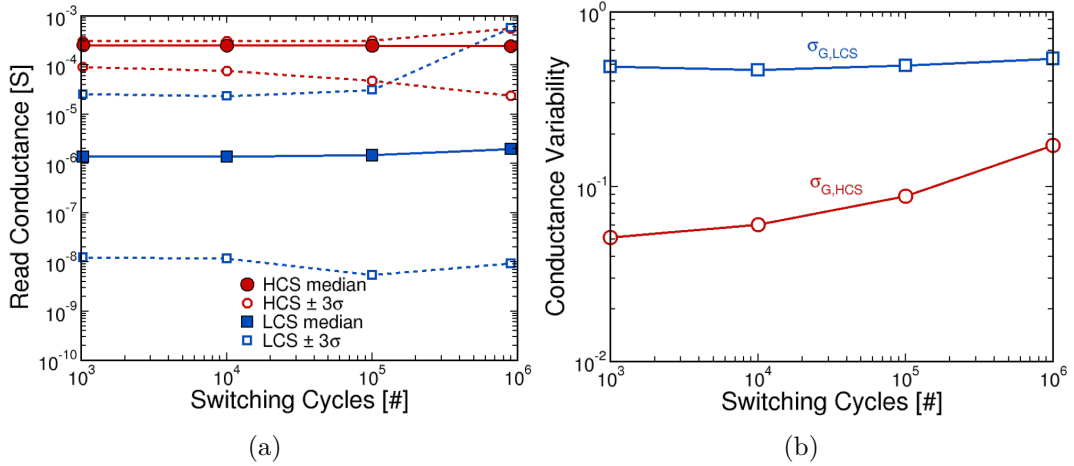


FIGURE 2.2.3: (a) Programming endurance characterisation with programming conditions A in TABLE 2.1. (b) Evolution of HCS and LCS conductance variability with programming conditions A during RRAM aging. Conductance variability is defined in EQUATION 2.2.2.

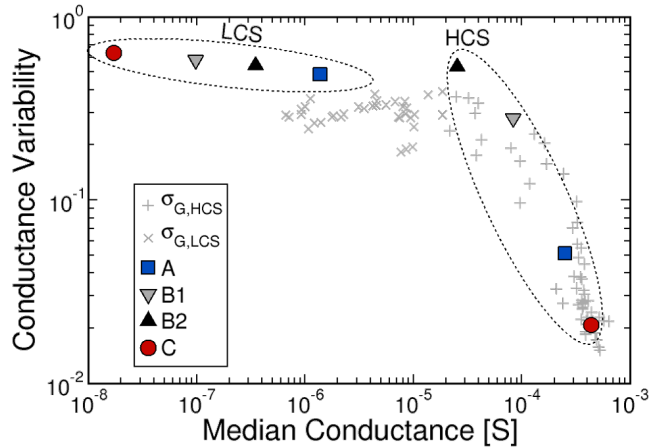


FIGURE 2.2.4: Conductance variability as a function of the median conductance value for different programming conditions. Conductance variability is defined in EQUATION 2.2.2.

the conductance should gradually increase or decrease, respectively [28, 33–35, 48, 53, 54, 56, 80, 82, 86]. FIGURE 2.2.5 reports the conductance response when a series of 20 identical Set and Reset pulses are applied on the 4-kbit RRAM array. Grey curves are the conductance response of single RRAM cells with an analog behaviour. Dotted black lines are the conductance response of single RRAM cells with a binary behaviour, *i.e.* an abrupt switching between the LCS and HCS is observed. Only ten single cells are plotted for the sake of clarity. Red and blue curves are the median conductance values extracted from 4 kbit cells during potentiation and depression, respectively. Low and high programming power conditions (B2 and A in TABLE 2.1) are used. For low programming power conditions (B2, FIGURE 2.2.5 (Left)), the evolution of the median conductance value shows an analog switching during depression. Unfortunately, this behaviour is difficult to control across a large array due to the strong device-to-device conductance variability. In addition, some cells present a binary behaviour and only switch between two distinct states (HCS and

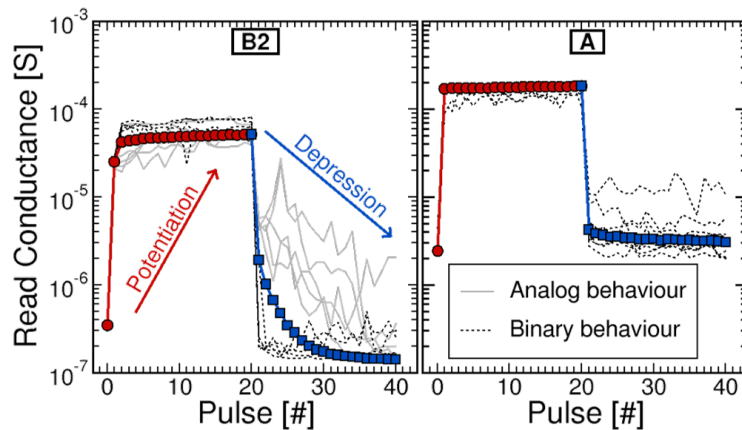


FIGURE 2.2.5: Conductance evolution during the application of a series of 20 identical Set pulses and Reset pulses with (Left) programming conditions B2 and (Right) programming conditions A of TABLE 2.1. Grey lines are representative of ten single cells behaving as analog devices (gradual increase or decrease of the conductance value). Black lines are representative of ten single cells behaving as binary devices (abrupt switching between the HCS and LCS). Red circles and blue squares correspond to the median conductance value calculated on 4 kbit cells during potentiation and depression, respectively. The pulse 0 is the conductance value before the first Set pulse.

LCS). Moreover, even in the cells presenting an analog-like switching behaviour, the amount of conductance increase (decrease) after a Set (Reset) pulse varies from device to device and pulse to pulse. In potentiation, the evolution of the median conductance value shows a binary switching. For high programming power conditions (FIGURE 2.2.5 (Right)), most of the RRAM cells (more than ninety percent) present a binary behaviour in both programming directions (potentiation and depression).

To overcome these limitations, we use a synaptic compound of multiple (n) RRAM cells connected in parallel associated with a probabilistic programming scheme [47, 49]. The circuit implementation is depicted in FIGURE 2.2.6 (a). Since parallel conductances add up, the equivalent synaptic weight spreads from the sum of n conductances in LCS to n conductances in HCS, with $n+1$ distinct intermediate conductance levels. In order to define the conductance state of each RRAM device (HCS or LCS), we associate this implementation with a stochastic Spike-Timing-Dependent Plasticity (STDP) learning rule [44] - a simplified form of the bio-inspired STDP rule [71, 72]. The learning rule is depicted in FIGURE 2.2.6 (b, left). When the presynaptic neuron spikes before the postsynaptic neuron spikes within a time window t_{STDP} , a Long-Term Potentiation (LTP) event occurs, and each RRAM of the synaptic compound has a probability p_{LTP} to switch to the HCS. Otherwise, a Long-Term Depression (LTD) event occurs, and each RRAM cell of the synaptic compound has a probability p_{LTD} to switch to the LCS. The switching probabilities, p_{LTP} and p_{LTD} , can be defined either with the use of an external Pseudo-Random Number Generator (PRNG) or the intrinsic RRAM switching probabilities [45, 92]. An external PRNG allows for a fine tuning of switching probabilities at the expense of circuitry overhead. Using the intrinsic RRAM switching probabilities reduces design complexity and

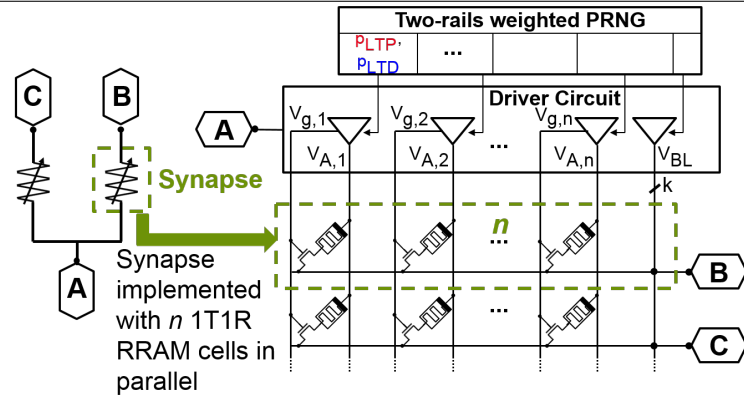
programming power consumption as lower Set and Reset voltage amplitudes and programming currents are used. However, this also leads to variability in switching probabilities [92]. In this section, switching probabilities are defined assuming the use of an external PRNG. The number of synaptic levels is defined by the number of RRAM cells operating in parallel as shown in FIGURE 2.2.6 (b, right). FIGURE 2.2.6 (c) shows the impact of the programming conditions on the conductance evolution of a synapse composed of 20 RRAM cells operating in parallel when 200 potentiation pulses followed by 200 depression pulses are applied. Grey lines show the conductance evolution of 100 different synapses, red circles and blue squares represent the median conductance over the 100 synapses in potentiation and depression, respectively. We observe a gradual increase (LTP) and decrease (LTD) of the conductance as a function of the number of pulses with any programming condition. The ratio between the median maximum conductance value (*i.e.* all the devices are in the HCS) and the initial conductance value (*i.e.* only one device is in the HCS while the others are in the LCS), $G_{\max,0\sigma}/G_{\text{init}}$, is similar for every programming condition since it mostly depends on the number of RRAM cells, n , per synapse (≈ 16 for A, ≈ 18 for C, ≈ 17 for B1 and B2).

2.2.2 Implications for a learning system: impact of binary RRAM-based synapse characteristics on the network performance

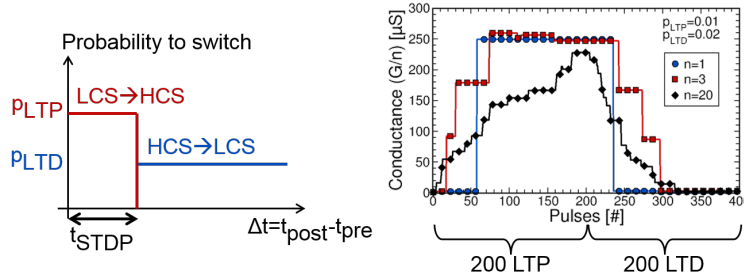
We now investigate the impact of RRAM conductance variability for two different applications implemented with Spiking Neural Networks (SNNs): a detection task [87] and a classification task [69]. The network performance is assessed by means of system-level simulations with the special-purpose neuromorphic hardware simulator N2D2 [73, 93]. The detailed RRAM physical characterisation presented in SECTION 2.2.1 has been implemented into physical models to understand how device properties translate in terms of learning. Variability effects due to peripheral circuits (such as neuron variability [33, 81]) are intentionally not taken into account. For each programming condition, the real conductance distributions measured on the 4-kbit array have been used to perform the simulations.

2.2.2.1 Network topology

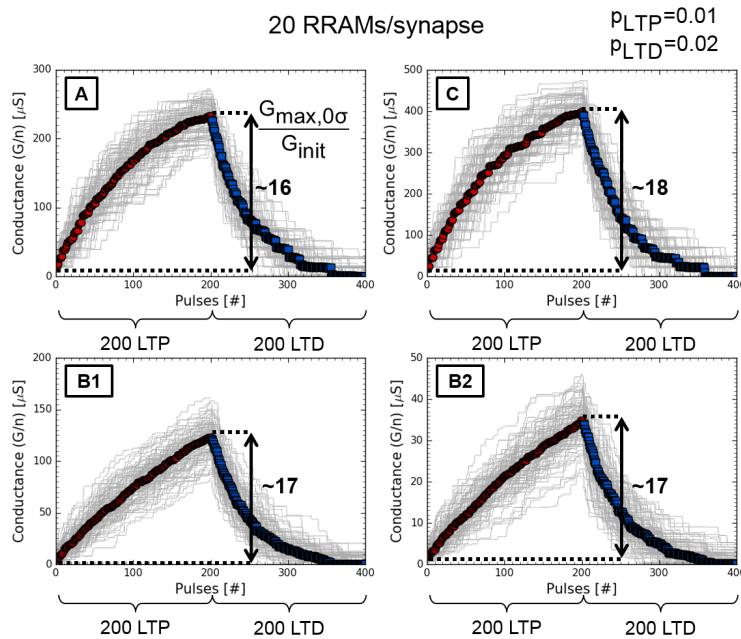
Both applications are based on a one-layer fully-connected feed-forward neural network topology: each neuron of the first layer (input layer) is connected to each neuron of the second layer (output layer) with a synaptic element. A detailed description of the simulated SNNs for detection and classification is provided in APPENDIX A. Neurons in the output layer are implemented with the Leaky Integrate-and-Fire (LIF) model [73]. For each simulation, all the output LIF neurons have the same firing threshold value which has been optimised to provide the best performance (see APPENDIX B). The other neuron parameters are kept constant and identical for all the simulations. For the car detection application (FIGURE 2.2.7 (a)), the input layer corresponds to an image sensor



(a)



(b)



(c)

FIGURE 2.2.6: (a) RRAM-based synapse implementation. The Pseudo-Random Number Generator (PRNG) is used to tune the switching probabilities. (b) Stochastic STDP rule and conductance evolution of a RRAM-based synapse composed of 1, 3, and 20 RRAM cells in parallel. 200 potentiation pulses followed by 200 depression pulses are applied. Condition A in TABLE 2.1 is used. (c) Conductance evolution of 100 RRAM-based synapses composed of 20 RRAM cells when 200 potentiation pulses and 200 depression pulses are applied. Condition A (Top left), C (Top right), B1 (Bottom left), and B2 (Bottom right) are used. Red and blue symbols represent the median conductance evolution; grey lines represent the evolution of each synapse.

composed of 128x128 spiking pixels. It is fully connected to an output layer of 60 neurons. The F1-score is used to assess network performance (see APPENDIX A). F1 ranges from 0 to 1 with 1 being the best detection score. For the digit classification application (FIGURE 2.2.7 (b)), the input layer corresponds to input images from the MNIST dataset [69] composed of 28x28 pixels. A spike frequency encoding is used to convert each input pixel into a spike train whose spike rate depends on the grey intensity level of the pixel. The input layer is fully connected to an output layer of 500 neurons. The Classification Rate (CR) is computed as the ratio between the number of successfully classified digits and the number of input digits presented. The full MNIST dataset (60 000 training digits, 10 000 testing digits) is used once. We implemented the synaptic elements with the RRAM-based synaptic compound presented in SECTION 2.2.1.2 (*cf* FIGURE 2.2.6 (a)). Networks are trained with the unsupervised stochastic STDP rule, extrinsic switching probabilities (an external PRNG is assumed), and lateral inhibition [33, 73].

2.2.2.2 Detection application

The impact of RRAM-based synapse characteristics (number of synaptic levels, RRAM conductance variability, memory window, and aging) on the learning performance of the network designed for detection is investigated. All results have been averaged over twenty simulations. Error bars represent the deviation at 1σ .

Impact of the RRAM memory window and conductance variability

The first step was to study the impact of the number of synaptic levels and the RRAM memory window on the network performance. To vary the number of synaptic levels, the number n of RRAMs per synapse is modified. FIGURE 2.2.8 (a) shows the F1-score as a function of the memory window at 3σ ($MW_{3\sigma}$, EQUATION 2.2.1) for different numbers of RRAMs per synapse. Each point has been averaged over twenty simulations. Error bars represent the deviation at 1σ . We used the LCS and HCS distributions measured under the programming conditions A (FIGURE 2.2.2 (a)). $MW_{3\sigma}$ is modified by a translation of the LCS distribution to higher (decrease of $MW_{3\sigma}$) or lower (increase of $MW_{3\sigma}$) conductance values with respect to the actual value measured under condition A (blue dashed line in FIGURE 2.2.8 (a)). This allows to vary the memory window while keeping the HCS and LCS conductance variability values constant (EQUATION 2.2.2), and it decouples the impact of $MW_{3\sigma}$ from the impact of conductance variability. Surprisingly, the SNN performance is independent of the number of devices per synapse: a binary synapse with only two distinct synaptic levels is sufficient for this type of application. We obtained the same result with the other LCS and HCS distributions from TABLE 2.1 (not shown). By contrast, the essential parameter to improve SNN performance is the $MW_{3\sigma}$: F1 increases with the $MW_{3\sigma}$, and it saturates at a F1-score of about 0.96 for a memory window at 3σ larger than 3.

Second, we studied the impact of the conductance variability. We simulated the proposed application with the four LCS and HCS distributions measured under the four programming conditions presented in TABLE 2.1. An artificial case of a

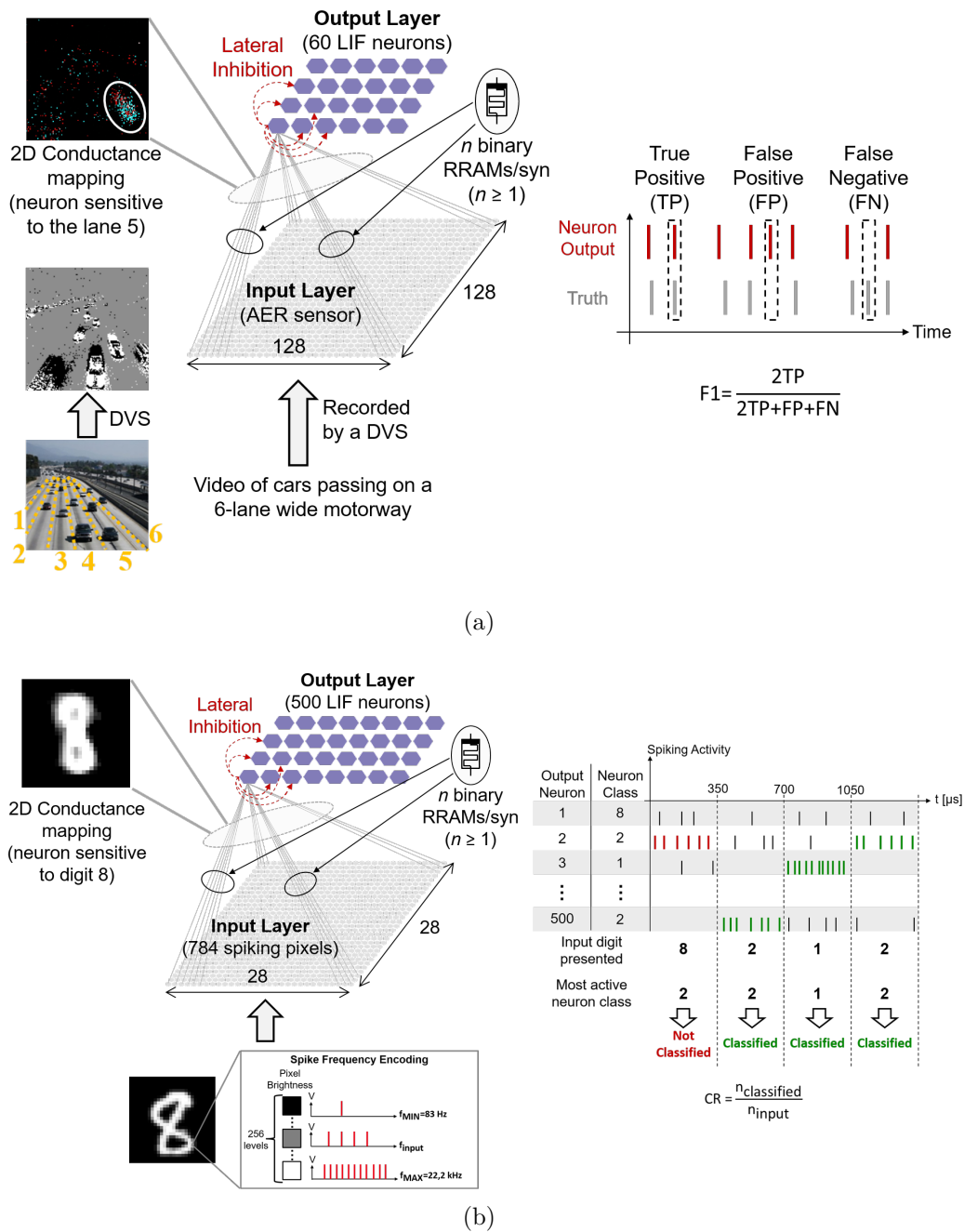


FIGURE 2.2.7: Simulated spiking neural networks used for (a) the car tracking and (b) the digit classification applications. The associated score definition to assess network performance is shown on the right-hand side of each network. See APPENDIX A for more details.

synapse with zero variability ($\sigma_{G,HCS}=0$ and $\sigma_{G,LCS}=0$) was also simulated for the sake of comparison. For each simulation, synaptic elements are implemented with only one RRAM device since increasing the number of synaptic levels has no impact on the network performance. FIGURE 2.2.8 (b) shows the simulated F1-score as a function of the memory window at 3σ , $MW_{3\sigma}$, for the different studied distributions. The different $MW_{3\sigma}$ values were obtained by translating the LCS distributions to lower or higher conductance values. This allows to decouple the impact of $MW_{3\sigma}$ from the conductance variability. The $MW_{3\sigma}$

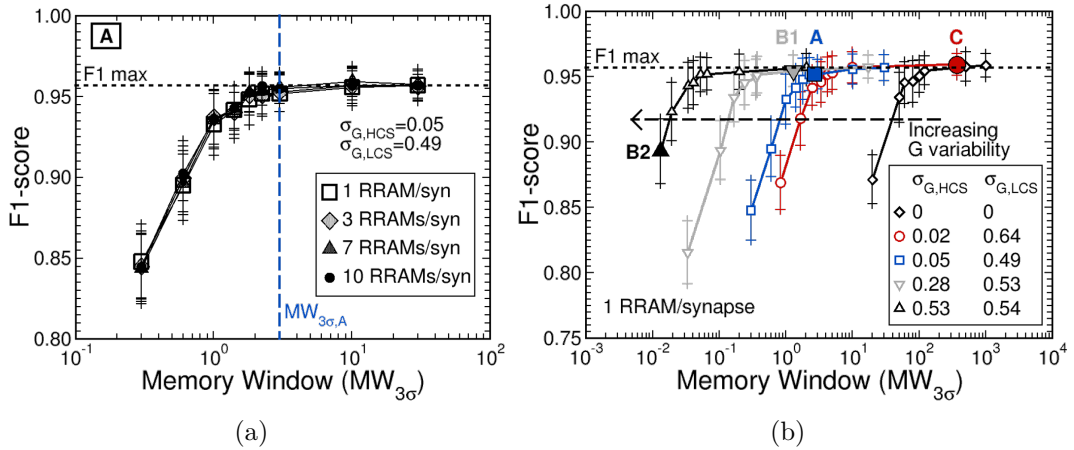


FIGURE 2.2.8: (a) F1-score as a function of the memory window at 3σ , $MW_{3\sigma}$ defined in EQUATION 2.2.1, for different numbers of RRAMs per synapse. The HCS and LCS distributions measured under the programming conditions A on the 4-kbit array are used (*cf* FIGURE 2.2.2 (a)). (b) F1-score as a function of the $MW_{3\sigma}$. One RRAM device per synapse is used. The HCS and LCS distributions measured on the 4-kbit array for the four conditions of TABLE 2.1 and an artificial case with zero variability are used. The $MW_{3\sigma}$ is varied by a translation of the LCS distributions to lower or higher conductance values.

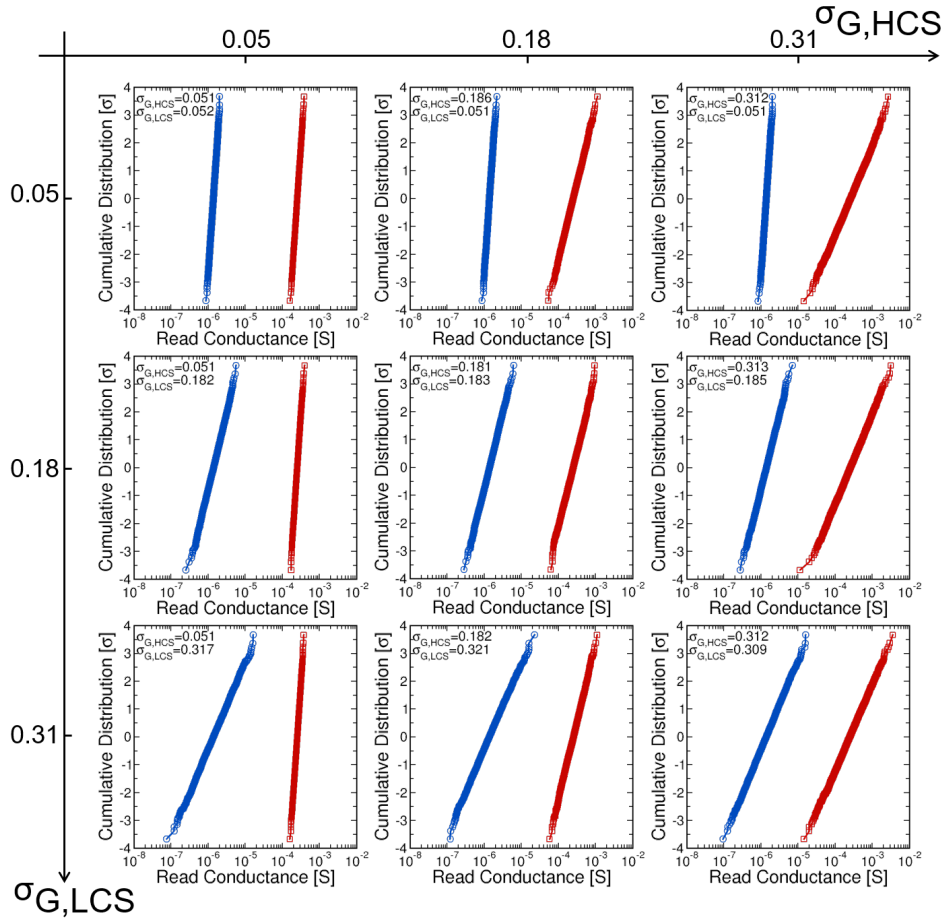
corresponding to the actual experimental results for each programming condition is highlighted by a filled symbol. For each case, F1 increases with the $MW_{3\sigma}$, and it saturates at a maximal F1-score of about 0.96 after a certain minimal $MW_{3\sigma}$. The higher the conductance variability, the lower the minimal $MW_{3\sigma}$ required to reach the maximum score F1 of 0.96. For $\sigma_{G,HCS}=0.53$ and $\sigma_{G,LCS}=0.54$ (condition B2, black triangle in FIGURE 2.2.8 (b)), a $MW_{3\sigma}$ larger than 0.5 is required to reach the maximum score, whereas with no variability (synapse with no conductance variability, black diamond in FIGURE 2.2.8 (b)), a $MW_{3\sigma}$ of at least 200 is necessary. We extended the simulated results obtained in FIGURE 2.2.8 (b) with nine different artificial combinations of HCS and LCS variability shown in FIGURE 2.2.9 (a). The HCS and LCS distributions of the nine artificial conditions follow a log-normal random law with different HCS and LCS conductance variability values. Synaptic elements are implemented with one RRAM device. We obtained the same result as in FIGURE 2.2.8 (b): for each of the nine combinations, a certain minimal $MW_{3\sigma}$ is required to reach the maximal F1-score of 0.96 (not shown). We plot in FIGURE 2.2.9 (b) the minimal memory window at 3σ , $MW_{3\sigma,min}$ (z-axis), required to reach the maximum F1-score of 0.96 as a function of the HCS (x-axis) and LCS (y-axis) conductance variability, for the four studied programming conditions (filled symbol), the synapse with no variability (black diamond), and the nine artificial log-normal HCS and LCS distributions (black circle). It is clear from FIGURE 2.2.8 (b) that increasing the RRAM synaptic variability, $\sigma_{G,HCS}$ and $\sigma_{G,LCS}$, is a way to relax the constraints on the minimal required $MW_{3\sigma}$: higher conductance variability values allow for a decrease of the minimal required $MW_{3\sigma}$. This can be explained by the increased dynamic range with higher conductance variability values, *i.e.* the increased range of synaptic values that are available during the learning phase.

After the learning phase with the STDP learning rule, potentiated synapses (RRAMs in HCS) represent relevant inputs, *i.e.* synapses transmitting spikes generated by a car passing on the motorway, and depressed synapses (RRAMs in LCS) represent noisy inputs. This is well illustrated on the two-dimensional conductance mapping example of one arbitrary output neuron after learning in FIGURE 2.2.10 (a). Each dot corresponds to one input synapse. Potentiated synapses (RRAMs in HCS) are represented by coloured dots (red, blue, and grey), depressed synapses (RRAMs in LCS) are represented by black dots (see APPENDIX A for more details). As a result of the learning phase, we can observe a pool of potentiated synapses (circled in white) denoting the sensitivity of this neuron to cars passing at this specific position on the motorway. As the size of a car is relatively small compared to the size of the video, the majority of the synaptic weights has to be weak (RRAMs in LCS) with a tail of stronger connections (RRAMs in HCS) in order to achieve high performance after the learning phase. In our simulations, high performance after the learning phase was reached ($F1 \approx 0.96$) when the sum of the synaptic weights of potentiated synapses was in average two hundred times as high as the sum of the synaptic weights of depressed synapses.

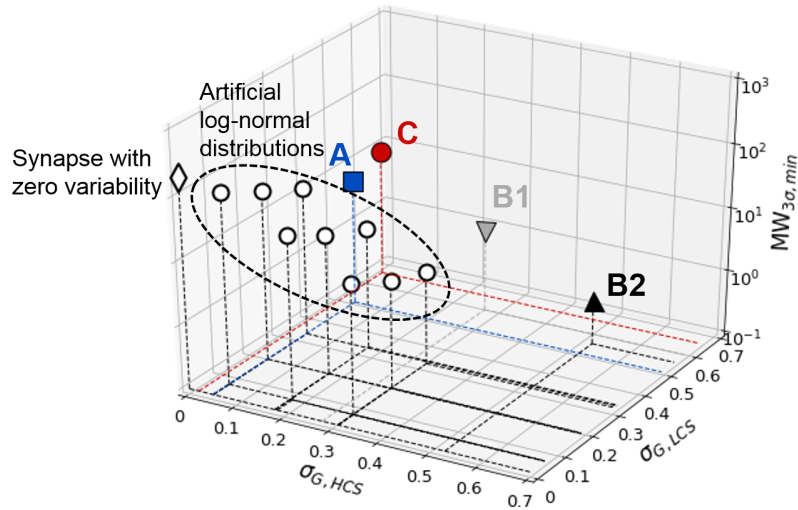
To quantify this result, we define the Synaptic Window (SW) as the ratio between the mean conductance value of synapses in HCS, $\overline{G_{HCS}}$, and the mean conductance value of synapses in LCS, $\overline{G_{LCS}}$:

$$SW = \frac{\overline{G_{HCS}}}{\overline{G_{LCS}}} \quad (2.2.4)$$

FIGURE 2.2.10 (b) shows the F1-score as a function of the synaptic window, SW , for the four experimental distributions in TABLE 2.1 plus the synapse with no variability. The actual experimental SW for each condition is highlighted by a filled symbol. As evidenced by FIGURE 2.2.10 (b), the network performance is independent of synaptic variability; F1 is defined by the synaptic window, SW . F1 saturates at 0.96 for synaptic windows larger than 200. This is illustrated in FIGURE 2.2.10 (c) which shows the synaptic weight distribution after the learning phase for the four programming conditions in TABLE 2.1. High performance after learning is reached when the ratio between the peaks of the HCS and LCS distributions is larger than 200. In that sense, the increase of both conductance variability and memory window allows for an increase of the ratio between the conductance values of potentiated synapses (red) and depressed synapses (blue). Potentiated synapses at the highest conductance tail of the distribution compensate for synapses at the lowest conductance tail. Similarly, depressed synapses at the lowest conductance tail of the distribution compensate for synapses at the highest conductance tail. This helps separate the peaks of both distributions leading to high network performance. It is worth noting that even with low programming power conditions (B1, $F1=0.96$) we have a score as good as with the high programming power condition (C, $F1=0.96$). The experimental condition B1 works well for neuromorphic applications whereas it cannot be used in a memory application due to its high HCS variability (no memory window). However, for the experimental condition B2, RRAM works neither for memory nor neuromorphic applications. A decrease in F1 is observed with the experimental condition A (optimised for standard memory applications)



(a)



(b)

FIGURE 2.2.9: (a) Cumulative distributions of the nine artificial log-normal distributions used to quantify the impact of synaptic variability. (b) Minimal memory window at 3σ , $MW_{3\sigma,min}$ (z-axis), required to reach the maximal F1-score of 0.96 as a function of the HCS (x-axis) and LCS (y-axis) conductance variability values. Higher conductance variability values allow to relax the constraints on $MW_{3\sigma,min}$.

but is still acceptable (F1=0.95) if we can tolerate a loss of performance for

an increase in endurance with respect to programming conditions C. FIGURE 2.2.10 (d) reports the simulated learning time as a function of the synaptic window. The learning time has been defined as the time at which F1 reaches its maximal value within a window of $\pm 1\%$, for a given RRAM programming condition. Learning time is degraded only for the condition B2 with a reduced synaptic window.

Impact of RRAM aging

Finally, we studied the impact of the RRAM aging with endurance on network performance. Both device-to-device and cycle-to-cycle variability are taken into account. We extracted the conductance distribution during cycling on the 4-kbit array up to one million cycles for the condition A (*cf* FIGURE 2.2.3 (a)), and we used these data to evaluate the impact of RRAM aging on the F1-score. The results are shown in FIGURE 2.2.11 (a). We can maintain a constant F1-score of 0.95 until 10^5 cycles despite the increase in conductance variability. At 10^6 cycles, F1 plummets (F1=0.92). The degradation of F1 after 10^6 cycles is not due to the increase in conductance variability and decrease of $MW_{3\sigma}$ but to the broken cells (RRAMs stuck in the HCS). Upon removal of the broken cells (1%) from the distribution (blue shaded square), it is possible to move back up to a score F1 of 0.95. FIGURE 2.2.11 (b) shows the average number of Set (red circle) and Reset (blue square) operations per RRAM device during the learning phase for each programming condition. Red and blue shaded areas represent the number of Set and Reset operations per RRAM device at $\pm 3\sigma$, respectively. Similar numbers of programming events are required for each programming condition. Considering the minimal learning time required to reach high performance simulated in FIGURE 2.2.10 (d) (165 s for conditions A, B1, and C), an average of about 0.1 Set operation and 10 Reset operations per device is required, up to a maximum of 20 Set and 40 Reset operations. Considering the programming endurance of this RRAM technology (*cf* TABLE 2.1), this makes possible the use of these programming conditions for learning.

2.2.2.3 Classification application

Impact of the RRAM memory window and conductance variability

A similar study on the impact of RRAM-based synapse characteristics on the fully-connected feed-forward spiking neural network designed for digit classification (see APPENDIX A) is performed. Each point has been averaged over twenty simulations. Error bars represent the deviation at 1σ . First, we investigated the impact of the number of synaptic levels and the RRAM memory window, then the conductance variability on the SNN performance. FIGURE 2.2.12 (a) reports the Classification Rate (CR) as a function of the memory window at 3σ , $MW_{3\sigma}$. The HCS and LCS distributions measured under the conditions A and B2 plus the synapse with zero variability ($\sigma_{G,HCS}=0$ and $\sigma_{G,LCS}=0$) are used. The different $MW_{3\sigma}$ values were obtained by translating the LCS distributions to lower or higher conductance values. The actual experimental $MW_{3\sigma}$ for each condition is highlighted by filled symbols. Each curve corresponds to a different number of RRAM devices per synapse. In contrast to the detection task, *the CR is independent of the $MW_{3\sigma}$* for all the studied distributions. The network

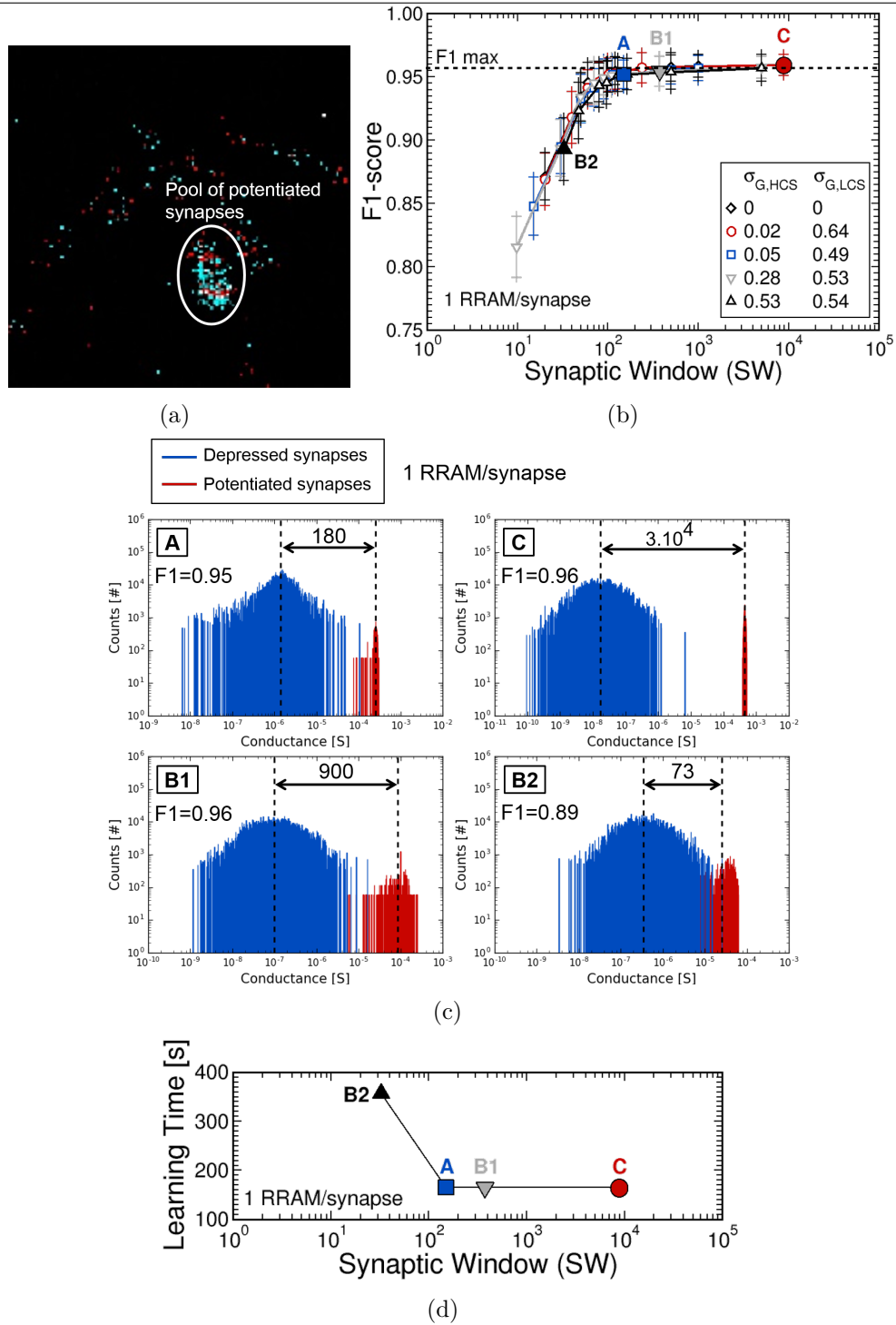


FIGURE 2.2.10: (a) Example of the two-dimensional conductance mapping of one arbitrary output neuron after learning. Potentiated synapses (RRAMs in HCS) are represented by coloured dots (red, blue, and grey); depressed synapses (RRAMs in LCS) are represented by black dots. (b) F1-score as a function of the synaptic window, SW, defined in EQUATION 2.2.4. (c) Synaptic weight distributions after the learning phase for the four programming conditions of TABLE 2.1. High performance after learning (F1=0.96) is reached when the ratio between the peaks of the HCS and LCS distributions is larger than 200. (d) Learning time as a function of the SW.

performance depends on:

- **Number of synaptic levels:** The CR increases with the number of

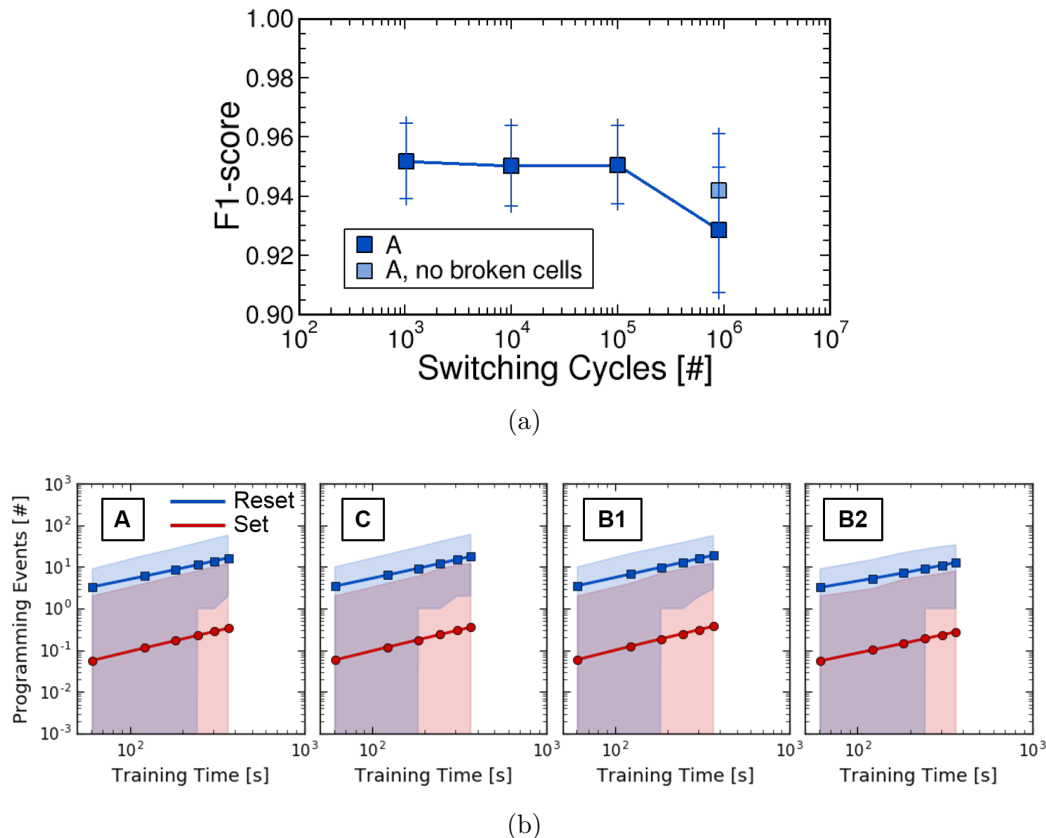


FIGURE 2.2.11: (a) Impact of the RRAM aging on the F1-score. Simulations have been calibrated using the data of FIGURE 2.2.3 (a) (condition A). Both device-to-device and cycle-to-cycle variability are taken into account. (b) Average number of Set (red circle) and Reset (blue square) operations during the learning phase for each programming condition. Red and blue shaded areas represent the evolution at $\pm 3\sigma$, respectively.

RRAMs per synapse, and it saturates after 10 RRAMs per synapse. This is in agreement with the studies performed in [33, 56, 82].

- **Synaptic variability:** The HCS and LCS distributions measured under the programming conditions A improve the performance with respect to the synapse with zero variability for a given number of RRAMs per synapse. However, similar performances are achieved with condition B2 (high conductance variability) and the synapse with zero variability.

To quantify and understand the impact of conductance variability, we simulated the SNN performance for the four programming conditions of TABLE 2.1 plus the synapse with zero variability. FIGURE 2.2.12 (b) plots the classification rate, CR, as a function of HCS variability, for a synapse composed by one and twenty RRAM cells. As the $MW_{3\sigma}$ has no impact on the CR, we simulated the proposed network calibrated on the experimental $MW_{3\sigma}$ for each programming condition and a $MW_{3\sigma}$ of 5 for the synapse with no variability. The network performance is maximal for $\sigma_{G,HCS} \approx 0.05$ (CR=81.81% for condition A and CR=81.78% for condition C, for 20 RRAMs per synapse), and it is degraded when HCS variability is too high (CR=78.65% for condition B2) or when there is no variability at all (CR=79.45% for the synapse with zero variability). In the case of one RRAM

device per synapse, the maximum score is 76.32% (condition A). These results are far from the best one (99.77%) obtained for the same dataset which used a supervised off-line learning approach and millions of adjustable parameters [94]. However, our results compare well to previously published scores - using a similar number of adjustable parameters and synaptic elements - with on-line supervised neural network with back-propagation (82.9%) [34] and on-line unsupervised learning (87.0% with the same network, homeostasis on neuron threshold, and presentation of the full MNIST dataset seven times for training compared to once in our case). To better understand the impact of synaptic variability, we plot in FIGURE 2.2.12 (c) the synaptic weight distribution after learning for the four experimental distributions plus the synapse with zero variability. We consider that a synapse is potentiated (red) if at least one of its RRAM devices is potentiated (RRAM in HCS). Otherwise, the synapse is depressed (blue, all RRAM devices in LCS). 20 RRAMs per synapse have been simulated which provides 21 different synaptic levels. In the case of the synapse with zero variability, we observe 21 distinct synaptic levels with a clear separation in between each level. As long as the HCS variability value remains low enough (conditions A and C), it is still possible to discriminate between the 21 synaptic levels. However, in the case of conditions B1 and B2, the increase in HCS conductance variability flattens the synaptic weight distributions. This decreases the number of distinguishable synaptic levels down to 9 and 7 for conditions B1 and B2, respectively. Consequently, the CR degrades. In the case of conditions A and C, the presence of conductance variability increases the range of synaptic weight values and enables synapses to access intermediate synaptic weights in between each synaptic level. This makes the transition more gradual between each level and permits a finer tuning of the synaptic weights. By contrast, in the case of the synapse with zero variability, synapses are constrained to the 21 different synaptic levels. This accounts for the improved CR for conditions A and C with respect to the synapse with zero variability.

Impact of RRAM aging

FIGURE 2.2.13 shows the impact of RRAM aging on the classification rate, CR. Both device-to-device and cycle-to-cycle variability are taken into account. Simulations have been calibrated using the data of FIGURE 2.2.3 (a) measured on the 4-kbit array with the programming conditions A. The HCS variability value during aging remains between 0.05 and 0.2 (*cf* FIGURE 2.2.3 (b)). As shown in FIGURE 2.2.12 (b), the CR varies little in that range. Therefore, with 20 RRAMs per synapse, we can sustain a constant score $CR \approx 81.5\%$ until 10^6 cycles. In addition, the network is proved to be robust to broken cells (1%) which have a minimal impact on the CR (blue shaded square). FIGURE 2.2.13 (b) shows the average number of Set (red circle) and Reset (blue square) operations per RRAM device during the learning phase for each programming condition. Red and blue shaded areas represent the number of Set and Reset operations per RRAM device at $\pm 3\sigma$, respectively. Similar numbers of programming events are required for each programming condition. Considering a learning time of 21 s (60 000 training digits), an average of about 1 Set operation and 10 Reset operations per device is required, up to a maximum of 60 Set and 100 Reset operations. Considering the programming endurance of this RRAM technology (*cf* TABLE 2.1), this makes possible the use of programming conditions A, B1,

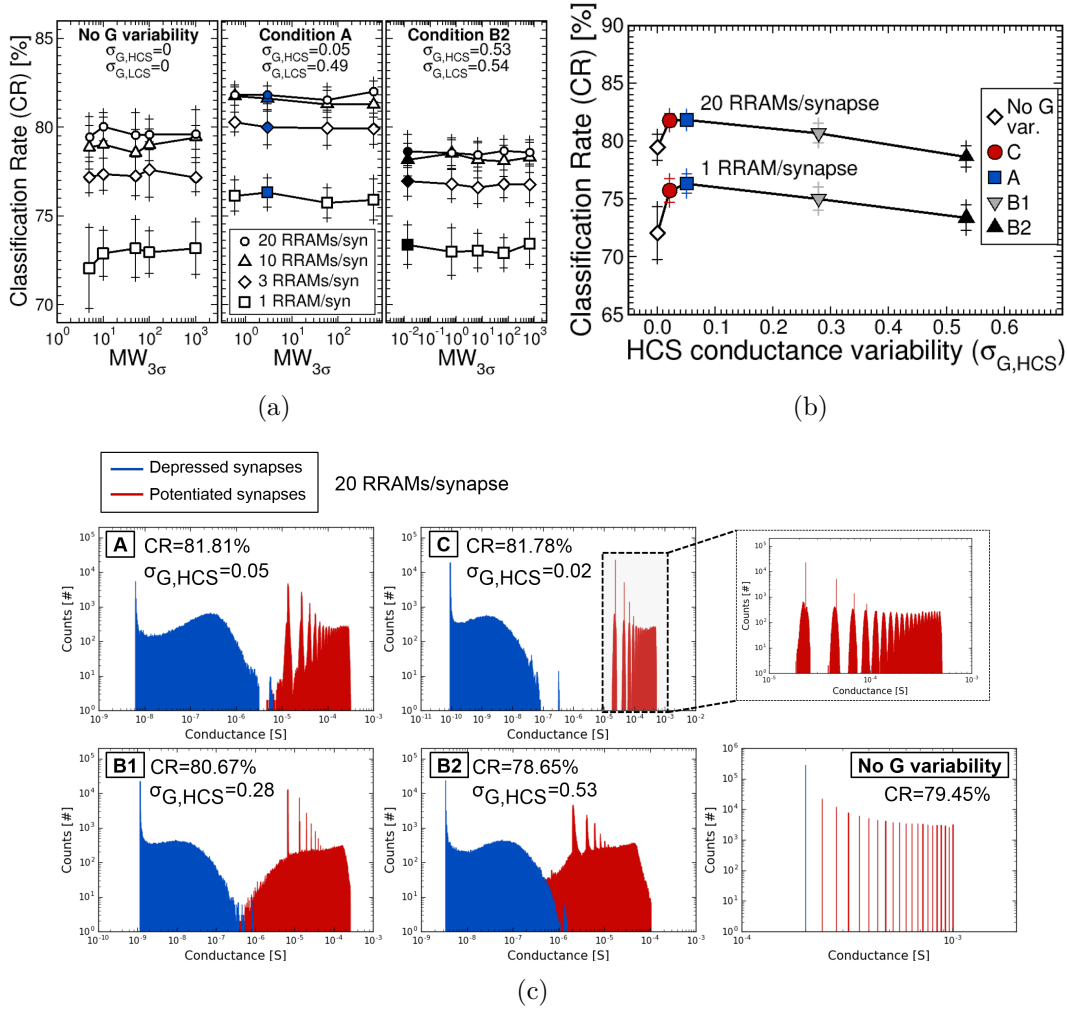


FIGURE 2.2.12: (a) Classification rate, CR, as a function of the memory window at 3σ , $MW_{3\sigma}$ defined in EQUATION 2.2.1, for different numbers of RRAMs per synapse. The HCS and LCS distributions measured on the 4-bit array for the conditions A and B2 of TABLE 2.1, and an artificial case with zero variability were used. The $MW_{3\sigma}$ was varied by a translation of the LCS distribution to lower or higher conductance values. (b) CR as a function of the conductance variability in HCS, $\sigma_{G,HCS}$, for 1 and 20 RRAMs per synapse. The four HCS and LCS distributions measured on the 4-kbit array for the conditions of TABLE 2.1 and an artificial case with zero variability were used. (c) Synaptic weight distributions after the learning phase for the four programming conditions of TABLE 2.1 and the synapse with zero variability. Higher performance after learning is reached with a small amount of HCS conductance variability (conditions A and C).

and B2 for learning. However, the low programming endurance of condition C can be detrimental.

Comparison between detection and classification tasks

We now explain the surprising different in how detection and classification tasks are affected by device characteristics. For the detection task, binary synapses are sufficient, and the maximal F1-score (0.96) is reached if, after learning, there are two synaptic populations: (i) potentiated synapses (RRAMs in HCS), and

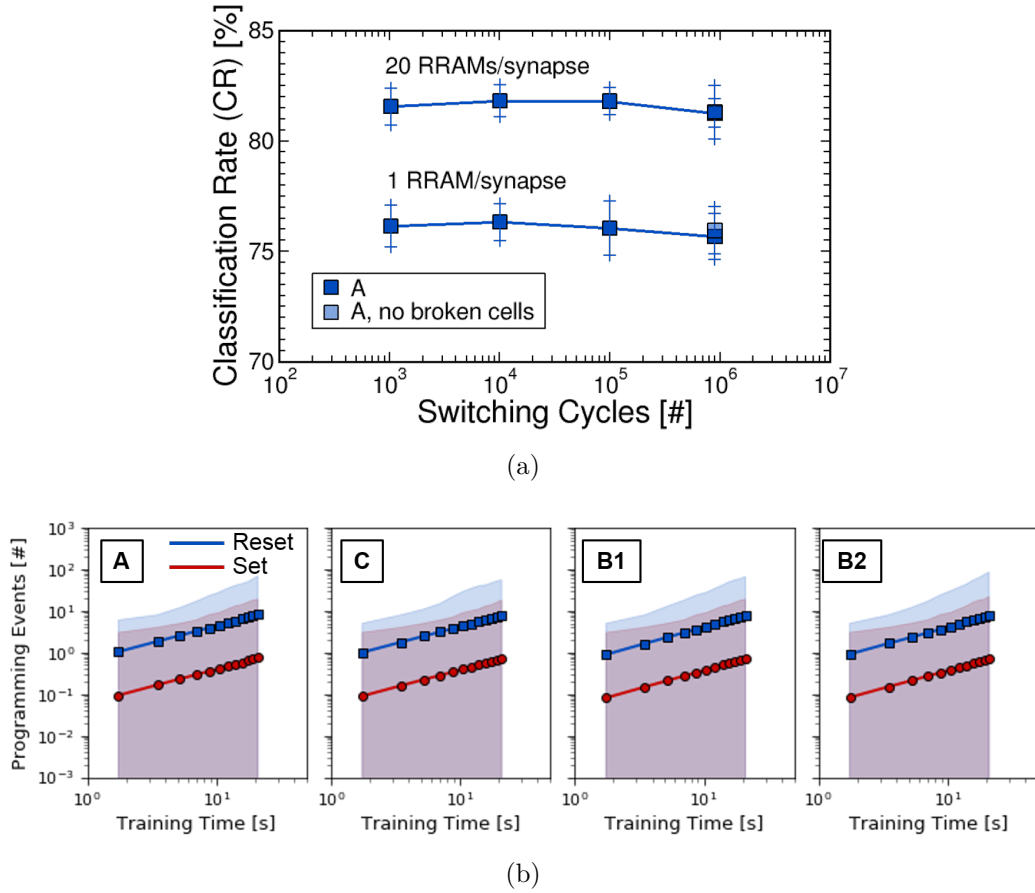


FIGURE 2.2.13: Impact of the RRAM aging on the classification rate, CR. Simulations have been calibrated using the data of FIGURE 2.2.3 (a) (condition A). Both device-to-device and cycle-to-cycle variability are taken into account. (b) Average number of Set (red circle) and Reset (blue square) operations during the learning phase for each programming condition. Red and blue shaded areas represent the evolution at $\pm 3\sigma$, respectively.

(ii) depressed synapses (RRAMs in LCS) (see FIGURE 2.2.10 (c) for conditions C and B1). The fundamental requirement is that a ratio higher than 200 exists between the peaks of the potentiated (HCS) and depressed (LCS) synaptic distributions. Therefore, both memory window and conductance variability are beneficial as they increase the dynamic range of synaptic weight values available during learning. This facilitates the separation of the HCS and LCS peaks after learning. For the classification task, multi-level conductance synapses are necessary to achieve the best performance. The number of RRAM cells per synapse defines the number of levels. As parallel conductances sum up, the equivalent synaptic weight is approximately $n_{HCS}HCS_{0\sigma}$, where n_{HCS} is the number of RRAMs in HCS, and $HCS_{0\sigma}$ is the median HCS conductance value. Unlike the detection task wherein the network exploits both HCS and LCS distributions, only the HCS distribution defines the synaptic weight value for the classification task. Consequently, the classification task is only sensitive to the HCS distribution; the LCS distribution and the memory window do not affect the network performance. To support this statement, we performed the same simulations as in FIGURE 2.2.12 (b) but this time with no variability in the LCS

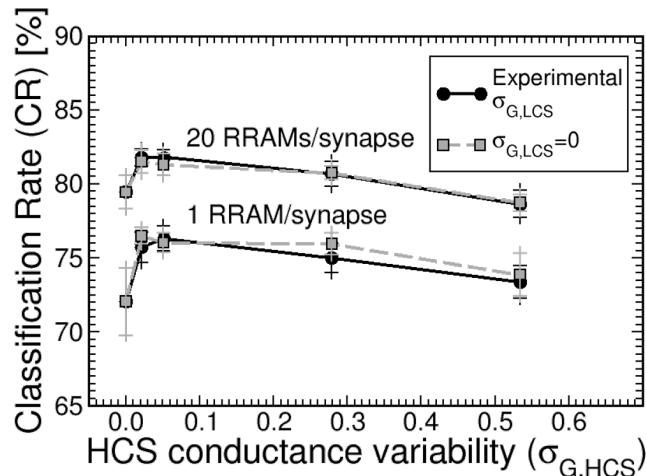


FIGURE 2.2.14: Classification rate, CR, as a function of the HCS conductance variability, $\sigma_{G,HCS}$, for the four programming conditions of TABLE 2.1. The CR has been calculated assuming the experimental LCS conductance variability (black line) and no LCS conductance variability (dotted grey line).

distribution, *i.e.* same $\sigma_{G,HCS}$ and $\sigma_{G,LCS}=0$ for each of the four experimental conditions (A, B1, B2, and C). The results are shown in FIGURE 2.2.14. We obtained similar network performance with (solid black line) and without (dotted grey line) LCS variability proving that only the HCS distribution affects the network performance. With the presence of a small amount of HCS variability ($\sigma_{G,HCS}\approx 0.05$, conditions A and C), a finer tuning of the synaptic weights is possible improving on the case with zero variability (see FIGURE 2.2.12 (c)). If the HCS variability is too high (condition B2), the synaptic weight distribution after the learning phase has only 7 distinguishable synaptic levels instead of the 21 levels achieved with conditions A and C as we showed in FIGURE 2.2.12 (c). TABLE 2.2 summarises learning performance and power required for learning for each programming condition, for the detection and classification tasks. Learning power has been calculated as:

$$\text{Learning power} = \frac{E_{\text{Set}} * \text{total Set pulses} + E_{\text{Reset}} * \text{total Reset pulses}}{\text{Learning time}} \quad (2.2.5)$$

with E_{Set} and E_{Reset} calculated in EQUATION 2.2.3, *total Set pulses* and *total Reset pulses* obtained by simulations in FIGURE 2.2.11 (b) and 2.2.13 (b) for the detection and the classification tasks, respectively, and *Learning time* defined as the minimal required time for the learning phase (simulated in FIGURE 2.2.10 (d) for the detection task and fixed at 21 s (60 000 training digits) for the classification task).

2.2.3 Conclusion

In this section, an extensive study of the conductance variability, power consumption, and aging of multi-kilobits RRAM array over the full operation range has been presented. The experimental results were used to perform system-level simulations of SNNs designed for (i) detection in dynamic patterns and (ii)

Condition		A	C	B1	B2
Detection Task (1 RRAM/synapse)	Performance (F1-score)	0.95	0.96	0.96	0.89
	Learning power [μW]	6.05	12.92	1.55	0.35
Classification Task (20 RRAMs/synapse)	Performance (CR) [%]	81.81	81.78	80.67	78.65
	Learning power [μW]	207.66	395.41	45.69	15.99

TABLE 2.2: Performance and power required for learning with each programming condition of TABLE 2.1, for the detection and classification tasks. The learning power has been calculated with EQUATION 2.2.5.

classification of static patterns applications. In comparison with previous studies [25, 28, 29, 33–35, 37, 41, 42, 45–47, 49, 52–54, 79–82], we demonstrate that SNNs are not only robust to synaptic variability but can also draw benefit from it. Variability can be beneficial as it increases the range of synaptic weight values available during learning. For detection applications, RRAM technology is well-suited to implement synaptic elements as only one RRAM device per synapse is needed (*i.e.* binary synapse with an abrupt switching between the HCS and LCS is sufficient), and their electrical characteristics enable to achieve maximal performance at low programming power consumption (less than 15 pJ/spike and 2 μW for learning). On the other hand, for classification applications, multi-level conductance synapses are necessary to achieve the best performance; a synaptic compound of at least ten RRAMs per synapse is required. The maximal performance was reached with a conductance variability in the HCS of roughly 0.05 that can be achieved with programming energy of about 50 pJ/spike. This study provides guidelines to optimise the programming conditions for RRAM-based synapses in SNNs capable of unsupervised learning by STDP. More importantly, it also highlights that memory devices for neuromorphic applications may be more optimally used in different physical regimes than for conventional memory applications and that RRAM requirements differ for memory and neuromorphic applications.

2.3 Analog devices

2.3.1 Goal of the section

As we evidenced in the previous section, the use of binary devices as synaptic elements can be detrimental for some applications [95], such as the classification task on the MNIST dataset (*cf* FIGURE 2.2.12 (a)). To overcome this issue, a synaptic compound of multiple binary devices operating in parallel associated with a stochastic programming [96] can be adopted as shown in the previous section. However, this increases silicon area consumption. Note that the proposed stochastic programming can be replaced by an incremental programming as demonstrated in [56]. A more efficient way is to implement the synaptic elements with single analog devices, *i.e.* devices intrinsically capable of multi-level conductance. However, such devices usually feature non-linear and asymmetric conductance response upon the application of identical pulses [30, 36, 48, 85] which are generally considered as non-idealities for neuromorphic computing

[28, 28, 34, 35, 48, 80, 83, 85]. Although many comprehensive studies on the impact of conductance response on learning performance have been reported, most of them are based on supervised learning algorithms [28, 34, 35, 48, 80, 83–85] like the gradient-descent back-propagation algorithm [67], and little has been done with unsupervised algorithms in spiking neural networks [33, 81, 82, 86]. In this section, we focus on analog devices to implement the synaptic elements. We investigate the impact of conductance response on learning performance of artificial Spiking Neural Network (SNN) systems with unsupervised learning by Spike-Timing-Dependent Plasticity (STDP). Learning performance is evaluated by means of system-level simulations of the handwritten digit classification task on the MNIST dataset [69]. We also evaluate the learning performance of the Phase-Change Memory (PCM) technology presented in [36] as a synaptic device, and we compare its performance with previously reported PCM-based synaptic implementations. The next section provides a brief overview of reported artificial synapses capable of analog conductance modulation.

2.3.2 Analog conductance modulation with non-volatile resistance-based memories

RRAM technologies have demonstrated multi-level conductance capability: when consecutive Set (potentiation) or Reset (depression) pulses are applied on the device, the conductance of the device gradually increases or decreases, respectively [28, 33–35, 48, 53, 54, 56, 80, 82, 86, 98]. A simple method to obtain analog conductance evolution with RRAMs is by increasing the compliance current during potentiation [23, 97–99] (*cf* FIGURE 2.3.1 (a)): when a synapse undergoes a series of potentiation events, the compliance current is increased from pulse to pulse. However, this brings circuitry overhead as the system needs to keep track of the history of each synapse. Optimally, analog modulation has to be obtained under *identical potentiation and depression pulses*. Unfortunately, most of RRAM technologies are intrinsically binary devices [42, 43, 45, 60, 100, 101] or gradual only in one programming direction [34, 37, 48] - generally in depression - as we demonstrated in SECTION 2.2. Phase-Change Memory (PCM) technology has attracted strong interest to implement electronic synapses [34, 35, 37, 40, 95, 102–104] due to its technological maturity [81, 105, 106]. One of the prominent features of PCM technology as artificial synapses is the gradual crystallisation process of the phase-change material: when a series of identical short Set pulses are applied on the PCM cell, only a small amount of the material is crystallised. This increases the device conductance in an analog fashion [34, 36, 37]. By contrast, the amorphisation is an abrupt process which is a critical limitation of PCM technology as artificial synapses. To overcome this issue, a stair-case programming scheme can be adopted wherein the amplitude of Set and Reset pulses is increased from pulse to pulse [103, 104] - similar to the use of increasing compliance currents for RRAMs - at the cost of increased circuit complexity. Another solution proposed by Suri et al. [37] is to implement one synaptic element with two PCM devices and benefit from the gradual crystallisation process in both potentiation and depression. This is the so-called *2-PCM synapse* that has been extensively used in neuromorphic systems [34, 37, 38, 107] (*cf*

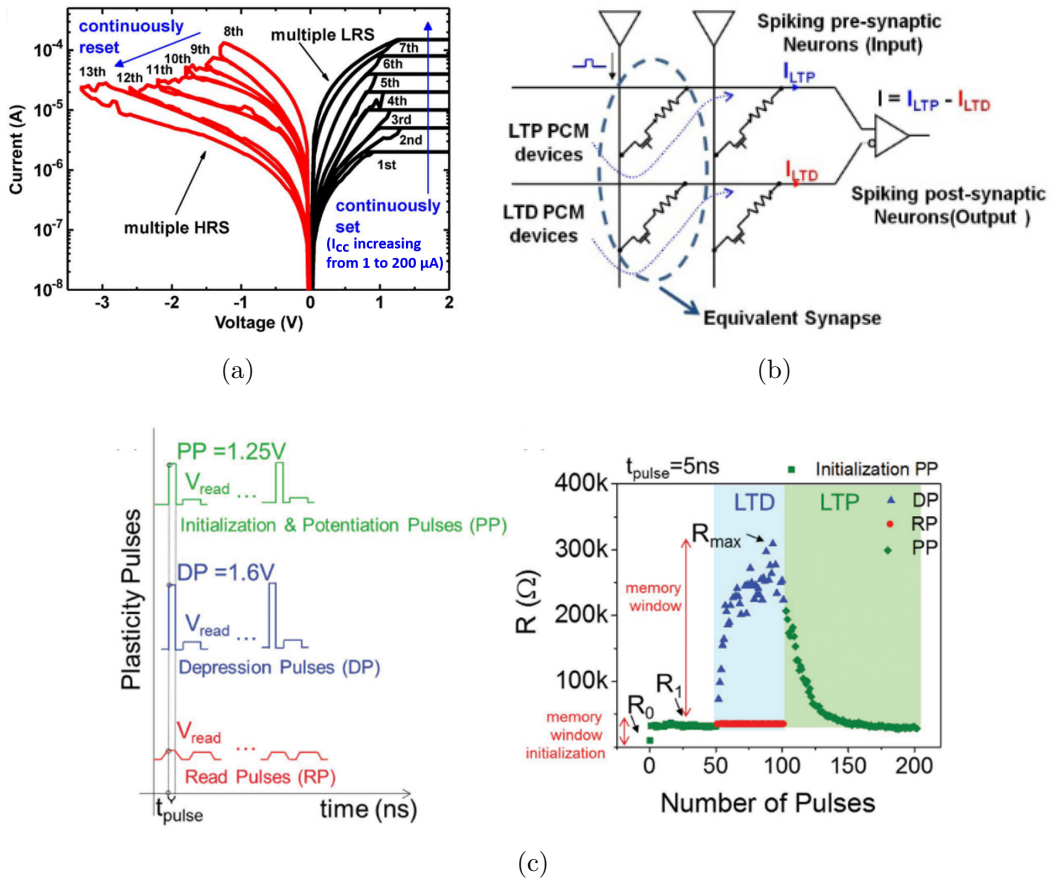


FIGURE 2.3.1: (a) I-V characteristics of a TiN/HfO_x/AlO_x/Pt RRAM device with increasing compliance current for each Set operation. The device conductance value (represented by the read current) increases with the compliance current. Reproduced from [97]. (b) 2-PCM synapse implementation. One synapse is implemented with two PCM devices in parallel (LTP and LTD PCMs) with opposite current contributions. Analog conductance modulation is obtained by exploiting the gradual crystallisation of both PCMs. Reproduced from [37]. (c) Programming strategy proposed in [36] to obtain analog conductance modulation with a single PCM device. In addition to the gradual crystallisation of PCM (potentiation, green area), initialising the PCM at an intermediate resistance value ($R_1 \approx 30$ k Ω) enables gradual amorphisation (depression, blue area) with short Reset pulses (<50 ns). Reproduced from [36].

FIGURE 2.3.1 (b)). The principle of the 2-PCM synapse is that one PCM device has a positive current contribution (LTP PCM), while the other one contributes negatively (LTD PCM) towards the post-synaptic neuron. As a result, a gradual crystallisation of the LTP PCM cell produces an *analog increase* of the synaptic weight, whereas a gradual crystallisation of the LTD cell produces an *analog decrease*. However, this halves synaptic density since two cells are required per synapse. Furthermore, periodical energy-hungry refresh process is mandatory to prevent conductance saturation [34, 35, 107]: when one of the two PCMs reaches its full crystallisation state, a Reset operation is required. The PCM device proposed by La Barbera et al. [36] overcomes this limitation thanks to the use of a new programming strategy that does not involve circuitry overhead and refresh

process. In addition to the natural gradual crystallisation process, gradual amorphisation is possible by initialising the cell at an intermediate resistance state and by applying short Reset pulses (<50 ns) that only amorphise a small amount of material at each pulse as shown in FIGURE 2.3.1 (c). In this section, we will compare learning performance of the PCM technology presented in [36] with that of conventional 2-PCM synapses.

2.3.3 Learning rule and synapse behavioural model

We consider synapses capable of continuous increase and decrease of their synaptic weight with consecutive potentiation and depression events, respectively. Synapses are trained using a *simplified Spike-Timing-Dependent Plasticity* (STDP) rule as depicted in FIGURE 2.3.2 (a). When the post-synaptic neuron spikes after the pre-synaptic neuron within a time window t_{STDP} , the corresponding synapse increases its synaptic weight by a quantity δw_+ . Otherwise, it decreases its synaptic weight by a quantity δw_- . Note that there are no switching probabilities with the simplified STDP rule. To model the synaptic weight increment and decrement in our system-level simulations, δw_+ and δw_- , we use the model introduced in [108]:

$$\begin{aligned}\delta w_+ &= \alpha_+ \exp\left(-\beta_+ \frac{w - W_{\min}}{W_{\max} - W_{\min}}\right) \\ \delta w_- &= \alpha_- \exp\left(-\beta_- \frac{W_{\max} - w}{W_{\max} - W_{\min}}\right)\end{aligned}\tag{2.3.1}$$

These equations allow to reproduce the conductance response of real devices upon the applications of a series of *identical potentiation and depression pulses*, such as PCM devices as in [37] and [81]. α_+ , β_+ , α_- , β_- , W_{\min} , and W_{\max} are fitting parameters that control the dynamics of the conductance response and depend on the device technology. w corresponds to the current synaptic weight. W_{\min} and W_{\max} represent the minimum and maximum conductance values of the device. β_+ and β_- inside the exponential factor control the linearity of the conductance response. In fact, as observed in most RRAM and PCM technologies, a given programming pulse has a reduced effect on the device conductance if applied several times [30, 36, 48, 85]. α_+ and α_- control the resolution of the device, *i.e.* the number of Set or Reset pulses required to go from the minimum boundary W_{\min} to the maximum boundary W_{\max} , and vice-versa, respectively. These six parameters are subject to cycle-to-cycle and device-to-device variability in real devices [33, 36, 81]. However, we assume no variability on these parameters in this section.

In order to investigate the impact of conductance response on SNN learning performance, we define different metrics. We define the *number of potentiation levels*, n_{pot} , as the number of potentiation pulses required to increase the synaptic weight from W_{\min} to W_{\max} . Similarly, we define the *number of depression levels*, n_{dep} , as the number of depression pulses required to decrease the synaptic weight from W_{\max} to W_{\min} . We define the *linearity factor in potentiation* as the parameter β_+ . Similarly, we define the *linearity factor in depression* as the parameter β_- . In this section, we investigate the impact of the number

of potentiation and depression levels, n_{pot} and n_{dep} , and the impact of the linearity factors, β_+ and β_- , on SNN learning performance. FIGURE 2.3.2 (b) shows the impact of the linearity factors, β_+ and β_- , on the conductance response for a fixed number of potentiation and depression levels ($n_{\text{pot}}=n_{\text{dep}}=200$). 200 potentiation pulses followed by 200 depression pulses are applied. The conductance response is linear for $\beta_+=\beta_-=0$. By contrast, it is less and less linear with higher β_+ and β_- , *i.e.* a given potentiation or depression pulse has a reduced effect on the conductance increment or decrement when applied several times, respectively. FIGURE 2.3.2 (c) shows the conductance response for different numbers of potentiation and depression levels, n_{pot} and n_{dep} , for a non-linear device ($\beta_+=\beta_-=3$). 500 potentiation pulses followed by 500 depression pulses are applied.

In real devices, the conductance response in potentiation and depression is not always symmetric [34–36, 48, 56, 80, 84] just like the RRAM technology presented in SECTION 2.2.1.2 (*cf* FIGURE 2.2.5 (Left)). FIGURE 2.3.2 (d) shows the fitting of the PCM device presented in [36] with EQUATION 2.3.1 (grey line). The PCM device in [36] features an analog conductance response in both potentiation (red circle) and depression (blue square). However, as evidenced by FIGURE 2.3.2 (d), it is asymmetric in potentiation and depression in both linearity ($\beta_+=3$ and $\beta_-=1$) and number of levels ($n_{\text{pot}}=200$ and $n_{\text{dep}}=30$). The 2-PCM synapse implementation does not exhibit this asymmetry as its conductance response involves the crystallisation process of two identical PCM devices. In that sense, the 2-PCM synapse implementation can be viewed as symmetric in potentiation and depression if variability is not considered. In this section, the impact of asymmetry in terms of number of levels and linearity is also assessed.

2.3.4 Impact of the conductance response on spiking neural network learning performance

We now investigate the impact of analog device conductance response on a classification task with Spiking Neural Networks (SNNs) (the same classification task as in SECTION 2.2.2.3 based on the handwritten digits MNIST dataset [69]). The network performance is assessed by means of system-level simulations with the special-purpose neuromorphic hardware simulator N2D2 [73, 93]. The behavioural model of the synaptic elements presented in SECTION 2.3.3 has been used to evaluate the impact of conductance response on learning performance. In this section, we consider no variability on the fitting parameters α_+ , β_+ , α_- , β_- , W_{min} , and W_{max} , just as we do not consider any variability effects due to peripheral circuits (such as neuron variability [33, 81]). For all the simulations, we fixed the minimum and maximum conductance values at $W_{\text{min}}=1 \mu\text{S}$ and $W_{\text{max}}=30 \mu\text{S}$, respectively. These values have been extracted from the measurements of the PCM technology presented in [36].

2.3.4.1 Network topology

The simulated classification application is based on a one-layer fully-connected feed-forward neural network topology (*cf* FIGURE 2.3.3). A detailed description

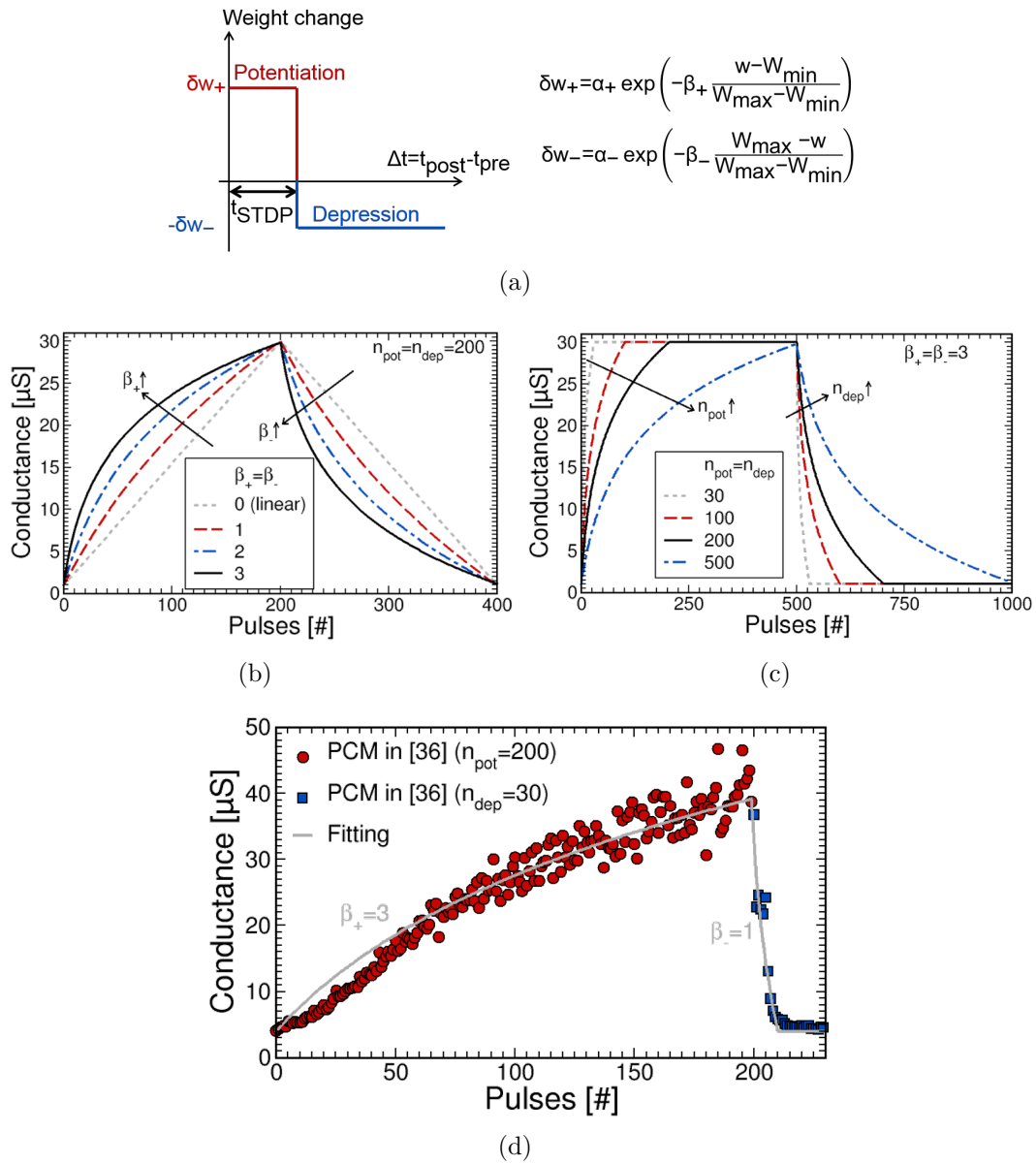


FIGURE 2.3.2: (a) Simplified STDP rule and synaptic weight increment and decrement equations. (b) Conductance response with the model described in EQUATION 2.3.1 for different linearity factors, β_+ and β_- . Potentiation and depression levels are fixed at 200 ($n_{\text{pot}} = n_{\text{dep}} = 200$). β_+ and β_- control the linearity of the conductance response. (c) Conductance response with the model described in EQUATION 2.3.1 for different numbers of potentiation and depression levels. Linearity factors are fixed at 3 ($\beta_+ = \beta_- = 3$). (d) Conductance response of the PCM technology presented in [36] (filled symbol) and fitting with EQUATION 2.3.1 (grey line).

of the simulated SNN is provided in APPENDIX A. Neurons in the output layer are implemented with the Leaky Integrate-and-Fire (LIF) model [73]. For each simulation, the output LIF neurons have the same firing threshold value which has been optimised to provide the best performance (see APPENDIX B). The other neuron parameters are kept constant and identical for all the simulations. The input layer corresponds to the 28x28 pixels input images and is fully connected to an output layer of n_{neurons} neurons. The Classification Rate

(CR) as defined in SECTION 2.2.2.1 is used to assess learning performance (see APPENDIX A). The network is trained with the unsupervised simplified STDP rule and lateral inhibition [33, 73]. All results have been averaged over five simulations.

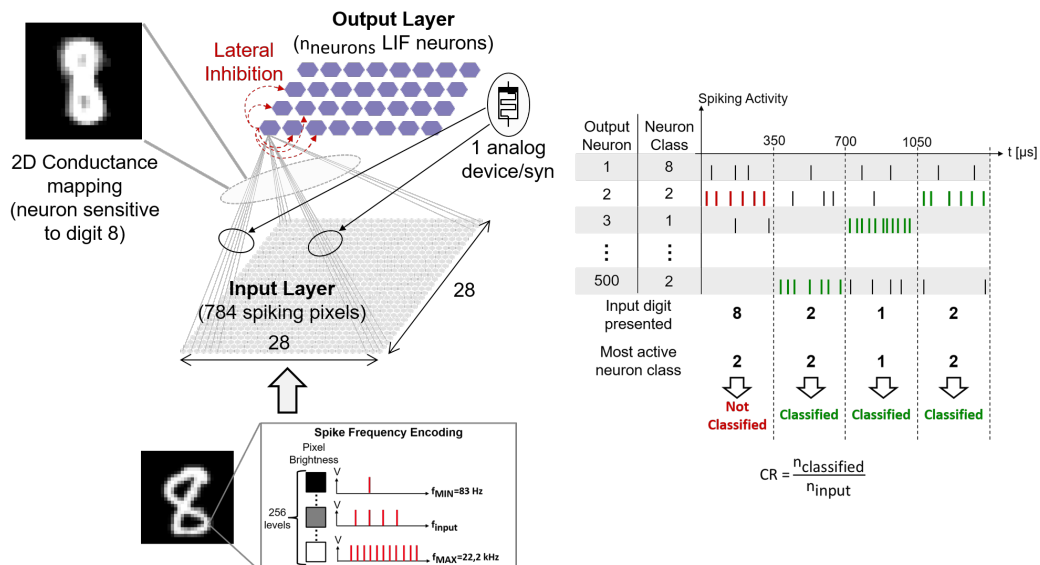


FIGURE 2.3.3: Simulated spiking neural networks used for the digit classification task with analog devices as synaptic elements. The associated score definition to assess network performance is shown on the right-hand side. See APPENDIX A for more details.

2.3.4.2 Impact of the number of potentiation and depression levels

We first studied the impact of the number of potentiation and depression levels, n_{pot} and n_{dep} , on learning performance. We consider the case of a non-linear, symmetric conductance response with linearity factors fixed at $\beta_+ = \beta_- = 3$ and $W_{\text{min}} = 1 \mu\text{S}$ and $W_{\text{max}} = 30 \mu\text{S}$ (*cf* the conductance responses in FIGURE 2.3.2 (c)). The first step was to optimise the network in terms of number of output neurons, n_{neurons} . As demonstrated by several works [33, 74, 81, 86], network performance improves with the number of output neurons as it allows the network to be sensitive to different handwritings of the same digits [33]. FIGURE 2.3.4 (a) shows the simulated classification rate, CR, as a function of the number of output neurons, n_{neurons} , for different numbers of potentiation and depression levels. In agreement with previous studies, the CR improves with an increasing number of output neurons and saturates after 500 output neurons. In the following, the proposed network is simulated with 500 output neurons.

From the results of FIGURE 2.3.4 (a), it can be seen that network performance also improves with an increasing number of potentiation and depression levels. This is in agreement with our study in SECTION 2.2.2.3. The CR increases with n_{pot} and n_{dep} , and it saturates after $n_{\text{pot}} = n_{\text{dep}} = 200$ levels. Note that these results are obtained for a symmetric conductance response, *i.e.* $n_{\text{pot}} = n_{\text{dep}}$. We then assessed the impact of an asymmetry in the number of potentiation and depression levels on network performance. We fixed the number of potentiation

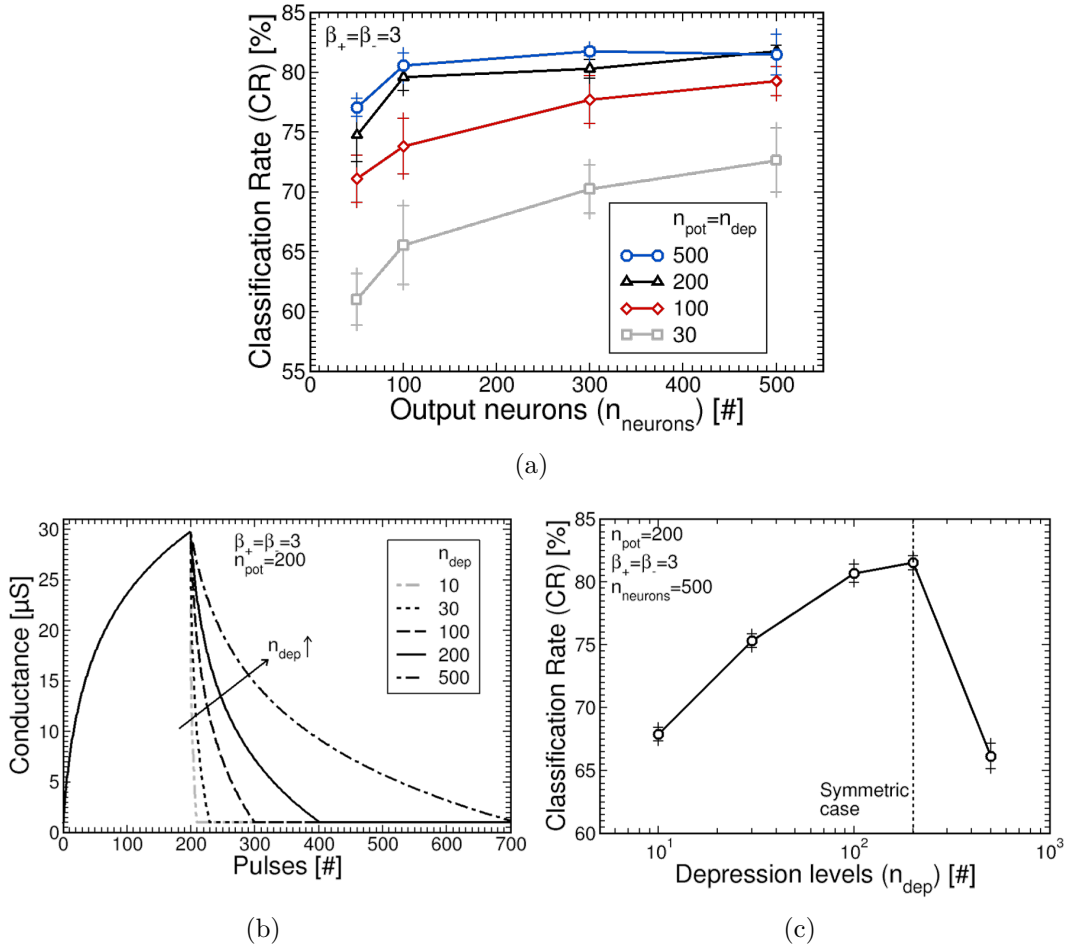


FIGURE 2.3.4: (a) Classification rate, CR, as a function of the number of output neurons, n_{neurons} . The conductance responses in FIGURE 2.3.2 (c) have been used for the simulations. (b) Conductance responses with EQUATION 2.3.1 for a fixed number of potentiation levels $n_{\text{pot}}=200$ and different numbers of depression levels, n_{dep} . (c) CR as a function of the number of depression levels, n_{dep} , for a fixed $n_{\text{pot}}=200$ (*cf* (b)).

levels at 200 levels ($n_{\text{pot}}=200$), $\beta_+ = \beta_- = 3$, and varied the number of depression levels, n_{dep} . FIGURE 2.3.4 (b) shows the simulated conductance responses of the synaptic elements. FIGURE 2.3.4 (c) shows the obtained performance for each conductance response, *i.e.* the CR as a function of the number of depression levels, n_{dep} , for a given $n_{\text{pot}}=200$ and 500 output neurons. Network performance is degraded when there is an asymmetry between potentiation and depression in terms of number of levels. Yet it can be noted that the network can accommodate a certain degree of asymmetry between the number of levels (CR=80.69% for $n_{\text{pot}}=200$ and $n_{\text{dep}}=100$ compared to CR=81.53% for $n_{\text{pot}}=n_{\text{dep}}=200$).

2.3.4.3 Impact of the linearity

We then assessed the impact of the linearity factors, β_+ and β_- , on network performance. We fixed the number of output neurons at $n_{\text{neurons}}=500$. In addition, the number of potentiation and depression levels, n_{pot} and n_{dep} , are fixed at 200 and 30, respectively. These numbers of levels correspond to the

experimental measurements obtained with the analog PCM presented in [36] (CF FIGURE 2.3.2 (d)). FIGURE 2.3.5 (Top) plots the CR as a function of the linearity factor in potentiation, β_+ . Open symbols correspond to symmetric conductance responses in linearity ($\beta_+=\beta_-$). The filled symbol corresponds to the PCM technology in [36] ($\beta_+=3$ and $\beta_-=1$). In this SNN based on unsupervised learning with STDP, the non-linearity of the conductance response remarkably improves system performance with respect to the linear case (CR=75.33% for $\beta_+=\beta_-=3$ compared to CR=68.60% for $\beta_+=\beta_-=0$). This is in sharp contrast with other types of neural networks based on supervised learning (*e.g.* based on back-propagation algorithm) wherein the non-linearity is strongly detrimental [28, 28, 34, 35, 48, 80, 83, 85]. To illustrate the origin of this result, we plot in FIGURE 2.3.5 (Bottom) the synaptic weight evolution of one hundred different synapses during the learning phase. Synapses with linear characteristics ($\beta_+=\beta_-=0$) converge to the minimum and maximum conductance values (FIGURE 2.3.5 (Bottom left)) after learning, whereas synapses with non-linear characteristics can also converge to intermediate conductance values between the minimum and maximum values (FIGURE 2.3.5 (Bottom right)). Therefore, they are able to fully exploit the analog synaptic behaviour. This result is well known in the literature [109, 110] and originates from the stabilising effect of *weight-dependent plasticity arising from the non-linearity*. With the STDP learning rule, strong synapses have a larger probability of being potentiated since the probability to induce a post-synaptic spike increases with the weight of the synapse. This is a *destabilising force* that makes strong synapses be more and more potentiated. However, stronger synapses experience smaller conductance increases at each potentiation event due to their non-linear conductance responses. Thus, the effect of potentiation events decreases with increasing weights, whereas the effect of depression events does not. This leads to a *stabilising force* that counteracts the destabilising force of the STDP learning rule and allows synaptic weights to remain at intermediate values. The same reasoning holds true for weak synapses: weak synapses have a larger probability of undergoing depression events, but they experience reduced conductance decreases at each event. In the case of linear conductance responses, conductance changes at each event are not weight-dependent. Therefore, the destabilising force of the STDP learning rule pushes synaptic weights to minimum and maximum boundaries. Finally, it is also worth noting that network performance is not affected by an asymmetry in linearity (CR=75.33% for $\beta_+=\beta_-=3$ compared to CR=76.24% for the experimental case with $\beta_+=3$ and $\beta_-=1$).

2.3.5 Discussion

In this section, the impact of conductance response on Spiking Neural Network (SNN) learning performance trained with the unsupervised Spike-Timing-Dependent Plasticity (STDP) learning paradigm was studied. While many studies on the impact of conductance response for neural networks trained with supervised algorithm have been reported [28, 34, 35, 48, 80, 83–85], only a few is available for neural networks trained with unsupervised algorithms [33, 81, 82, 86]. For neural networks trained with supervised algorithms, it is now widely accepted

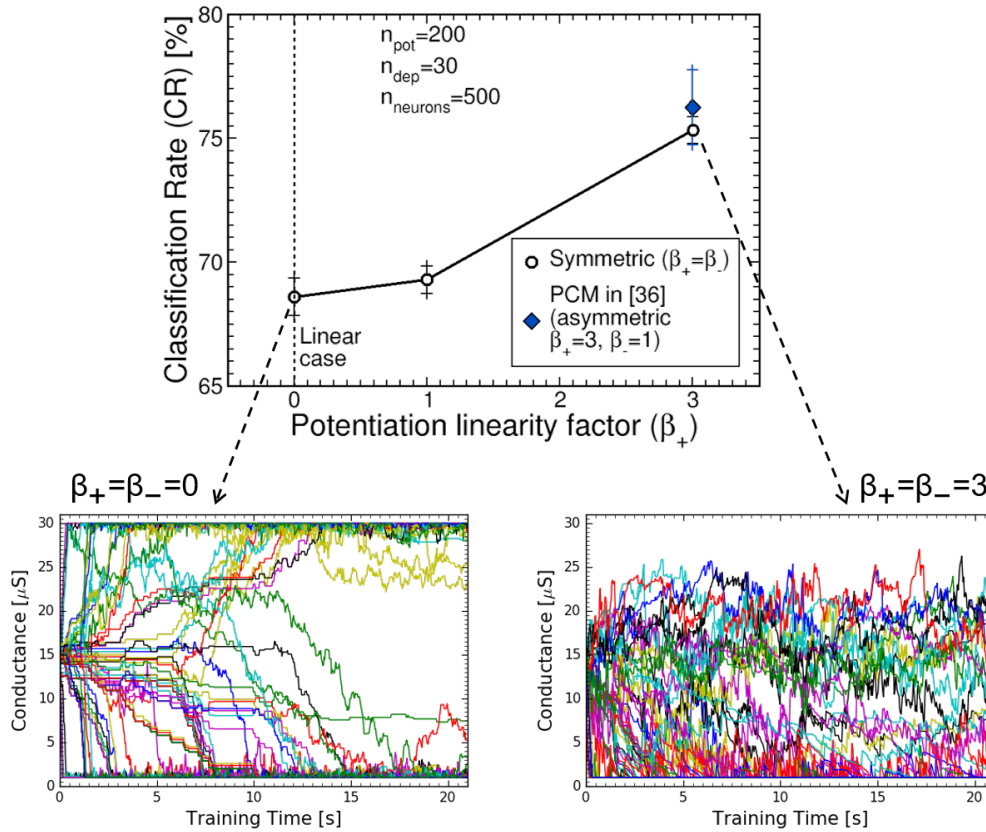


FIGURE 2.3.5: (Top) Classification Rate, CR, as a function of the linearity factor in potentiation, β_+ , for 200 potentiation levels and 30 depression levels. Open symbols correspond to symmetric responses in linearity ($\beta_+=\beta_-$). The filled symbol corresponds to an asymmetric case ($\beta_+=3$ and $\beta_-=1$) fitted with the PCM technology presented in [36] (*cf* FIGURE 2.3.2 (d)). (Bottom) Synaptic weight evolution of 100 synapses during the training phase for (Left) a linear conductance response and (Right) a non-linear conductance response.

that the ideal synapse has a linear, symmetric conductance response with a high number of levels in potentiation and depression [28, 28, 34, 35, 48, 80, 83, 85]. For SNNs trained with an unsupervised simplified STDP rule, the ideal synapse also needs a symmetric conductance response with a high number of levels in potentiation and depression - at least 200 levels in this study (*cf* FIGURE 2.3.4 (a) and (c)), although this number is application-dependent and might be lower or higher for other applications. However, network performance improves with a non-linear conductance response (*cf* FIGURE 2.3.5 (Top)). This is in sharp contrast with neural networks trained with supervised algorithms; SNNs trained with unsupervised algorithms benefit from the natural non-linearity of many technologies [30, 36, 48, 85], such as the PCM technology presented in [36] and used in this study. This comes from the fact that non-linear synapses are able to exploit the full range of synaptic weight values during the learning phase (*cf* FIGURE 2.3.5 (Bottom right)), whereas linear synapses tend to converge to the lower and upper boundaries (*cf* FIGURE 2.3.5 (Bottom left)). This is a significant advantage of unsupervised over supervised learning upon employing PCM-based synapses since it avoids the use of complex synaptic design and algorithm to mitigate PCM non-linearity - for instance in [35] wherein synaptic

elements are implemented with eight transistors, two PCM devices, and two capacitors.

As evidenced by the study, an asymmetry in the number of levels in potentiation and depression can be detrimental for network learning performance (*cf* FIGURE 2.3.4 (c)). Since the 2-PCM synaptic implementation in [37] exploits the crystallisation process of two identical PCM devices for gradual potentiation and depression, it does not suffer from this asymmetry at the cost of a reduced synaptic density and the need of energy-hungry refresh process [34, 52]. Considering 200 levels of potentiation and depression ($n_{\text{pot}}=n_{\text{dep}}=200$), a classification rate of 81.53% can be reached (FIGURE 2.3.4 (c)) with synaptic elements implemented with the 2-PCM scheme. By contrast, the asymmetry of the analog PCM technology of [36] ($n_{\text{pot}}=200$ and $n_{\text{dep}}=30$) deteriorates network performance, and the classification rate degrades down to 76.24%. To mitigate this asymmetry, it is necessary to balance the weight increments and decrements to virtually symmetrise the numbers of levels. One demonstration is the fully-connected neural network implemented by Fumarola et al. in [48] with analog Al/Mo/PCMO RRAMs as synaptic elements and trained with the supervised back-propagation algorithm. They showed that the asymmetry between potentiation and depression can be corrected by defining two different learning rates for Set and Reset operations. As a result, the output neuron fires more pulses when performing a Set operation with respect to Reset operations which allows to lessen the asymmetry. In a similar manner, the network implemented in [56] with a synaptic compound of several RRAMs per synapse utilises additional potentiation and depression counters in order to modulate the frequency of weight increase and decrease. This allows to compensate for asymmetric conductance changes and could be implemented in our network. Another solution would be to replace the simplified STDP rule used in this section with the stochastic STDP rule presented in SECTION 2.2.1.2 (see FIGURE 2.2.6 (b)). By doing so, the frequency of weight increase and decrease can be controlled by the switching probabilities. Another example is the work in [84] wherein they accumulate weight increments or decrements before a potentiation or depression event actually occurs, respectively. When the accumulation reaches a certain threshold, the system undergoes a potentiation or a depression event. By defining two different thresholds for weight increment and decrement, this allows to cope with asymmetric conductance changes. However, all the solutions presented come at the expense of peripheral circuitry overhead. Finally, we did not consider in this study any variability in the fitting parameters of EQUATION 2.3.1. Nonetheless, this variability would have a limited impact on network performance. Indeed, it has been demonstrated in [81] that the SNN used in this study is robust to more than 25% of variability in the fitting parameters.

References: Chapter 2

- [1] Yosef Yarom and Jorn Hounsgaard. “Voltage fluctuations in neurons: Signal or noise?”. *Physiological Reviews*, 91(3):917–929, 2011. ISSN 00319333. doi: 10.1152/physrev.00019.2010.
- [2] A. Aldo Faisal, Luc P.J. Selen, and Daniel M. Wolpert. “Noise in the nervous system”. *Nature Reviews Neuroscience*, 9(4):292–303, apr 2008. ISSN 1471003X. doi: 10.1038/nrn2258.
- [3] Neal A. Hessler, Aneil M. Shirke, and Roberto Malinow. “The probability of transmitter release at a mammalian central synapse”. *Nature*, 366(6455):569–572, 1993. ISSN 00280836. doi: 10.1038/366569a0.
- [4] Silvio O. Rizzoli and William J. Betz. “Synaptic vesicle pools”. *Nature Reviews Neuroscience*, 6(1):57–69, jan 2005. ISSN 1471003X. doi: 10.1038/nrn1583.
- [5] J. Gerard G Borst. “The low synaptic release probability in vivo”. *Trends in Neurosciences*, 33(6):259–266, jun 2010. ISSN 01662236. doi: 10.1016/j.tins.2010.03.003.
- [6] Rasmus M. Birn. “The Behavioral Significance of Spontaneous Fluctuations in Brain Activity”. *Neuron*, 56(1):8–9, oct 2007. ISSN 08966273. doi: 10.1016/j.neuron.2007.09.021.
- [7] Michael D. Fox, Abraham Z. Snyder, Justin L. Vincent, and Marcus E. Raichle. “Intrinsic Fluctuations within Cortical Systems Account for Intertrial Variability in Human Behavior”. *Neuron*, 56(1):171–184, oct 2007. ISSN 08966273. doi: 10.1016/j.neuron.2007.08.023.
- [8] Peter G.H. Clarke. “The limits of brain determinacy”. *Proceedings of the Royal Society B: Biological Sciences*, 279(1734):1665–1674, may 2012. ISSN 14712970. doi: 10.1098/rspb.2011.2629.
- [9] Anthony Randal McIntosh, Natasa Kovacevic, and Roxane J. Itier. “Increased brain signal variability accompanies lower behavioral variability in development”. *PLoS Computational Biology*, 4(7), jul 2008. ISSN 1553734X. doi: 10.1371/journal.pcbi.1000106.
- [10] John J Hopfield and D. W. Tank. “”Neural” computation of decisions in optimization problems”. *Biological Cybernetics*, 52(3):141–152, 1985. ISSN 03401200. doi: 10.1007/BF00339943.
- [11] Mark D. McDonnell and Derek Abbott. “What is stochastic resonance? Definitions, misconceptions, debates, and its relevance to biology”. *PLoS Computational Biology*, 5(5), may 2009. ISSN 1553734X. doi: 10.1371/journal.pcbi.1000348.
- [12] G. Bard Ermentrout, Roberto F. Galán, and Nathaniel N. Urban. “Reliability, synchrony and noise”. *Trends in Neurosciences*, 31(8):428–434, aug 2008. ISSN 01662236. doi: 10.1016/j.tins.2008.06.002.

- [13] David C. Knill and Alexandre Pouget. “The Bayesian brain: The role of uncertainty in neural coding and computation”. *Trends in Neurosciences*, 27(12):712–719, 2004. ISSN 01662236. doi: 10.1016/j.tins.2004.10.007.
- [14] Edwin R Lewis, Kenneth R Henry, and Walter M Yamada. “Essential roles of noise in neural coding and in studies of neural coding”. *BioSystems*, 58(1):109–115, 2000. ISSN 03032647. doi: 10.1016/S0303-2647(00)00113-1.
- [15] Christina Allen and Charles F Stevens. “An evaluation of causes for unreliability of synaptic transmission”. *Proceedings of the National Academy of Sciences of the United States of America*, 91(22):10380–10383, 1994. ISSN 00278424. doi: 10.1073/pnas.91.22.10380.
- [16] Mark S Goldman. “Enhancement of information transmission efficiency by synaptic failures”. *Neural Computation*, 16(6):1137–1162, 2004. ISSN 08997667. doi: 10.1162/089976604773717568.
- [17] Stefan Habenschuss, Zeno Jonke, and Wolfgang Maass. “Stochastic Computations in Cortical Microcircuit Models”. *PLoS Computational Biology*, 9(11), nov 2013. ISSN 1553734X. doi: 10.1371/journal.pcbi.1003311.
- [18] Uri Rokni, Andrew G. Richardson, Emilio Bizzi, and H. Sebastian Seung. “Motor Learning with Unstable Neural Representations”. *Neuron*, 54(4):653–666, may 2007. ISSN 08966273. doi: 10.1016/j.neuron.2007.04.030.
- [19] G. Indiveri, F. Corradi, and N. Qiao. “Neuromorphic architectures for spiking deep neural networks”. In *2015 IEEE International Electron Devices Meeting (IEDM)*, pages 4.2.1–4.2.4. IEEE, 2015. doi: 10.1109/IEDM.2015.7409623.
- [20] Alice Mizrahi, Tifenn Hirtzlin, et al. “Neural-like computing with populations of superparamagnetic basis functions”. *Nature Communications*, 9(1), dec 2018. ISSN 20411723. doi: 10.1038/s41467-018-03963-w.
- [21] Jingrui Wang and Fei Zhuge. “Memristive Synapses for Brain-Inspired Computing”. *Advanced Materials Technologies*, 4(3):1800544, mar 2019. ISSN 2365709X. doi: 10.1002/admt.201800544.
- [22] Shimeng Yu. “Neuro-Inspired Computing with Emerging Nonvolatile Memory”. *Proceedings of the IEEE*, 106(2):260–285, feb 2018. ISSN 00189219. doi: 10.1109/JPROC.2018.2790840.
- [23] Duygu Kuzum, Shimeng Yu, and H. S. Philip Wong. “Synaptic electronics: Materials, devices and applications”. *Nanotechnology*, 24(38):382001, sep 2013. ISSN 09574484. doi: 10.1088/0957-4484/24/38/382001.
- [24] Changhyuck Sung, Hyunsang Hwang, and In Kyeong Yoo. “Perspective: A review on memristive hardware for neuromorphic computation”. *Journal of Applied Physics*, 124(15):151903, oct 2018. ISSN 10897550. doi: 10.1063/1.5037835.
- [25] Sungmin Hwang, Hyungjin Kim, et al. “System-level simulation of hardware spiking neural network based on synaptic transistors and if neuron circuits”. *IEEE Electron Device Letters*, 39(9):1441–1444, sep 2018. ISSN 07413106. doi: 10.1109/LED.2018.2853635.
- [26] Chiara Bartolozzi and Giacomo Indiveri. “Synaptic Dynamics in Analog VLSI”. *Neural Computation*, 19(10):2581–2603, 2007. doi: doi:10.1162/neco.2007.19.10.2581.
- [27] Sung Yun Woo, Kyu Bong Choi, et al. “Synaptic device using a floating fin-body MOSFET with memory functionality for neural network”. *Solid-State Electronics*, 156: 23–27, jun 2019. ISSN 00381101. doi: 10.1016/j.sse.2019.02.011.

-
- [28] Suhwan Lim, Jong Ho Bae, et al. “Hardware-based Neural Networks using a Gated Schottky Diode as a Synapse Device”. In *Proceedings - IEEE International Symposium on Circuits and Systems*, volume 2018-May, 2018. ISBN 9781538648810. doi: 10.1109/ISCAS.2018.8351152.
- [29] Mohammad Mahvash and Alice C. Parker. “Synaptic variability in a cortical neuromorphic circuit”. *IEEE Transactions on Neural Networks and Learning Systems*, 24(3): 397–409, 2013. ISSN 2162237X. doi: 10.1109/TNNLS.2012.2231879.
- [30] Sungho Kim, Jinsu Yoon, Hee Dong Kim, and Sung Jin Choi. “Carbon Nanotube Synaptic Transistor Network for Pattern Recognition”. *ACS Applied Materials and Interfaces*, 7(45):25479–25486, oct 2015. ISSN 19448252. doi: 10.1021/acsami.5b08541.
- [31] Adrien F. Vincent, Jérôme Larroque, et al. “Spin-transfer torque magnetic memory as a stochastic memristive synapse for neuromorphic systems”. *IEEE Transactions on Biomedical Circuits and Systems*, 9(2):166–174, apr 2015. ISSN 19324545. doi: 10.1109/TBCAS.2015.2414423.
- [32] Elena Ioana Vatajelu and Lorena Anghel. “Fully-connected single-layer STT-MTJ-based spiking neural network under process variability”. In *Proceedings of the IEEE/ACM International Symposium on Nanoscale Architectures, NANOARCH 2017*, pages 21–26, 2017. ISBN 9781509060368. doi: 10.1109/NANOARCH.2017.8053727.
- [33] Damien Querlioz, Olivier Bichler, Adrien Francis Vincent, and Christian Gamrat. “Bioinspired Programming of Memory Devices for Implementing an Inference Engine”. *Proceedings of the IEEE*, 103(8):1398–1416, aug 2015. ISSN 00189219. doi: 10.1109/JPROC.2015.2437616.
- [34] Geoffrey W. Burr, Robert M. Shelby, et al. “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, jul 2015. ISSN 00189383. doi: 10.1109/TED.2015.2439635.
- [35] Giorgio Cristiano, Massimo Giordano, et al. “Perspective on training fully connected networks with resistive memories: Device requirements for multiple conductances of varying significance”. *Journal of Applied Physics*, 124(15), oct 2018. ISSN 10897550. doi: 10.1063/1.5042462.
- [36] Selina La Barbera, Denys R.B. Ly, et al. “Narrow Heater Bottom Electrode-Based Phase Change Memory as a Bidirectional Artificial Synapse”. *Advanced Electronic Materials*, 4(9), sep 2018. ISSN 2199160X. doi: 10.1002/aelm.201800223.
- [37] Manan Suri, Olivier Bichler, et al. “Phase change memory as synapse for ultra-dense neuromorphic systems: Application to complex visual pattern extraction”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 4.4.1–4.4.4. IEEE, 2011. ISBN 9781457705052. doi: 10.1109/IEDM.2011.6131488.
- [38] Olivier Bichler, Manan Suri, et al. “Visual pattern extraction using energy-efficient ”2-PCM synapse” neuromorphic architecture”. *IEEE Transactions on Electron Devices*, 59(8):2206–2214, 2012. ISSN 00189383. doi: 10.1109/TED.2012.2197951.
- [39] M. Suri, O. Bichler, et al. “CBRAM devices as binary synapses for low-power stochastic neuromorphic systems: Auditory (Cochlea) and visual (Retina) cognitive processing applications”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 10.3.1–10.3.4, 2012. ISBN 9781467348706. doi: 10.1109/IEDM.2012.6479017.
- [40] M. Suri, D. Garbin, et al. “Impact of PCM resistance-drift in neuromorphic systems and drift-mitigation strategy”. In *2013 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*, pages 140–145, 2013. doi: 10.1109/NanoArch.2013.6623059.

- [41] Giacomo Pedretti, Valerio Milo, et al. “Stochastic Learning in Neuromorphic Hardware via Spike Timing Dependent Plasticity with RRAM Synapses”. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 8(1):77–85, mar 2018. ISSN 21563357. doi: 10.1109/JETCAS.2017.2773124.
- [42] T. Werner, D. Garbin, et al. “Real-time decoding of brain activity by embedded Spiking Neural Networks using OxRAM synapses”. In *IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 2318–2321, 2016. ISBN 9781479953417. doi: 10.1109/ISCAS.2016.7539048.
- [43] E. Vianello, T. Werner, et al. “Resistive memories for spike-based neuromorphic circuits”. In *2017 IEEE 9th International Memory Workshop, IMW 2017*, pages 1–6, 2017. ISBN 9781509032723. doi: 10.1109/IMW.2017.7939100.
- [44] Daniele Garbin, Quentin Raffay, et al. “Modeling of OxRAM variability from low to high resistance state using a stochastic trap assisted tunneling-based resistor network”. In *EUROSOI-ULIS 2015 - 2015 Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon*, pages 125–128, 2015. ISBN 9781479969111. doi: 10.1109/ULIS.2015.7063789.
- [45] T. Werner, E. Vianello, et al. “Experimental demonstration of short and long term synaptic plasticity using OxRAM multi k-bit arrays for reliable detection in highly noisy input data”. In *IEEE International Electron Devices Meeting (IEDM)*, pages 16.6.1–16.6.4, 2016. ISBN 9781509039029. doi: 10.1109/IEDM.2016.7838433.
- [46] M. Suri, V. Parmar, A. Kumar, D. Querlioz, and F. Alibart. “Neuromorphic hybrid RRAM-CMOS RBM architecture”. In *15th Non-Volatile Memory Technology Symposium (NVMTS)*, pages 1–6, 2015. ISBN 9781509021260. doi: 10.1109/NVMTS.2015.7457484.
- [47] D. Garbin, E. Vianello, et al. “On the impact of OxRAM-based synapses variability on convolutional neural networks performance”. In *Proceedings of the 2015 IEEE/ACM International Symposium on Nanoscale Architectures, NANOARCH 2015*, pages 193–198, 2015. ISBN 9781467378482. doi: 10.1109/NANOARCH.2015.7180611.
- [48] Alessandro Fumarola, Severin Sidler, et al. “Bidirectional Non-Filamentary RRAM as an Analog Neuromorphic Synapse, Part II: Impact of Al/Mo/Pr 0.7 Ca 0.3 MnO 3 Device Characteristics on Neural Network Training Accuracy”. *IEEE Journal of the Electron Devices Society*, 6(1):169–178, 2018. ISSN 21686734. doi: 10.1109/JEDS.2017.2782184.
- [49] D. Garbin, O. Bichler, et al. “Variability-tolerant Convolutional Neural Network for Pattern Recognition applications based on OxRAM synapses”. In *Technical Digest - International Electron Devices Meeting, IEDM*, volume 2015-Febru, pages 28.4.1–28.4.4, 2015. ISBN 9781479980017. doi: 10.1109/IEDM.2014.7047126.
- [50] Caidie Cheng, Yiqing Li, et al. “Bipolar to unipolar mode transition and imitation of metaplasticity in oxide based memristors with enhanced ionic conductivity”. *Journal of Applied Physics*, 124(15), oct 2018. ISSN 10897550. doi: 10.1063/1.5037962.
- [51] Takeo Ohno, Tsuyoshi Hasegawa, et al. “Short-term plasticity and long-term potentiation mimicked in single inorganic synapses”. *Nature Materials*, 10(8):591–595, 2011. ISSN 14764660. doi: 10.1038/nmat3054.
- [52] Anakha V. Babu, Sandip Lashkare, Udayan Ganguly, and Bipin Rajendran. “Stochastic learning in deep neural networks based on nanoscale PCMO device characteristics”. *Neurocomputing*, 321:227–236, dec 2018. ISSN 18728286. doi: 10.1016/j.neucom.2018.09.019.
- [53] Erika Covi, Stefano Brivio, et al. “Analog memristive synapse in spiking networks implementing unsupervised learning”. *Frontiers in Neuroscience*, 10(OCT), oct 2016. ISSN 1662453X. doi: 10.3389/fnins.2016.00482.

-
- [54] Alexander Serb, Johannes Bill, et al. “Unsupervised learning in probabilistic neural networks with multi-state metal-oxide memristive synapses”. *Nature Communications*, 7, sep 2016. ISSN 20411723. doi: 10.1038/ncomms12611.
- [55] Konstantin Zarudnyi, Adnan Mehonic, et al. “Spike-timing dependent plasticity in unipolar silicon oxide RRAM devices”. *Frontiers in Neuroscience*, 12(FEB), feb 2018. ISSN 1662453X. doi: 10.3389/fnins.2018.00057.
- [56] Irem Boybat, Cecilia Giovinazzo, et al. “Multi-ReRAM synapses for artificial neural network training”. In *Proceedings - IEEE International Symposium on Circuits and Systems*, volume 2019-May, pages 1–5, 2019. ISBN 9781728103976. doi: 10.1109/ISCAS.2019.8702714.
- [57] Haitong Li, Kai Shin Li, et al. “Four-layer 3D vertical RRAM integrated with FinFET as a versatile computing unit for brain-inspired cognitive information processing”. In *Digest of Technical Papers - Symposium on VLSI Technology*, volume 2016-Septe, pages 1–2, 2016. ISBN 9781509006373. doi: 10.1109/VLSIT.2016.7573431.
- [58] G. Piccolboni, G. Molas, et al. “Investigation of the potentialities of Vertical Resistive RAM (VRRAM) for neuromorphic applications”. In *Technical Digest - International Electron Devices Meeting, IEDM*, volume 2016-Febru, pages 17.2.1–17.2.4. IEEE, 2015. ISBN 9781467398930. doi: 10.1109/IEDM.2015.7409717.
- [59] L. Goux, A. Fantini, et al. “Ultralow sub-500nA operating current high-performance TiN/Al₂O₃/HfO₂/Hf/TiN bipolar RRAM achieved through understanding-based stack-engineering”. In *Digest of Technical Papers - Symposium on VLSI Technology*, pages 159–160, 2012. ISBN 9781467308458. doi: 10.1109/VLSIT.2012.6242510.
- [60] A. Grossi, E. Nowak, et al. “Fundamental variability limits of filament-based RRAM”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 4.7.1–4.7.4, 2016. ISBN 9781509039012. doi: 10.1109/IEDM.2016.7838348.
- [61] E. Vianello, G. Molas, et al. “Sb-doped GeS₂ as performance and reliability booster in Conductive Bridge RAM”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 31.5.1–31.5.4, 2012. ISBN 9781467348706. doi: 10.1109/IEDM.2012.6479145.
- [62] Tz Yi Liu, Tian Hong Yan, et al. “A 130.7-mm² 2-layer 32-gb reram memory device in 24-nm technology”. *IEEE Journal of Solid-State Circuits*, 49(1):140–153, jan 2013. ISSN 00189200. doi: 10.1109/JSSC.2013.2280296.
- [63] Giacomo Indiveri, Bernabe Linares-Barranco, et al. “Neuromorphic silicon neuron circuits”. *Frontiers in Neuroscience*, 5(73):1–23, 2011. ISSN 16624548. doi: 10.3389/fnins.2011.00073.
- [64] Syed Ahmed Aamir, Paul Muller, Andreas Hartel, Johannes Schemmel, and Karlheinz Meier. “A highly tunable 65-nm CMOS LIF neuron for a large scale neuromorphic system”. In *European Solid-State Circuits Conference*, volume 2016-October, pages 71–74, 2016. ISBN 9781509029723. doi: 10.1109/ESSCIRC.2016.7598245.
- [65] Antoine Joubert, Bilel Belhadj, and Rodolphe Héliot. “A robust and compact 65 nm LIF analog neuron for computational purposes”. In *2011 IEEE 9th International New Circuits and Systems Conference, NEWCAS 2011*, pages 9–12. IEEE, 2011. ISBN 9781612841359. doi: 10.1109/NEWCAS.2011.5981206.
- [66] Ning Qiao and Giacomo Indiveri. “Scaling mixed-signal neuromorphic processors to 28 nm FD-SOI technologies”. *Proceedings - 2016 IEEE Biomedical Circuits and Systems Conference, BioCAS 2016*, pages 552–555, 2016. doi: 10.1109/BioCAS.2016.7833854.

- [67] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. “Learning representations by back-propagating errors”. *Nature*, 323(6088):533–536, 1986. ISSN 00280836. doi: 10.1038/323533a0.
- [68] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. *Commun. ACM*, 60(6):84—90, 2017. doi: 10.1145/3065386.
- [69] Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. “Gradient-based learning applied to document recognition”. *Proceedings of the IEEE*, 86(11):2278–2323, 1998. ISSN 00189219. doi: 10.1109/5.726791.
- [70] Daniel E. Feldman. “The Spike-Timing Dependence of Plasticity”, 2012. ISSN 08966273.
- [71] Henry Markram, Joachim Lübke, Michael Frotscher, and Bert Sakmann. “Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs”. *Science*, 275(5297): 213–215, 1997. ISSN 00368075. doi: 10.1126/science.275.5297.213.
- [72] Guo Qiang Bi and Mu Ming Poo. “Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type”. *Journal of Neuroscience*, 18(24):10464–10472, 1998. ISSN 02706474. doi: 10.1523/jneurosci.18-24-10464.1998.
- [73] Olivier Bichler, Damien Querlioz, Simon J. Thorpe, Jean Philippe Bourgoin, and Christian Gamrat. “Unsupervised features extraction from asynchronous silicon retina through spike-timing-dependent plasticity”. In *Proceedings of the International Joint Conference on Neural Networks*, pages 859–866. IEEE, 2011. ISBN 9781457710865. doi: 10.1109/IJCNN.2011.6033311.
- [74] Peter U. Diehl and Matthew Cook. “Unsupervised learning of digit recognition using spike-timing-dependent plasticity”. *Frontiers in Computational Neuroscience*, 9 (AUGUST), aug 2015. ISSN 16625188. doi: 10.3389/fncom.2015.00099.
- [75] P. Gonon, C. Vallee, C. Mannequin, M. Saadi, and .F Jomni. “Mechanisms of resistance switching in nanometric metal oxides and their dependence on electrodes”. In *Proceedings of the IEEE International Conference on Properties and Applications of Dielectric Materials*, volume 2015-October, pages 56–59, 2015. ISBN 9781479989034. doi: 10.1109/ICPADM.2015.7295207.
- [76] Yoshifumi Nishi, Ulrich Bottger, Rainer Waser, and Stephan Menzel. “Crossover from Deterministic to Stochastic Nature of Resistive-Switching Statistics in a Tantalum Oxide Thin Film”. *IEEE Transactions on Electron Devices*, 65(10):4320–4325, oct 2018. ISSN 00189383. doi: 10.1109/TED.2018.2866127.
- [77] Christopher H Bennett, Diana Garland, Robin B Jacobs-Gedrim, Sapan Agarwal, and Matthew J Marinella. “Wafer-Scale TaOx Device Variability and Implications for Neuromorphic Computing Applications”. In *IEEE International Reliability Physics Symposium Proceedings*, pages 1–4, 2019. ISBN 9781538695043. doi: 10.1109/IRPS.2019.8720596.
- [78] C. Cagli, G. Piccolboni, et al. “About the intrinsic resistance variability in HfO₂-based RRAM devices”. In *Joint International EUROSOL Workshop and International Conference on Ultimate Integration on Silicon-ULIS, EUROSOL-ULIS 2017 - Proceedings*, pages 31–34, 2017. ISBN 9781509053131. doi: 10.1109/ULIS.2017.7962593.
- [79] Sheng Sung Yang, Chia Lu Ho, and Sammy Siu. “Computing and analyzing the sensitivity of MLP due to the errors of the i.i.d. inputs and weights based on CLT”. *IEEE Transactions on Neural Networks*, 21(12):1882–1891, dec 2010. ISSN 10459227. doi: 10.1109/TNN.2010.2077681.

- [80] Robert M. Shelby, Geoffrey W. Burr, Irem Boybat, and Carmelo Di Nolfo. “Non-volatile memory as hardware synapse in neuromorphic computing: A first look at reliability issues”. In *IEEE International Reliability Physics Symposium Proceedings*, volume 2015-May, pages 6A11–6A16, 2015. ISBN 9781467373623. doi: 10.1109/IRPS.2015.7112755.
- [81] Damien Querlioz, Olivier Bichler, Philippe Dollfus, and Christian Gamrat. “Immunity to device variations in a spiking neural network with memristive nanodevices”. *IEEE Transactions on Nanotechnology*, 12(3):288–295, 2013. ISSN 1536125X. doi: 10.1109/TNANO.2013.2250995.
- [82] Johannes Bill and Robert Legenstein. “A compound memristive synapse model for statistical learning through STDP in spiking neural networks”. *Frontiers in Neuroscience*, 8(DEC), 2014. ISSN 1662453X. doi: 10.3389/fnins.2014.00412.
- [83] Pritish Narayanan, Geoffrey W. Burr, Stefano Ambrogio, and Robert M. Shelby. “Neuromorphic technologies for next-generation cognitive computing”. In *2017 IEEE 9th International Memory Workshop, IMW 2017*, pages 1–4, 2017. ISBN 9781509032723. doi: 10.1109/IMW.2017.7939095.
- [84] S. R. Nandakumar, Manuel Le Gallo, et al. “Mixed-precision architecture based on computational memory for training deep neural networks”. In *Proceedings - IEEE International Symposium on Circuits and Systems*, volume 2018-May, pages 1–5, 2018. ISBN 9781538648810. doi: 10.1109/ISCAS.2018.8351656.
- [85] Sang Gyun Gi, Injune Yeo, et al. “Modeling and System-Level Simulation for Nonideal Conductance Response of Synaptic Devices”. *IEEE Transactions on Electron Devices*, 65(9):3996–4003, sep 2018. ISSN 00189383. doi: 10.1109/TED.2018.2858762.
- [86] Sungho Kim, Meehyun Lim, Yeamin Kim, Hee Dong Kim, and Sung Jin Choi. “Impact of Synaptic Device Variations on Pattern Recognition Accuracy in a Hardware Neural Network”. *Scientific Reports*, 8(1), dec 2018. ISSN 20452322. doi: 10.1038/s41598-018-21057-x.
- [87] Tobi Delbruck. “Frame-free dynamic digital vision”. In *Intl. Symp. on Secure-Life Electronics, Advanced Electronics for Quality Life and Society*, pages 21–26, 2008. doi: <http://dx.doi.org/10.5167/uzh-17620>.
- [88] E. Vianello, O. Thomas, et al. “Resistive Memories for Ultra-Low-Power embedded computing design”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 6.3.1–6.3.4, 2014. ISBN 9781479980017. doi: 10.1109/IEDM.2014.7046995.
- [89] Alessandro Grossi, Elisa Vianello, et al. “Experimental Investigation of 4-kb RRAM Arrays Programming Conditions Suitable for TCAM”. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 26(12):2599–2607, dec 2018. ISSN 10638210. doi: 10.1109/TVLSI.2018.2805470.
- [90] L Perniola, G Molas, et al. “Universal Signatures from Non-Universal Memories: Clues for the Future.”. In *2016 IEEE 8th International Memory Workshop, IMW 2016*, pages 1–3, 2016. ISBN 9781467388313. doi: 10.1109/IMW.2016.7495295.
- [91] H.-S. P. Wong, C. Ahn, et al. “Stanford Memory Trends”, 2018. URL <https://nano.stanford.edu/stanford-memory-trends/>.
- [92] Thomas Dalgaty, Melika Payvand, et al. “Hybrid neuromorphic circuits exploiting non-conventional properties of RRAM for massively parallel local plasticity mechanisms”. *APL Materials*, 7(8):081125, aug 2019. ISSN 2166532X. doi: 10.1063/1.5108663.
- [93] O. Bichler, D. Roclin, C. Gamrat, and D. Querlioz. “Design exploration methodology for memristor-based spiking neuromorphic architectures with the Xnet event-driven simulator”. In *Proceedings of the 2013 IEEE/ACM International Symposium on Nanoscale Architectures, NANOARCH 2013*, pages 7–12, 2013. ISBN 9781479908738. doi: 10.1109/NanoArch.2013.6623029.

- [94] Dan Ciregan, Ueli Meier, and Jurgen Schmidhuber. “Multi-column deep neural networks for image classification”. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3642–3649, 2012. ISBN 9781467312264. doi: 10.1109/CVPR.2012.6248110.
- [95] Abu Sebastian, Tomas Tuma, et al. “Temporal correlation detection using computational phase-change memory”. *Nature Communications*, 8(1), 2017. ISSN 20411723. doi: 10.1038/s41467-017-01481-9.
- [96] Daniele Garbin, Elisa Vianello, et al. “HfO₂-Based OxRAM Devices as Synapses for Convolutional Neural Networks”. *IEEE Transactions on Electron Devices*, 62(8): 2494–2501, 2015. ISSN 00189383. doi: 10.1109/TED.2015.2440102.
- [97] Shimeng Yu, Yi Wu, Rakesh Jeyasingh, Duygu Kuzum, and H. S.Philip Wong. “An electronic synapse device based on metal oxide resistive switching memory for neuromorphic computation”. *IEEE Transactions on Electron Devices*, 58(8):2729–2737, 2011. ISSN 00189383. doi: 10.1109/TED.2011.2147791.
- [98] Daniele Ielmini. “Resistive switching memories based on metal oxides: Mechanisms, reliability and scaling”. *Semiconductor Science and Technology*, 31(6):063002, 2016. ISSN 13616641. doi: 10.1088/0268-1242/31/6/063002.
- [99] Thomas Dalgaty, Elisa Vianello, et al. “Insect-inspired elementary motion detection embracing resistive memory and spiking neural networks”. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 115–128, 2018. ISBN 9783319959719. doi: 10.1007/978-3-319-95972-6_13.
- [100] M. Suri, D. Querlioz, et al. “Bio-Inspired Stochastic Computing Using Binary CBRAM Synapses”. *IEEE Transactions on Electron Devices*, 60(7):2402–2409, 2013. doi: 10.1109/TED.2013.2263000.
- [101] S Ambrogio, S Balatti, et al. “Neuromorphic Learning and Recognition With One-Transistor-One-Resistor Synapses and Bistable Metal Oxide RRAM”. *IEEE Transactions on Electron Devices*, 63(4):1508–1515, 2016. ISSN 1557-9646. doi: 10.1109/TED.2016.2526647.
- [102] Geoffrey W Burr, Robert M Shelby, et al. “Neuromorphic computing using non-volatile memory”. *Advances in Physics: X*, 2(1):89–124, 2017. ISSN 23746149. doi: 10.1080/23746149.2016.1259585.
- [103] D Kuzum, R G D Jeyasingh, and H . P Wong. “Energy efficient programming of nano-electronic synaptic devices for large-scale implementation of associative and temporal sequence learning”. In *2011 International Electron Devices Meeting*, pages 30.3.1–30.3.4, 2011. doi: 10.1109/IEDM.2011.6131643.
- [104] Duygu Kuzum, Rakesh G.D. Jeyasingh, Byoungil Lee, and H. S.Philip Wong. “Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing”. *Nano Letters*, 12(5):2179–2186, 2012. ISSN 15306984. doi: 10.1021/nl201040y.
- [105] STMicroelectronics. “Phase-Change Memory”, 2019. URL https://www.st.com/content/st_{_}com/en/about/innovation---technology/PCM.html.
- [106] F. Arnaud, P. Zuliani, et al. “Truly Innovative 28nm FDSOI Technology for Automotive Micro-Controller Applications embedding 16MB Phase Change Memory”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 18.4.1–18.4.4. IEEE, 2018. ISBN 9781728119878. doi: 10.1109/IEDM.2018.8614595.
- [107] A Fumarola, P Narayanan, et al. “Accelerating machine learning with Non-Volatile Memory: Exploring device and circuit tradeoffs”. In *2016 IEEE International Conference on Rebooting Computing (ICRC)*, pages 1–8, oct 2016. doi: 10.1109/ICRC.2016.7738684.

- [108] Damien Querlioz, Philippe Dollfus, Olivier Bichler, and Christian Gamrat. “Learning with memristive devices: How should we model their behavior?”. In *Proceedings of the 2011 IEEE/ACM International Symposium on Nanoscale Architectures, NANOARCH 2011*, pages 150–156, 2011. ISBN 9781457709944. doi: 10.1109/NANOARCH.2011.5941497.
- [109] Sen Song, Kenneth D Miller, and L F Abbott. “Competitive Hebbian learning through spike-timing-dependent synaptic plasticity”. *Nature Neuroscience*, 3(9):919–926, 2000. doi: 10.1038/78829.
- [110] M C W Van Rossum, G Q Bi, and G G Turrigiano. “Stable Hebbian Learning from Spike Timing-Dependent Plasticity”. *Journal of Neuroscience*, 20(23):8812–8821, 2000. doi: 10.1523/JNEUROSCI.20-23-08812.2000.

Synaptic routing reconfigurability of spiking neural networks with resistive memory-based ternary content-addressable memory systems

Contents

3.1	Content-addressable memory systems	100
3.1.1	Basics on content-addressable memories	100
3.1.2	Motivations for the implementation of resistive memory-based ternary content-addressable memories	101
3.1.3	Examples of ternary content-addressable memory applications	103
3.1.4	Goal of this chapter	107
3.2	Characterisation of resistive memory-based ternary content-addressable memories	108
3.2.1	Fabricated resistive memory-based ternary content-addressable memory circuits	108
3.2.2	Search operation principle	109
3.2.3	Common 2T2R TCAM circuit characterisation . . .	110
3.2.4	Novel 1T2R1T TCAM circuit characterisation . . .	122

3.1 Content-addressable memory systems

3.1.1 Basics on content-addressable memories

CONTENT-Addressable Memories (CAMs) are specialised hardware capable of performing high-speed in-memory search and pattern matching. They allow to search a data in a memory table of pre-stored entries. In classic Random Access Memory (RAM) systems a stored data is accessed by its address as shown in FIGURE 3.1.1 (a). On the other hand, CAM systems proceed the other way around: a stored data is accessed by its content rather than by its address. FIGURE 3.1.1 (b) shows a simplified schematic of a CAM system storing four words of 4 bits each. An input searched data is broadcast to the table of stored data through the search lines. Each stored word compares its content with this input searched data. The result of the comparison is returned by the Match Line (ML) of the stored word. If the search and stored data are identical, the corresponding match line returns a *match case*. If at least one bit is mismatching, a *mismatch case* is returned. Match lines are then fed into an encoder. The encoder outputs the address location of the match line in the match state. The main advantage of CAM systems over conventional memory systems is that they can perform the search in parallel over the whole memory table [1]. While data are sequentially accessed and compared in RAM systems, the input searched data in CAM systems is simultaneously broadcast to every stored word. This allows to perform the search in a single clock cycle which offers a significant advantage in terms of speed.

CAM systems can be divided into Binary CAMs (BCAMs) and Ternary CAMs (TCAMs). While BCAM bitcells can only store the values '0' and '1', TCAM bitcells allow for storing a third value called *don't care* state and denoted as 'X'. An 'X' state returns a match whatever the input data. FIGURE 3.1.2 shows a block diagram of a TCAM system. In BCAM systems where a single match is

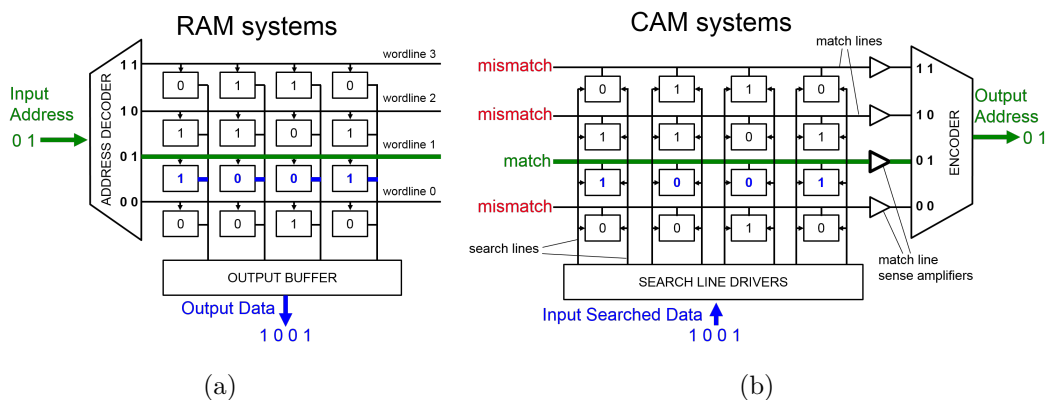


FIGURE 3.1.1: Block diagram of (a) a Random Access Memory (RAM) system and (b) a Content-Addressable Memory (CAM) system. In RAM systems, stored data are accessed by their physical address location, and the system outputs the stored content. In CAM systems, stored data are accessed by their content rather than by their address, and the system outputs the address of the searched data.

expected, an encoder is used. In TCAM systems where more than one word may match, a priority encoder is used instead of a simple encoder. The priority encoder selects the matching address based on the highest priority matching location that can be for instance words in lower address locations or words with the most matching bits that are not in the 'X' states [2]. The advantage of TCAMs over BCAMs is that they are capable of storing ranges of data which can be useful to save entries for applications where an exact match is not necessary. In the example of FIGURE 3.1.2, the word stored at the address '1 0' can match with four different input patterns, whereas in a BCAM it would be necessary to use four distinct words.

In this chapter we only focus on TCAM systems. In particular, we will only focus on TCAM bitcell implementation and consider neither search line or match line power reduction scheme [2–5] nor priority encoder implementation [6].

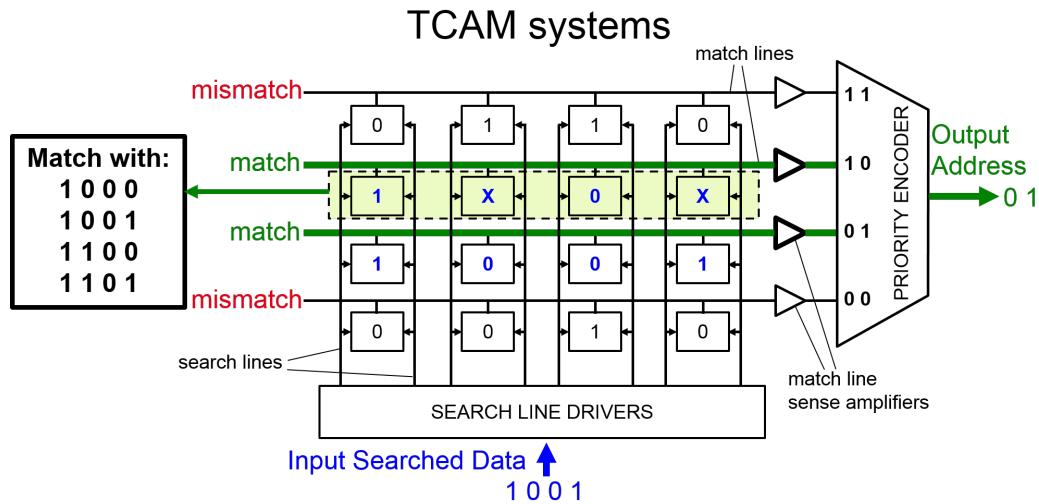


FIGURE 3.1.2: Block diagram of a Ternary Content-Addressable Memory (TCAM) system. The use of the *don't care* state, 'X', allows to perform local masking and store data ranges. As more than one word may match, a priority encoder is used instead of an encoder. A single address is output based on the highest priority matching location (*e.g.* lowest address location, most matching bits that are not in the 'X' states, ...).

3.1.2 Motivations for the implementation of resistive memory-based ternary content-addressable memories

One of the first hardware demonstration of a CAM system can be dated back in 1956 [7]. It made use of cryotron memories, a type of memory based on superconductivity. When immersed in liquid helium, cryotron memories could switch between a superconductive and a resistive state depending on the injected current. However, due to the cost of refrigeration and maintenance needed for this technology, efforts were rather focused on the use of non-cryogenic components [8]. It was not before 1970 that the first practical integrated CAM circuit was demonstrated by Koo [9] using CMOS or bipolar transistors. Since

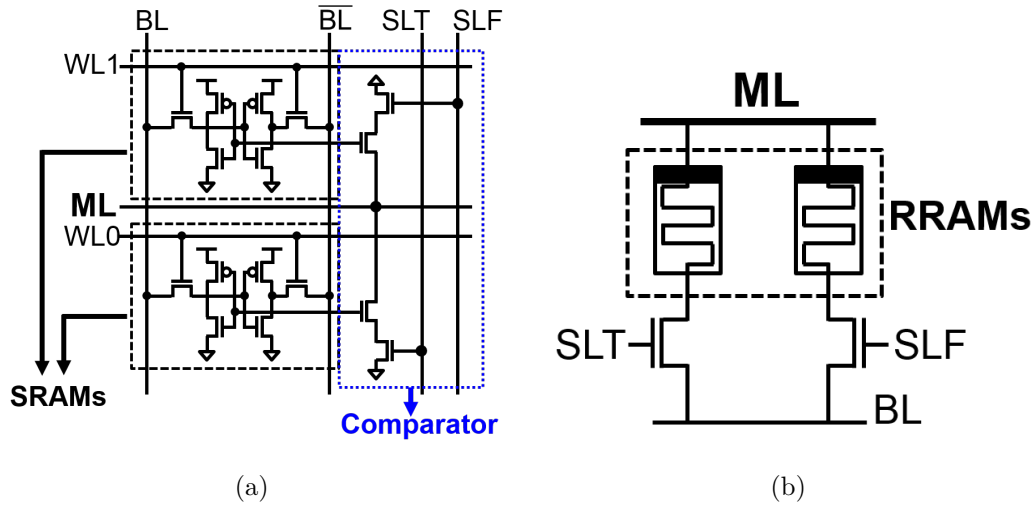


FIGURE 3.1.3: (a) Conventional sixteen-transistors (16T) SRAM-based TCAM. (b) Common two-transistors/two-RRAMs (2T2R) RRAM-based TCAM.

then, many CAM implementations with CMOS transistors have been proposed [10]. Today, conventional CMOS-based TCAMs are implemented with sixteen transistors (16T) with the structure proposed in [11] and depicted in FIGURE 3.1.3 (a). This structure uses two six-transistors Static Random Access Memory (SRAM) cells to store data and encode the *don't care* state, and a comparator circuit to compare the stored and searched data. In the following, this structure is referred to as SRAM-based TCAM.

The SRAM-based TCAM implementation presents inherent disadvantages. First, SRAMs are volatile. This leads to static power consumption due to the need to restore data every time the system is turned OFF and ON. Second, the use of two SRAM cells and a comparator circuit per bit strongly limits the storage density to tens of megabit [5, 12, 13]. As a rule of thumb, the largest available TCAM array is usually twice or three times as small as the largest available SRAM array [2]. This has motivated the replacement of SRAM cells with Resistive Memory (RRAM) technology. Many RRAM-based TCAM bitcells have been proposed, each implemented with different numbers of transistors (T) and RRAMs (R): 12T2R [14], 8T4R [15], 6T2R [16], 5T2R [17–19], 4T2R [20–22], 3T2R [6, 23, 24], 3T1R [25, 26], 2.5T1R [27], and 2T2R [28–32]. The use of RRAMs allows for a reduction in the number of transistors per bitcell from sixteen for the conventional SRAM-based TCAM down to two transistors [28, 30, 32] (FIGURE 3.1.3 (b)). FIGURE 3.1.4 (a) shows reported RRAM- (blue diamond) and SRAM-based (black circle) TCAM bitcell size as a function of the technology node. RRAM-based TCAM bitcells allow for a gain in silicon area with respect to SRAM-based TCAM bitcells, while reaching similar performance in terms of search time and search energy (FIGURE 3.1.4 (b) and (c)). A few silicon-proven RRAM-based TCAM circuits have already been presented in the literature [6, 19, 21, 26, 27, 29]. One of the main drawbacks of RRAMs with respect to SRAMs is the low ratio between the ON and OFF current of the memory elements ($>10^5$ for SRAMs, 10-100 for RRAMs). This poses serious challenges for designing RRAM-based TCAMs as performance and reliability

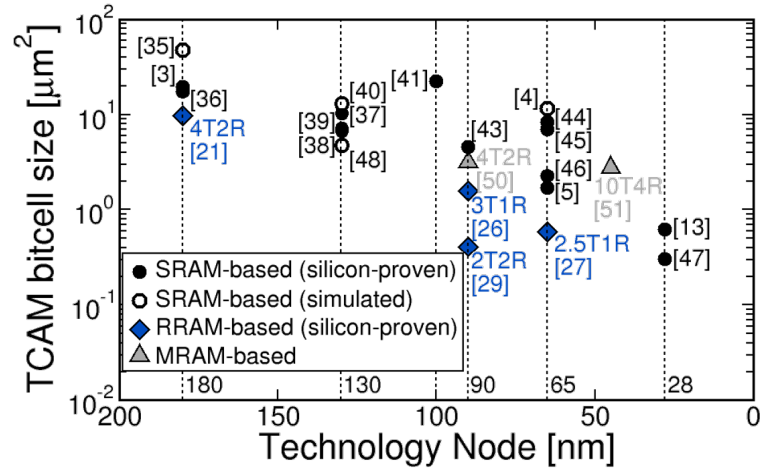
strongly depend on this ratio. Leakage currents in case of a match can overwhelm the system and become comparable to a mismatching current if a sufficient ON/OFF ratio is not ensured. This limits the sensing margin that allows to discriminate between a match and mismatch case, and it constraints the maximum number of bits per word. Up to date, the impact of ON/OFF ratio on RRAM-based TCAMs has only been studied in simulation [18, 21, 29, 31, 32]. SRAM-based TCAMs have been proven to be functional for more than 640 bits [33], whereas silicon-proven RRAM-based TCAM circuits do not exceed 256 bits [27]. This can be limiting for applications requiring long pattern matching, such as Internet Protocol packet routing or active control list management [34].

Another drawback of RRAMs is their limited endurance (between 10^6 and 10^9 write cycles [52]) with respect to SRAMs ($>10^{16}$ cycles [53]). However, the endurance for SRAMs refers to both programming and search operations, whereas in RRAM-based TCAMs programming and search (read) operations are two distinguished operations, and they must be characterised separately. No characterisation of the impact of search operations on RRAM-based TCAM reliability has been reported so far.

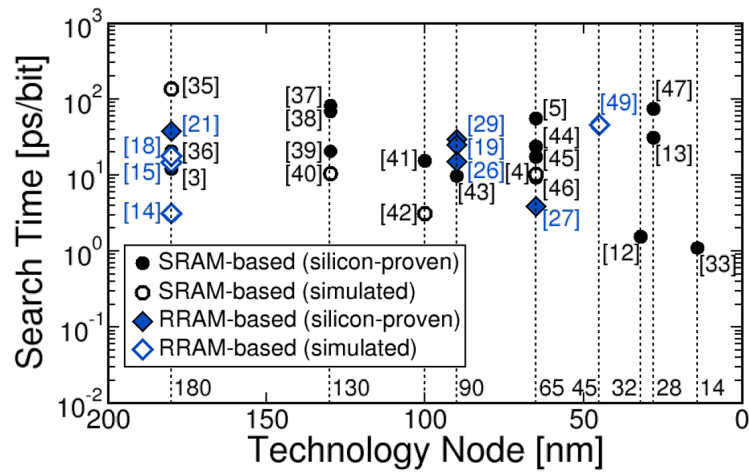
3.1.3 Examples of ternary content-addressable memory applications

3.1.3.1 Conventional ternary content-addressable memory applications

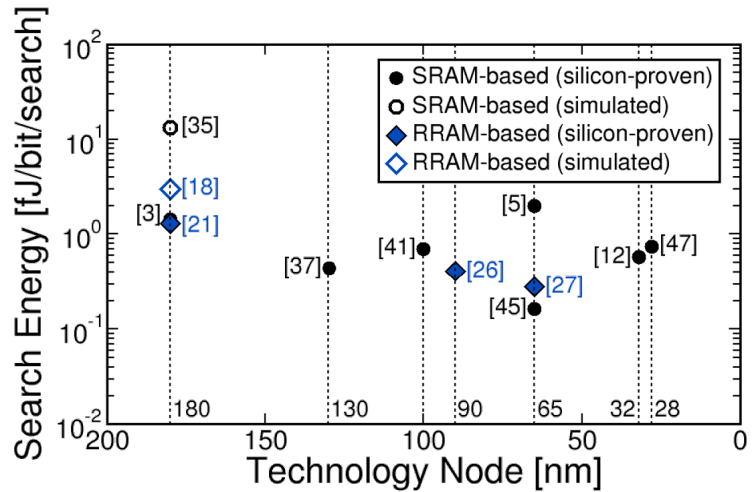
TCAMs find application in domains that require pattern matching [8], such as branch prediction tables or cache controllers in processors [54], network intrusion detection system [55], alignment of DNA sequences [56], in-memory computing applications [57], and lookup tables [58]. The primary commercial application of TCAM systems is to classify and forward Internet Protocol (IP) packets in network routers [34]. Network routers use address lookup tables to forward IP packets from an incoming port to an output port. FIGURE 3.1.5 shows a TCAM-based implementation of address lookup tables [2]. Each IP packet contains its destination address. When a router receives an IP packet, the destination address is searched in the TCAM table. The address of the matching data is then fed into a RAM system. The RAM system contains the table of output ports corresponding to each destination address. Nowadays, with the rapid growth of the Internet of Things (IoT), it is critical for network equipment to handle the explosion of data and to meet performance requirements in terms of bandwidth. It is now common that all packet-processing functions, like parsing, classification, and forwarding have to be completed within times as short as 2.5 ns [34] in order to ensure no latency. Thus, TCAM systems appear as ideal since they can process packets with latencies below nanoseconds.



(a)



(b)



(c)

FIGURE 3.1.4: Reported SRAM-based (black circle) [3–5, 12, 13, 33, 35–48] and RRAM-based (blue diamond) [14, 15, 18, 19, 21, 26, 27, 29, 49] (a) TCAM bitcell size, (b) TCAM search time, and (c) TCAM search energy as a function of technology node. Search times in (b) have been normalised by the number of bits per TCAM word to provide a fair comparison. TCAM bitcell size with Magnetic Memories (MRAMs, grey triangle) [50, 51] have been plotted for comparison.

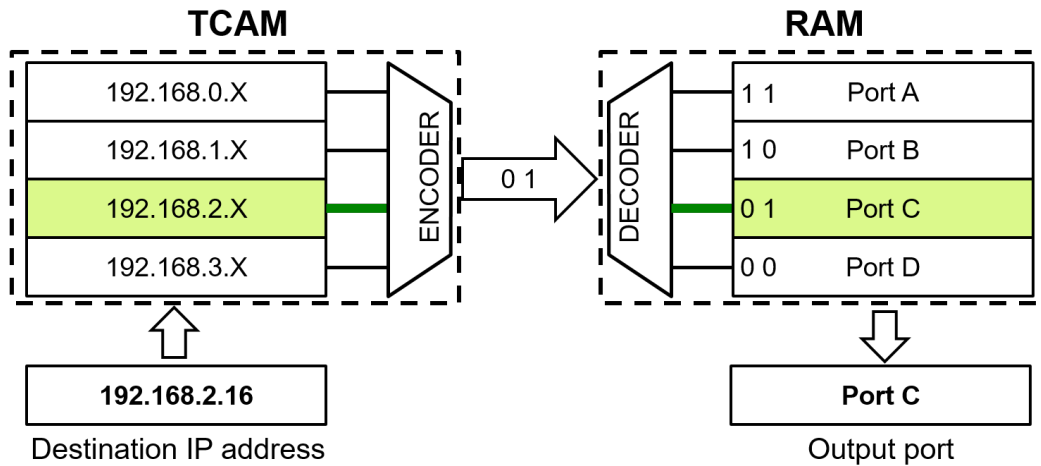


FIGURE 3.1.5: TCAM-based implementation of a network router address lookup table. Reproduced from [2].

3.1.3.2 Ternary content-addressable memories to enable spiking neural network reconfigurability

Recently, TCAMs have also been used to enable dynamic reconfigurability of the synaptic connections in Spiking Neural Networks (SNNs) [59–62]. SNNs use spikes for communication, and part of neural computation is achieved by the connectivity between neurons, *i.e.* the network topology. Therefore, one of the challenges in designing SNN hardware circuits is to properly route spiking events between neurons, *i.e.* to define the appropriate synaptic network topology. This can generally be realised either with dedicated routing, *i.e.* neurons are hardwired together, or with synaptic Lookup Table (LUT) routing scheme [62]. The main advantage of the dedicated routing scheme is that its parallel event routing protocol supports the real-time operation of SNNs. However, it suffers from limited reconfigurability unless using additional memory, *i.e.* N^2 programmable synapses for N neurons [62, 63]. The LUT routing scheme addresses this issue by enabling any synapse in the synaptic array to be used for any arbitrary pair of neurons with much fewer built-in synapses than N^2 . When a neuron fires an event, the event is sent to a synaptic routing LUT storing the entire neuronal connections of the circuit, and then it is transmitted to the appropriate neurons following the LUT. Most of spike-based reconfigurable neural networks [59–61, 64–67] use the Address Event Representation (AER) [68, 69] as data representation and communication protocol. AER protocol allows for asynchronous communication between neurons: each neuron has a unique address that is encoded as a digital word, and it transmits its address to every output neuron as soon as it produces an event. The AER protocol naturally permits to implement synaptic reconfigurability since the addresses of neurons connected together can be directly stored in the LUTs.

The major downside of the LUT routing scheme is that the use of LUTs involves additional delays in the handling of spiking events which can create traffic congestion in case of slow LUTs. This limits the maximum number of neurons and synapses in neuromorphic cores [62]. In [62], the authors provide a comparative study of four different LUT implementations and their impact

on neuromorphic cores in terms of maximum number of neurons and synapses, and maximum fan-in/fan-out for each neuron without traffic congestion. As the study shows, LUTs can be efficiently implemented with TCAMs as they allow for the highest number of synapses and synaptic operations per second (routing speed) compared to the other implementations. As an example, the number of synapses with TCAM-based LUTs surpasses by more than three orders of magnitude that of RAM-based LUTs and by two orders of magnitude the routing speed owing to the fast parallel search capability of TCAMs. The Dynamic Neuromorphic Asynchronous Processors (DYNAPs) presented in [59] are a case in point of multi-core neuromorphic processors using CAM-based LUTs for dynamic reconfigurability. In DYNAPs, neurons are all connected to each other through different levels of routers. The routing has been optimised with the use of shared addresses and tags between neurons in order to minimise memory requirements without any loss of generality. Each core in DYNAPs is composed of 16x16 computing nodes. FIGURE 3.1.6 (a) shows a simplified block diagram of one computing node composed of a CAM table, Pulse Generators (PGs), a Differential-Pair Integrator (DPI) synapse [70], and a CMOS-based leaky Integrate-and-Fire (IF) neuron. CAM tables each contain sixty-four words (CAM size=64 rows) of 10 bits. They implement asynchronous synaptic routing tables by storing neuron addresses and are directly integrated within computing nodes for better energy-efficiency - SpiNNaker [61] and HiAER [60] store their synaptic routing tables in external DRAM chips which results in frequent off-chip communications. FIGURE 3.1.6 (b) illustrates an example of the working principle of DYNAPs. For the sake of simplicity, we only represent six computing nodes, each composed of a CAM table of size 2 and a leaky IF neuron (grey circle). The connections between CAM tables and neurons encompass the pulse generator and DPI synapse circuits. Each computing node stores in its CAM table the addresses of pre-synaptic neurons it is virtually connected to. Each neuron has an address which is placed on a shared digital bus as soon as it spikes. When a neuron spikes (*e.g.* the neuron with the address 6), its address (6) is broadcast to every computing node (including itself). Each computing node compares this address with the addresses stored in its CAM table. If the address of the spiking neuron is stored within a CAM table (green rows), the corresponding match line activates the pulse generator which transmits a spike to the leaky IF neuron along the DPI synapse (neurons 2 and 6 in the example of FIGURE 3.1.6 (b)). Therefore, appropriate programming of the CAM tables defines the network topology (FIGURE 3.1.6 (b, right)), and network topology can be reconfigured on-the-fly. Note that, in the case of DYNAPs, CAM tables are implemented with Binary CAMs (BCAMs). Therefore, each BCAM row can only store one neuron address. Consequently, the fan-in of each neuron, *i.e.* the number of pre-synaptic neurons connected to this neuron, corresponds to the CAM size (64 for DYNAPs). The use of TCAMs instead of BCAMs allows to store several pre-synaptic neuron addresses in each TCAM row, thus *increasing the fan-in of each neuron as well as the total number of synapses operating in parallel*. On the other hand, each neuron can be connected to any other neuron (fan-out).

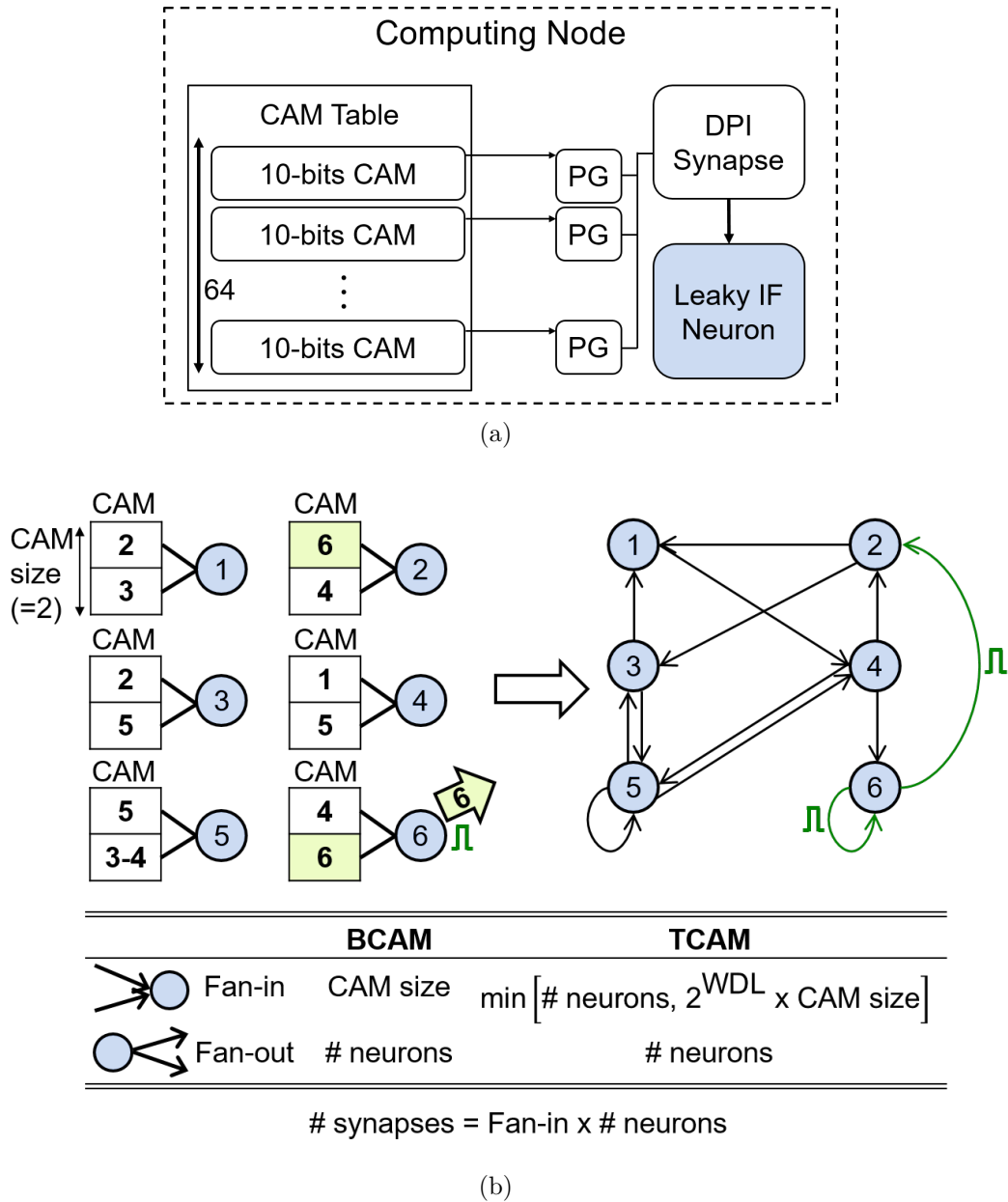


FIGURE 3.1.6: (a) Simplified block diagram of one computing node in DYNAPs [59] composed of a 64x10 bits BCAM table, 64 Pulse Generators (PGs), a Differential-Pair Integrator (DPI) synapse circuit [70], and a CMOS-based leaky Integrate-and-Fire (IF) neuron. (b) Working principle of DYNAPs. When the neuron 6 spikes, its address, 6, is broadcast to every other neuron (including itself). As its address is stored in the CAM table of the computing nodes 2 and 6 (green rows), a pulse is locally generated and transmitted to the corresponding leaky IF neuron circuits. The use of TCAMs instead of BCAMs allows to increase the fan-in of each neuron. Adapted from [59].

3.1.4 Goal of this chapter

Main challenges in designing TCAMs are energy, area, speed, and reliability [54]. While reported RRAM-based TCAMs have already demonstrated a decrease in TCAM bitcell size with respect to the conventional SRAM-based TCAM bitcell

at similar speed and energy (*cf* FIGURE 3.1.4), reliability has not been addressed yet. In particular, the impact of RRAM electrical properties on TCAM reliability has only been assessed in simulation. In addition, TCAM reliability in terms of endurance has never been studied.

In this chapter, we present an extensive electrical characterisation study of two different TCAM structures: (i) the most common type of two-transistors/two-RRAMs (2T2R) RRAM-based TCAM, and (ii) a new TCAM bitcell composed of two transistors and two RRAMs in a one-transistor/two-RRAMs/one-transistor (1T2R1T) configuration with a sensing margin insensitive to the High Resistance State (HRS) and the Low Resistance State (LRS) RRAM resistance ratio, HRS/LRS, and variability.

3.2 Characterisation of resistive memory-based ternary content-addressable memories

Two different 3x128-bits RRAM-based TCAM circuits have been fabricated and tested. In both circuits, TCAM bitcells are implemented with two transistors and two HfO₂-based RRAMs. The first structure is the most common two-transistors/two-RRAMs (2T2R) TCAM implemented with a pair of access transistors and RRAMs [28–32]. The second structure is implemented in a new one-transistor/two-RRAMs/one-transistor (1T2R1T) configuration. Extensive electrical characterisations have been performed, in particular the impact of RRAM electrical properties on performance, reliability, and endurance is quantified.

3.2.1 Fabricated resistive memory-based ternary content-addressable memory circuits

Both TCAM circuits have their own peripheral circuitry which is identical in both cases. FIGURE 3.2.1 (a) presents the schematic of the two fabricated RRAM-based TCAM circuits. Each TCAM circuit is composed of a search word register, a TCAM array, and a read circuit (Sense Amplifier, SA). The search word register outputs the 128-bits searched data to every search line (SLT and SLF) of the TCAM array. The TCAM array comprises three 128-bits TCAM rows. *Only the TCAM array is different between both circuits.* For each TCAM circuit, *all the measurements are performed on the middle TCAM* whose Match Line (ML) is connected to the SA. FIGURE 3.2.1 (b) shows die pictures of the 2T2R and (c) the 1T2R1T TCAM circuits fabricated using the same 130-nm CMOS process and RRAM technologies. TiN/HfO₂/Ti/TiN (100 nm/10 nm/100 nm/100 nm) RRAMs are integrated in the back-end-of-line on top of the fourth metal layer (Cu) (FIGURE 3.2.1 (d)).

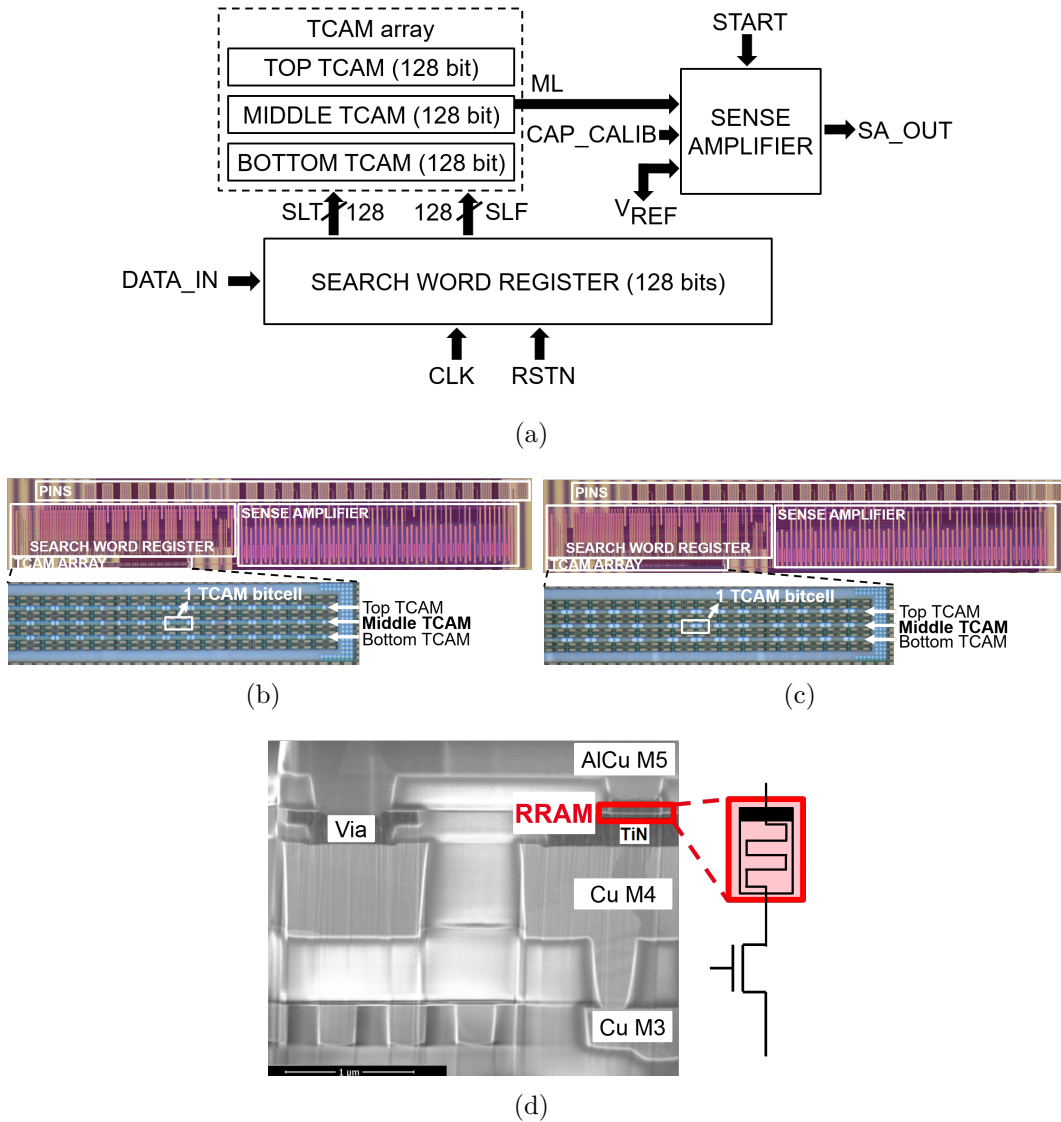


FIGURE 3.2.1: (a) Common block diagram of the fabricated 2T2R and 1T2R1T TCAM circuits. Only the TCAM array is different between both circuits. (b) Die picture of the fabricated 2T2R and (c) 1T2R1T circuits. (d) Scanning electron microscope cross-section of the integrated HfO₂-based RRAMs.

3.2.2 Search operation principle

In both circuits, the search operation relies on the discharge of a pre-charged Match Line (ML). The ML is first pre-charged at a voltage V_{DD_ML} (*ML pre-charge phase*). The ML is then left floating and starts discharging through each TCAM cell (*ML sensing phase*). If the data stored in the TCAM word matches with the searched data (*match case*, FIGURE 3.2.2 (top)), the ML slowly discharges through leakage currents, I_{match} . If at least one bit of the stored data mismatches with the searched data (*mismatch case*, FIGURE 3.2.2 (bottom)), the ML quickly discharges through the mismatching cells with a high discharge current, $I_{mismatch}$.

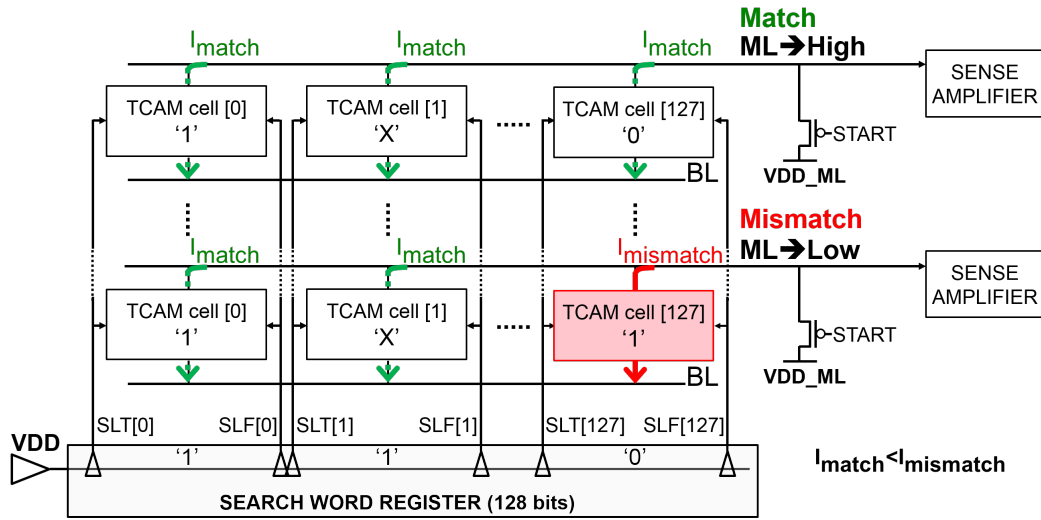


FIGURE 3.2.2: Example of the search operation principle. The Match Line (ML) is first pre-charged at VDD_ML , then it is left floating. (Top) In a match case, the ML stays high. (Bottom) In the mismatch case, the ML is pulled down to a low level.

3.2.3 Common 2T2R TCAM circuit characterisation

3.2.3.1 2T2R TCAM bitcell working principle

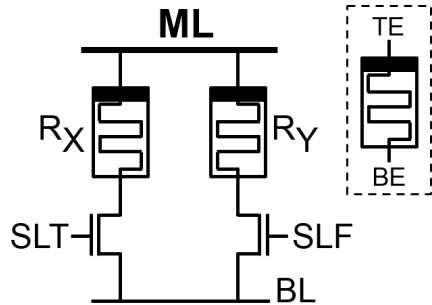


FIGURE 3.2.3: Common 2T2R TCAM bitcell schematic. Top (TE) and Bottom (BE) Electrodes are indicated with the black rectangle.

Stored data	R_X	R_Y
'0'	LRS	HRS
'1'	HRS	LRS
'X'	HRS	HRS

TABLE 3.1: RRAM state definition as a function of the stored data.

Searched data	SLT	SLF
'0'	0	VDD
'1'	VDD	0

TABLE 3.2: SLT and SLF voltages as a function of the searched data.

FIGURE 3.2.3 depicts the common 2T2R TCAM bitcell implementation [28–32]. It is composed of a pair of access transistors and RRAMs. Depending on the stored data, RRAMs are programmed either in the Low Resistance State (LRS) or High Resistance State (HRS) as summarised in TABLE 3.1. Forming, Set, Reset, and read operations are performed as in single 1T1R RRAM cells. Required voltages on top and bottom electrodes are applied on the ML and BL, respectively. The transistor in series with the RRAM to be programmed is activated via the search line SLT (resp. SLF), while the complementary search

line SLF (resp. SLT) is kept at 0 V in order to activate each 1T1R structure independently. The gate voltage is applied on the activated search line.

During a search operation, one of the two transistors is activated via SLT or SLF, while the complementary search line is set at '0'. TABLE 3.2 shows SLT and SLF voltages as a function of the searched data. In the ML pre-charge phase (FIGURE 3.2.4 (a)), the ML is pre-charged high at a voltage V_{DD_ML} with a PMOS transistor. The voltage applied on the 1T1R during the search operation (applied on the ML during the pre-charge phase) is referred to as the *search voltage*, V_{search} . In the ML sensing phase (FIGURE 3.2.4 (b)), the ML is left floating and starts discharging through each TCAM bitcell. The discharge follows that of a RC circuit [2]. The capacitance C_{ML} consists of the ML wiring capacitance and depends on ML length. The equivalent resistance R_{ML} depends on the state of each RRAM, *i.e.* their resistance value. If the stored and searched data match (FIGURE 3.2.5 (a)), the activated transistor is in series with a RRAM programmed in HRS. The equivalent resistance R_{ML} is high, the ML slowly discharges through leakage currents, I_{match} , and stays high. If at least one bitcell of the stored data mismatches with the searched data (FIGURE 3.2.5 (b)), the activated transistors of the mismatching bitcells are in series with RRAMs in LRS. The equivalent resistance R_{ML} is low, the ML quickly discharges through mismatching currents, $I_{mismatch}$, and the ML is pulled down to a low level. R_{ML} decreases with more mismatching bits which accelerates the discharge.

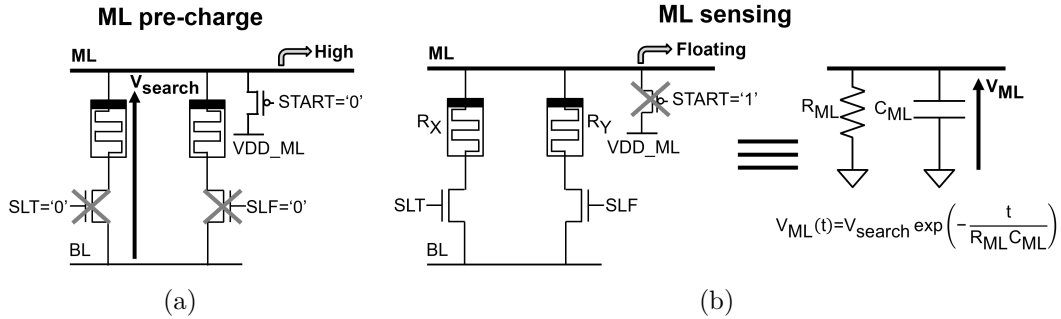


FIGURE 3.2.4: (a) In the Match Line (ML) pre-charge phase, the ML is pre-charged high at a voltage V_{search} . (b) In the ML sensing phase, the ML is left floating and discharges through each TCAM cell. The discharge follows that of a RC circuit.

3.2.3.2 Performance and reliability metrics definition

During a search operation, the ML voltage is compared to a reference voltage, V_{REF} , by a Sense Amplifier (SA) circuit. The SA returns the comparison result with the signal SA_OUT. FIGURE 3.2.6 (Top) sketches an example of ML voltage evolution during a search operation in a match (green) and mismatch of 1 bit (red) case. FIGURE 3.2.6 (Bottom) shows the corresponding measured SA_OUT waveforms. During the ML pre-charge phase, SA_OUT is at '1'. SA_OUT goes to '0' at the beginning of the sensing phase and remains at '0' while the ML voltage is higher than V_{REF} . SA_OUT goes back to '1' once the ML voltage goes

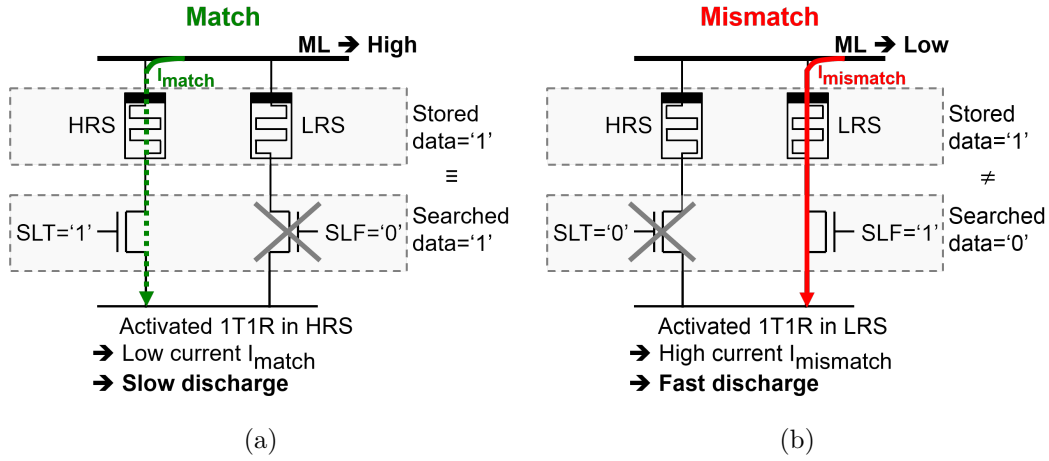


FIGURE 3.2.5: (a) In a match case, the activated transistor is in series with a RRAM in the High Resistance State (HRS). (b) In a mismatch case, the activated transistor is in series with a RRAM in the Low Resistance State (LRS).

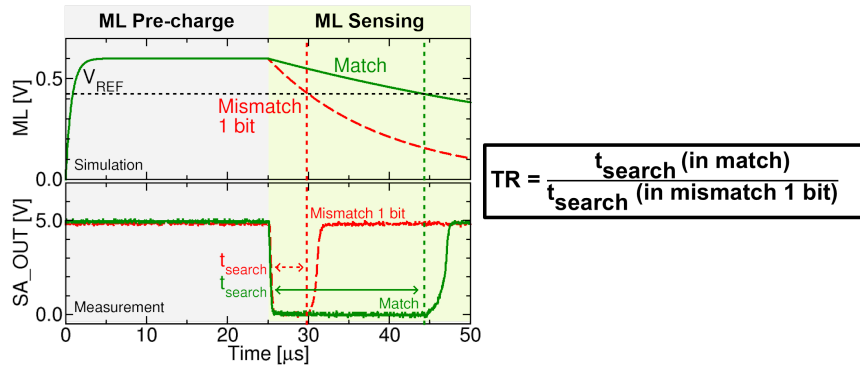


FIGURE 3.2.6: (Top) Example of the Match Line (ML) voltage evolution during a search operation in the case of match (green) and 1-bit mismatch (red). (Bottom) Corresponding measured waveforms output by the sense amplifier, SA_OUT. The duration for which SA_OUT stays at '0' defines the ML discharge time, t_{search} .

below V_{REF} , *i.e.* when the ML can be considered as discharged. The duration for which SA_OUT stays at '0' corresponds to the ML discharge time, t_{search} . Capacitances can be added on the ML with the signal CAP_CALIB (see FIGURE 3.2.1 (a)) to slow down ML discharge and facilitate measurements.

We define the *search time* as the minimal time required to discriminate between a match and a mismatch, *i.e.* the discharge time in the worst-case scenario when only one bit is mismatching (hardest mismatch to detect as it is the slowest mismatch case). Lower the search time, better the TCAM performance. To assess reliability, we characterised the sensing margin and the maximum number of cycles during searching (TCAM read) and programming (TCAM write). To assess the sensing margin, *i.e.* the ability of the TCAM to discriminate between a match and a mismatch, three metrics can be adopted: the resistance-based sensing margin (ML resistance ratio between the match and 1-bit mismatch cases [29, 32]), the voltage-based sensing margin (ML voltage difference between the match and 1-bit mismatch cases at a certain t_{search} [26]), and the time-based

sensing margin. Here, we adopt the third one since we can only measure ML discharge times. We define the *Time Ratio* (TR) as the t_{search} ratio between the match and 1-bit mismatch cases:

$$\text{TR} = \frac{t_{\text{search}} (\text{in match})}{t_{\text{search}} (\text{in mismatch 1 bit})} \quad (3.2.1)$$

TR has to be maximised in order to guarantee a sufficient sensing margin to improve the parallel search capability. Second, we fully characterised the *search endurance*, *i.e.* the maximum number of search operations before disturbing TCAM bitcells (RRAM read disturb). Third, we measured the *programming endurance*, *i.e.* the maximum number of programming operations before the TCAM bitcells break (RRAM breakdown). TABLE 3.3 summarises the different metrics used to assess TCAM performance and reliability.

	Name	Definition	Requirement
Performance	Search Time	Discharge time in the 1-bit mismatch state	Lower is better
	Time Ratio (TR)	Sensing margin, discharge time ratio between the match and 1-bit mismatch states	
Reliability	Search Endurance	Maximum number of search operations before disturbing TCAM cells	Higher is better
	Programming Endurance	Maximum number of programming operations before TCAM cells break	

TABLE 3.3: Performance and reliability metric definition used in this work.

3.2.3.3 Circuit basic functionality: match line discharge time characterisation

To assess the impact of RRAM electrical properties on TCAM, RRAMs have been programmed with different programming conditions. FIGURE 3.2.7 shows the Low Resistance State (LRS), High Resistance State (HRS), and pristine cumulative distributions directly measured on the TCAM bitcells. TABLE 3.11 shows the associated programming conditions. HRS distributions have been obtained using the Soft and Strong HRS conditions. Here, the Memory Window (MW) is defined as the ratio between the HRS and LRS values at -2σ and $+2\sigma$ of the distributions, respectively. Using Strong HRS instead of Soft HRS allows to increase the MW from 27 (Soft HRS) to 230 (Strong HRS) at the cost of a decrease in programming endurance [31]. The pristine resistance distribution can be used if the TCAM is programmed only once.

We first verified the basic functionality of the circuit. We applied a search voltage, V_{search} , of 0.6 V (voltage across ML and BL during the pre-charge phase, *cf* FIGURE 3.2.4 (a)), and we varied the ML capacitance (signal CAP_CALIB, *cf* FIGURE 3.2.1 (a)). As V_{search} is applied directly across RRAMs, it has to be kept relatively low not to disturb the RRAM states. FIGURE 3.2.8 (a) shows the discharge time, t_{search} , as a function of ML capacitance in the case of match (green) and mismatch of 1 bit and 128 bits (red). RRAMs are programmed using the LRS, (Left) Soft HRS, and (Middle) Strong HRS programming conditions of FIGURE 3.2.7, or (Right) kept in pristine state. Increasing ML capacitance

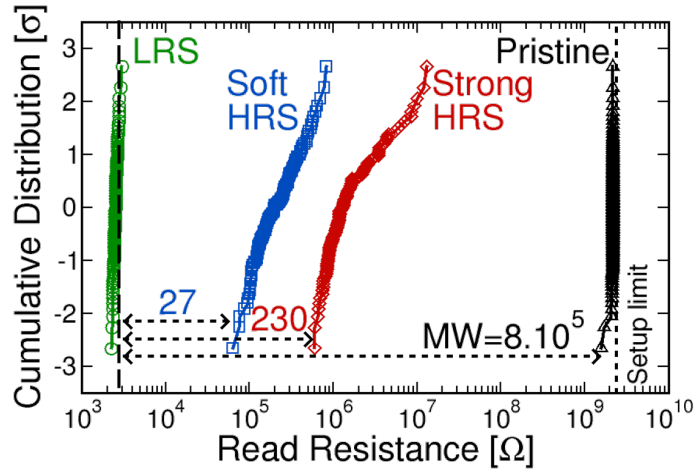


FIGURE 3.2.7: Low Resistance State (LRS), High Resistance State (HRS), and pristine resistance cumulative distributions directly measured on the TCAM cells. HRS resistance distribution can be obtained using the Soft HRS or Strong HRS programming conditions.

	LRS	Soft HRS	Strong HRS	Pristine
V_{top} (VDD_ML)	2.0 V	GND	GND	-
V_{bottom} (BL)	GND	2.5 V	2.5 V	-
I_{prog}	200 μA	-	-	-
V_{gate} (SLT or SLF)	-	3.0 V	3.5 V	-
R_{median}	2.52 k Ω	198 k Ω	1.27 M Ω	2.4 G Ω (limit)
$R_{\pm 2\sigma}$	2.79 k Ω	76.1 k Ω	654 k Ω	2.1 G Ω
MW (@ 2σ)		27	230	8.10^5
Programming endurance		10^6	10^4	1

TABLE 3.4: Programming conditions used for the characterisation of the 2T2R structure.

slows down ML discharge as expected from a RC circuit. Due to equipment limitations, we could not measure t_{search} without additional capacitances on the ML as the discharge was too fast. In the following, measurements are performed with an additional capacitance of 315 pF.

We then characterised the impact of the search voltage, V_{search} , on t_{search} when the RRAMs are programmed in Soft HRS, Strong HRS, or kept in pristine state (FIGURE 3.2.8 (b)). To vary V_{search} , we kept the pre-charge ML voltage, VDD_ML, constant at 2.6 V, and varied BL voltage (*cf* FIGURE 3.2.4 (a)). Increasing V_{search} decreases t_{search} . In addition, t_{search} decreases with more mismatching bits or when the TCAM is programmed in Soft HRS instead of Strong HRS since the equivalent ML resistance, R_{ML} , is decreased.

The measurements performed prove the functionality of the circuit. The search time, *i.e.* the discharge time when only 1 bit mismatches (slowest mismatch case) has been measured. In order to improve performance (decrease the search time), the following strategies can be adopted:

- Decrease the match line capacitance, C_{ML} , *i.e.* avoid long match lines.
- Decrease the match line equivalent resistance, R_{ML} , by using lower LRS

resistance values, *i.e.* higher programming current to program the RRAM cells.

- Increase the search voltage, V_{search} .

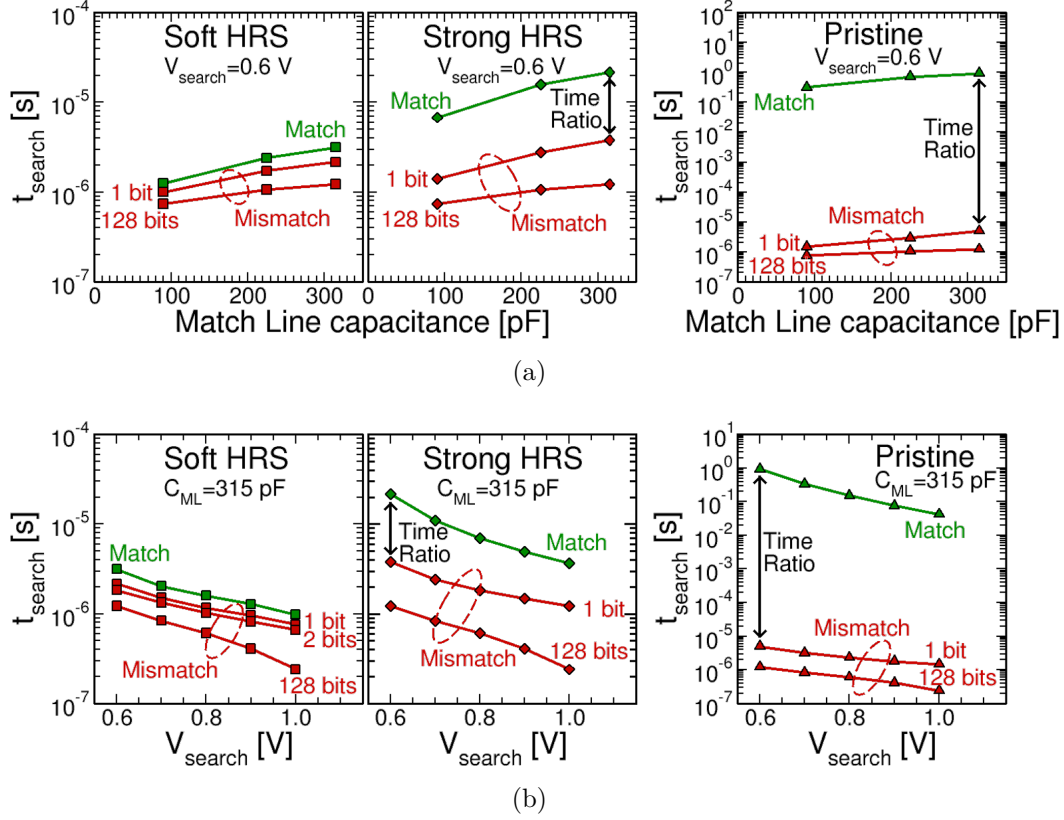


FIGURE 3.2.8: Discharge time, t_{search} , as a function of (a) Match Line (ML) capacitance (the search voltage, V_{search} , is fixed at 0.6 V), and (b) V_{search} (ML capacitance is fixed at 315 pF).

3.2.3.4 Sensing margin and search capacity

In this section, we characterise TCAM reliability in terms of sensing margin using the Time Ratio (TR) defined in EQUATION 3.2.1. As it is not possible to perform measurements without additional capacitances, we first studied the impact of ML capacitance on the TR. FIGURE 3.2.9 (a) shows the TR as a function of ML capacitance when the TCAM is programmed in Soft (square) and Strong (diamond) HRS. TR is almost constant with ML capacitance. Thus, the results we obtained on TR hold true whether or not an additional capacitance is added. In the following, an additional capacitance of 315 pF is used.

The ideal TCAM should minimise the search time (discharge time in the 1-bit mismatch state) while maximising the TR. FIGURE 3.2.9 (b) shows the impact of the search voltage, V_{search} , on TR when bitcells are programmed in Soft HRS (square), Strong HRS (diamond), or kept in pristine state (triangle). First, TR slightly decreases when V_{search} increases. Second, increasing HRS resistance values from Soft HRS to Strong HRS ($\approx 7x$) improves TR by 4x. Indeed, in the case of match, the ML discharges through leakage currents flowing through

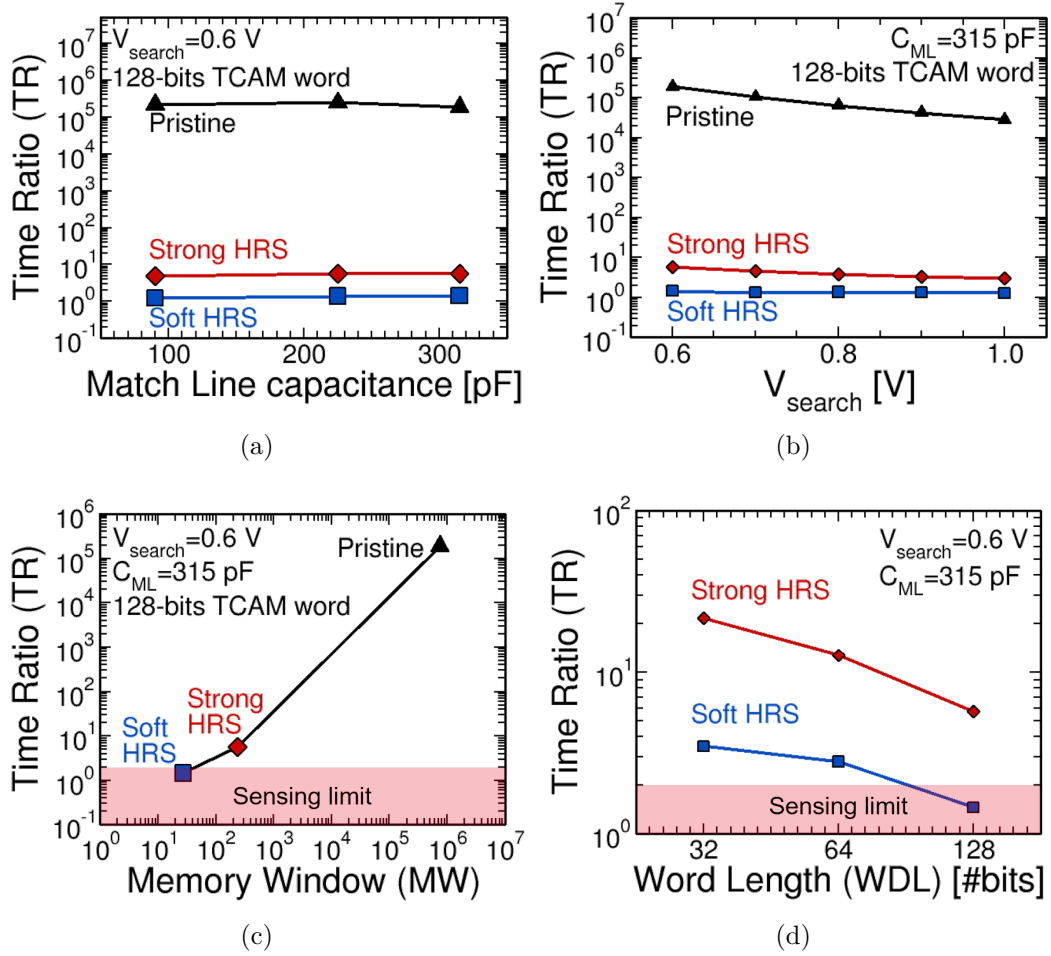


FIGURE 3.2.9: Measured Time Ratio (TR) as a function of (a) match line capacitance, (b) search voltage V_{search} , (c) memory window, and (d) TCAM word length.

RRAMs in HRS. If HRS resistance values are too low, the sum of leakage currents can be comparable to a mismatching current. This leads to a loss of the TR. To evaluate the impact of HRS resistance values on TR, we plot in FIGURE 3.2.9 (c) the TR as a function of the Memory Window (MW) for a given V_{search} of 0.6 V. MW is varied by varying HRS resistance values, LRS resistance values being the same for each condition. In accordance with reported simulations [31, 32], the TR increases with MW. The same result is obtained for any V_{search} from 0.6 V to 1.0 V (not shown). If we consider a minimal TR of 2 to guarantee reliable searches, a minimal MW of 50 is required with 128-bit TCAM words for a given V_{search} of 0.6 V. Finally, we studied the impact of TCAM word length on the sensing margin. Indeed, increasing the number of bitcells per TCAM word increases leakage currents. FIGURE 3.2.9 (d) shows the TR as a function of TCAM Word Length (WDL) for Soft and Strong HRS and for a given V_{search} of 0.6 V. To emulate smaller TCAM words, certain TCAM cells were completely deactivated by keeping both transistors OFF and RRAMs in pristine state. Therefore, these bitcells behave as matching bitcells with negligible impact on ML discharge with respect to bitcells with RRAMs in LRS or HRS. As expected, the TR decreases with word length as this increases the sum of leakage currents in the match state. The same result is obtained for any

V_{search} from 0.6 V to 1.0 V (not shown). This limits the length of TCAM words to 97 bits when RRAMs are programmed in Soft HRS for a given V_{search} of 0.6 V. Therefore, Soft HRS cannot be used for 128-bits TCAM words, and stronger programming conditions (Strong HRS) are required.

To sum up, reliability in terms of sensing margin can be improved by:

- Decreasing the search voltage, V_{search} .
- Increasing RRAM memory window.
- Decreasing TCAM word length.

3.2.3.5 Search endurance characterisation

During a search operation, a positive voltage V_{search} is applied on RRAM top electrodes (R_X for search '1', R_Y for search '0') in the same polarity as a Set operation as sketched in FIGURE 3.2.10 (a). Therefore, in the match state, undesired switchings from HRS to LRS can occur after a certain number of search operations. This causes a match failure. In the mismatch state, undesired switchings from HRS to LRS accelerates ML discharge. The system remains in its mismatch state, there is no mismatch failure. Therefore, undesired switchings are detrimental only in the match state. Here, we define the search endurance as the maximum number of search operations we can perform before at least one RRAM switches from HRS to LRS when the system is initially in a match state, *i.e.* before we lose the match configuration.

We characterised the search endurance of the system by applying a series of search operations when the TCAM is initially in a match configuration, *i.e.* the search voltage, V_{search} , is applied across RRAMs programmed in Soft or Strong HRS. FIGURE 3.2.10 (b) shows the measured discharge time, t_{search} , after a series of search operations. Discharge times, t_{search} , have been normalised by the discharge time at the first search operation. We first studied the impact of programming conditions in FIGURE 3.2.10 (b, top). V_{search} is fixed at 0.6 V and ML capacitance, C_{ML} , at 315 pF. When programmed in Soft HRS (square), no RRAM switches before $9 \cdot 10^4$ searches. After $9 \cdot 10^4$ searches one RRAM switched from HRS to LRS (not shown). Programming in Strong HRS (diamond) instead of Soft HRS allows to improve the search endurance up to $4 \cdot 10^5$ searches. This is in accordance with [71] wherein the authors demonstrate that RRAMs with higher HRS resistance values require longer pulses to switch to the LRS.

Another way to improve the search endurance is by decreasing the search voltage, *i.e.* reducing the stress applied on RRAMs. FIGURE 3.2.10 (b, middle) shows the impact of the search voltage, V_{search} , on the search endurance. The TCAM is programmed in Soft HRS, and C_{ML} is fixed at 315 pF. Decreasing the search voltage, V_{search} , from 0.6 V (filled symbol) down to 0.4 V (open symbol) allows to improve the search endurance from $9 \cdot 10^4$ to $4.5 \cdot 10^5$ searches.

We finally studied the impact of ML capacitance. Indeed, these results are obtained in the worst-case scenario since the 315-pF ML capacitance artificially increases the search time, *i.e.* artificially increases the stress applied on RRAMs. FIGURE 3.2.10 (c, bottom) shows the impact of ML capacitance, C_{ML} , on the search endurance. The TCAM is programmed in Strong HRS, and V_{search} is

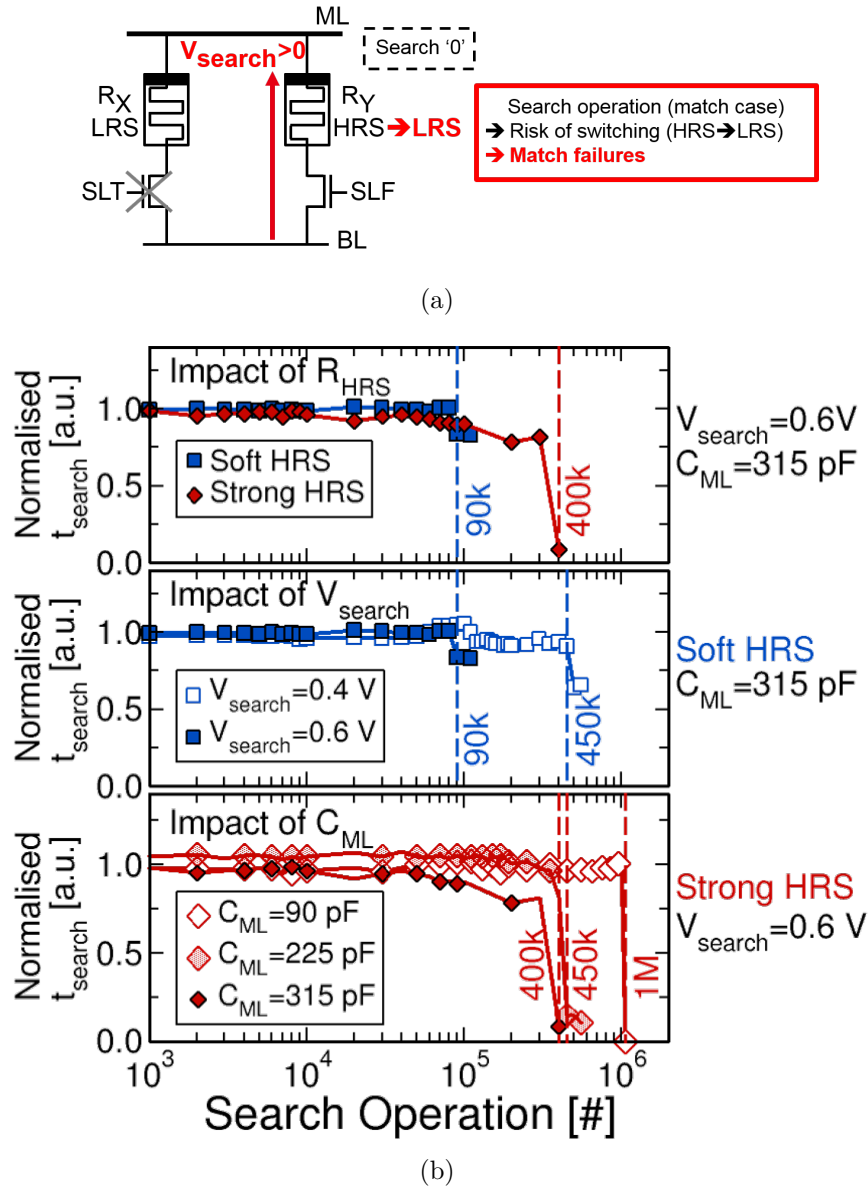


FIGURE 3.2.10: (a) During a search operation, a positive voltage is applied on RRAM top electrodes in the same configuration as a Set operation. (b) Characterisation of the search endurance. Measured discharge times, t_{search} , as a function of the number of search operations are reported.

fixed at 0.6 V. As expected, we observe an improvement in search endurance when ML capacitance is decreased from 315 pF (filled symbol) down to 90 pF (open symbol). At $C_{\text{ML}} = 90\text{ pF}$, a search endurance of $1 \cdot 10^6$ searches is reached. To sum up, reliability in terms of search endurance can be improved by:

- Increasing HRS resistance values, *i.e.* programming in Strong HRS instead of Soft HRS.
- Decreasing the search voltage, V_{search} .
- Decreasing match line capacitance, C_{ML} .

3.2.3.6 Programming endurance

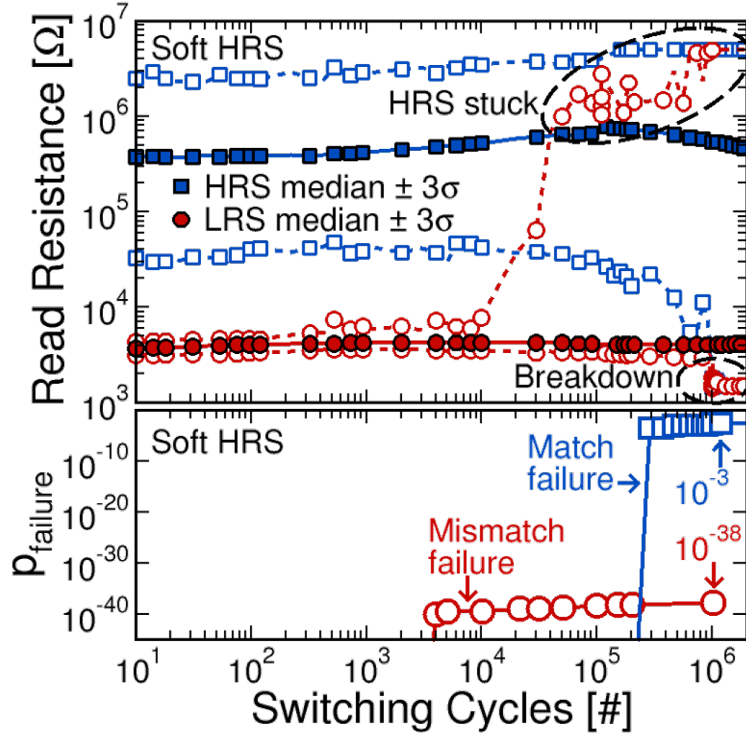


FIGURE 3.2.11: (Top) Soft HRS programming endurance characterisation. (Bottom) Probabilities of match (square) and mismatch (circle) failures as a function of the number of Set/Reset switching cycles.

FIGURE 3.2.11 (Top) shows a programming endurance characterisation measured on a 4-kbit RRAM 1T1R array (with the same stack as our RRAM cells) with Soft HRS programming conditions. After 10^4 Set/Reset switching cycles some cells remain stuck in HRS with a probability $p_{\text{HRS stuck}}$. After 10^6 Set/Reset switching cycles breakdown failures occur with a probability $p_{\text{breakdown}}$, and broken cells are stuck in LRS. Probabilities $p_{\text{HRS stuck}}$ and $p_{\text{breakdown}}$ have been extracted after different programming cycles from the endurance measurement, and their impact on TCAM circuits is evaluated. For a TCAM circuit, HRS stuck failures artificially increase the discharge time. Thus, they have no impact on matches. However, they can lead to mismatch failures. In a mismatch case with m bits mismatching, mismatch failures occur if all m bits remain stuck in HRS. This has a probability $p_{\text{HRS stuck}}^m$ to happen. Considering all the possible combinations leading to a mismatch of m bits, for m ranging from 1 to WDL (WDL being the TCAM word length), mismatch failures can occur with a probability $p_{\text{mismatch, failure}}$:

$$p_{\text{mismatch, failure}} = \frac{(1 + p_{\text{HRS stuck}})^{\text{WDL}} - 1}{2^{\text{WDL}} - 1} \quad (3.2.2)$$

with WDL=128 bits in our TCAM circuit. $p_{\text{mismatch, failure}}$ depends on the word length.

On the other hand, breakdown failures decrease the discharge time. Therefore, they have no impact on mismatches. However they lead to match failures with

a probability $p_{\text{match, failure}}$ if at least one matching cell is impacted, *i.e.*:

$$P_{\text{match, failure}} = P_{\text{breakdown}} \quad (3.2.3)$$

$P_{\text{match, failure}}$ is independent of the word length. FIGURE 3.2.11 (Bottom) shows probabilities of match (blue) and mismatch (red) failures as a function of the number of Set/Reset switching cycles. As probability of mismatch failures remains negligible after every programming operation ($< 10^{-38}$), a programming endurance of 10^6 cycles can be reached with Soft HRS. If the TCAM is programmed in Strong HRS, programming endurance is degraded down to 10^4 cycles [31]. Therefore, for 128-bits TCAM words, a programming endurance of 10^4 cycles can be reached since Soft HRS cannot be used.

3.2.3.7 Extrapolated figures of merit

Due to equipment limitations, all the measurements were performed with additional Match Line (ML) capacitances. We extrapolated performance and reliability of the system without any additional capacitance, *i.e.* considering only the intrinsic ML capacitance. The ML has an intrinsic capacitance of 4 pF, measured on the TCAM circuit and consistent with simulation extraction. We fitted measured t_{search} as a function of ML capacitance from FIGURE 3.2.8 (a) with a linear function as the discharge behaves as a RC circuit. FIGURE 3.2.12 (a) shows the extrapolated discharge time, t_{search} , as a function of ML capacitance (dotted line) in the case of match (green) and mismatch of 1 bit (red). Extrapolations have been confirmed by simulations (open symbol). Considering the intrinsic ML capacitance of 4 pF, a search time (t_{search} in the case of 1-bit mismatch) of 90 ns and a time ratio, TR, of 3.1 are obtained (for $V_{\text{search}}=0.6\text{V}$ and Strong HRS). Finally, we plot in FIGURE 3.2.12 (b) the search endurance as a function of the ML capacitance. The search endurance improves when ML capacitance decreases. Less stress is applied on RRAMs when ML capacitance decreases since the discharge time decreases.

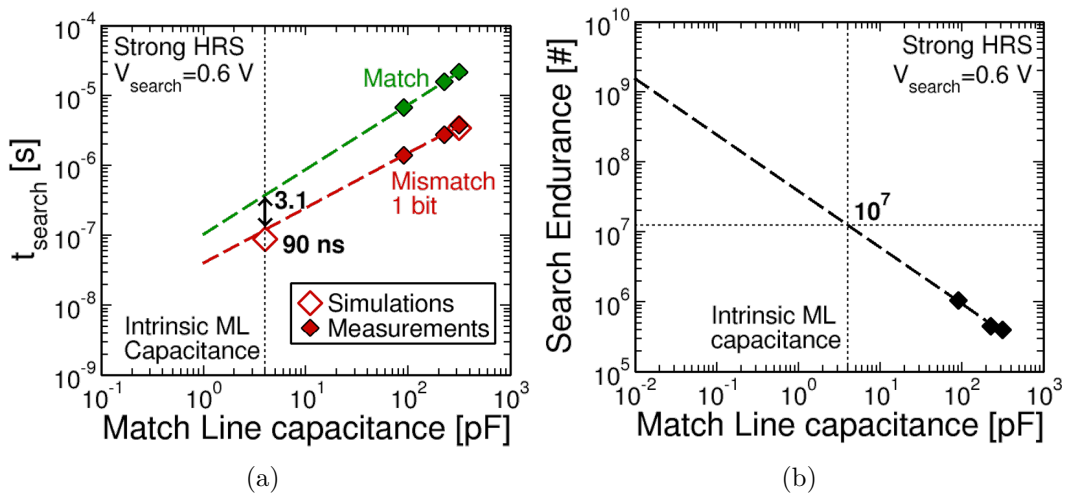


FIGURE 3.2.12: (a) Extrapolated t_{search} as a function of the match line capacitance. (b) Extrapolated search endurance as a function of the match line capacitance.

		Search voltage V_{search} [V] 0.6 \longrightarrow 1.0	Memory window (HRS/LRS) ~30 \longrightarrow ~100	Match Line capacitance [pF] 90 \longrightarrow 315
Performance (Lower is better)	Search Time	\downarrow (/3)	\uparrow ($\times 1.5$)	\downarrow (/3)
Reliability (Higher is better)	Sensing Margin (Time Ratio)	\downarrow (/1.5)	\uparrow ($\times 4$)	-
	Search Endurance	\downarrow (N/A)	\uparrow ($\times 4$)	\downarrow (/2.5)
	Prog. Endurance	-	\downarrow (/100)	-
Capacity (Higher is better)	Word Length max	\downarrow (/1.3)	\downarrow (/3)	-

TABLE 3.5: Summary of the characterisation performed in this section on the common 2T2R TCAM bitcell.

3.2.3.8 Discussion and conclusion

TABLE 3.5 summarises the results of characterisation of the 2T2R structure. Performance and reliability have been characterised as a function of RRAM electrical properties (*i.e.* HRS value), the voltage V_{search} applied across RRAMs during a search operation, and Match Line (ML) capacitance. We demonstrated that a trade-off exists between performance and reliability: shorter search times can be obtained by either increasing V_{search} or programming the RRAM cells in Soft HRS at the expense of a degraded sensing margin (Time Ratio) and search endurance. In addition, we showed that decreasing ML capacitance has little impact on the sensing margin (FIGURE 3.2.9 (a)), while improving the search time (FIGURE 3.2.8 (a)). Using scaled CMOS technology nodes lowers the ML capacitance, therefore an improvement in performance and search endurance can be expected at similar sensing margin and programming endurance.

A critical limitation of this structure is the strong dependency of the sensing margin on the Memory Window (MW), *i.e.* with RRAM programming conditions (FIGURE 3.2.9 (c), summarised in TABLE 3.5). This puts constraints on the maximum permitted length of TCAM words (FIGURE 3.2.9 (d)). As characterisation results evidenced, the TCAM is limited to 97 bits per word when it is programmed in Soft HRS, and it is mandatory to program in Strong HRS to enable the use of longer words. However, this leads to a decrease in programming endurance by 100x. This is due to the fact that, in the case of match, the ML still discharges due to leakage currents flowing through each RRAM in HRS. Therefore, larger MWs are required for longer TCAM words in order to limit leakage currents. This can be detrimental for classic applications, such as internet protocol packet routing [34] which requires pattern matching with words longer than 128 bits and extremely low search times (below nanoseconds). However, multi-core neuromorphic computing architectures would be little affected by these problems. Indeed, they usually implement CAM tables small in word length and can tolerate longer search times (*e.g.* 10 bits of word length and search times in the order of tens to hundreds of nanoseconds for the DYNAPs [59]). In terms of programming endurance, CAM tables are programmed only during network topology configuration. Therefore, RRAM programming endurance is not a key metric. On the other hand, search endurance requirement is application-dependent. For instance, a total of 0.5

million events has to be searched for the card classification application tested with DYNAPs. A search endurance of 10^6 search operations is sufficient for the classification of 52 cards (one deck), but it may be too low if more inputs need to be classified. For the car detection application studied in CHAPTER 2 [72], each neuron receives in average an input event every 364 ms. For a search endurance of 10^6 search operations, the network can continuously operate without any search failure for about four days.

A possible way to improve the sensing margin and enable the use of this TCAM structure for more conventional applications is to use the two-bits encoding scheme from Li et al. [29]. FIGURE 3.2.13 depicts the principle of the two-bits encoding scheme. In the conventional encoding scheme that has been used here (FIGURE 3.2.13 (Left)), one bit is encoded with one TCAM bitcell. During a search operation, one transistor out of two is activated, *i.e.* 128 transistors out of 256 for a 128-bits TCAM word. In the case of match, leakage currents, I_{match} , flow through 128 1T1R structures in HRS. The principle of the two-bit encoding scheme is that, instead of coding one bit with one TCAM bitcell, two bits are encoded with an association of two TCAM bitcells (FIGURE 3.2.13 (Right)). During a search operation, this allows to activate only one transistor out of four, *i.e.* 64 transistors out of 256 for a 128-bits TCAM word. This improves the sensing margin as leakage currents, I_{match} , are halved. Note that we do not lose any storage capacity, *i.e.* the TCAM has the same word length with and without the two-bits encoding scheme. We measured the Time Ratio (TR) as a function of ML capacitance for our 128-bits TCAM circuit using the two-bits encoding scheme of [29]. The results are shown in FIGURE 3.2.14. The TCAM is programmed in Strong HRS and $V_{\text{search}}=0.6$ V. The TR improves by 3.8x with the two-bits encoding scheme for a 128-bits TCAM word. As a result, longer TCAM words can be implemented with a lower MW. In particular, this enables the use of Soft HRS programming conditions for 128-bits TCAM words. A drawback of this technic is that it degrades performance by 3%, *i.e.* it increases search times since half as many transistors are turned ON. However, this is acceptable if the slight loss of performance can be tolerated in order to improve the reliability.

3.2.4 Novel 1T2R1T TCAM circuit characterisation

3.2.4.1 1T2R1T TCAM bitcell working principle

In this section, a novel RRAM-based TCAM bitcell was designed, integrated in a 3x128 bits TCAM circuit, and fabricated using the same CMOS and RRAM technologies as the common 2T2R TCAM structure characterised in the previous section (*c.f.* SECTION 3.2.1). FIGURE 3.2.15 depicts the new TCAM bitcell composed of two transistors and two RRAMs in a 1T2R1T configuration. Two RRAMs (2R) compose a voltage divider that biases the transistor gate of N2 (1T). An additional transistor N1 (1T) works as an access transistor to program the RRAMs. To store a data, RRAMs are programmed either in HRS or LRS following the same combinations as the common 2T2R TCAM bitcell (*cf* TABLE 3.8). Forming, Set, and Reset operations are performed by applying the required

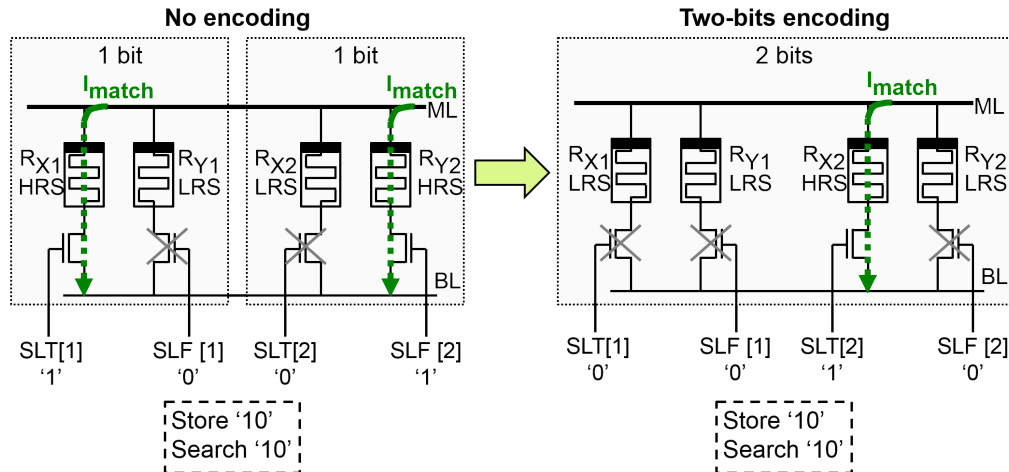


FIGURE 3.2.13: Two-bits encoding principle. (Left) When no encoding is used, two bits are encoded with two distinct TCAM bitcells. During a search operation, two out of the four transistors are turned ON. In the case of match, leakage currents, I_{match} , flow through two 1T1R structures in HRS. (Right) When the two-bits encoding scheme is used, two bits are encoded with an association of two TCAM bitcells. During a search operation, only one out of four transistors is turned ON. In the case of match, leakage currents are halved with respect to the case of no encoding as leakage currents only flow in one 1T1R structure in HRS. Reproduced from [30].

Stored data	R_{X1}	R_{Y1}	R_{X2}	R_{Y2}
'00'	HRS	LRS	LRS	LRS
'01'	LRS	HRS	LRS	LRS
'10'	LRS	LRS	HRS	LRS
'11'	LRS	LRS	LRS	HRS
'0X'	HRS	HRS	LRS	LRS
'1X'	LRS	LRS	HRS	HRS
'X0'	HRS	LRS	HRS	LRS
'X1'	LRS	HRS	LRS	HRS
'XX'	HRS	HRS	HRS	HRS

TABLE 3.6: RRAM state definition as a function of the stored data for the two-bits encoding scheme. Reproduced from [30].

Searched data	SLT[1]	SLF[1]	SLT[2]	SLF[2]
'00'	VDD	0	0	0
'01'	0	VDD	0	0
'10'	0	0	VDD	0
'11'	0	0	0	VDD

TABLE 3.7: SLT and SLF voltages as a function of the searched data for the two-bits encoding scheme. Reproduced from [30].

programming voltages on SLT, SLF, WL, and BL. To read RRAM resistance

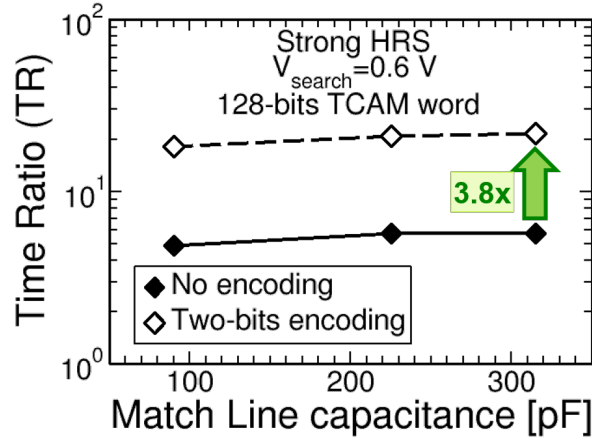


FIGURE 3.2.14: Measured Time Ratio (TR) as a function of the match line capacitance using the two-bits encoding scheme of [29]. The TR improves by 3.8x with the two-bits encoding scheme.

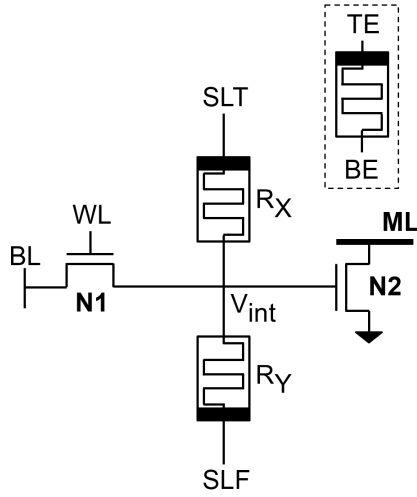


FIGURE 3.2.15: Proposed 1T2R1T TCAM bitcell schematic. Top (TE) and Bottom (BE) Electrodes are indicated with the black rectangle.

Stored data	R_X	R_Y
'0'	LRS	HRS
'1'	HRS	LRS
'X'	HRS	HRS

TABLE 3.8: RRAM state definition as a function of the stored data.

Searched data	SLT	SLF
'0'	0	V_{search}
'1'	V_{search}	0

TABLE 3.9: SLT and SLF voltages as a function of the searched data.

values, read voltage is applied on either SLT or SLF, and the current is read from BL. TABLE 3.10 summarises the voltages applied during a programming or read operation.

During a search operation, a voltage V_{search} is applied across the RRAM voltage divider, *i.e.* between SLT and SLF. TABLE 3.9 shows SLT and SLF voltages as a function of the searched data. In the ML pre-charge phase, the ML is pre-charged high at $V_{\text{DD_ML}}$. In the ML sensing phase, the ML is left floating, and the voltage V_{search} is applied between SLT and SLF. If the stored and searched data match (FIGURE 3.2.16 (a)), the internal node V_{int} in the RRAM voltage divider is kept at 0 V if the resistance ratio between R_X and R_Y is sufficiently high. Transistor N2 is OFF, and the ML stays high. If the stored and searched data mismatch (FIGURE 3.2.16 (b)), V_{int} is almost equal to V_{search} . This turns ON the transistor N2 if V_{search} is higher than the threshold voltage, $V_{\text{th,N2}}$, of the

	SLT	SLF	WL	BL
Forming R_X	V_{forming}	GND	$V_{\text{gate,forming}}$	GND
Set R_X	V_{set}	GND	$V_{\text{gate,set}}$	GND
Reset R_X	GND	V_{reset}	$V_{\text{gate,reset}}$	V_{reset}
Read R_X	V_{read}	GND	VDD	GND
Forming R_Y	GND	V_{forming}	$V_{\text{gate,forming}}$	GND
Set R_Y	GND	V_{set}	$V_{\text{gate,set}}$	GND
Reset R_Y	V_{reset}	GND	$V_{\text{gate,reset}}$	V_{reset}
Read R_Y	GND	V_{read}	VDD	GND

TABLE 3.10: Programming scheme for the proposed 1T2R1T TCAM.

transistor N2, and the ML is pulled down to a low level.

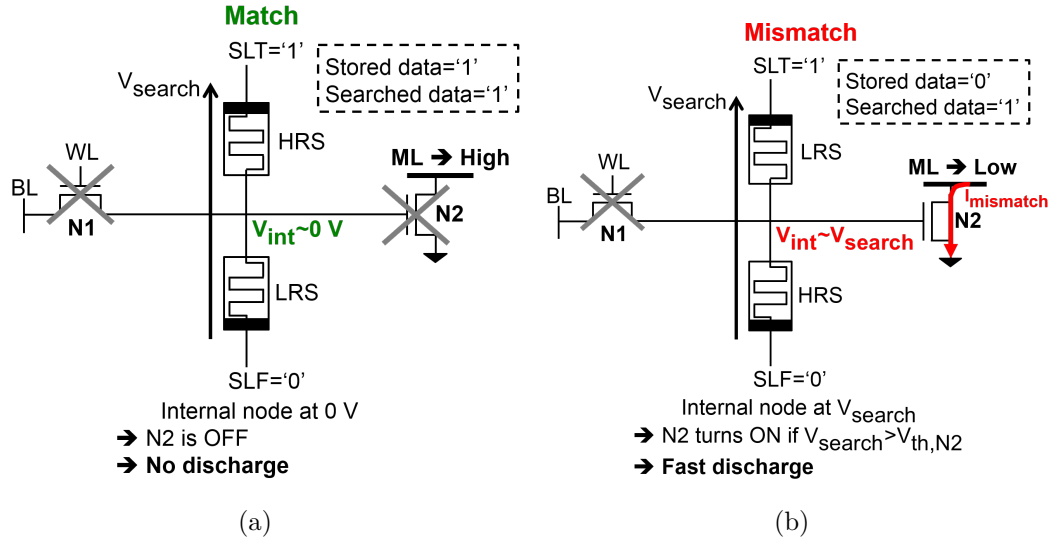


FIGURE 3.2.16: (a) In a match case, the internal node voltage, V_{int} , in the RRAM voltage divider is kept at 0 V, the transistor N2 is OFF. (b) In a mismatch case, V_{int} is almost equal to V_{search} , the transistor N2 turns ON if V_{search} is higher than the threshold voltage of transistor N2, $V_{\text{th,N2}}$.

3.2.4.2 Comparison between the two TCAM structures

In the common 2T2R RRAM-based TCAM [28–32] characterised in the previous section, the top electrodes of both RRAMs are connected to the ML. During the ML sensing phase, current flows in the 1T1R branches with the selector transistor in the ON state which discharges the ML. In the case of match the ML slowly discharges through RRAMs in HRS as shown in FIGURE 3.2.17 (a, top). In the case of mismatch (FIGURE 3.2.17 (a, bottom)) the ML discharges through RRAMs in LRS. Since the leakage currents, I_{match} , of the TCAM cells on the same ML add together, the low resistance ratio between HRS and LRS (memory window) degrades the sensing margin as we proved in the previous section (FIGURE 3.2.9 (c), time ratio as a function of the memory window). We showed that this limits the maximum length of TCAM words (FIGURE 3.2.9 (d), time ratio as a function of the TCAM word length). In the proposed 1T2R1T

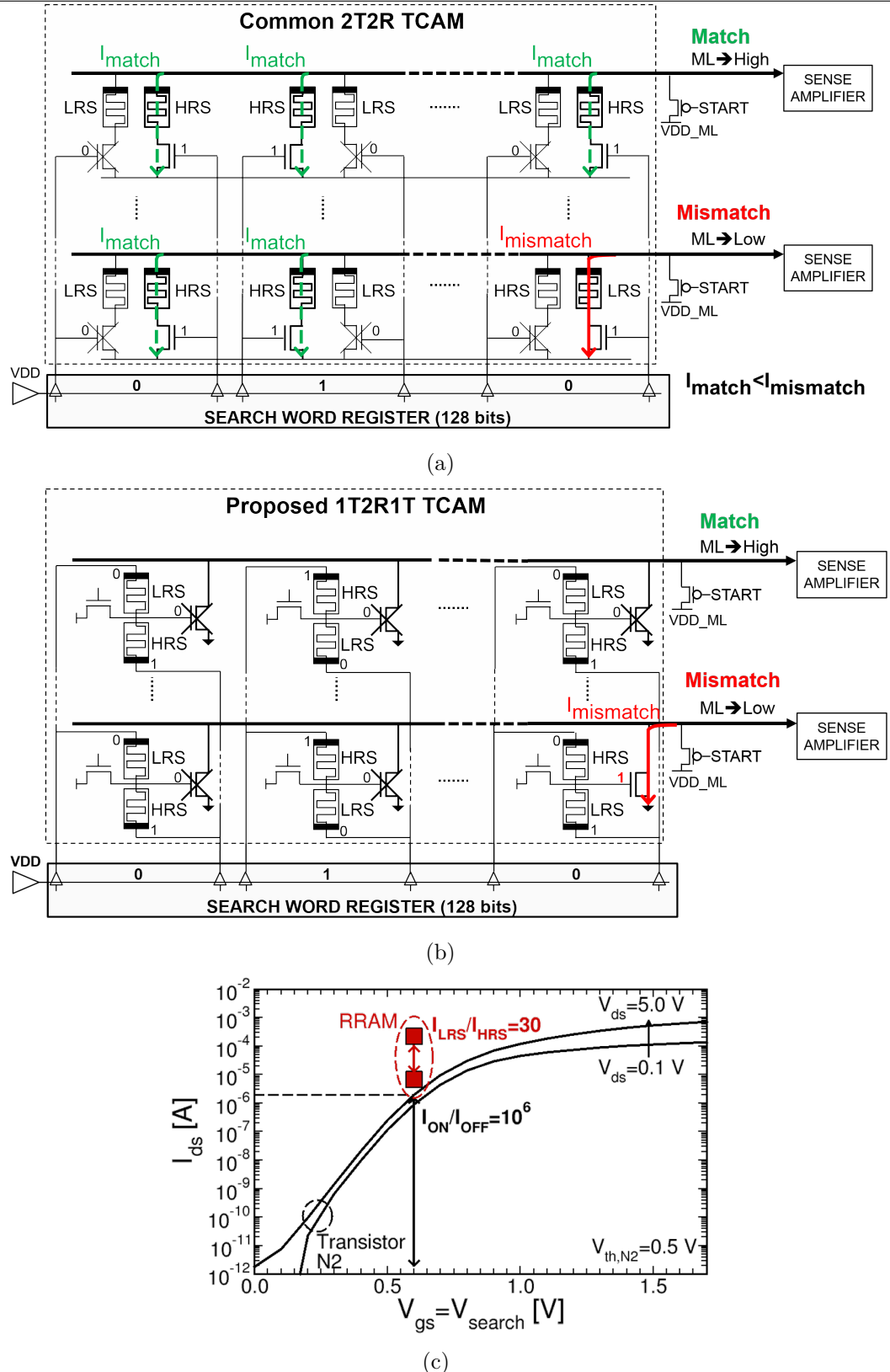


FIGURE 3.2.17: Match and mismatch cases for (a) the common 2T2R and (b) the proposed 1T2R1T structures. The sensing margin ($\approx I_{mismatch}/\sum I_{match}$) of the common 2T2R structure depends on the memory window, whereas for the proposed 1T2R1T structure, it depends on the transistor N2 characteristic. (c) Measured I_{ds} - V_{gs} characteristic of transistors N2. In the case of match, $V_{gs}=0$ V. In the case of mismatch, $V_{gs}=V_{search}$.

structure the ML is connected to transistors (transistor N2) controlled by the RRAM voltage divider. In the case of match (FIGURE 3.2.17 (b, top)) the ML slowly discharges through NMOS transistors in the OFF state. In the case of mismatch (FIGURE 3.2.17 (b, bottom)) the ML discharges through NMOS transistors in the ON state. Therefore, the sensing margin no longer depends on the RRAM memory window (≈ 30 for Soft HRS, ≈ 100 for Strong HRS) but on the MOSFET current ratio between I_{ON} and I_{OFF} . FIGURE 3.2.17 (c) shows the measured I_{ds} - V_{gs} characteristic of transistors N2. The threshold voltage, $V_{th,N2}$, is 0.5 V. At a search voltage, V_{search} , of 0.6 V, a ratio between I_{ON} and I_{OFF} of 10^6 is obtained in the proposed 1T2R1T structure.

In the previous section, we used the Time Ratio (TR) as a metric to assess the sensing margin. However, in the proposed 1T2R1T structure, the discharge time in the match case is longer than the limit of measurement of one second as shown in the measured SA_OUT waveforms in FIGURE 3.2.18. For the sake of a fair comparison, we keep the TR to assess the sensing margin by fixing t_{search} in match at 1 s for the proposed 1T2R1T TCAM.

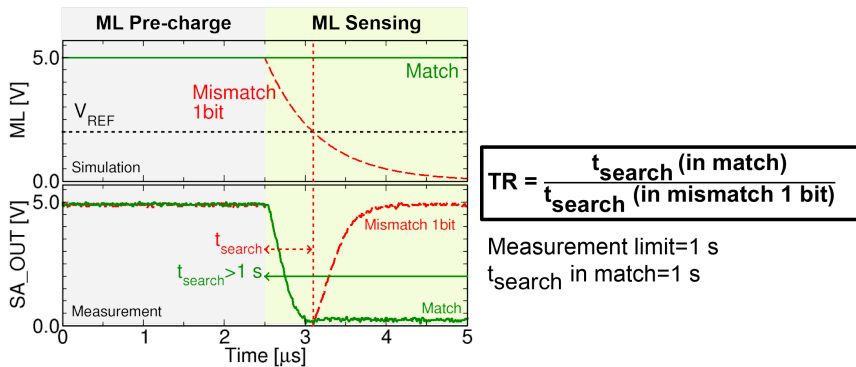


FIGURE 3.2.18: (Top) Example of the Match Line (ML) voltage evolution during a search operation in the case of match (green) and 1-bit mismatch (red). The ML does not discharge in the match case. (Bottom) Corresponding measured waveforms output by the sense amplifier, SA_OUT. The duration for which SA_OUT stays at '0' defines the ML discharge time, t_{search} . t_{search} is longer than the measurement limit (one second) in the match case.

3.2.4.3 Circuit basic functionality: match line discharge time characterisation

We first verified the basic functionality of the circuit. RRAMs have been programmed with the same programming conditions as those of the previous 2T2R structure (TABLE 3.11). FIGURE 3.2.19 shows the LRS, Soft HRS, Strong HRS, and pristine cumulative distributions directly measured on the 1T2R1T TCAM bitcells with their associated programming conditions in TABLE 3.11. In the measurements performed in this section, we did not add any additional capacitance on the ML as the discharge was slow enough to be measured: in the 1-bit mismatch state, at $V_{search}=0.6$ V, the ML discharges through a transistor with an equivalent resistance of about 1 M Ω , whereas in the previous structure

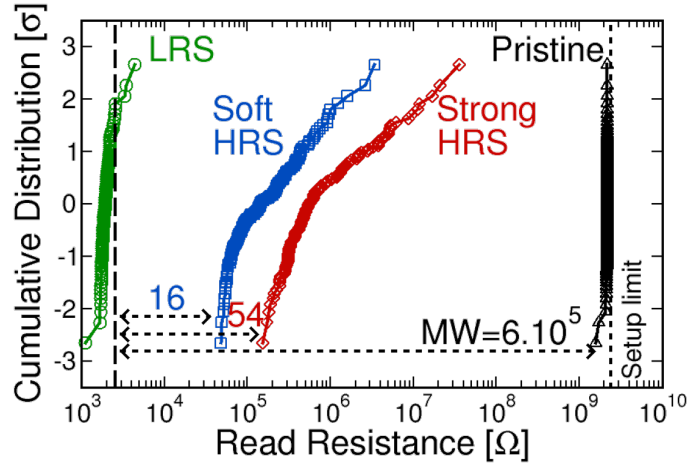


FIGURE 3.2.19: Low Resistance State (LRS), High Resistance State (HRS), and pristine resistance cumulative distributions directly measured on the TCAM cells. HRS resistance distribution can be obtained using the Soft HRS or Strong HRS programming conditions.

	LRS	Soft HRS	Strong HRS	Pristine
V_{prog} (set or reset)	2.0 V	2.5 V	2.5 V	-
I_{prog}	200 μA	-	-	-
V_{gate} (WL)	-	3.0 V	3.5 V	-
R_{median}	1.93 k Ω	143 k Ω	545 k Ω	2.4 G Ω (limit)
$R_{\pm 2\sigma}$	3.31 k Ω	51.7 k Ω	180 k Ω	2.1 G Ω
MW (@ 2σ)		16	54	$6 \cdot 10^5$
Programming endurance		10^6	10^4	1

TABLE 3.11: Programming conditions used for the characterisation of the 1T2R1R structure.

the equivalent resistance of the mismatching TCAM cell was about 3 k Ω . FIGURE 3.2.20 shows the measured discharge times, t_{search} , as a function of the search voltage, V_{search} (voltage applied across the RRAM voltage divider), in the case of match (green) and mismatch of 1 bit and 128 bits (red). RRAMs have been programmed either in Soft HRS (square), Strong HRS (diamond), or kept in pristine state (triangle). In the match case, t_{search} is higher than one second with any programming conditions as the ML does not discharge. In the 1-bit mismatch state, the ML discharges if V_{search} is higher than $V_{\text{th,N2}}=0.5$ V as transistor N2 of the mismatching cell turns ON. The higher V_{search} , the shorter the search time (discharge time in the 1-bit mismatch state) since this lowers the equivalent resistance of transistor N2. In addition, t_{search} is almost independent of RRAM programming conditions since it mostly depends on the current flowing through the transistor N2; a minimal resistance ratio between the RRAMs, R_X and R_Y , has to be ensured in order to activate N2 with the RRAM voltage divider, and Soft HRS programming conditions are sufficient. The search time can be improved by:

- Decreasing the match line equivalent resistance, *i.e.* by increasing the search voltage, V_{search} , and/or using transistors N2 with lower equivalent ON resistance values (*i.e.* using a scaled technology node).

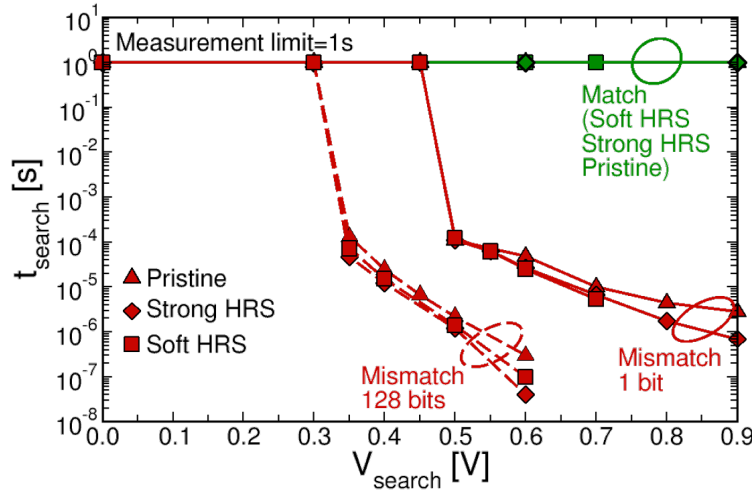


FIGURE 3.2.20: Discharge time, t_{search} , as a function of the search voltage, V_{search} , in the case of match (green) and mismatch (red) of 1 bit and 128 bits. t_{search} is almost independent of RRAM programming conditions.

- Decreasing the match line capacitance, *i.e.* avoid long match line.

3.2.4.4 Sensing margin and search capacity

The sensing margin (Time Ratio, TR) has to be maximised in order to improve the maximum search capacity, *i.e.* the maximum TCAM word length. FIGURE 3.2.21 (a) compares the TR as a function of the search voltage, V_{search} , for the common 2T2R structure characterised in the previous section (open symbol) and the proposed 1T2R1T structure (filled symbol). The TR improves by $>2000\times$ / $>5000\times$ for the Strong/Soft HRS programming conditions, respectively, thanks to the improved current ratio between the match state, I_{match} , and the 1-bit mismatch state, I_{mismatch} (10^6 for the proposed structure compared to roughly 100 for the previous one). Therefore, this structure enables the use of Soft HRS for 128-bits TCAM words at any V_{search} , unlike the previous structure. This allows to improve the programming endurance by 100x (*cf* TABLE 3.11). The TR of the proposed 1T2R1T TCAM increases with V_{search} since the ML discharges faster in mismatch while it does not discharge in match.

FIGURE 3.2.21 (b) compares the TR as a function of RRAM Memory Window (MW) for the common 2T2R TCAM (open symbol) and the proposed 1T2R1T (filled symbol) for a given V_{search} of 0.6 V. The sensing margin of the proposed 1T2R1T TCAM is insensitive to the MW, whereas that of the 2T2R TCAM could not operate for memory windows below 50. This is due to the fact that the discharge of ML for the 1T2R1T TCAM only depends on the current flowing through transistors N2. The same result is obtained for any V_{search} from 0.5 V to 0.7 V (not shown). We finally studied the impact of TCAM Word Length (WDL) on the sensing margin. In order to emulate smaller TCAM words, certain TCAM cells were deactivated by applying no voltage between their RRAMs, *i.e.* $V_{\text{search}}=0$ V (SLT=SLF=0). FIGURE 3.2.21 (c) compares the impact of TCAM Word Length (WDL) on the TR for the common 2T2R (open symbol) and the proposed 1T2R1T (filled symbol) structures. The plot has been obtained

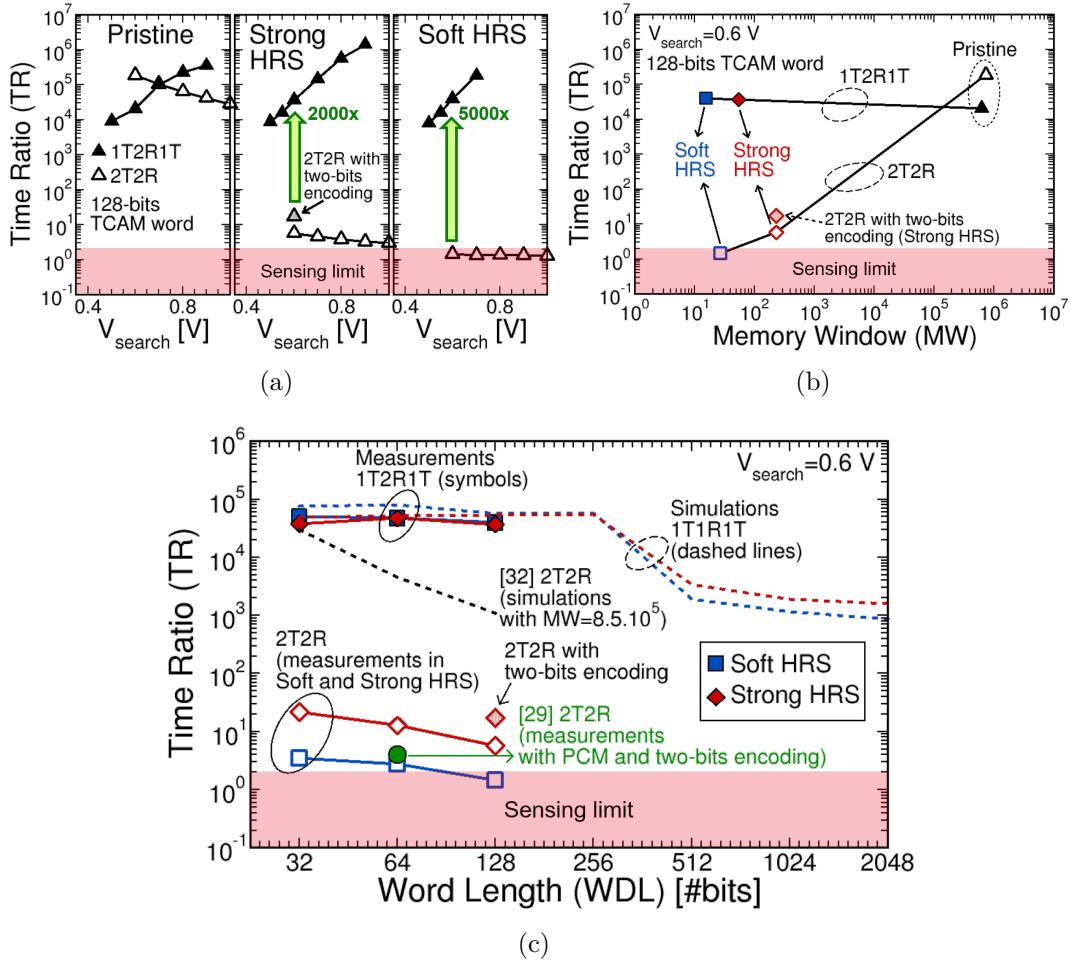


FIGURE 3.2.21: Measured time ratio for the proposed 1T2R1T structure (filled symbol) and the common 2T2R structure measured in the previous section (open symbol) as a function of (a) the search voltage V_{search} , (b) the memory window, and (c) the TCAM word length. The measurements performed on the 2T2R structure using the two-bits encoding scheme of Li et al. [29] is represented by the shaded red diamond.

with measured results for Soft HRS and Strong HRS (blue and red symbol, respectively) and extended with simulations for TCAM words longer than 128 bits (blue and red dotted line for Soft HRS and Strong HRS, respectively). The sensing margin measured in [29] (green circle, 2T2R TCAM with Phase-Change Memory (PCM) technology) and sensing margin simulated in [32] (black dotted line, simulations performed assuming a memory window of $8.5 \cdot 10^5$ with a 2T2R TCAM) are also reported for comparison. Unlike the common 2T2R structure whose TR decreases with WDL, WDL has a minimal impact on the TR of the proposed 1T2R1T structure up to 256 bits. For longer words (>512 bits), the ML starts discharging in the match state through all the transistors N2. However, the TR remains higher than the sensing limit (the TR is still higher than 10^3). Therefore, the proposed 1T2R1T structure allows for an increase in the maximum word length (more than 2 kbits) for both Soft and Strong HRS with respect to the common 2T2R structure.

3.2.4.5 Search endurance characterisation

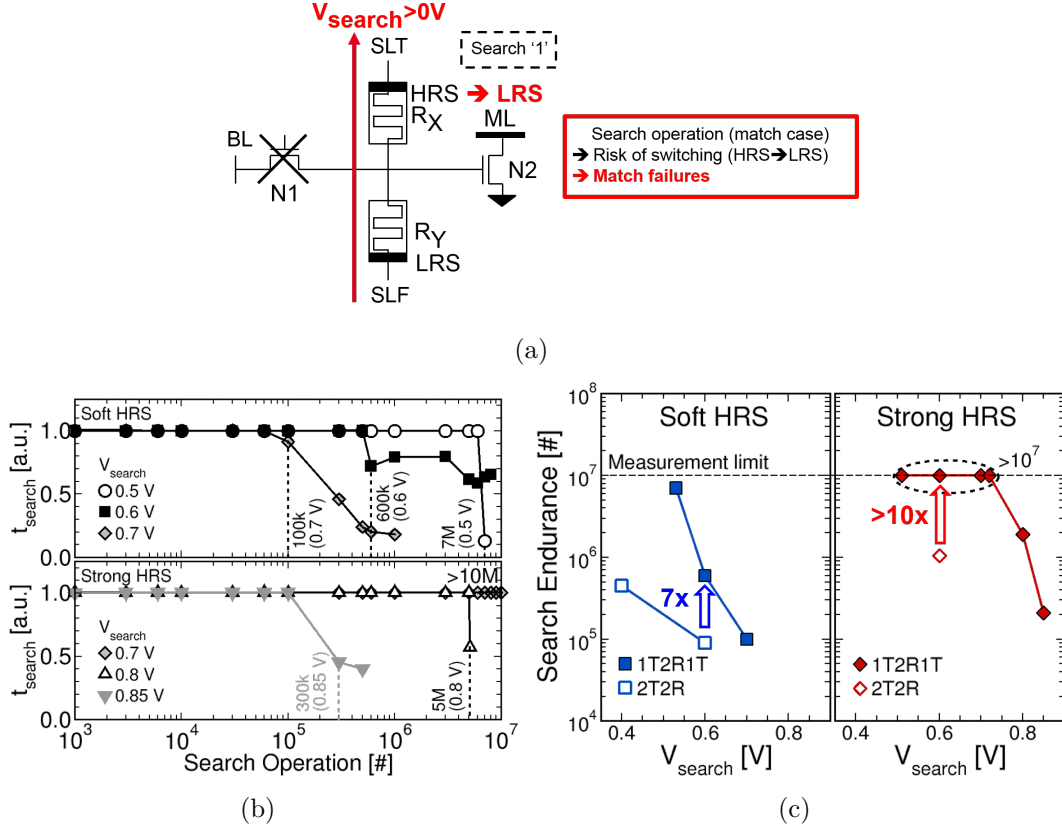


FIGURE 3.2.22: (a) During a search operation, a positive voltage is applied on the top electrode of one RRAM cell (here R_X for a search '1') in the same configuration as a Set operation. (b) Search endurance characterisation for the proposed 1T2R1T structure. (c) Comparison between the search endurance of the common 2T2R TCAM (open symbol) and the 1T2R1T TCAM (filled symbol) as a function of V_{search} for Soft and Strong HRS.

During a search operation, a positive voltage is applied on the top electrode of one RRAM cell (R_X for search '1', R_Y for search '0') in the same polarity as a Set operation as sketched in FIGURE 3.2.22 (a). As in the common 2T2R, undesired switchings from HRS to LRS can occur leading to a match failure in the case of a match configuration. We characterised the search endurance of the proposed 1T2R1T structure for Soft and Strong HRS and different search voltages, V_{search} , in FIGURE 3.2.22 (b). As with the previous structure, we measured the discharge time, t_{search} , after a series of search operations when the TCAM is initially in a match state. However, contrary to the previous structure whose ML discharges in the match state, we limited here the duration of each search operation. The duration of each search operation has been chosen as at least twice the discharge time, t_{search} , in the 1-bit mismatch state. TABLE 3.12 shows the duration of search operations for each search endurance characterisation. Note that t_{search} in the 1-bit mismatch state only depends on the applied V_{search} and not the programming conditions. t_{search} has been normalised here by the duration of each search operation. In accordance with

the previous 2T2R TCAM, the search endurance degrades with higher V_{search} or when the TCAM is programmed in Soft HRS instead of Strong HRS. FIGURE 3.2.22 (c) compares the search endurance of the common 2T2R TCAM (open symbols) and the proposed 1T2R1T TCAM (filled symbols) as a function of V_{search} . The proposed 1T2R1T TCAM improves the search endurance in both Soft and Strong HRS. At $V_{\text{search}}=0.6$ V, we observe no degradation for more than 10^7 search operations (limit of measurement), improving on the previous structure by $>10x$. The improvement in search endurance with the proposed 1T2R1T TCAM can be accounted for by the fact that the search voltage, V_{search} , is applied on RRAM cells only during the match line sensing phase. By contrast, in the 2T2R TCAM, the search voltage is already applied on RRAM cell top electrodes during the match line pre-charge phase before the match line sensing phase.

V_{search}	t_{search} (1-bit mismatch state)	Duration of search operations
0.5 V	$\sim 120 \mu\text{s}$	500 μs
0.6 V	$\sim 25 \mu\text{s}$	55 μs
0.7 V	$\sim 6 \mu\text{s}$	15 μs
0.8 V	$\sim 2 \mu\text{s}$	15 μs
0.85 V	$\sim 1 \mu\text{s}$	15 μs

TABLE 3.12: Duration of each search operation as a function of the search voltage, V_{search} , for the search endurance characterisation. Note that t_{search} in the 1-bit mismatch state only depends on V_{search} , and it is similar whether RRAMs are programmed in Soft or Strong HRS.

3.2.4.6 Search time and search energy consumption

To reduce the search time (discharge time in the 1-bit mismatch state), the ML sensing circuit has to be as fast as possible. This is the reason why we used an analog circuit to sense the ML voltage. The analog circuit senses the ML voltage and compares it to a reference voltage, V_{REF} . Here, we defined the search time as the time taken to discharge the ML of a given voltage (ΔV) from the pre-charged value ($V_{\text{DD_ML}}$) (FIGURE 3.2.23 (a, top)). Note that with the proposed 1T2R1T structure, the sensing circuit can also be simplified by the use of a digital inverter (FIGURE 3.2.23 (a, bottom)), thereby reducing design complexity. This is possible because the ML discharges only in the mismatch state. In the case where an inverter is used, the search time would be defined as the time required to fully discharge the ML, *i.e.* when the output of the inverter switches. Therefore, using an inverter instead of an analog circuit comes at the expense of longer search times as well as higher search energy consumption (since we also need to fully charge the ML again at each search operation).

Here, we consider an analog circuit for the sensing circuit. FIGURE 3.2.23 (b) shows measured (symbol) and simulated (line) discharge times as a function of V_{search} for (Left) a mismatch state of 128 bits and (Right) 1 bit. The TCAM is programmed in Strong HRS. In the fabricated circuit, ΔV is fixed at 3.0 V (solid line in FIGURE 3.2.23 (b)). As it can be seen, measurements and simulations are in good agreement. A first possibility to reduce the search time

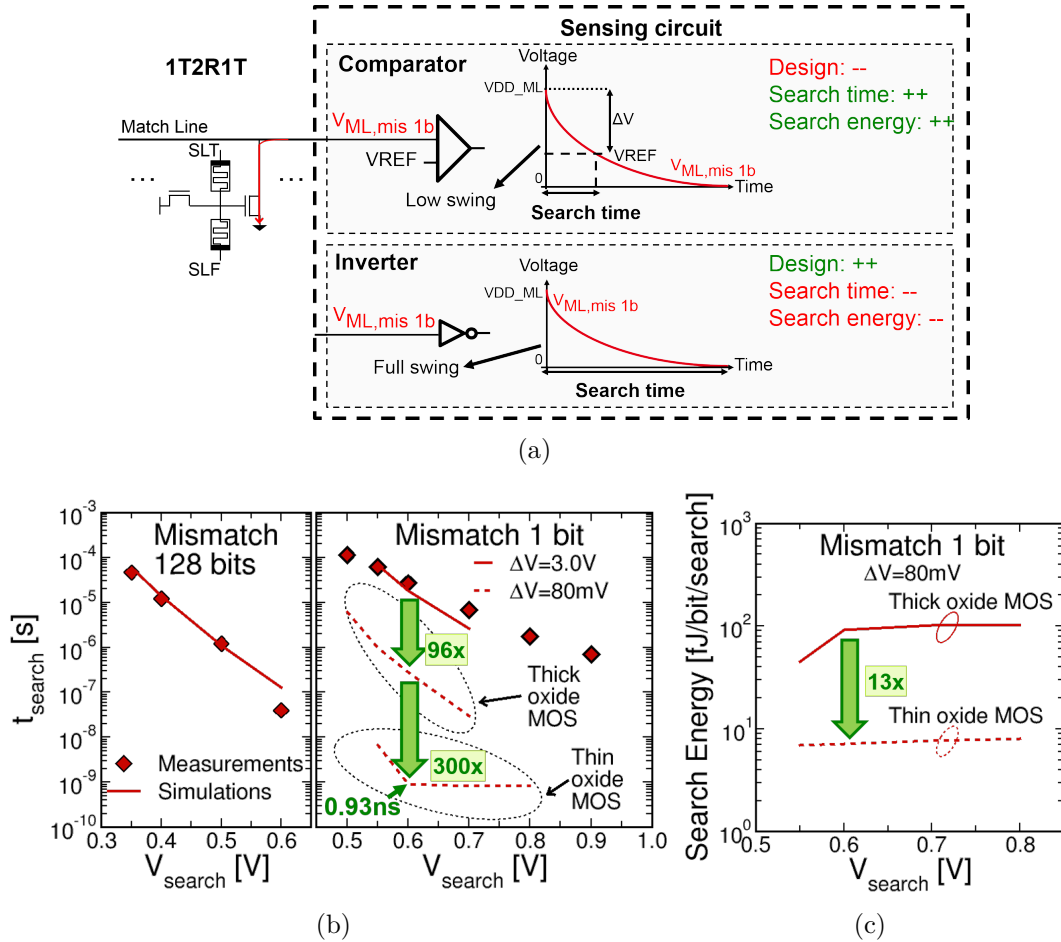


FIGURE 3.2.23: (a) With the proposed 1T2R1T structure, the sensing circuit can be implemented either with (Top) a comparator circuit (low swing) or (Bottom) a digital inverter (full swing). (b) Measured (symbol) and simulated (line) discharge times, t_{search} , as a function of the search voltage, V_{search} , in the (Left) 128-bits and (Right) 1-bit mismatch states. (c) Simulated search energy consumption as a function of V_{search} in the 1-bit mismatch state when transistors N2 are implemented with thick oxide MOS (solid line) and thin oxide MOS (dotted line).

is by decreasing ΔV from 3.0 V down to 80 mV (dotted line, voltage which still ensures an accurate switching of the comparator). Simulations show that it improves the search time by 96x. Another possibility is to replace the thick oxide MOS used for transistors N2 by a thin oxide MOS. Indeed, since the transistor N2 is not involved in RRAM programming operations, a thin oxide transistor with minimum permitted gate length can be adopted. With the same W/L ratio as before (with $W=1.56 \mu\text{m}$ and $L=130 \text{ nm}$), this speeds up searches by 300x. At $V_{\text{search}}=0.6 \text{ V}$, we reach a search time of 0.93 ns.

We finally simulated the energy consumption during a search operation. The search energy is calculated as the integral of power consumed by the match line (pre-charge and discharge) and search lines over the search time. Considering an analog circuit for the sensing circuit with $\Delta V=80 \text{ mV}$, the implementation of transistors N2 with thin oxide MOS improves search energy consumption by 13x thanks to the lower transistor activation energy. At $V_{\text{search}}=0.6 \text{ V}$ and the same W/L ratio as before ($W/L=12$ with $W=1.56 \mu\text{m}$ and $L=130 \text{ nm}$),

the TCAM consumes 7.18 fJ/bit/search. Note that using minimum permitted gate length and width ($W=150$ nm and $L=130$ nm), the TCAM consumes 1.24 fJ/bit/search.

3.2.4.7 Discussion

In this section, we proposed a new TCAM bitcell composed of two transistors and two RRAMs in a 1T2R1T configuration. It is based on a RRAM voltage divider (2R) biasing a transistor gate (1T). An additional transistor is used for RRAM programming operations (1T). The proposed TCAM cell addresses the main limitations of the most common 2T2R RRAM-based TCAM [28–32], namely the strong dependency on RRAM memory window and the limited word length. TABLE 3.13 compares the characterisation results obtained for both structures. We experimentally demonstrated that the proposed 1T2R1T structure is insensitive to the memory window (FIGURE 3.2.21 (b)) and can operate with standard programming conditions for the RRAM memories (Soft HRS). Following the same reasoning on the programming endurance explained in SECTION 3.2.3.6 (which is still valid for the proposed 1T2R1T TCAM circuit), this leads to a programming endurance of 10^6 cycles for the RRAM cells, improving on the common 2T2R structure by 100x. More importantly, we experimentally proved a large sensing margin with the proposed structure which is comparable to that of SRAM-based TCAMs ($>10^4$). This allows a large volume of data to be searched in parallel thanks to the long word length (>2 kbits, FIGURE 3.2.21 (c)). This makes this bitcell suitable for applications requiring long pattern matching, such as internet protocol (IP) v6 packet routing, DNA sequence matching, or active control list management [34]. Neuromorphic multi-core architectures would also benefit from this structure, thanks to the improved performance, reliability, and especially better search endurance. Finally, this structure also allows for more relaxed design constraints. First, only one transistor is required for programming operations (transistor N1), the other one is only involved in search operations (transistor N2). As a result, the transistor N2 can be implemented with thin oxide MOS with minimum permitted gate length. This improves the search time, search energy efficiency, and TCAM bitcell size. In addition, the lower threshold voltage of such transistors allows to operate the TCAM in a low-voltage regime during search operations, hence significantly improving search endurance. This is more challenging with the common 2T2R structure as both transistors are required for programming operations. Second, as the match line does not discharge in the match state, we also have more flexibility in the design of the sensing circuit. Whereas a comparator circuit is mandatory for the 2T2R structure, a digital inverter can be used for the 1T2R1T structure. This reduced design complexity at the expense of longer search times and higher search energy consumption. TABLE 3.14 compares the measured performance and reliability metrics of both structures with reported silicon-proven RRAM-based TCAM circuits [19, 21, 26, 27, 29]. We also report in FIGURE 3.2.24 the TCAM bitcell size, search time, and search energy of our circuits with reported silicon-proven SRAM- (circle) and RRAM-based (diamond) TCAM circuits.

The main drawbacks of this structure are the more complex programming scheme

3.2. CHARACTERISATION OF RRAM-BASED TCAMS

		2T2R (in Strong HRS)		1T2R1T (in Strong HRS)		2T2R (@ $V_{\text{search}}=0.6\text{ V}$)		1T2R1T (@ $V_{\text{search}}=0.6\text{ V}$)	
		V_{search}		V_{search}		HRS/LRS		HRS/LRS	
		0.6 V	1.0 V	0.6 V	0.8 V	~30	~100	~30	~100
Performance (Lower is better)	Search Time	90 ns	~30 ns	0.93 ns*	0.82 ns*	~51 ns	90 ns	0.84 ns*	0.93 ns*
	Sensing Margin	5.7	3.0	4.10^4	6.10^5	1.5	5.7	4.10^4	4.10^4
Reliability (Higher is better)	Search Endurance	9.10^4	N/A	$>10^7$	5.10^6	9.10^4	10^6	6.10^5	$>10^7$
	Prog. endurance	10^4	10^4	10^4	10^4	10^6	10^4	10^6	10^4
Capacity (Higher is better)	Word Length max	~256 bits	~197 bits	>2 kbits	>2 kbits	97 bits	~256 bits	>2 kbits	>2 kbits

*With $\Delta V=80\text{mV}$ and thin oxide MOS

TABLE 3.13: Comparison of the characterisation results obtained with the common 2T2R and the proposed 1T2R1T structure.

	[29] 2T2R	[21] 4T2R	[26] 3T1R	[27] 2.5T1R	[19] 5T2R	This work 2T2R	This work 1T2R1T
Technology node	90 nm	180 nm	90 nm	65 nm	90 nm	130 nm	
RRAM technology	GST PCM	HfO-based	HfO-based	N/A	HfO-based	HfO ₂ -based	
TCAM capacity	16k×64 bits	128×32 bits	128×64 bits	64×256 bits	128×64 bits	3×128bits	
Prog. Endurance [#cycles]	-	-	-	-	-	10⁴	10⁶
Search Endurance [#searches]	-	-	-	-	-	10⁶ ($V_{\text{search}}=0.6\text{V}$)	>10⁷ ($V_{\text{search}}=0.6\text{V}$)
Word Length max	-	-	-	-	-	97 bits (Soft HRS)	>2 kbits (Soft HRS)
Normalized Search Time [ps/bit]	30	38	15	4	25	700 (thick oxide MOS)	2180* (thick oxide MOS) 7.3* (thin oxide MOS)

*With $\Delta V=80\text{mV}$

TABLE 3.14: Comparison with silicon-proven RRAM-based TCAM circuits presented in the literature [19, 21, 26, 27, 29]. Search times have been normalised by the TCAM word length.

and sensing in the *don't care* state ('X' state). Indeed, search lines (SLT and SLF) are shared among all the cells in the same column. Thus, programming a cell can disturb other cells in the same column as programming pulses are applied directly on SLT and SLF. The 4T2R TCAM bitcell proposed in [21] solves this problem by using two additional transistors per cell, one for each search line (SLT and SLF). This permits to independently select each TCAM cell in a column. However, this increases the cell size. In our structure, an appropriate programming scheme has to be implemented in order to limit the voltage drop across each RRAM in the same column. The second drawback is the sensing in the 'X' state. When a TCAM cell is programmed in the 'X' state, *i.e.* both RRAMs are in the HRS, the gate of transistor N2 is biased at $\approx V_{\text{search}}/2$ during a search operation. This can lead to a discharge of the ML in the match state if transistors N2 leak too much. It is then required to use search voltages, V_{search} , as low as possible. Ideally, the intermediate node capacitance of the RRAM voltage divider should be high enough to keep the voltage close to 0 V for a sufficiently long time (*i.e.* longer than the search time). The 3T2R TCAM proposed in [23] addresses this limitation by using an additional transistor whose source and drain are connected between the internal node of the RRAM voltage divider and the gate of transistor N2. This transistor

transmits the voltage of the RRAM voltage divider to the gate of transistor N2 only in the mismatch state, *i.e.* if the voltage exceeds a minimal voltage. Otherwise no voltage is transmitted.

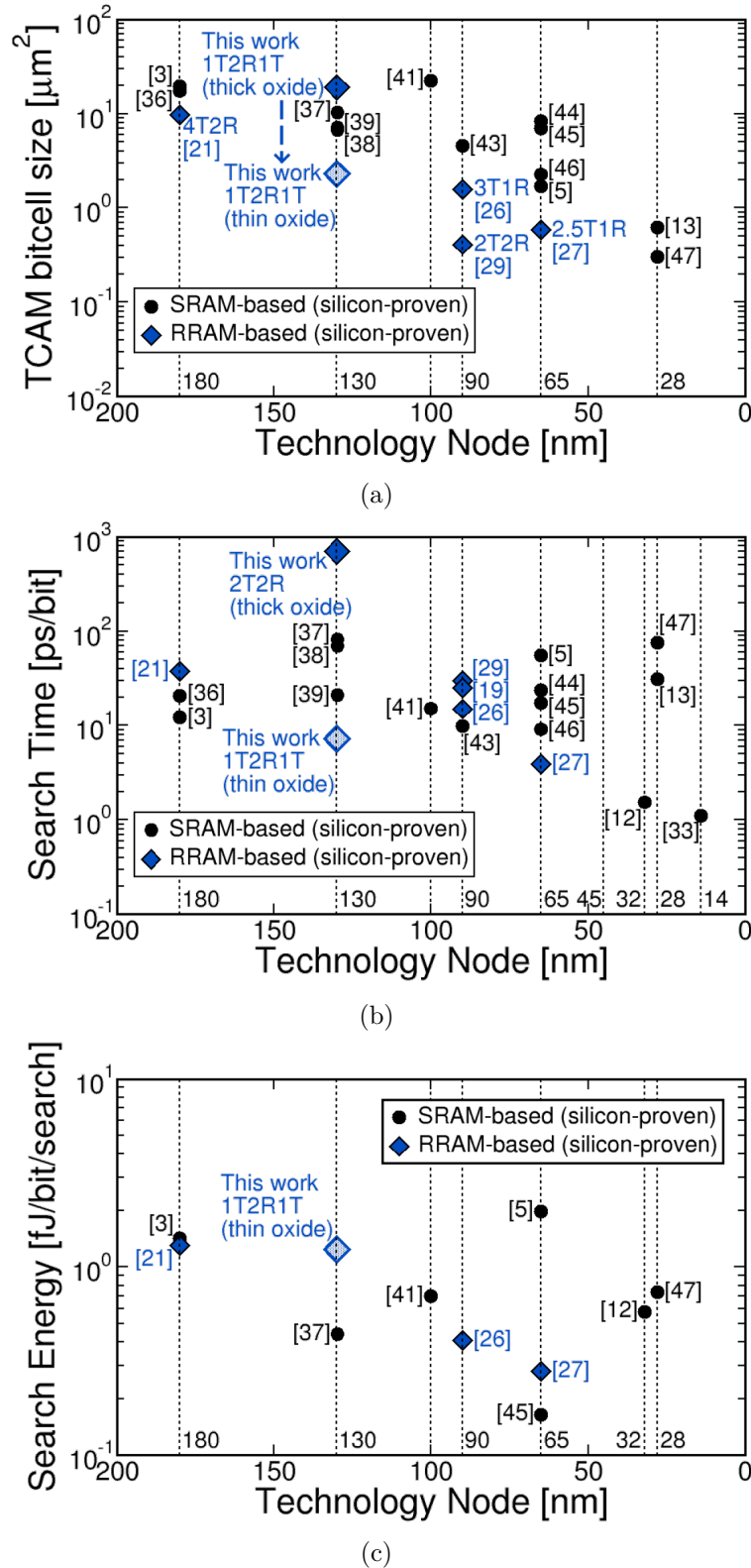


FIGURE 3.2.24: Comparison in terms of (a) TCAM bitcell size, (b) search time, and (c) search energy as a function of technology node with reported silicon-proven SRAM- (black circle) [3, 5, 12, 13, 33, 36–39, 41, 43–47] and RRAM-based (blue diamond) [19, 21, 26, 27, 29] TCAM circuits.

References: Chapter 3

- [1] A. V. Campi, R. M. Dunn, and B. H. Gray. “Content Addressable Memory System Concepts”. *IEEE Transactions on Aerospace and Electronic Systems*, AES-1(2):168–173, 1965. doi: 10.1109/TAES.1965.4501678.
- [2] Kostas Pagiamtzis and Ali Sheikholeslami. “Content-addressable memory (CAM) circuits and architectures: A tutorial and survey”. *IEEE Journal of Solid-State Circuits*, 41(3): 712–727, mar 2006. ISSN 00189200. doi: 10.1109/JSSC.2005.864128.
- [3] Chao Ching Wang, Jinn Shyan Wang, and Chingwei Yeh. “High-speed and low-power design techniques for TCAM macros”. *IEEE Journal of Solid-State Circuits*, 43(2): 530–540, feb 2008. ISSN 00189200. doi: 10.1109/JSSC.2007.914330.
- [4] Woong Choi, Kyeongho Lee, and Jongsun Park. “Low Cost Ternary Content Addressable Memory Using Adaptive Matchline Discharging Scheme”. In *Proceedings - IEEE International Symposium on Circuits and Systems*, volume 2018-May. Institute of Electrical and Electronics Engineers Inc., apr 2018. ISBN 9781538648810. doi: 10.1109/ISCAS.2018.8351461.
- [5] Isamu Hayashi, Teruhiko Amano, et al. “A 250-MHz 18-Mb Full Ternary CAM with Low-Voltage Matchline Sensing Scheme in 65-nm CMOS”. *IEEE Journal of Solid-State Circuits*, 48(11):2671–2680, 2013. ISSN 00189200. doi: 10.1109/JSSC.2013.2274888.
- [6] Cheol Kim, Sung Gi Ahn, Jisu Min, and Kee Won Kwon. “Power efficient and reliable nonvolatile TCAM with Hi-PFO and semi-complementary driver”. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 66(2):605–615, feb 2019. ISSN 15498328. doi: 10.1109/TCSI.2018.2867005.
- [7] A. E. Slade and H. O. McMahon. “A cryotron catalog memory system”. In *Proceedings of the Eastern Joint Computer Conference: New Developments in Computers, AIEE-IRE 1956*, pages 115–120, New York, 1956. doi: 10.1145/1455533.1455560.
- [8] A. G. Hanlon. “Content-Addressable and Associative Memory Systems”. *IEEE Transactions on Electronic Computers*, EC-15(4):509–521, 1966. ISSN 03677508. doi: 10.1109/PGEC.1966.264358.
- [9] James T Koo. “Integrated-Circuit Content-Addressable Memories”. *IEEE Journal of Solid-State Circuits*, SC-5(5):208–215, 1970. ISSN 1558173X. doi: 10.1109/JSSC.1970.1050115.
- [10] Kenneth J. Schultz. “Content-addressable memory core cells a survey”. *Integration, the VLSI Journal*, 23(2):171–188, 1997. ISSN 01679260. doi: 10.1016/S0167-9260(97)00021-7.
- [11] S. R. Ramirez-Chavez. “Encoding Don’t Cares in Static and Dynamic Content-Addressable Memories”. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 39(8):575–578, 1992. ISSN 10577130. doi: 10.1109/82.168950.
- [12] Igor Arsovski, Travis Hebig, Daniel Dobson, and Reid Wistort. “A 32 nm 0.58-fJ/Bit/Search 1-GHz ternary content addressable memory compiler using silicon-aware

- early-predict late-correct sensing with embedded deep-trench capacitor noise mitigation”. *IEEE Journal of Solid-State Circuits*, 48(4):932–939, 2013. ISSN 00189200. doi: 10.1109/JSSC.2013.2239092.
- [13] Koji Nii, Teruhiko Amano, et al. “A 28nm 400MHz 4-parallel 1.6Gsearch/s 80Mb ternary CAM”. In *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, pages 240–241, 2014. ISBN 9781479909186. doi: 10.1109/ISSCC.2014.6757417.
- [14] Shawana Tabassum, Farhana Parveen, and A. B.M.Harun Ur Rashid. “Low power high speed ternary content addressable memory design using MOSFET and memristors”. In *2014 International Conference on Electronics and Communication Systems, ICECS 2014*, pages 1–6. Institute of Electrical and Electronics Engineers Inc., 2014. ISBN 9781479923205. doi: 10.1109/ECS.2014.6892672.
- [15] Shawana Tabassum, Farhana Parveen, and A. B.M.Harun Ur Rashid. “Low power high speed Ternary Content Addressable Memory design using 8 MOSFETs and 4 memristors - Hybrid structure”. In *8th International Conference on Electrical and Computer Engineering: Advancing Technology for a Better Tomorrow, ICECE 2014*, pages 168–171. Institute of Electrical and Electronics Engineers Inc., jan 2015. ISBN 9781479941667. doi: 10.1109/ICECE.2014.7026989.
- [16] Pilin Junsangsri, Fabrizio Lombardi, and Jie Han. “A memristor-based TCAM (Ternary Content Addressable Memory) cell”. In *Proceedings of the 2014 IEEE/ACM International Symposium on Nanoscale Architectures, NANOARCH 2014*, pages 1–6, 2014. ISBN 9781479963836. doi: 10.1109/NANOARCH.2014.6880478.
- [17] Le Zheng, Sangho Shin, and Sung Mo Steve Kang. “Memristors-based Ternary Content Addressable Memory (mTCAM)”. In *Proceedings - IEEE International Symposium on Circuits and Systems*, pages 2253–2256. Institute of Electrical and Electronics Engineers Inc., 2014. ISBN 9781479934324. doi: 10.1109/ISCAS.2014.6865619.
- [18] Le Zheng, Sangho Shin, et al. “RRAM-based TCAMs for pattern search”. In *Proceedings - IEEE International Symposium on Circuits and Systems*, volume 2016-July, pages 1382–1385. Institute of Electrical and Electronics Engineers Inc., jul 2016. ISBN 9781479953400. doi: 10.1109/ISCAS.2016.7527507.
- [19] Meng Fan Chang, Ching Hao Chuang, et al. “Designs of emerging memory based non-volatile TCAM for Internet-of-Things (IoT) and big-data processing: A 5T2R universal cell”. In *Proceedings - IEEE International Symposium on Circuits and Systems*, volume 2016-July, pages 1142–1145. Institute of Electrical and Electronics Engineers Inc., jul 2016. ISBN 9781479953400. doi: 10.1109/ISCAS.2016.7527447.
- [20] Li Yue Huang, Meng Fan Chang, et al. “ReRAM-based 4T2R nonvolatile TCAM with 7x NVM-stress reduction, and 4x improvement in speed-wordlength-capacity for normally-off instant-on filter-based search engines used in big-data processing”. In *IEEE Symposium on VLSI Circuits, Digest of Technical Papers*. Institute of Electrical and Electronics Engineers Inc., 2014. ISBN 9781479933273. doi: 10.1109/VLSIC.2014.6858404.
- [21] Meng Fan Chang, Lie Yue Huang, et al. “A ReRAM-Based 4T2R Nonvolatile TCAM Using RC-Filtered Stress-Decoupled Scheme for Frequent-OFF Instant-ON Search Engines Used in IoT and Big-Data Processing”. *IEEE Journal of Solid-State Circuits*, 51(11): 2786–2789, nov 2016. ISSN 00189200. doi: 10.1109/JSSC.2016.2602218.
- [22] Catherine E. Graves, Martin Foltin, et al. “Regular Expression Matching with Memristor TCAMs for Network Security”. In *Proceedings of the 14th IEEE/ACM International Symposium on Nanoscale Architectures*, pages 65–71. Association for Computing Machinery (ACM), dec 2018. doi: 10.1145/3232195.3232201.
- [23] C. Kim and K. W. Kwon. “3T-2R non-volatile TCAM with voltage limiter and self-controlled bias circuit”. *Electronics Letters*, 53(13):837–839, may 2017. ISSN 00135194. doi: 10.1049/el.2017.1027.

-
- [24] Cheol Kim, Rak Joo Sung, Sung Gi Ahn, Jisu Min, and Kee Won Kwon. “Low Power Search Engine using Non-volatile Memory based TCAM with Priority Encoding and Selective Activation of Search Line and Match Line”. In *Proceedings - IEEE International Symposium on Circuits and Systems*, volume 2018-May. Institute of Electrical and Electronics Engineers Inc., apr 2018. ISBN 9781538648810. doi: 10.1109/ISCAS.2018.8351237.
- [25] Meng Fan Chang, Chien Chen Lin, et al. “A 3T1R nonvolatile TCAM using MLC ReRAM with Sub-1ns search time”. In *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, pages 318–319, 2015. ISBN 9781479962235. doi: 10.1109/ISSCC.2015.7063054.
- [26] Meng Fan Chang, Chien Chen Lin, et al. “A 3T1R Nonvolatile TCAM Using MLC ReRAM for Frequent-Off Instant-On Filters in IoT and Big-Data Processing”. *IEEE Journal of Solid-State Circuits*, 52(6):1664–1679, jun 2017. ISSN 00189200. doi: 10.1109/JSSC.2017.2681458.
- [27] Chien Chen Lin, Jui Yu Hung, et al. “A 256b-wordlength ReRAM-based TCAM with 1ns search-time and 14x improvement in wordlength-energyefficiency-density product using 2.5T1R cell”. In *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, pages 136–137. Institute of Electrical and Electronics Engineers Inc., feb 2016. ISBN 9781467394666. doi: 10.1109/ISSCC.2016.7417944.
- [28] Bipin Rajendran, Roger W. Cheek, et al. “Demonstration of CAM and TCAM using phase change devices”. In *2011 3rd IEEE International Memory Workshop, IMW 2011*, 2011. ISBN 9781457702266. doi: 10.1109/IMW.2011.5873229.
- [29] Jing Li, Robert K. Montoye, Masatoshi Ishii, and Leland Chang. “1 Mb 0.41 μm^2 2T-2R cell nonvolatile TCAM with two-bit encoding and clocked self-referenced sensing”. *IEEE Journal of Solid-State Circuits*, 49(4):896–907, 2014. ISSN 00189200. doi: 10.1109/JSSC.2013.2292055.
- [30] J. Li, R. Montoye, et al. “1Mb 0.41 μm^2 2T-2R cell nonvolatile TCAM with two-bit encoding and clocked self-referenced sensing”. In *Digest of Technical Papers - Symposium on VLSI Technology*, pages 104–105. IEEE, 2013. ISBN 978-1-4673-5226-0.
- [31] Alessandro Grossi, Elisa Vianello, et al. “Experimental Investigation of 4-kb RRAM Arrays Programming Conditions Suitable for TCAM”. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 26(12):2599–2607, dec 2018. ISSN 10638210. doi: 10.1109/TVLSI.2018.2805470.
- [32] Rui Yang, Haitong Li, et al. “Ternary content-addressable memory with MoS₂ transistors for massively parallel data search”. *Nature Electronics*, 2(3):108–114, mar 2019. ISSN 25201131. doi: 10.1038/s41928-019-0220-7.
- [33] Igor Arsovski, Akhilesh Patil, et al. “1.4Gsearch/s 2-Mb/mm² TCAM Using Two-Phase-Pre-Charge ML Sensing and Power-Grid Pre-Conditioning to Reduce Ldi/dt Power-Supply Noise by 50%”. *IEEE Journal of Solid-State Circuits*, 53(1):155–163, jan 2018. ISSN 00189200. doi: 10.1109/JSSC.2017.2739178.
- [34] Dennis Dudeck and Lisa Minwell. “Why is TCAM Essential for the Cloud?”. Technical report, 2014. URL www.esilicon.com.
- [35] M J Akhbarizadeh, M Nourani, D S Vijayasarathi, and T Balsara. “A nonredundant ternary CAM circuit for network search engines”. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 14(3):268–278, 2006. ISSN 1557-9999. doi: 10.1109/TVLSI.2006.871760.
- [36] Igor Arsovski, Trevis Chandler, and Ali Sheikholeslami. “A ternary content-addressable memory (TCAM) based on 4T static storage and including a current-race sensing scheme”. *IEEE Journal of Solid-State Circuits*, 38(1):155–158, 2003. ISSN 00189200. doi: 10.1109/JSSC.2002.806264.

- [37] C Wang, C Cheng, T Chen, and J Wang. “An Adaptively Dividable Dual-Port BiTCAM for Virus-Detection Processors in Mobile Devices”. *IEEE Journal of Solid-State Circuits*, 44(5):1571–1581, 2009. ISSN 1558-173X. doi: 10.1109/JSSC.2009.2017009.
- [38] G Kasai, Y Takarabe, K Furumi, and M Yoneda. “200MHz/200MSPS 3.2W at 1.5V Vdd, 9.4Mbits ternary CAM with new charge injection match detect circuits and bank selection scheme”. In *Proceedings of the IEEE 2003 Custom Integrated Circuits Conference, 2003.*, pages 387–390, 2003. doi: 10.1109/CICC.2003.1249424.
- [39] Alan Roth, Dick Foss, Robert McKenzie, and Douglas Perry. “Advanced ternary CAM circuits on 0.13um logic process technology”. In *Proceedings of the Custom Integrated Circuits Conference*, pages 465–468, 2004. doi: 10.1109/cicc.2004.1358852.
- [40] M Sultan, M Siddiqui, Sonika, and G S Visweswaran. “A low-power ternary content addressable memory (TCAM) with segmented and non-segmented matchlines”. In *TENCON 2008 - 2008 IEEE Region 10 Conference*, pages 1–5, nov 2008. doi: 10.1109/TENCON.2008.4766746.
- [41] Sungdae Choi, K Sohn, and Hoi-Jun Yoo. “A 0.7-fJ/bit/search 2.2-ns search time hybrid-type TCAM architecture”. *IEEE Journal of Solid-State Circuits*, 40(1):254–260, jan 2005. ISSN 1558-173X. doi: 10.1109/JSSC.2004.837979.
- [42] Bruce Gamache, Zachary Pfeffer, and Sunil P. Khatri. “A fast ternary CAM design for IP networking applications”. In *Proceedings - International Conference on Computer Communications and Networks, ICCCN*, pages 434–439. Institute of Electrical and Electronics Engineers Inc., 2003. ISBN 0780379454. doi: 10.1109/ICCCN.2003.1284205.
- [43] Igor Arsovski and Rahul Nadkarni. “Low-noise embedded CAM with reduced slew-rate match-lines and asynchronous search-lines”. In *Proceedings of the Custom Integrated Circuits Conference*, volume 2005, pages 440–443, 2005. ISBN 0780390237. doi: 10.1109/CICC.2005.1568702.
- [44] Ki Chan Woo and Byung Do Yang. “Low-Area TCAM using a don’t care reduction scheme”. *IEEE Journal of Solid-State Circuits*, 53(8):2427–2433, aug 2018. ISSN 00189200. doi: 10.1109/JSSC.2018.2822696.
- [45] P Huang and W Hwang. “A 65 nm 0.165 fJ/Bit/Search 256x144 TCAM Macro Design for IPv6 Lookup Tables”. *IEEE Journal of Solid-State Circuits*, 46(2):507–519, 2011. ISSN 1558-173X. doi: 10.1109/JSSC.2010.2082270.
- [46] I Arsovski and R Wistort. “Self-referenced sense amplifier for across-chip-variation immune sensing in high-performance Content-Addressable Memories”. In *IEEE Custom Integrated Circuits Conference 2006*, pages 453–456, 2006. doi: 10.1109/CICC.2006.320819.
- [47] Supreet Jeloka, Naveen Bharathwaj Akesh, Dennis Sylvester, and David Blaauw. “A 28 nm Configurable Memory (TCAM/BCAM/SRAM) Using Push-Rule 6T Bit Cell Enabling Logic-in-Memory”. *IEEE Journal of Solid-State Circuits*, 51(4):1009–1021, apr 2016. ISSN 00189200. doi: 10.1109/JSSC.2016.2515510.
- [48] H Noda, K Inoue, H J Mattausch, T Koide, and K Arimoto. “A cost-efficient dynamic Ternary CAM in 130 nm CMOS technology with planar complementary capacitors and TSR architecture”. In *2003 Symposium on VLSI Circuits. Digest of Technical Papers (IEEE Cat. No.03CH37408)*, pages 83–84, 2003. doi: 10.1109/VLSIC.2003.1221168.
- [49] Yuanfan Yang, Jimson Mathew, Marco Ottavi, Salvatore Pontarelli, and Dhiraj K. Pradhan. “2T2M memristor based TCAM cell for low power applications”. In *Proceedings - 2015 10th IEEE International Conference on Design and Technology of Integrated Systems in Nanoscale Era, DTIS 2015*. Institute of Electrical and Electronics Engineers Inc., jun 2015. ISBN 9781479919994. doi: 10.1109/DTIS.2015.7127379.

-
- [50] Shoun Matsunaga, Sadahiko Miura, et al. “A 3.14 μm^2 4T-2MTJ-cell fully parallel TCAM based on nonvolatile logic-in-memory architecture”. In *IEEE Symposium on VLSI Circuits, Digest of Technical Papers*, pages 44–45, 2012. ISBN 9781467308458. doi: 10.1109/VLSIC.2012.6243781.
- [51] Byungkyu Song, Taehui Na, Jung Pill Kim, Seung H. Kang, and Seong Ook Jung. “A 10T-4MTJ Nonvolatile Ternary CAM Cell for Reliable Search Operation and a Compact Area”. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 64(6):700–704, 2017. ISSN 15497747. doi: 10.1109/TCSII.2016.2594827.
- [52] H.-S. P. Wong, C. Ahn, et al. “Stanford Memory Trends”, 2018. URL <https://nano.stanford.edu/stanford-memory-trends/>.
- [53] Hongsik Jeong and Luping Shi. “Memristor devices for neural networks”. *Journal of Physics D: Applied Physics*, 52(2):023003, 2019. ISSN 13616463. doi: 10.1088/1361-6463/aae223.
- [54] Robert Karam, Ruchir Puri, Swaroop Ghosh, and Swarup Bhunia. “Emerging Trends in Design and Applications of Memory-Based Computing and Content-Addressable Memories”. *Proceedings of the IEEE*, 103(8):1311–1330, aug 2015. ISSN 00189219. doi: 10.1109/JPROC.2015.2434888.
- [55] Naoya Onizawa, Warren J. Gross, and Takahiro Hanyu. “A low-energy variation-tolerant asynchronous TCAM for network intrusion detection systems”. In *Proceedings - International Symposium on Asynchronous Circuits and Systems*, pages 8–15, 2013. doi: 10.1109/ASYNC.2013.16.
- [56] Diego Valverde Garro, Claudio Viquez Calderon, and Christopher Simon Yeung. “Using a programmable network switch TCAM to find the best alignment of two DNA sequences”. In *2016 IEEE 36th Central American and Panama Convention, CONCAPAN 2016*, pages 1–5, 2016. ISBN 9781467395786. doi: 10.1109/CONCAPAN.2016.7942372.
- [57] Yue Zha, Etienne Nowak, and Jing Li. “Liquid Silicon: A Nonvolatile Fully Programmable Processing-In-Memory Processor with Monolithically Integrated ReRAM for Big Data/Machine Learning Applications”. In *IEEE Symposium on VLSI Circuits, Digest of Technical Papers*, pages C206–C207, 2019. ISBN 9784863487185. doi: 10.23919/VLSIC.2019.8778064.
- [58] Anthony J. McAuley and Paul Francis. “Fast routing table lookup using CAMs”. In *Proceedings - IEEE INFOCOM*, volume 3, pages 1382–1891. Institute of Electrical and Electronics Engineers (IEEE), mar 1993. ISBN 0818635800. doi: 10.1109/infcom.1993.253403.
- [59] Saber Moradi, Ning Qiao, Fabio Stefanini, and Giacomo Indiveri. “A Scalable Multi-core Architecture with Heterogeneous Memory Structures for Dynamic Neuromorphic Asynchronous Processors (DYNAPs)”. *IEEE Transactions on Biomedical Circuits and Systems*, 12(1):106–122, feb 2018. ISSN 19324545. doi: 10.1109/TBCAS.2017.2759700.
- [60] Jongkil Park, Theodore Yu, Siddharth Joshi, Christoph Maier, and Gert Cauwenberghs. “Hierarchical Address Event Routing for Reconfigurable Large-Scale Neuromorphic Systems”. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10):2408–2422, oct 2017. ISSN 21622388. doi: 10.1109/TNNLS.2016.2572164.
- [61] Steve B. Furber, David R. Lester, et al. “Overview of the SpiNNaker system architecture”. *IEEE Transactions on Computers*, 62(12):2454–2467, 2013. ISSN 00189340. doi: 10.1109/TC.2012.142.
- [62] Vladimir Kornijcuk, Jongkil Park, et al. “Reconfigurable Spike Routing Architectures for On-Chip Local Learning in Neuromorphic Systems”. *Advanced Materials Technologies*, 4(1), jan 2019. ISSN 2365709X. doi: 10.1002/admt.201800345.

- [63] David H Goldberg, Gert Cauwenberghs, and Andreas G Andreou. “Analog VLSI spiking neural network with address domain probabilistic synapses”. In *ISCAS 2001. The 2001 IEEE International Symposium on Circuits and Systems (Cat. No.01CH37196)*, pages 241–244, 2001. ISBN 0780366859.
- [64] Ben Varkey Benjamin, Peiran Gao, et al. “Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations”. *Proceedings of the IEEE*, 102(5):699–716, 2014. ISSN 00189219. doi: 10.1109/JPROC.2014.2313565.
- [65] Mike Davies, Narayan Srinivasa, et al. “Loihi: A Neuromorphic Manycore Processor with On-Chip Learning”. *IEEE Micro*, 38(1):82–99, 2018. ISSN 02721732. doi: 10.1109/MM.2018.112130359.
- [66] Paul A. Merolla, John V. Arthur, et al. “A million spiking-neuron integrated circuit with a scalable communication network and interface”. *Science*, 345(6197):668–673, 2014. ISSN 10959203. doi: 10.1126/science.1254642.
- [67] G. Indiveri, F. Corradi, and N. Qiao. “Neuromorphic architectures for spiking deep neural networks”. In *2015 IEEE International Electron Devices Meeting (IEDM)*, pages 4.2.1–4.2.4. IEEE, 2015. doi: 10.1109/IEDM.2015.7409623.
- [68] Steve Deiss, Rodney Douglas, Mike Fischer, Misha Mahowald, and Tony Matthews. “Address-Event Asynchronous Local Broadcast Protocol”, 1994. URL <https://www.ini.uzh.ch/~jamw/scx/aeprotocol.html>.
- [69] Kwabena A. Boahen. “Point-to-point connectivity between neuromorphic chips using address events”. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 47(5):416–434, 2000. ISSN 10577130. doi: 10.1109/82.842110.
- [70] Chiara Bartolozzi and Giacomo Indiveri. “Synaptic dynamics in analog VLSI”. *Neural Computation*, 19(10):2581–2603, 2007. ISSN 08997667. doi: 10.1162/neco.2007.19.10.2581.
- [71] Yoshifumi Nishi, Ulrich Bottger, Rainer Waser, and Stephan Menzel. “Crossover from Deterministic to Stochastic Nature of Resistive-Switching Statistics in a Tantalum Oxide Thin Film”. *IEEE Transactions on Electron Devices*, 65(10):4320–4325, oct 2018. ISSN 00189383. doi: 10.1109/TED.2018.2866127.
- [72] Tobi Delbruck. “Frame-free dynamic digital vision”. In *Intl. Symp. on Secure-Life Electronics, Advanced Electronics for Quality Life and Society*, pages 21–26, 2008. doi: <http://dx.doi.org/10.5167/uzh-17620>.

Three-dimensional monolithic integration of two layers of high-performance CMOS transistors with one layer of resistive memory devices

Contents

4.1	Goal of this chapter	144
4.2	Three-dimensional monolithic co-integration of resistive memories and CMOS transistors	144
4.2.1	CoolCube™ technology	144
4.2.2	Resistive memory integration	146
4.3	Electrical characterisation of the three-dimensional monolithic integration of two tiers of NMOS transistors with a tier of resistive memory devices	148
4.3.1	Basic functionality of bottom and top transistors . .	148
4.3.2	Characterisation of 1T1R structures	149
4.4	Discussion and conclusion	153

4.1 Goal of this chapter

THIS chapter concludes the study presented in this dissertation by demonstrating the co-integration of Resistive Memories (RRAMs) with CMOS transistors in a three-dimensional (3D) sequential integration, also called *monolithic integration*. Two layers (or *tiers*) of transistors are fabricated in a 3D monolithic integration of a 65-nm design rules Silicon On Insulator (SOI) CMOS over CMOS process, *i.e.* the top tier is fabricated directly on top of the bottom tier on a new active area. This allows alignment accuracy at the transistor scale. RRAMs are then integrated in the Back-End-Of-Line (BEOL) of the process on top of transistor contact plugs. As a proof-of-concept, we show the functionality of two one-transistor/one-RRAM (1T1R) structures in parallel, *i.e.* one RRAM is connected to a NMOS transistor from the bottom tier and one RRAM is connected to a NMOS transistor from the top tier. This enables to further benefit from the third dimension (*i.e.* the vertical axis) by both stacking several tiers of CMOS transistors on top of each other as well as RRAMs in the BEOL, and it opens up new perspectives for neuromorphic applications. For instance, this can potentially provide better synaptic density on a given silicon chip area - by vertically stacking several layers of RRAM-based 1T1R synapses such as the ones studied in CHAPTER 2 - and better synaptic routing density - by using Ternary Content-Addressable Memory (TCAM) circuits such as the ones studied in CHAPTER 3. We first describe the integration process, then we present the electrical characterisation results that demonstrate the basic functionality of the integration.

4.2 Three-dimensional monolithic co-integration of resistive memories and CMOS transistors

4.2.1 CoolCubeTMtechnology

Top and bottom CMOS transistors have been integrated with CoolCubeTMtechnology developed by CEA-Leti [1]. FIGURE 4.2.1 summarises the process flow. A first layer of CMOS transistors - the *bottom tier* or bottom level - is fabricated in a 65-nm Silicon On Insulator (SOI) CMOS process with high-k dielectric metal gate and raised source and drain junctions with epitaxy (FIGURE 4.2.1 (Left)) [1]. Then, a silicon layer is transferred on top of the bottom tier which corresponds to the top active area (FIGURE 4.2.1 (Middle)). In order to guarantee a top active area with electrical properties compatible with industrial requirements [2], a new SOI substrate layer is transferred by oxide-oxide (SiO₂/SiO₂) bonding on top of the bottom tier. The substrate is then thinned down by grinding and etching. Finally, the top tier is fabricated directly on the new transferred Si substrate (FIGURE 4.2.1 (Right)).

The main challenge of monolithic integration is to preserve performance of bottom transistors during the fabrication of the top tier. Indeed, it has been shown

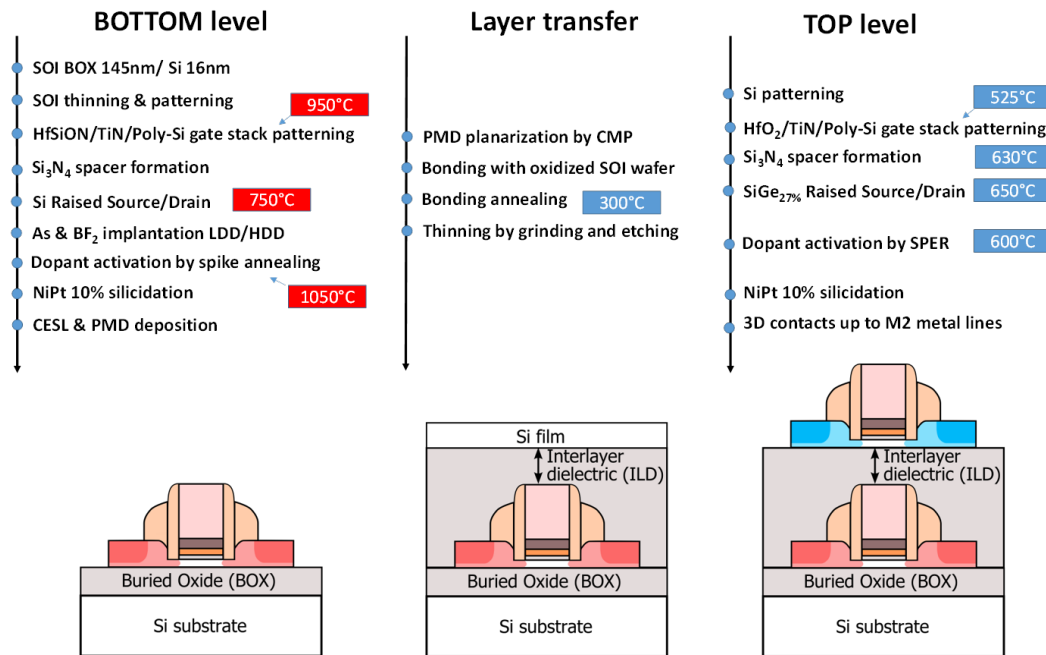


FIGURE 4.2.1: Process flow scheme of 3D CoolCube™ integration. (Left) The bottom level is fabricated at high temperature in a conventional 65-nm SOI CMOS process. (Middle) A new SOI wafer is transferred on top of the bottom level by oxide bonding, and it represents the top active area. (Right) The top level is fabricated at low thermal budget directly on top of the bottom level. Reproduced from [1].

that MOSFET performance are ensured up to 500°C for a few hours [3]. The degradation at higher temperature is mainly attributed to the deterioration of $\text{Ni}_{0.9}\text{Pt}_{0.1}$ silicide used to enhance access conductivity [3–5]. At high temperature, Transmission Electron Microscopy (TEM) observations show silicide clustering effects (for NMOS transistors) or silicide spread (for PMOS transistors) that degrade MOSFET performance [3, 5]. However, as shown in FIGURE 4.2.1 (Left), temperature can go up to 1050°C during transistor fabrication. In order to enable the fabrication of the top tier while maintaining bottom tier performance, Thermal Budget (TB) - *i.e.* temperature and process time - of the four critical steps (dopant activation, gate oxide stabilisation, source and drain epitaxy, and spacer deposition) during the top tier fabrication has been reduced. This has been made possible by the use and development of different processes, for instance Solid Phase Epitaxy Regrowth (SPER) or laser annealing [4–6]. All the technics employed to enable low-temperature process will not be reviewed here. It is also worth noting that all this process has been realised within an ultra-clean environment respecting hard contamination constraints [1, 7]. For future works, solutions to further decrease TB have been demonstrated, such as the use of low-k spacers deposited at temperatures below 500°C and low-temperature epitaxy down to 550°C [2]. Another alternative is to improve silicide stability of the bottom tier. For instance, the use of $\text{Ni}_{0.9}\text{Co}_{0.1}$ with Si-capping instead of $\text{Ni}_{0.9}\text{Pt}_{0.1}$ allows for a thermal stability up to 600°C for 2 hours [8]. Another significant advantage of CoolCube™ technology is that it is compatible with industrial requirements thanks to the use of conventional foundry process.

4.2.2 Resistive memory integration

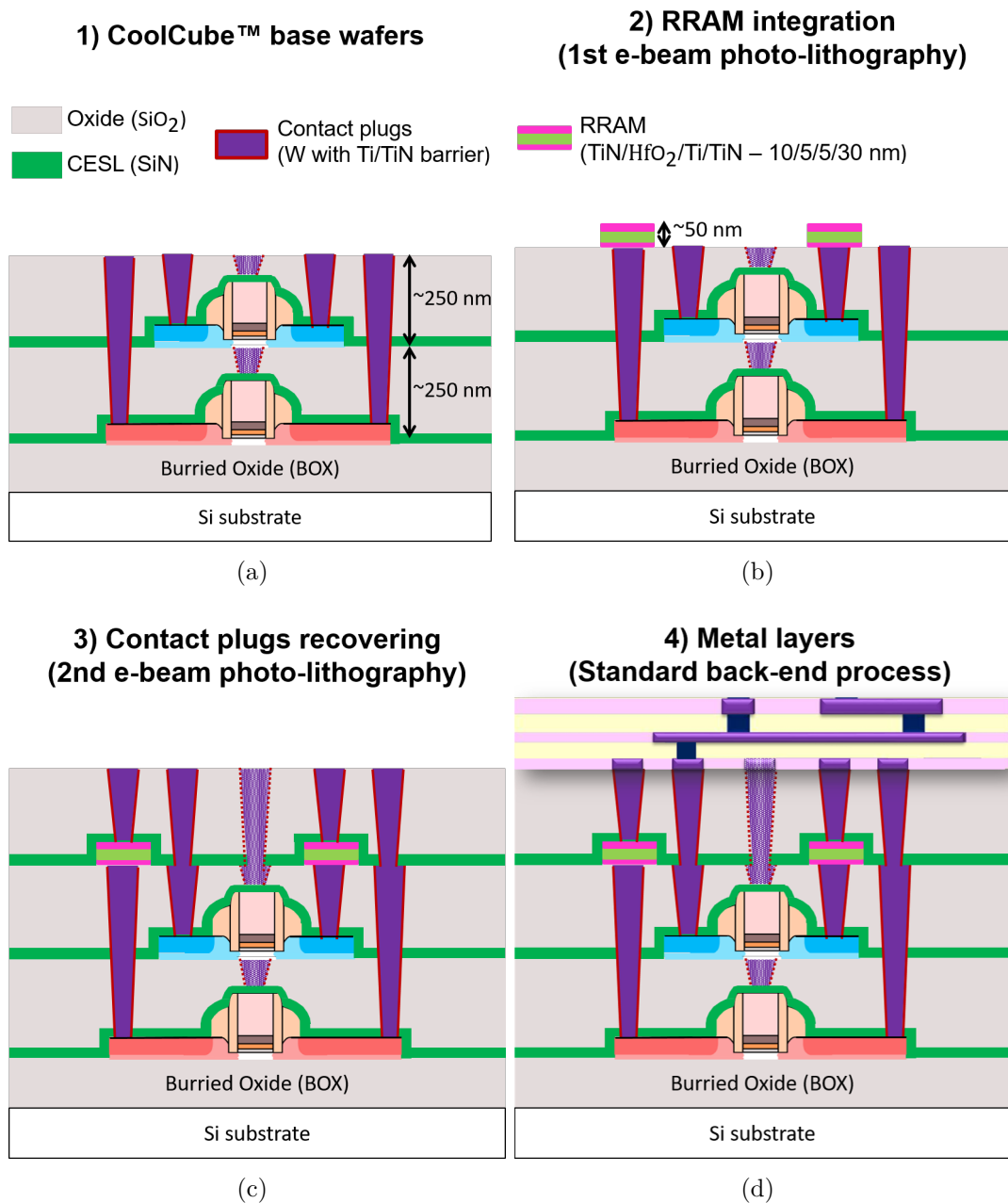


FIGURE 4.2.2: (a) Schematic illustration of CoolCube™ wafers before RRAM integration. Gate contact plugs (shaded purple area) are deposited. (b) TiN/HfO₂/Ti/TiN (10 nm/5 nm/5 nm/30 nm) RRAM devices are fabricated directly on top of contact plugs by a first e-beam photo-lithography. (c) Oxide and Contact Etch Stop Layer (CESL) are deposited on top of RRAM devices. Then, contact plugs are recovered by a second e-beam photo-lithography. (d) Integration is finished by standard CoolCube™ back-end-of-line process.

The goal of this work is to integrate HfO₂-based RRAMs - the same technology as in CHAPTER 2 and 3 - on top of CMOS transistors fabricated with CoolCube™ integration presented in the previous section. FIGURE 4.2.2 (a) depicts an example of two levels of NMOS transistors fabricated with Cool-

Cube™, just before the integration of RRAM devices. In this work, we retrieved CoolCube™ wafers before the first level of metal lines. RRAM devices have been integrated on top of CMOS contacts by a first e-beam photo-lithography step (FIGURE 4.2.2 (b)). The RRAM devices are composed of a TiN/HfO₂/Ti/TiN stack where layers are 10 nm/5 nm/5 nm/30 nm thick. Then, a second e-beam photo-lithography step has been added to recover contacts plugs on top of RRAM devices after oxide and Contact Etch Stop Layer (CESL) deposition (FIGURE 4.2.2 (c)). Finally, the integration has been finished by adding metal levels with the standard CoolCube™ back-end-of-line process (FIGURE 4.2.2 (d)). FIGURE 4.2.3 shows a TEM observation of the integrated RRAM devices on top of two levels of NMOS transistors in a 65-nm SOI CMOS process on 300-mm (12-inch) SOI wafers. This represents two one-transistor/one-RRAM (1T1R) structures in parallel, *i.e.* one RRAM is controlled by the transistor of the bottom tier and one RRAM is controlled by the transistor of the top tier. Ideally, a level of RRAM should be integrated in between the two NMOS levels to benefit as much as possible from the third dimension. However, for the sake of simplicity and to prevent any contamination issue, we integrated RRAM devices only on top of the second NMOS level. In the next section, we will show the functionality of this first-in-world co-integration.

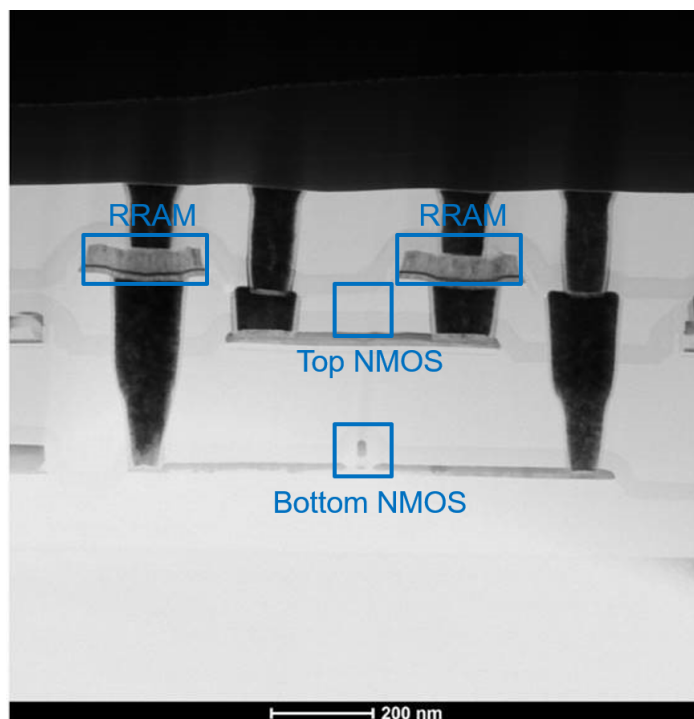


FIGURE 4.2.3: Transmission electron microscopy of the co-integration of HfO₂-based RRAMs on top of two NMOS transistors fabricated with CoolCube™ technology.

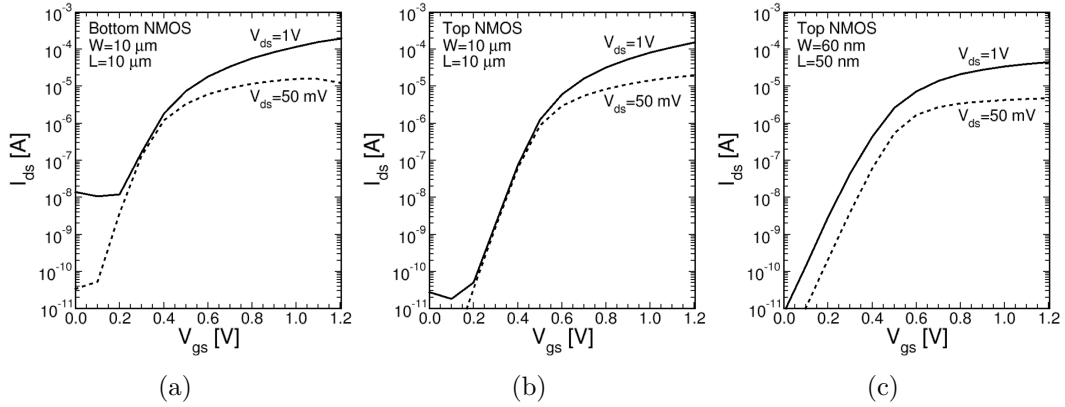


FIGURE 4.3.1: I_{ds} - V_{gs} characteristics measured on (a) a bottom NMOS transistor with $W=L=10 \mu\text{m}$, (b) a top NMOS transistor with $W=L=10 \mu\text{m}$, and (c) a top NMOS transistor with $W=60\text{ nm}$ and $L=50\text{ nm}$. Characteristics have been measured at $V_{ds}=50\text{ mV}$ (dotted line) and $V_{ds}=1\text{ V}$ (solid line).

4.3 Electrical characterisation of the three-dimensional monolithic integration of two tiers of NMOS transistors with a tier of resistive memory devices

In this section, we present electrical characterisation results showing the basic functionality of RRAMs co-integrated with 3D monolithic NMOS transistors. We will refer to transistors of the bottom tier as *bottom transistors* and to transistors of the top tier as *top transistors*.

4.3.1 Basic functionality of bottom and top transistors

We first tested the functionality of bottom and top NMOS transistors. For this purpose, devices with only the bottom NMOS transistor (without the top NMOS transistor) and devices with only the top NMOS transistor (without the bottom NMOS transistor) have been fabricated on the same die. FIGURE 4.3.1 shows I_{ds} - V_{gs} characteristic for (a) a bottom and (b) a top transistor with $W=L=10 \mu\text{m}$ at $V_{ds}=50\text{ mV}$ (dotted line) and $V_{ds}=V_{dd}=1.0\text{ V}$ (solid line). The characteristics show control of the current with V_{gs} for currents up to $300 \mu\text{A}$ (setup limit). This is an essential requirement to limit current during programming of RRAMs. In addition, transistors must drive enough current to enable switching of the RRAMs - usually in the order of hundreds of microamperes. Smaller top transistors have also been proven to be functional (FIGURE 4.3.1 (c), with $W=60\text{ nm}$ and $L=50\text{ nm}$).

We demonstrated here the basic functionality of top and bottom transistors in order to program RRAM devices in 1T1R structures. We will now demonstrate the functionality of the 1T1R structures.

4.3.2 Characterisation of 1T1R structures

The device under test is equivalent to two one-transistor/one-RRAM (1T1R) in parallel (NMOS and HfO₂-based RRAM). Both bottom and top NMOS transistors feature a gate width $W=1\ \mu\text{m}$ and length $L=100\ \text{nm}$. Both RRAMs are rectangular with an area of $0.25\ \mu\text{m}^2$ ($0.25\times 1\ \mu\text{m}$). One RRAM is connected to the bottom NMOS and one RRAM is connected to the top NMOS. In the following, we will refer to the 1T1R with the RRAM connected to the bottom NMOS as the *bottom 1T1R* and to the 1T1R with the RRAM connected to the top NMOS as the *top 1T1R*.

4.3.2.1 Quasi-static and pulsed measurements

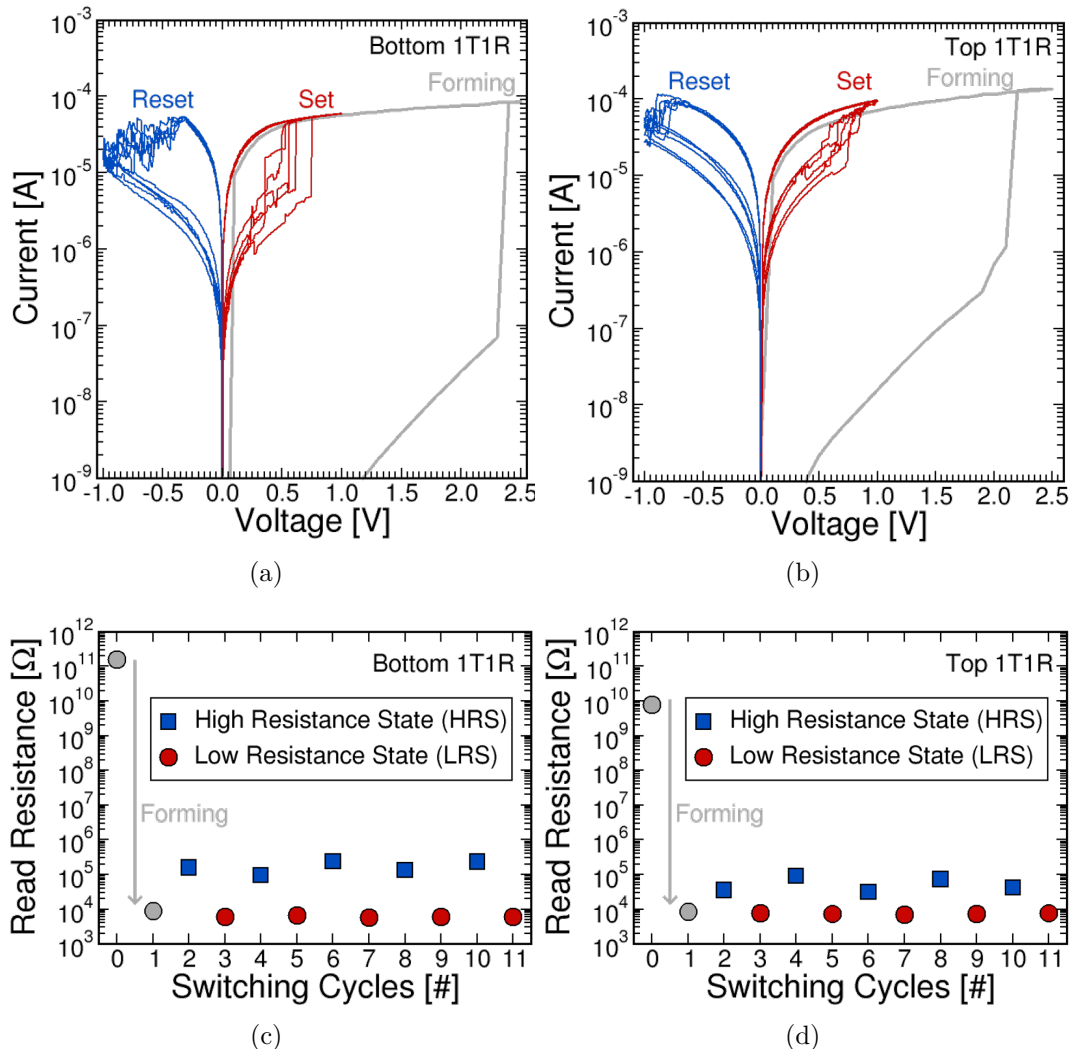


FIGURE 4.3.2: (a) Butterfly I-V curves measured on the bottom 1T1R and (b) the top 1T1R. Forming, then five Reset-Set cycles have been performed with the programming conditions in TABLE 4.1. (c) Read resistance values measured after each switching operation on the bottom 1T1R and (d) on the top 1T1R.

We first characterised the bottom and top 1T1R in quasi-static measurements.

FIGURE 4.3.2 shows classic butterfly I-V curve measurements performed on (a) the bottom 1T1R and (b) the top 1T1R. Forming, then five Reset-Set switching cycles in quasi-static have been performed on the bottom 1T1R (FIGURE 4.3.2 (a)). Then, the same protocol (Forming, five Reset-Set cycles) has been performed on the top 1T1R (FIGURE 4.3.2 (b)). TABLE 4.1 summarises the programming conditions used for the measurements. A *positive voltage* is applied on the top electrode during Forming and Set operations, while a *negative voltage* is applied on the top electrode during Reset operations. Resistance values of the RRAMs have been read after each programming operation with a read voltage $V_{\text{read}}=0.1$ V and $V_{\text{gs}}=1.2$ V - in order to fully activate the transistor. FIGURE 4.3.2 (c) and FIGURE 4.3.2 (d) show read resistance values after each programming operation for the bottom and top 1T1R, respectively. As evidenced here, forming occurs on the bottom 1T1R at $V_{\text{forming}}=2.4$ V, and the RRAM switches from its pristine state to a Low Resistance State (LRS) ($R \approx 8.9$ k Ω). Similarly, the top 1T1R is formed at $V_{\text{forming}}=2.2$ V with a LRS resistance value after forming of ≈ 8.5 k Ω . Subsequent Reset and Set operations make RRAM devices switch between the High Resistance State (HRS) and the LRS. Switching to HRS is gradual and depend on the applied voltage, while switching to LRS is abrupt. This is consistent with previous works reported on this technology [9]. In addition, switching voltages, $V_{\text{switching}}$, obtained with our RRAM devices are in agreement with reported data on the same RRAM technology - in particular same oxide thickness [10–12]. We then demonstrated there was no crosstalk between the two RRAMs. We performed Set-Reset operations on one of the two 1T1R, and we verified the resistance value of the other 1T1R after each programming operation. The results are shown in FIGURE 4.3.3. During Set and Reset operations performed on the top 1T1R (FIGURE 4.3.3 (a)), the bottom 1T1R resistance state is not altered after each switching operation. Similarly, performing Set and Reset operations on the bottom 1T1R (FIGURE 4.3.3 (b)) does not alter the top 1T1R resistance state. This confirms that there is no crosstalk between the two 1T1R since programming one of the two 1T1R does not affect the resistance state of the other 1T1R.

We then characterised the bottom and top 1T1R in pulsed measurements. TABLE 4.2 summarises the programming conditions of each operation. After forming of the bottom and top 1T1R, 10^5 Set-Reset cycles are applied on the bottom 1T1R (*cf* FIGURE 4.3.4 (a)). Then, 10^5 Set-Reset cycles are applied on the top 1T1R (*cf* FIGURE 4.3.4 (b)). We can sustain a ratio between the HRS and LRS resistance values of about 15 and 12 for the bottom 1T1R and the top 1T1R for 10^5 cycles, respectively.

4.3.2.2 Impact of programming current

We finally verified the possibility to control LRS resistance values with the programming current, I_{prog} , *i.e.* by biasing transistor gates at different gate voltages, V_{gs} . As shown in FIGURE 4.3.5 (a) and (b), it is possible to program the RRAM devices at several LRS resistance values by fixing different I_{prog} ((a) and (b) for the bottom and top 1T1R, respectively). FIGURE 4.3.5 (c) reports LRS resistance values as a function of I_{prog} for the bottom 1T1R (red diamond) and the top 1T1R (blue triangle). For the sake of comparison, reported data on

4.3. ELECTRICAL CHARACTERISATION

	Sweep Voltage	Gate Voltage V_{gs}	Switching Voltage $V_{switching}$	Read Resistance
<i>Bottom 1T1R</i>				
Forming	$0 \rightarrow 2.5 \text{ V} \rightarrow 0$	0.6 V ($I_{cc} \approx 100 \mu\text{A}$)	2.4 V	8.9 k Ω
Set	$0 \rightarrow 1.0 \text{ V} \rightarrow 0$	0.6 V ($I_{cc} \approx 100 \mu\text{A}$)	$\approx 0.60 \text{ V}$	$\approx 6.1 \text{ k}\Omega$
Reset	$0 \rightarrow -1.0 \text{ V} \rightarrow 0$	1.2 V	$\approx -0.34 \text{ V}$	$\approx 172 \text{ k}\Omega$
<i>Top 1T1R</i>				
Forming	$0 \rightarrow 2.5 \text{ V} \rightarrow 0$	0.6 V ($I_{cc} \approx 100 \mu\text{A}$)	2.2 V	8.5 k Ω
Set	$0 \rightarrow 1.0 \text{ V} \rightarrow 0$	0.85 V ($I_{cc} \approx 130 \mu\text{A}$)	$\approx 0.71 \text{ V}$	$\approx 7.5 \text{ k}\Omega$
Reset	$0 \rightarrow -1.0 \text{ V} \rightarrow 0$	1.35 V	$\approx -0.82 \text{ V}$	$\approx 54.4 \text{ k}\Omega$

TABLE 4.1: Programming conditions used for the bottom and top 1T1R measurements in quasi-static mode (butterfly I-V curves). For forming and Set operations, the switching voltage $V_{switching}$ corresponds to the voltage required to abruptly switch from the High Resistance State (HRS) to the Low Resistance State (LRS). For Reset operations, it corresponds to the voltage at which the resistance value of RRAM devices starts to decrease (onset of the switching from LRS to HRS). $V_{switching}$ and read resistance values have been averaged over five Set or Reset operations.

	Programming Voltage V_{Prog}	Gate Voltage V_{gs}	Pulse Width
<i>Bottom 1T1R</i>			
Forming	2.75 V	1.2 V ($I_{cc} \approx 200 \mu\text{A}$)	1 ms
Set	1.6 V	1.2 V ($I_{cc} \approx 200 \mu\text{A}$)	1 μs
Reset	-2.0 V	1.2 V	1 μs
<i>Top 1T1R</i>			
Forming	2.75 V	1.2 V ($I_{cc} \approx 200 \mu\text{A}$)	1 ms
Set	1.6 V	1.35 V ($I_{cc} \approx 220 \mu\text{A}$)	1 μs
Reset	-2.0 V	1.35 V	1 μs

TABLE 4.2: Programming conditions used for the bottom and top 1T1R measurements in pulsed mode.

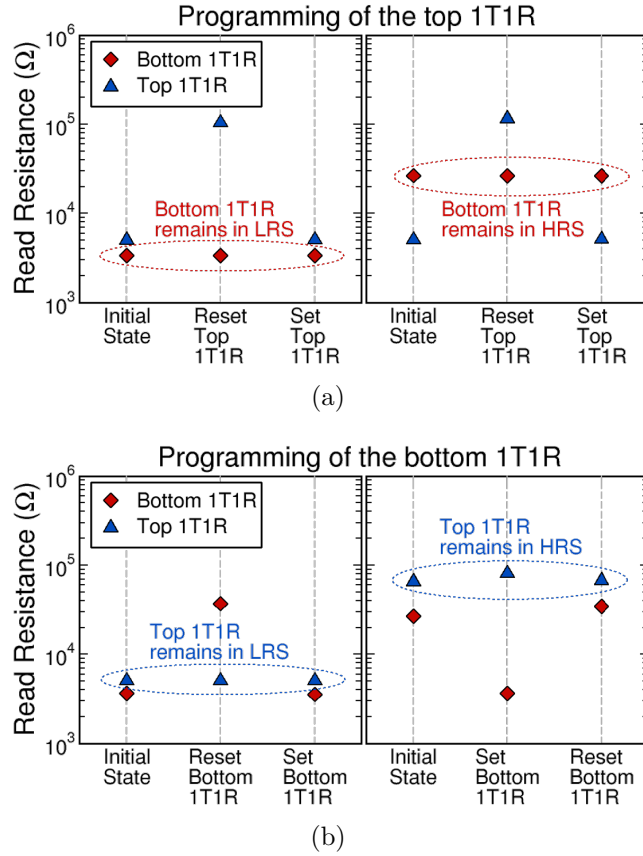


FIGURE 4.3.3: Demonstration of the absence of crosstalk between the bottom and top 1T1R. (a) Set and Reset operations have been performed on the top 1T1R, and the resistance values of the bottom and top 1T1R have been read after each switching operation. Bottom 1T1R resistance states remain unaltered after each switching operation on the top 1T1R. (b) Similarly, Set and Reset operations have been performed on the bottom 1T1R. Top 1T1R resistance states remain unaltered after each switching operation on the bottom 1T1R. Programming conditions in TABLE 4.1 have been used.

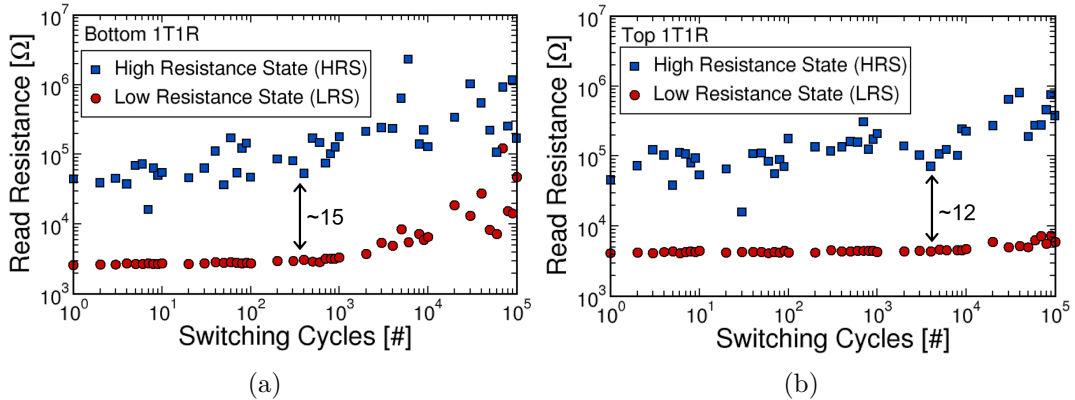


FIGURE 4.3.4: Endurance characterisations performed on (a) the bottom 1T1R and (b) the top 1T1R for 10^5 switching cycles. Measurements have been performed with the programming conditions in TABLE 4.2.

different RRAM technologies are also shown [9]. In agreement with previous

works, LRS resistance values decrease when I_{prog} increases. LRS resistance values exhibit the same power law relationship with I_{prog} as reported RRAM technologies. This proves the complete functionality of the integration.

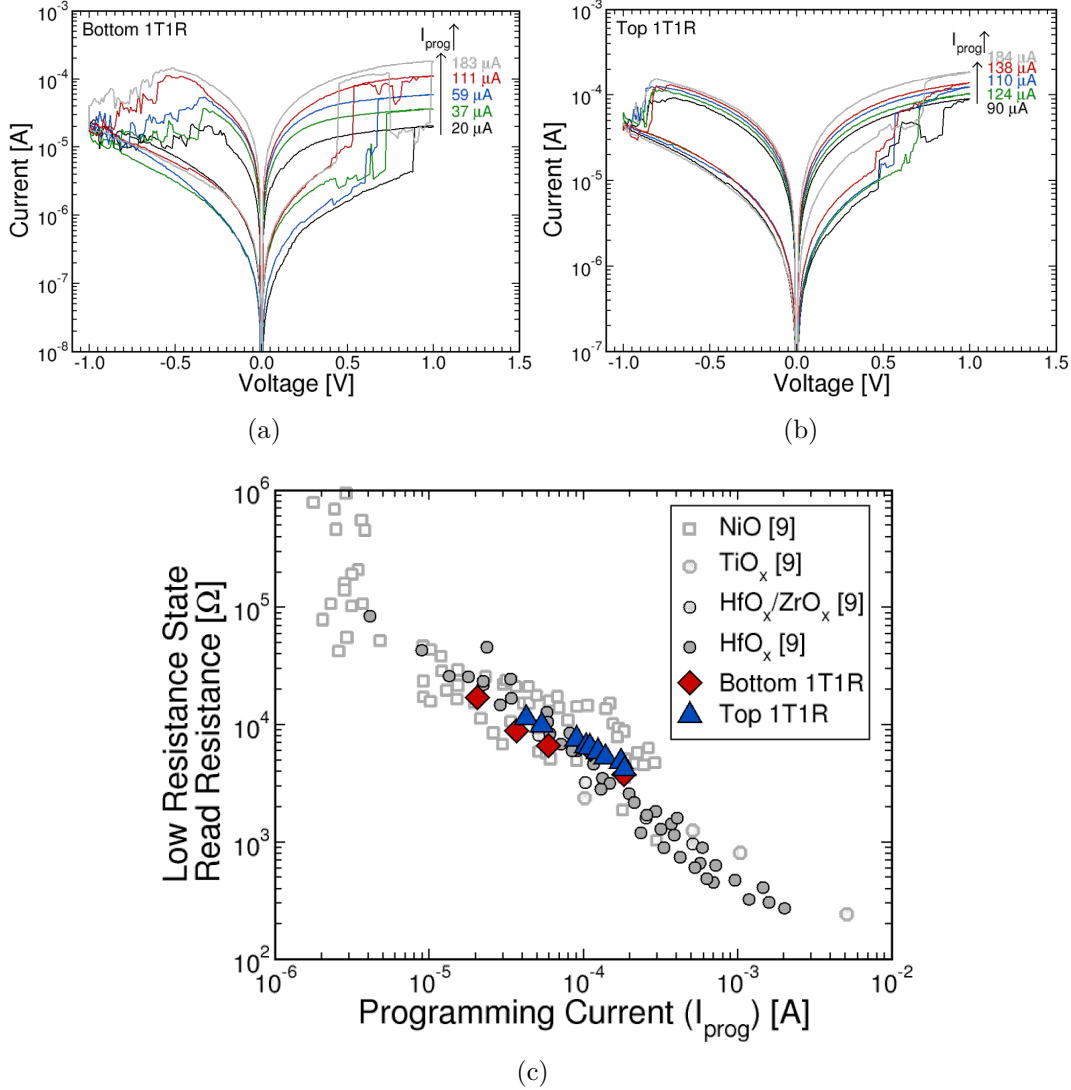


FIGURE 4.3.5: (a) Butterfly I-V curves measured on the bottom 1T1R and (b) the top 1T1R with different programming currents, I_{prog} , for each Set operation. Increasing I_{cc} allows to decrease resistance values after each Set operation. (c) Low Resistance State (LRS) resistance values as a function of programming current, I_{prog} , for the bottom 1T1R (red diamond) and the top 1T1R (blue triangle). Data on different RRAM technologies reproduced from [9] are reported for comparison. Measured data of the bottom and top 1T1R are in good agreement with previous works.

4.4 Discussion and conclusion

In this chapter we demonstrated the feasibility of three-dimensional sequential (3D monolithic) integration of CMOS transistors with Resistive Memories

(RRAMs). Two levels (*tiers*) of CMOS transistors are vertically stacked on top of each other with CoolCube™ technology [1], and HfO₂-based RRAMs have been integrated in the Back-End-Of-Line (BEOL) of the process. This results in two one-transistor/one-RRAM (1T1R) structures in parallel. This work advances the state-of-the-art by demonstrating a *full 3D monolithic integration of two tiers of high-performance CMOS transistors with a tier of RRAMs*. Indeed, 3D monolithic integration of CMOS transistors have already been demonstrated - for instance in [13–15]. Also, heterogenous 3D integration of one tier of CMOS with other technologies, such as carbon nanotubes and RRAMs, have also been presented in [16, 17]. However, previous works often rely on complex and expensive fabrication process, for instance with the use of III-V materials [5, 14, 18] or polysilicon channel integration resulting in lower electrical performance of top tiers [13, 15, 17]. A significant advantage of CoolCube™ technology [1] over other 3D monolithic technologies is the fabrication of a top level compatible with state-of-the-art high-performance Fully-Depleted SOI process requirements, such as high-k/metal gate or raised source and drain [1]. In addition, CoolCube™ technology is compatible with industrial requirements thanks to the use of conventional foundry process, in particular in terms of hard contamination constraints. Similarly, RRAM devices are fabricated using mature and extensively studied RRAM technology - TiN/HfO₂/Ti/TiN RRAM stack - compatible with industrial CMOS BEOL process [10, 11, 19]. Future works on CoolCube™ technology will permit the introduction of intermediate metallic lines (*intermediate BEOL*) in between CMOS tiers by decontamination and encapsulation of the wafer bevel edge as described in [20]. Intermediate metallic lines are mandatory to avoid routing congestion and allow for the integration of highly interconnected systems to fully benefit from the third dimension. For this purpose, intermediate ULK/copper with standard BEOL will be introduced in future works. In addition, this also provides opportunities to integrate RRAM devices in between CMOS tiers (inside the intermediate metal layers) and to bring memory cells as close as possible to processing units for in-memory computing architectures [21, 22].

The major challenge of 3D monolithic integration - in this case of CoolCube™ integration - is the fabrication of top tiers at low thermal budget, while ensuring high-performance for every CMOS tier. In this work, performance of bottom and top transistors differ since this is still a preliminary work, and the integration was not fully optimised in the light of recent advances [5, 6, 8, 20]. Indeed, recent works on 3D monolithic integration have allowed to address the main technological roadblocks, and similar performance are expected to be obtained for every CMOS tier [5, 6, 8, 20]. Still, we experimentally proved in this chapter that current electrical performance of bottom and top transistors is sufficient to program RRAMs. That is, enough current can be driven during programming operations to trigger the switching between High and Low Resistance States (HRS and LRS) (see FIGURE 4.3.4) and to control LRS resistance values (see FIGURE 4.3.5).

Benefits have been envisioned with 3D monolithic technologies [23–29]. The major advantage of such technologies is the gain in area for a given chip area. Depending on the partitioning method, gain is expected to be about 2x for cell-on-cell partitioning and less (≈ 1.67) for transistor-on-transistor level parti-

tioning [23]. Indeed, we need more room for the bottom active area in order to etch contacts with respect to the planar case [27] (see the transmission electron microscopy in FIGURE 4.2.3, a certain distance has to be guaranteed between top and bottom contact plugs, hence increasing the bottom active area). In terms of cost (that takes area gain into account), cost benefits of the order of 50% or more for large wafers are expected [23]. 3D monolithic integration can also provide benefits in terms of circuit performance as it results in smaller cell size and generally in shorter interconnections [26, 27]. As a result, we can expect gains in speed and energy consumption as it has been simulated for instance in [21] or [24]. In CHAPTER 2, we studied the implementation of artificial synapses with 1T1R synapses using the same technology as the one studied in this chapter. Spiking neural networks can naturally benefit from 3D monolithic integration [25]. As an example, the different layers of neuron circuits can be fabricated on each tier, while RRAM-based synapses are physically located in between each layer (inside the intermediate metal layers) to interconnect neurons in a truly brain-like architecture. In addition, alignment accuracy of 3D monolithic integration allows for higher density of interconnections with respect to 3D parallel integration [1]. This is an important requirement to implement hardware neuromorphic systems since biological brains contain more than 10^{14} synapses, and synapses outnumber neurons by four orders of magnitude. Finally, we demonstrated in CHAPTER 2 that RRAM resistance variability can be beneficial for spiking neural networks trained with unsupervised learning rules. As a result, even though electrical performance of top transistors may be degraded, this would have a minimal impact on network performance since it only affects RRAM during programming operations (*i.e.* RRAMs programmed with top transistors probably exhibit higher resistance variability). In CHAPTER 3, we studied the implementation of Ternary Content-Addressable Memories (TCAMs) with RRAMs as synaptic routing lookup tables for multi-core neuromorphic processor reconfigurability. As we experimentally proved, RRAM-based TCAMs considerably benefit from the gain in area with 3D monolithic integration. Indeed, this allows for a decrease in TCAM bitcell size, thus shorter match lines [28, 29]. Consequently, search times, search endurance, and search energy consumption improve. However, the work presented in this chapter remains a proof-of-concept of the 3D monolithic integration of CMOS and RRAMs, and previous statements are perspectives to be explored. Recent advances in 3D monolithic technology will permit to achieve similar performance for every fabricated tier. Therefore, future works need to investigate more in depth the pros and cons of 3D monolithic integration in order to optimise and fully benefit from such integration. In particular, a thorough evaluation of the advantages of 3D monolithic integration over more conventional integrations, such as planar or 3D parallel integrations, is essential.

References: Chapter 4

- [1] L. Brunet, P. Batude, et al. “First demonstration of a CMOS over CMOS 3D VLSI CoolCube™ integration on 300mm wafers”. *Digest of Technical Papers - Symposium on VLSI Technology*, 2016-Sept:11–12, 2016. ISSN 07431562. doi: 10.1109/VLSIT.2016.7573428.
- [2] Perrine Batude. *Integration a trois dimensions sequentielle : Etude, fabrication et caracterisation*. PhD thesis, Institut National Polytechnique de Grenoble - INPG, 2009.
- [3] C. Fenouillet-Beranger, B. Previtali, et al. “FDSOI bottom MOSFETs stability versus top transistor thermal budget featuring 3D monolithic integration”. In *European Solid-State Device Research Conference*, pages 110–113, 2014. ISBN 9781479943784. doi: 10.1109/ESSDERC.2014.6948770.
- [4] C. Fenouillet-Beranger, B. Mathieu, et al. “New insights on bottom layer thermal stability and laser annealing promises for high performance 3D VLSI”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 27.5.1–27.5.4, 2014. ISBN 9781479980017. doi: 10.1109/IEDM.2014.7047121.
- [5] Cao-minh Lu. *Fabrication de CMOS à basse temperature pour l'integration 3D sequentielle*. PhD thesis, Universite Grenoble Alpes, 2017.
- [6] Jessy Micout. *Fabrication et caracterisation de transistor realisee a basse temperature pour l'integration 3D sequentielle*. PhD thesis, Universite Grenoble Alpes, 2019.
- [7] C. Fenouillet-Beranger, S. Kerdiles, et al. “W and Copper Interconnection Stability for 3D VLSI CoolCube Integration”. In *International Conference on Solid State Devices and Materials*, 2015. doi: 10.7567/ssdm.2015.k-4-3.
- [8] Fabien Deprat, Fabrice Nemouchi, et al. “Technological enhancers effect on Ni_{0.9}Co_{0.1} silicide stability for 3D sequential integration”. *Physica Status Solidi (C) Current Topics in Solid State Physics*, 13(10-12):760–765, 2016. ISSN 16101642. doi: 10.1002/pssc.201600043.
- [9] Daniele Ielmini. “Resistive switching memories based on metal oxides: Mechanisms, reliability and scaling”. *Semiconductor Science and Technology*, 31(6), 2016. ISSN 13616641. doi: 10.1088/0268-1242/31/6/063002.
- [10] H. Y. Lee, P. S. Chen, et al. “Low power and high speed bipolar switching with a thin reactive ti buffer layer in robust HfO₂ based RRAM”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 1–4, 2008. ISBN 9781424423781. doi: 10.1109/IEDM.2008.4796677.
- [11] Erika Covi, Stefano Brivio, et al. “Analog memristive synapse in spiking networks implementing unsupervised learning”. *Frontiers in Neuroscience*, 10(OCT), oct 2016. ISSN 1662453X. doi: 10.3389/fnins.2016.00482.
- [12] Jacopo Frascaroli, Stefano Brivio, Erika Covi, and Sabina Spiga. “Evidence of soft bound behaviour in analogue memristive devices for neuromorphic computing”. *Scientific Reports*, 8(1):1–12, 2018. ISSN 20452322. doi: 10.1038/s41598-018-25376-x.

- [13] Soon Moon Jung, Hoon Lim, et al. “High speed and highly cost effective 72M bit density S3 SRAM technology with doubly stacked Si layers, peripheral only CoSix layers and Tungsten shunt W/L scheme for standalone and embedded memory”. In *Digest of Technical Papers - Symposium on VLSI Technology*, pages 82–83, 2007. ISBN 9784900784031. doi: 10.1109/VLSIT.2007.4339736.
- [14] T. Irisawa, K. Ikeda, et al. “Demonstration of ultimate CMOS based on 3D stacked InGaAs-OI/SGOI wire channel MOSFETs with independent back gate”. In *Digest of Technical Papers - Symposium on VLSI Technology*, pages 1–2, 2014. ISBN 9781479933310. doi: 10.1109/VLSIT.2014.6894395.
- [15] Chang Hong Shen, Jia Min Shieh, et al. “Heterogeneously integrated sub-40nm low-power epi-like Ge/Si monolithic 3D-IC with stacked SiGeC ambient light harvester”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 3.6.1–3.6.4, 2014. ISBN 9781479980017. doi: 10.1109/IEDM.2014.7046975.
- [16] M. M. Shulaker, K. Saraswat, H. . P. Wong, and S. Mitra. “Monolithic three-dimensional integration of carbon nanotube FETs with silicon CMOS”. In *Digest of Technical Papers - Symposium on VLSI Technology*, pages 1–2, 2014. ISBN 9780979806476. doi: 10.1017/S1431927607073709.
- [17] Max M. Shulaker, Gage Hills, et al. “Three-dimensional integration of nanotechnologies for computing and data storage on a single chip”. *Nature*, 547(7661):74–78, 2017. ISSN 14764687. doi: 10.1038/nature22994.
- [18] V. Deshpande, V. Djara, et al. “Advanced 3D Monolithic hybrid CMOS with Sub-50 nm gate inverters featuring replacement metal gate (RMG)-InGaAs nFETs on SiGe-OI Fin pFETs”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 8.8.1–8.8.4, 2015. ISBN 9781467398930. doi: 10.1109/IEDM.2015.7409658.
- [19] A. Grossi, E. Nowak, et al. “Fundamental variability limits of filament-based RRAM”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 4.7.1–4.7.4, 2016. ISBN 9781509039012. doi: 10.1109/IEDM.2016.7838348.
- [20] L Brunet, P Batude, et al. “Breakthroughs in 3D Sequential technology”. In *2018 IEEE International Electron Devices Meeting (IEDM)*, pages 7.2.1–7.2.4. IEEE, 2018. ISBN 9781728119878.
- [21] M. M. Sabry Aly, M. Gao, et al. “Energy-Efficient Abundant-Data Computing: The N3XT 1,000x”. *Computer*, 48(12):24–33, 2015. doi: 10.1109/MC.2015.376.
- [22] Hongsik Jeong and Luping Shi. “Memristor devices for neural networks”. *Journal of Physics D: Applied Physics*, 52(2):023003, 2019. ISSN 13616463. doi: 10.1088/1361-6463/aae223.
- [23] Daniel Gitlin, Maud Vinet, and Fabien Clermidy. “Cost model for monolithic 3D integrated circuits”. In *2016 SOI-3D-Subthreshold Microelectronics Technology Unified Conference, S3S 2016*, pages 1–2, 2016. ISBN 9781509043903. doi: 10.1109/S3S.2016.7804408.
- [24] Ogun Turkyilmaz, Gerald Cibrario, Olivier Rozeau, Perrine Batude, and Fabien Clermidy. “3D FPGA using high-density interconnect Monolithic Integration”. In *Proceedings - Design, Automation and Test in Europe, DATE*, pages 1–4, 2014. ISBN 9783981537024. doi: 10.7873/DATE2014.351.
- [25] Hongyu An, Zhen Zhou, and Yang Yi. “Memristor-based 3D neuromorphic computing system and its application to associative memory learning”. In *2017 IEEE 17th International Conference on Nanotechnology, NANO 2017*, pages 555–560, 2017. ISBN 9781509030286. doi: 10.1109/NANO.2017.8117459.

- [26] W. Rhett Davis, Eun Chu Oh, Ambarish M. Sule, and Paul D. Franzon. “Application exploration for 3-D integrated circuits: TCAM, FIFO, and FFT case studies”. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 17(4):496–506, 2009. ISSN 10638210. doi: 10.1109/TVLSI.2008.2009352.
- [27] Y Lee, P Morrow, and S K Lim. “Ultra high density logic designs using transistor-level monolithic 3D integration”. In *2012 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 539–546, nov 2012.
- [28] Mingjie Lin, Jianying Luo, and Ma Yaling. “A low-power monolithically stacked 3D-TCAM”. In *Proceedings - IEEE International Symposium on Circuits and Systems*, pages 3318–3321, 2008. ISBN 9781424416844. doi: 10.1109/ISCAS.2008.4542168.
- [29] Eun Chu Oh and Paul D. Franzon. “Design Considerations and Benefits of Three-Dimensional Ternary Content Addressable Memory”. In *Proceedings of the IEEE 2007 Custom Integrated Circuits Conference, CICC 2007*, pages 591–594, 2007. ISBN 1424407869. doi: 10.1109/CICC.2007.4405801.

Conclusion and perspectives

THE objective of this work was to study the role of resistive memories and three-dimensional monolithic technologies to enable the hardware implementation of compact, energy-efficient reconfigurable multi-core neuromorphic processors. We investigated the *impact of Resistive Memory (RRAM) electrical characteristics on performance and reliability of Spiking Neural Networks (SNNs)*. To this end, we provided comprehensive studies of the impact of RRAM resistance ratio (memory window), resistance variability, and programming endurance on RRAM-based synaptic arrays and synaptic routing tables in CHAPTER 2 and CHAPTER 3, respectively. Finally, we opened up technological perspectives to further improve neuromorphic core area-efficiency by demonstrating the feasibility of a complete three-dimensional sequential integration (3D monolithic integration) of two levels of CMOS transistors associated with a level of RRAM devices in CHAPTER 4.

The main building blocks of SNN neuromorphic cores are: (i) synaptic arrays with adjustable synaptic weights, (ii) synaptic routing tables for on-the-fly network topology reconfigurability, (iii) neuron circuits with adjustable parameters, and (iv) on-line learning circuitry. In this work, we *solely focused on (i) synaptic arrays with adjustable synaptic weights, and (ii) synaptic routing tables, both implemented with RRAMs*. Many challenges related to RRAM electrical characteristics are yet to be addressed to utilise RRAMs for standard memory applications: the rather low memory window (≈ 10 -100) and high cycle-to-cycle and device-to-device resistance variability in order to obtain programming endurance as long as that of flash memories ($\approx 10^6$ Set/Reset cycles). While these limitations hinder the integration of large memory arrays with RRAMs, we argued that RRAM requirements for neuromorphic cores remarkably differ from those of memory systems. We demonstrated that the aforementioned issues are not detrimental for the implementation of RRAM-based synaptic circuits and RRAM-based synaptic routing tables.

In CHAPTER 2, we studied the implementation of RRAM-based synapses and investigated the impact of RRAM electrical properties on SNN performance trained with the unsupervised spike-timing-dependent plasticity learning paradigm. A systematic study has been carried out by means of extensive system-level simulations calibrated on experimental electrical characterisation of 4-kbit RRAM arrays. Two applications have been simulated: (i) detection in dynamic patterns (car tracking [1]), and (ii) classification of static patterns (handwritten digits of MNIST [2]). For this purpose, SNNs were based on canonical fully-connected feed-forward neural network topology and trained with the unsupervised Spike-Timing-Dependent Plasticity (STDP) learning rule. Synaptic elements were implemented with simple one-transistor/one-RRAM (1T1R) structures. First, we considered binary RRAM devices, *i.e.* RRAMs exhibiting abrupt switching between two distinct states: the Low Resistance State (LRS) and High Resistance State (HRS). In order to obtain analog synaptic weight evolution using binary RRAMs, the synaptic compound associated with a stochastic STDP learning rule proposed in [3] has been used wherein each synaptic element is implemented by several RRAM devices operating in parallel. For detection applications, we demonstrated that both the Memory Window (MW, ratio between HRS and LRS resistance values) and resistance variability (cycle-to-cycle and device-to-device) improve SNN performance after training. Consequently, resistance variability is beneficial as it can compensate for low memory windows. For classification applications, we demonstrated that the MW has no impact on SNN performance. However, a certain amount of resistance variability improves network performance with respect to the case of zero variability. The results obtained in this work provide valuable insights for the use of RRAMs as artificial synapses in SNNs with unsupervised learning. Indeed, this allows to program RRAMs in a low-current regime during training. This improves power consumption during training by a factor 4x with respect to the use of programming conditions optimised for standard memory applications. Furthermore, the use of low-energy programming pulses also improves RRAM programming endurance. This is a significant advantage since we observed that broken cells due to RRAM aging are detrimental for SNN performance. While it has been shown numerous times that SNNs trained with unsupervised learning are robust to RRAM resistance variability, we clarified here the role of RRAM resistance variability. We proved that RRAM resistance variability is actually beneficial to reach high performance as it increases the dynamic range of synaptic weights available during training. The performance improvement lies in the unsupervised nature of the learning rule: as in biological brains, RRAM resistance variability provides more synaptic weights to synapses to be explored during training, and it prevents the system from being stuck in sub-optimal solutions. Second, we considered analog devices, *i.e.* devices capable of gradual modulation of their conductance value upon the application of identical potentiation or depression pulses. In particular, we evaluated the potential of the PCM technology presented in [4] wherein analog conductance modulation is obtained in both crystallisation and amorphisation. The crucial result of this study was that the natural non-linearity of conductance response of PCM technology as well as many RRAM technologies improves SNN performance. This is in sharp contrast with neural networks trained with supervised learning

rules, such as the back-propagation algorithm, wherein it has been shown many times that the non-linear conductance response of PCM and RRAM technologies is a critical drawback for the implementation of electronic synapses based on these technologies.

The results presented in CHAPTER 2 highlight that RRAM requirements for neuromorphic cores heavily differ from those of standard memory arrays. Yet RRAM requirements for RRAM-based synapses remain application-dependent, and a systematic study is mandatory for each application. Therefore, a better understanding of RRAM physics is still needed to further optimise RRAM programming conditions for neuromorphic applications and to benefit as much as possible from RRAM physical and electrical properties. Finally, we only considered in this work RRAM-based synapses implemented with the simple 1T1R structure, and we did not address STDP learning circuitry. Future works must find efficient designs for synaptic elements associated with appropriate learning circuitry in order to enable real-time on-line learning for neuromorphic cores.

In CHAPTER 3, we studied the implementation of RRAM-based Ternary Content-Addressable Memories (TCAMs) to realise synaptic routing lookup tables for network topology reconfigurability. This has been performed by means of extensive electrical characterisations measured on multi-bits (3x128 bits) RRAM-based TCAM arrays. We investigated two different RRAM-based TCAM structures. Both were made of two NMOS transistors and two RRAM cells which currently allows for the smallest TCAM bitcell size. The first TCAM structure is the most common 2T2R TCAM composed of two one-transistor/one-RRAM structures in parallel. We experimentally proved that this structure is heavily constrained by the rather low RRAM Memory Window (MW) and high RRAM resistance variability. These two issues degrade the sensing margin of 2T2R TCAM arrays and limit the maximum number of bits per TCAM word. In order to use 2T2R TCAM for classic applications, such as Internet Protocol (IP) packet routing, RRAMs need to be programmed with stronger programming conditions than those of used for standard memory applications. This allows for a programming endurance of 10^4 cycles and a search endurance higher than 10^6 search operations. While the programming endurance may be sufficient for classic TCAM applications, the search endurance is probably too low. For synaptic routing tables in neuromorphic cores, neuron addresses are usually short in size (below 32 bits). This allows to use standard RRAM programming conditions featuring a programming endurance of 10^6 cycles which is sufficient for neuromorphic applications. The search endurance is enough for simple neuromorphic applications, but it still needs to be improved. We then proposed a new RRAM-based TCAM structure composed of two NMOS transistors and two RRAM cells in a 1T2R1T structure. One NMOS transistor (N1) is involved in RRAM programming operations, while the second NMOS transistor (N2) is used during search operations. Search operations rely on a voltage divider between the two RRAM devices. During search operations, the transistor N2 is turned ON by the RRAM voltage divider (mismatch case) or kept OFF (match case). This structure has the significant advantage of decoupling RRAM electrical characteristics from TCAM performance and reliability. As expected, we experimentally demonstrated that this structure is insensitive to

RRAM memory window and resistance variability, and it can be integrated in long TCAM words ($>2\text{kbits}$) even with standard programming conditions (programming endurance of 10^6 cycles). In addition, this TCAM structure also improves search endurance ($>10^7$ search operations) with respect to the common 2T2R structure because the search voltage is applied on RRAMs only during the search operation. Therefore, we argued that this new 1T2R1T TCAM structure is suitable for both classic TCAM applications, like IP packet routing - thanks to the possibility of searching long TCAM words - and neuromorphic cores - owing to the improved search endurance. Moreover, the new 1T2R1T TCAM structure relaxes design constraints. Transistors N2 of each TCAM bitcell can be implemented with thin oxide CMOS transistors since these transistors are involved only in search operations, unlike the 2T2R TCAM where both transistors are involved in RRAM programming operations. This drastically improves TCAM performance.

Finally, in CHAPTER 4, we concluded the work presented in this dissertation by presenting a proof-of-concept of three-dimensional sequential (3D monolithic) integration of two levels of CMOS transistors with a level of RRAM devices. We showed the possibility to integrate two 1T1R structures in parallel with the two NMOS transistors and RRAM devices vertically stacked on top of each other. We experimentally demonstrated the functionality of the integration at the device-level. The results obtained in this chapter open up technological perspectives to further improve the works presented in the two previous chapter - the implementation of RRAM-based synapses and RRAM-based synaptic routing tables. In the case of RRAM-based synapses, 3D monolithic integration can roughly double synaptic density. In addition, future works on 3D monolithic integration will enable the fabrication of RRAMs directly in between levels of CMOS. This will allow to integrate synaptic elements in between CMOS-based neuron circuits in a truly brain-like architecture. Yet this needs to be fabricated and tested. In the case of RRAM-based synaptic routing tables, 3D monolithic integration makes possible to decrease TCAM bitcell size. This is beneficial as it decreases silicon area consumption as well as improves TCAM performance and search endurance - search times decrease with bitcell size which improves performance and search endurance. In the future, electrical measurements of the integration at the array-level are necessary to fully validate the functionality of the concept. Furthermore, there is still a serious need for a thorough and clear evaluation of the advantages and drawbacks of 3D monolithic integration over more conventional integrations, such as planar or 3D parallel integrations. In particular, since each tier fabricated with 3D monolithic integration may differ in terms of performance, it is essential to optimise the partitioning method.

To summarise, future works should cover the following subjects:

- Better understanding of RRAM physics to benefit as much as possible from RRAMs electrical properties in both standard memory applications and neuromorphic cores.
- Identification of efficient designs of synaptic arrays associated with appropriate learning circuitry for scalable hardware implementation and real-time on-line learning.

-
- Development of a suitable RRAM programming scheme and circuitry for the 1T2R1T TCAM structure in order to facilitate its integration in real circuits.
 - Thorough evaluation of the opportunities offered by 3D monolithic integration and related costs to fully benefit from the integration at a circuit-level.
 - Identification of efficient designs for neuron circuits since current CMOS-based neurons still consume substantial chip area in neuromorphic cores. This issue has not been discussed in this work, however preliminary works from our group have been initiated to address this problem [5].

References: Conclusion

- [1] Tobi Delbruck. “Frame-free dynamic digital vision”. In *Intl. Symp. on Secure-Life Electronics, Advanced Electronics for Quality Life and Society*, pages 21–26, 2008. doi: <http://dx.doi.org/10.5167/uzh-17620>.
- [2] Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. “Gradient-based learning applied to document recognition”. *Proceedings of the IEEE*, 86(11):2278–2323, 1998. ISSN 00189219. doi: 10.1109/5.726791.
- [3] Daniele Garbin, Elisa Vianello, et al. “HfO₂-Based OxRAM Devices as Synapses for Convolutional Neural Networks”. *IEEE Transactions on Electron Devices*, 62(8):2494–2501, 2015. ISSN 00189383. doi: 10.1109/TED.2015.2440102.
- [4] Selina La Barbera, Denys R.B. Ly, et al. “Narrow Heater Bottom Electrode-Based Phase Change Memory as a Bidirectional Artificial Synapse”. *Advanced Electronic Materials*, 4(9):1800223, 2018. ISSN 2199160X. doi: 10.1002/aelm.201800223.
- [5] Thomas Dalgaty, Melika Payvand, et al. “Hybrid neuromorphic circuits exploiting non-conventional properties of RRAM for massively parallel local plasticity mechanisms”. *APL Materials*, 7(8), 2019. ISSN 2166532X. doi: 10.1063/1.5108663.

Author's publications

- **D.R.B. Ly**, J-P. Noel, B. Giraud, P. Royer, E. Esmanhotto, N. Castellani, T. Dalgaty, J-F. Nodin, C. Fenouillet-Beranger, E. Nowak, and E. Vianello. "Novel 1T2R1T RRAM-based Ternary Content Addressable Memory for Large Scale Pattern Recognition". In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 35.5.1-35.5.4, 2019.
- T. Dalgaty, M. Payvand, F. Moro, **D.R.B. Ly**, F. Pebay-Peyroula, J. Casas, G. Indiveri, and E. Vianello. "Hybrid neuromorphic circuits exploiting non-conventional properties of RRAM for massively parallel local plasticity mechanisms". In *APL Materials*, 7(8):081125, 2019.
- E. Vianello, **D.R.B Ly**, S. La Barbera, T. Dalgaty, N. Castellani, G. Navarro, G. Bourgeois, A. Valentian, E. Nowak, and D. Querlioz. "Metal Oxide Resistive Memory (OxRAM) and Phase Change Memory (PCM) as Artificial Synapses in Spiking Neural Networks". In *IEEE International Conference on Electronics Circuits and Systems, ICECS*, pages 561-564, 2018.
- **D.R.B. Ly**, B. Giraud, J-P. Noel, A. Grossi, N. Castellani, G. Sassine, J-F. Nodin, G. Molas, C. Fenouillet-Beranger, G. Indiveri, E. Nowak, and E. Vianello. "In-depth Characterization of Resistive Memory-Based Ternary Content Addressable Memories". In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 20.3.1-20.3.4, 2018.
- L. Anghel, **D.R.B. Ly**, G. Di Natale, B. Miramond, E. I. Vatajelu, and E. Vianello. "Neuromorphic Computing - From Robust Hardware Architectures to Testing Strategies". In *IEEE/IFIP International Conference on VLSI and System-on-Chip, VLSI-SoC*, pages 176-179, 2018.
- **D.R.B. Ly**, A. Grossi, C. Fenouillet-Beranger, E. Nowak, D. Querlioz, and E. Vianello. "Role of synaptic variability in resistive memory-based spiking neural networks with unsupervised learning". In *Journal of Physics D: Applied Physics*, 51(44):444002, 2018.
- T. Dalgaty, E. Vianello, **D.R.B. Ly**, G. Indiveri, E. Nowak, and J. Casas. "Insect-inspired elementary motion detection embracing resistive memory and spiking neural networks". In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 115-128, 2018.

- S. La Barbera, **D.R.B. Ly**, G. Navarro, N. Castellani, O. Cueto, G. Bourgeois, B. De Salvo, E. Nowak, D. Querlioz, and E. Vianello. "Narrow Heater Bottom Electrode-Based Phase Change Memory as a Bidirectional Artificial Synapse". In *Advanced Electronic Materials*, 4(9):1800223, 2018.
- **D.R.B. Ly**, A. Grossi, T. Werner, T. Dalgaty, C. Fenouillet-Beranger, E. Vianello, and E. Nowak. "Role of synaptic variability in spike-based neuromorphic circuits with unsupervised learning". In *Proceedings - IEEE International Symposium on Circuits and Systems, ISCAS*, pages 1-5, 2018.

Appendices



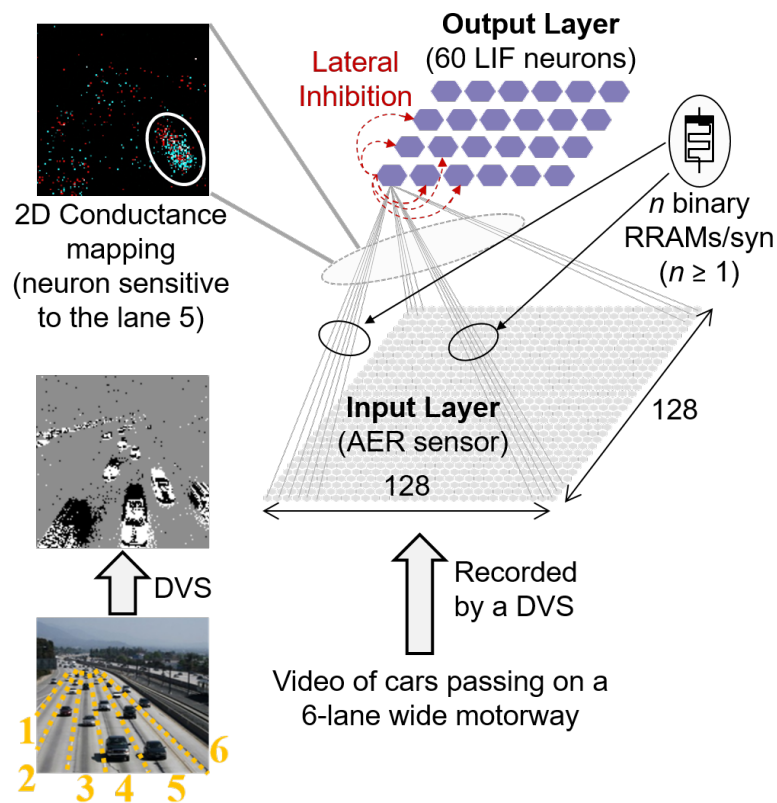
Impact of resistive memory-based synapses on spiking neural network performance: Network topology

IN CHAPTER 2, we studied the impact of binary and analog Resistive Memories (RRAMs) on Spiking Neural Network (SNN) performance with RRAM-based synaptic elements (SECTION 2.2 and 2.3, respectively). To this end, we simulated two applications: (i) a detection application based on a car tracking task [1], and (ii) a character classification based on the handwritten digit dataset MNIST [2]. Both applications rely on a one-layer fully-connected feed-forward neural network topology: each input neuron is connected to each output neuron with a synaptic element. This appendix describes the SNN topology simulated for each application.

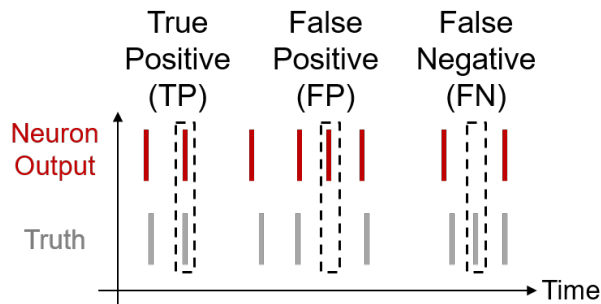
A.1 Network topology with binary devices

A.1.1 Car tracking

FIGURE A.1.1 (a) presents the network simulated for the detection task. A video of cars passing on a six-lanes wide motorway is recorded using the Address Event Representation (AER) [3, 4] format by a Dynamic Vision Sensor (DVS) with 128x128 pixels [5], and it represents the input data [1]. An input pixel generates a spike each time there is a change of luminosity at its location in the visual field. Each input pixel is connected with two synapses to every output neuron to denote an increase (ON synapse) or decrease (OFF synapse) in illumination, respectively. The input layer is composed of $128 \times 128 = 16\,384$ input neurons, and the output layer is composed of 60 output neurons. A similar network has



(a)



$$F1 = \frac{2TP}{2TP + FP + FN}$$

(b)

FIGURE A.1.1: (a) Simulated spiking neural network for the car tracking application with binary devices, trained with an unsupervised stochastic Spike-Timing-Dependent Plasticity (STDP) learning rule and lateral inhibition. (b) Example of spiking activity of one output neuron (red) and the actual traffic (a grey spike corresponds to a car passing on the lane). True Positive (TP) events, False Positive (FP) events, and False Negative (FN) events are put in evidence. The F1-score is used to assess network performance.

been implemented in [6] and [7] exploiting multi-level phase-change memory and binary conductive-bridge RAM synapses, respectively. In this work, we adopted the binary RRAM technology presented in CHAPTER 2 (SECTION 2.2.1),

and synaptic elements are implemented with the synaptic compound of [8] - a synapse is composed of n RRAM devices operating in parallel. The total number of RRAM devices is $128 \times 128 \times 2 \times 60 \times n = 1\,966\,080n$, where n is the number of RRAM cells per synapse. Output neurons are implemented with the Leaky Integrate-and-Fire (LIF) model [9, 10], with a leak time constant $\tau_{\text{leak}} = 105.5$ ms. Note that after an output neuron fires a spike, it cannot integrate any incoming spikes for a refractory period $t_{\text{refrac}} = 218$ ms. It also prevents all the other neurons of the layer from integrating incoming spikes for a period $t_{\text{inhibit}} = 29.9$ ms, referred to as *lateral inhibition*. These parameters have been obtained by a genetic algorithm.

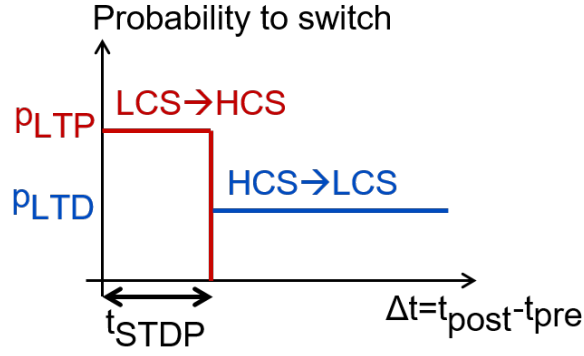
The network is trained with the unsupervised stochastic Spike-Timing-Dependent Plasticity (STDP) rule presented in CHAPTER 2 (SECTION 2.2.1.2), with $p_{\text{LTP}} = 0.11$, $p_{\text{LTD}} = 0.20$, and a STDP time window $t_{\text{STDP}} = 16.0$ ms (see FIGURE A.1.2). TABLE A.1 summarises the main parameters of the network. After a training phase, every output neuron becomes sensitive to a specific lane. An example of the 2D conductance mapping of one output neuron after training is shown in the top left of FIGURE A.1.1 (a). A potentiated ON synapse (resp. OFF synapse) of an input pixel is represented by a red (resp. blue) dot. When both ON and OFF synapses are potentiated, the resulting color is grey. When both ON and OFF synapses are depressed, the resulting color is black. As a result of the training phase, we can observe a pool of potentiated synapses (circled in white) denoting the sensitivity of this neuron to a car at this specific position on the motorway: when a car passes at that position, the neuron spikes. In this example, the output neuron is sensitive to the lane 5; the neuron spikes whenever a car passes on that lane. Figure A.1.1 (b) sketches the spiking activity of one output neuron (red) and the actual traffic (a grey spike corresponds to a car passing on the lane). If the neuron detects a car, we have a True Positive (TP) event. If it spikes with no car passing, we have a False Positive (FP) event. If it misses a car, we have a False Negative (FN) event. We use the F1-score as a metric to assess network performance:

$$F1 = \frac{2TP}{2TP + FN + FP} \quad (\text{A.1.1})$$

F1 ranges from 0 to 1 with $F1 = 1$ being the best performance. Each output neuron becomes sensitive to one lane. Since there are 60 output neurons and only 6 lanes, several neurons become sensitive to the same lane. As more cars pass on the lanes 4 and 5, more neurons are sensitive to these lanes than to the lane 6, the least active lane. To assess network performance, *only the most sensitive neuron for each lane is considered*.

A.1.2 Digit classification

FIGURE A.1.3 (a) presents the network simulated for the classification task. The Mixed National Institute of Standards and Technology (MNIST) dataset is used for the training and testing, with 60 000 training digits and 10 000 testing digits [2]. Each digit is composed of 28×28 pixels. The input layer converts the input digit with a spike frequency encoding: each input neuron generates a spike train with a spiking rate f_{input} proportional to the grey level of the corresponding



(a)

FIGURE A.1.2: Stochastic Spike-Timing-Dependent Plasticity (STDP) learning rule. If the post-synaptic neuron spikes after the pre-synaptic neuron within a time window t_{STDP} (the STDP time window), the synapse undergoes a potentiation event. Otherwise, it undergoes a depression event. At each potentiation (resp. depression) event, each RRAM device has a probability p_{LTP} (resp. p_{LTD}) to switch to the High Conductance State, HCS (resp. Low Conductance State, LCS).

input pixel. f_{input} ranges from $f_{\text{MIN}}=83$ Hz to $f_{\text{MAX}}=22.2$ kHz, with a total of 256 different grey levels. Each input digit is presented to the network for $350 \mu\text{s}$ during the training phase. The input layer is composed of 28×28 input neurons, the output layer is composed of 500 LIF output neurons with a leak time constant $\tau_{\text{leak}}=120.0 \mu\text{s}$. Synaptic elements are implemented with n RRAMs in parallel. The total number of RRAM devices is $28 \times 28 \times 500 \times n = 392\,000n$. The network is trained with the unsupervised stochastic STDP rule (see FIGURE A.1.2), with $p_{\text{LTP}}=0.010$, $p_{\text{LTD}}=0.020$, a STDP time window $t_{\text{STDP}}=60.0 \mu\text{s}$, $t_{\text{refrac}}=1$ ns, and lateral inhibition with $t_{\text{inhibit}}=10 \mu\text{s}$. TABLE A.1 summarises the main parameters of the network.

During the training phase, each output neuron becomes sensitive to a specific class of digit, for example the output neuron 94 becomes sensitive to the class of digit ‘8’ as illustrated in the 2D conductance mapping of FIGURE A.1.3 (a). After training, each output neuron is associated with the digit it is the most sensitive to - this represents the class of the neuron. To assess network performance during the testing phase, the Classification Rate (CR) is computed as shown in FIGURE A.1.3 (b). Each input digit is presented to the network for $350 \mu\text{s}$, and the output neuron that spikes the most within this time window - the most active neuron - corresponds to the network response. If the class of this most active neuron coincides with the input digit, the digit is successfully classified (green spikes in FIGURE A.1.3 (b)). If its class is different from the input digit, the digit is not classified (red spikes in FIGURE A.1.3 (b)). The CR is calculated as the ratio between the number of successfully classified digits, $n_{\text{classified}}$, and the number of input digits, n_{input} :

$$\text{CR} = \frac{n_{\text{classified}}}{n_{\text{input}}} \quad (\text{A.1.2})$$

As there are multiple ways to handwrite the same digit, increasing the number of output neurons allows for an improvement of network performance as

	Car Tracking	Digit Classification
τ_{leak}	105.5 ms	120.0 μs
t_{refrac}	218 ms	1 ns
t_{inhibit}	29.9 ms	10 μs
t_{STDP}	16.0 ms	60.0 μs
PLTP	0.11	0.010
PLTD	0.20	0.020

TABLE A.1: Spiking neural network parameters used for the simulations of the car tracking and digit classification applications with binary devices. All the parameters have been obtained with a genetic algorithm.

demonstrated in [10]. Indeed, this enables the network to have at its disposal several neurons specialised to the same digit, and more precisely to have neurons specialised in different handwritings of the same digit. As shown in [10], the increase of CR with the number of output neurons saturates after 500 output neurons.

A.2 Network topology with analog devices

The impact of conductance response of analog devices on SNN performance has been assessed in CHAPTER 2 - 2.3. To this end, the MNIST dataset has been simulated with a similar network as the one presented in the previous section. FIGURE A.2.1 (a) shows the network topology simulated for this study. Each synaptic element is implemented with one analog device, and the number of output neurons, n_{neurons} , was varied ($n_{\text{neurons}}=50, 100, 300, \text{ and } 500$). The network is trained with the simplified STPD learning rule presented in CHAPTER 2 (SECTION 2.3.3, see FIGURE A.2.1 (b)): at each potentiation (resp. depression) event, the synaptic weight increases (resp. decreases) by a quantity δw_+ (resp. δw_-). Note that with this learning rule, there is no probability p_{LTP} and p_{LTD} . Parameters τ_{leak} , t_{refrac} , t_{inhibit} , and t_{STDP} are the same as previously (see TABLE A.1).

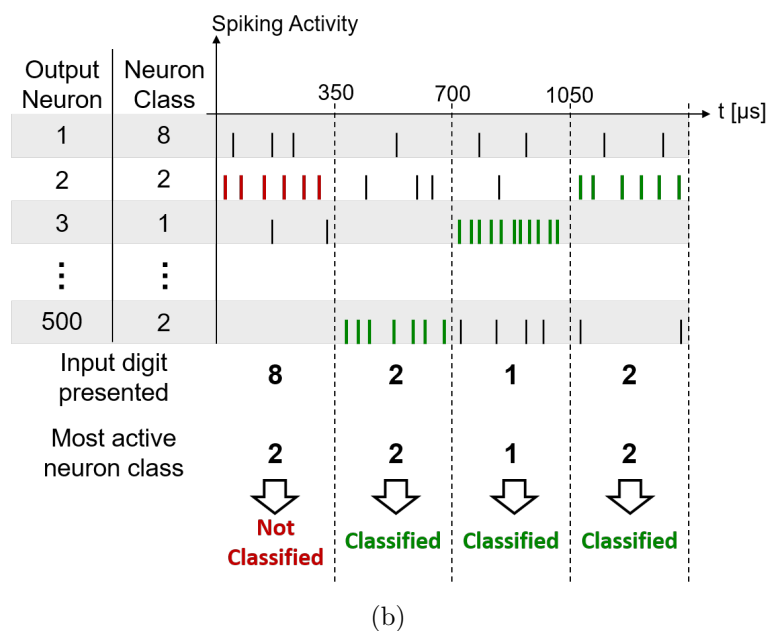
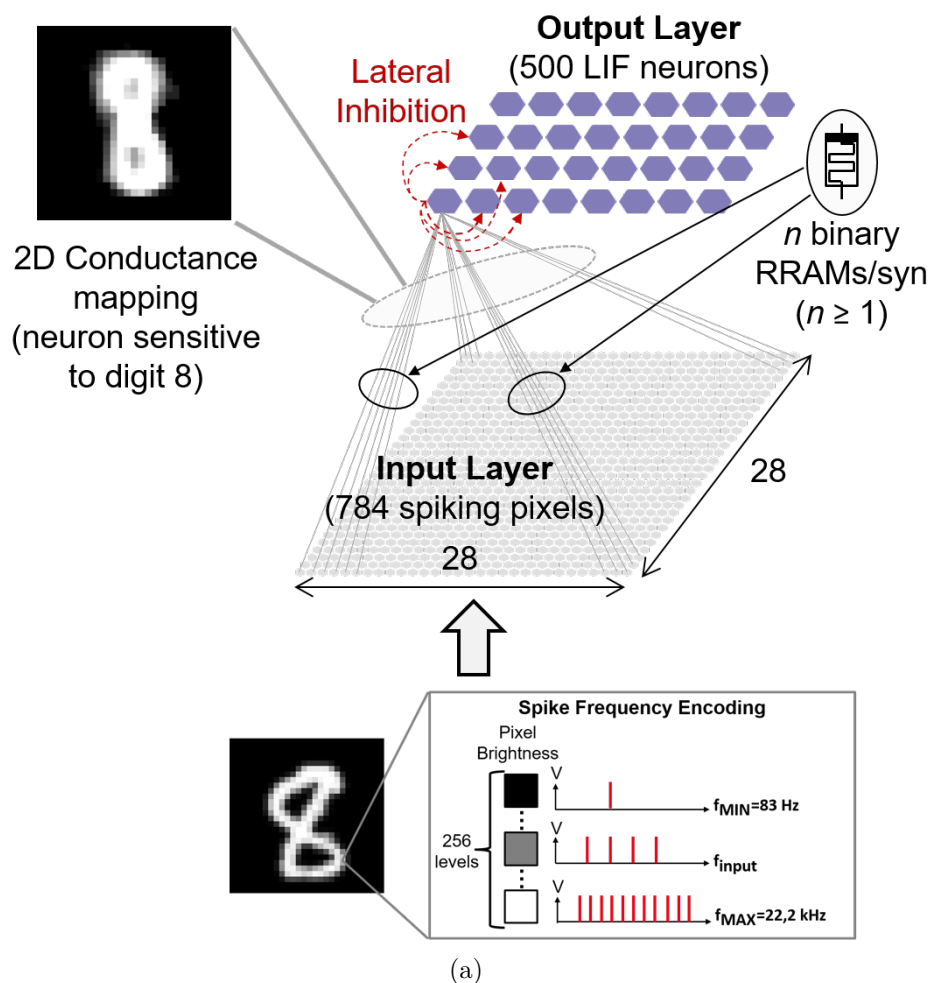


FIGURE A.1.3: (a) Simulated spiking neural network for the digit classification application with binary devices, trained with a stochastic STDP learning rule and lateral inhibition. (b) Example of spiking activity of four output neurons when four different input digits are presented. If the class of the most active neuron corresponds to the input digit, the digit is successfully classified (green), otherwise the digit is not classified (red).

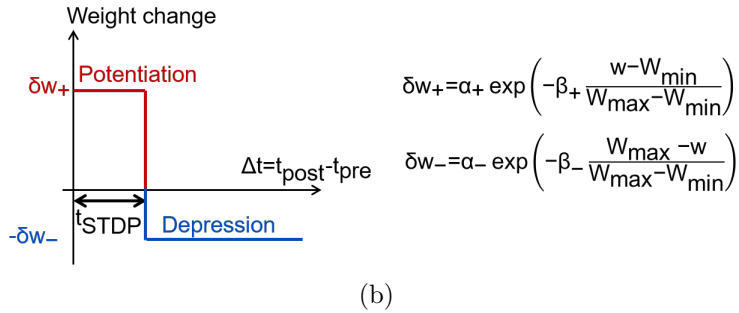
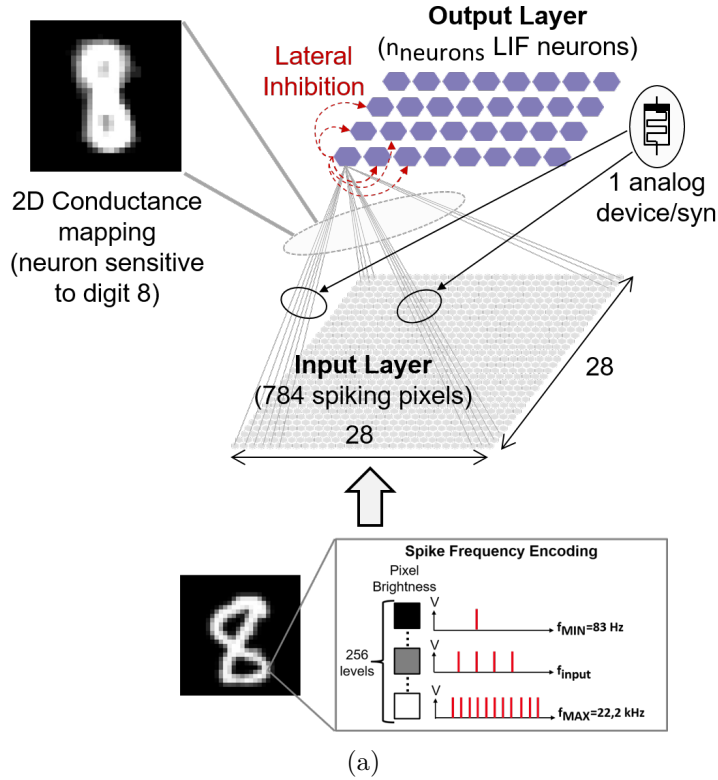


FIGURE A.2.1: (a) Simulated spiking neural network for the digit classification application with analog devices, trained with a simplified Spike-Timing-Dependent Plasticity (STDP) learning rule and lateral inhibition. (b) Simplified STDP learning rule. If the post-synaptic neuron spikes after the pre-synaptic neuron within a time window t_{STDP} (the STDP time window), the synapse undergoes a potentiation event. Otherwise, it undergoes a depression event. At each potentiation (resp. depression) event, the synaptic weight increases by a quantity δw_+ (resp. δw_-). α_+ , α_- , β_+ , β_- , W_{MIN} , and W_{MAX} are fitting parameters of the conductance response of synaptic elements.

Impact of leaky integrate-and-fire neuron threshold value on spiking neural network performance

IN CHAPTER 2, we studied the impact of Resistive Memories (RRAMs) electrical properties on Spiking Neural Network (SNNs) performance with RRAM-based synaptic elements. To this end, we simulated two applications: (i) a detection application based on a car tracking task [1], and (ii) a character classification based on the handwritten digit dataset MNIST [2]. Both applications rely on a one-layer fully-connected feed-forward neural network topology: each input neuron is connected to each output neuron with a synaptic element (*cf* APPENDIX A). Output neurons are implemented with the Leaky Integrate-and-Fire (LIF) model [9, 10]: output LIF neurons integrate input currents coming from input synapses, and they emit a spike when the integration value reaches a certain *firing threshold value*, I_{th} . In this work, *all the output neurons have the same firing threshold value, I_{th} , for a given simulation which is kept constant throughout the simulation*. In order to maximise network performance after learning, we observed that *I_{th} has to be carefully tuned for each simulation*. The optimised I_{th} value mostly depends on synaptic parameters, in particular synaptic dynamic range. In this appendix, we will show the dependency of network performance on LIF neuron firing threshold value. Note that a possible solution to be less dependent on I_{th} would be to implement homeostasis algorithms [10, 11]. Homeostasis allows for continuous adaptation of the threshold value for each individual neuron of the network during the learning phase. This prevents output neurons from over-firing or under-firing by decreasing or increasing their firing threshold value, respectively. This has been proven to improve network performance at the cost of increased circuit complexity. In this work, *we did not implement homeostasis, and we kept neuron threshold value constant during the learning phase*.

B.1 Car tracking

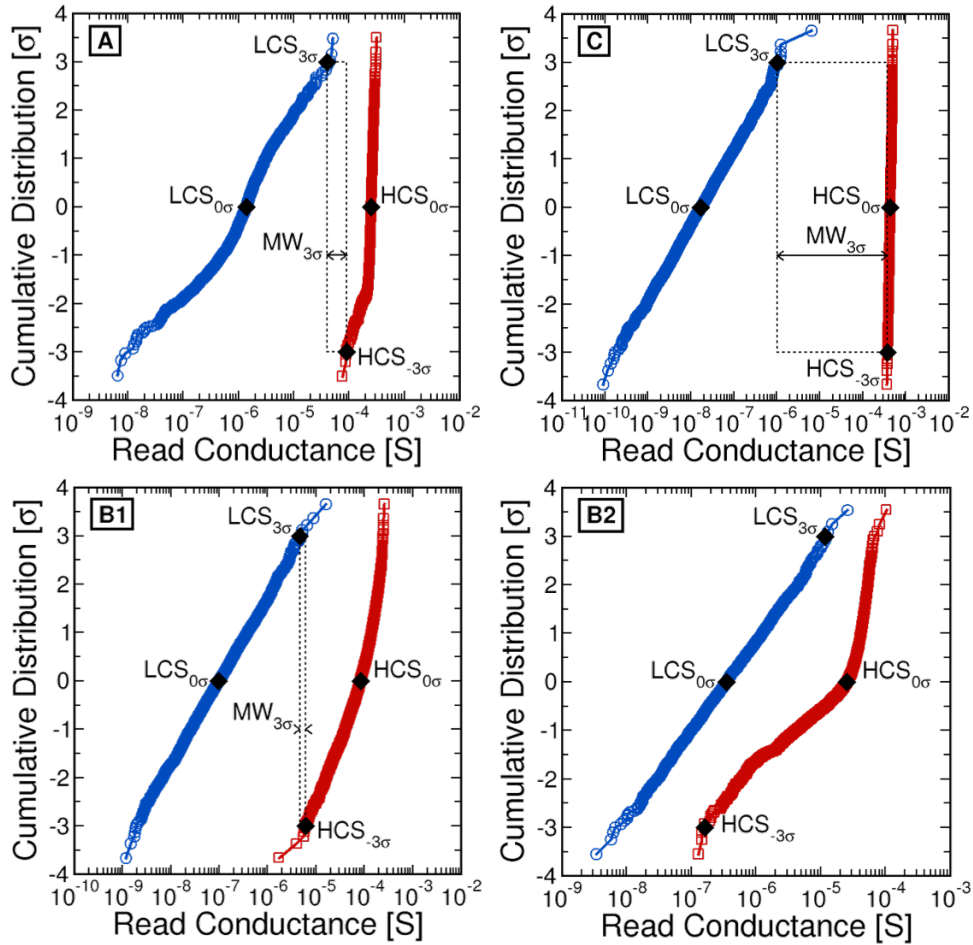


FIGURE B.1.1: Cumulative distributions of the Low Conductance State (LCS) and High Conductance State (HCS) distributions measured on 4-kbit RRAM arrays (Top left) after 1000 switching cycles with condition A, (Top right) with condition C, (Bottom left) with condition B1, and (Bottom right) with condition B2 (see CHAPTER 2 - SECTION 2.2).

In CHAPTER 2, we studied the impact of binary Resistive Memory (RRAM) electrical properties on Spiking Neural Network (SNN) performance with RRAM-based synapses trained with the unsupervised Spike-Timing-Dependent Plasticity (STDP) rule (*cf* CHAPTER 2 - SECTION 2.2). We first simulated a detection application based on a car tracking task [1]. The simulated SNN is provided in APPENDIX A (SECTION A.1.1), and the F1-score is used to assess network performance (F1 ranges from 0 to 1, with 1 being the best detection score). Synaptic elements are implemented with n binary RRAMs operating in parallel. Here, we consider the case $n=1$. RRAM model is calibrated on experimental results of 4-kbit RRAM arrays measured with four different programming conditions (*cf* FIGURE B.1.1, conditions A, B1, B2, and C). FIGURE B.1.2 (a) shows the impact of the firing threshold value, I_{th} , on F1 when the network is calibrated with one of the four RRAM programming conditions. Each result has been averaged over twenty simulations, and error bars represent the deviation at $\pm 1\sigma$. For all the simulations, we used the same network parameters (*cf*

TABLE A.1 in APPENDIX A). Only RRAM programming conditions and firing threshold value, I_{th} , differ (with the same I_{th} for every output neuron that is kept constant throughout the simulation). Note that, in a real SNN, an input event corresponds to a voltage spike which is converted to an input current by Ohm’s law, then integrated by output neurons. Therefore, I_{th} represents a current value. However, in our simulations, an input event is represented by a simple binary ‘1’ transmitted along weighted synapses. As a result, I_{th} represents here a conductance value (*i.e.* a current value normalised by the input voltage spike). As evidenced by FIGURE B.1.2 (a), F1 is maximal for a certain optimised I_{th} value, and this optimised value depends on synaptic parameters (*i.e.* RRAM programming conditions). The decrease in F1-score around the optimised firing threshold value, $I_{th,opt}$, can be explained by examining in FIGURE B.1.2 (b) the False Negative (FN, red square) and False Positive (FP, blue square) rates as a function of I_{th} . A FN event corresponds to a car missed by the network - *i.e.* not detected -; a FP event corresponds to a car detected by the network, whereas no car is passing on the motorway (see APPENDIX A). For low I_{th} , output neurons require fewer input spikes to reach their threshold value. Consequently, output neurons spike more often and are more sensitive to background noise (*e.g.* input noise, cars passing on other lanes, ...). This results in an increase in the number of FP events (high FP rate hence lower F1). For high I_{th} , the network is more robust to background noise, hence a lower FP rate. Yet if I_{th} is too high, the network is more likely to miss cars passing on the motorway (higher FN rate hence lower F1). Indeed, weighted input spikes coming from a car passing on the road may no longer be sufficient to make output neurons spike. Therefore, the optimised threshold value, $I_{th,opt}$, generally comes from a trade-off between the FP and FN rates. In FIGURE B.1.3 we plotted the optimised firing threshold value, $I_{th,opt}$, as a function of the mean High Conductance State (HCS) conductance value of each programming condition. For this application, $I_{th,opt}$ is proportional to the mean HCS conductance value.

B.2 Digit classification

The same study has been carried out on a classification application based on handwritten digit classification (MNIST dataset [2]). The simulated SNN is provided in APPENDIX A (SECTION A.1.2), and the Classification Rate (CR) is used to assess network performance. Synaptic elements are implemented with n binary RRAMs operating in parallel. Here, we consider the case $n=20$. FIGURE B.2.1 (a) shows the impact of the firing threshold value, I_{th} , on the CR when the network is calibrated with one of the four RRAM programming conditions (A, B1, B2, and C). Each result has been averaged over twenty simulations, and error bars represent the deviation at $\pm 1\sigma$. As previously, the CR is maximal for a certain optimised firing threshold value, $I_{th,opt}$. FIGURE B.2.1 (b) shows the optimal firing threshold, $I_{th,opt}$, as a function of the mean High Conductance State (HCS) conductance value of each programming condition. $I_{th,opt}$ is proportional to the mean HCS conductance value.

B.3 Impact of firing threshold variability

In this section we investigate the robustness of SNN performance to firing threshold variability. For this purpose, we simulated the car tracking application (*cf* SECTION B.1) with different firing threshold variability values. Synaptic elements are implemented with one binary RRAM device and are calibrated on the experimental programming conditions A (*cf* FIGURE B.1.1 (Top left)). Each output LIF neuron has a different firing threshold value I_{th} . The distribution of I_{th} values follows a *gaussian distribution with a mean value $I_{th,mean}$ and a standard deviation $\sigma(I_{th})$ (representing threshold variability)*. $\sigma(I_{th})$ is expressed here as a percentage of the mean value, $I_{th,mean}$. The particular case $\sigma(I_{th})=0\%$ corresponds to the situation wherein all the output LIF neurons have the same firing threshold value (*i.e.* the situation in the previous section). FIGURE B.3.1 (a) shows the F1-score as a function of firing threshold variability, $\sigma(I_{th})$. Each result has been averaged over twenty simulations, and error bars denote the deviation at $\pm 1\sigma$. The mean firing threshold value, $I_{th,mean}$, has been optimised by simulations for each simulated threshold variability value (*cf* FIGURE B.3.1 (b)). For $\sigma(I_{th})=10\%$, network performance is degraded by only 0.5%. For 35% of threshold variability, network performance is only degraded by 2.3%. We plotted in FIGURE B.3.1 (b) the F1-score as a function of the mean firing threshold value, $I_{th,mean}$, for different threshold variability values, $\sigma(I_{th})$. Each result has been averaged over twenty simulations. For the sake of clarity, we do not show error bars at $\pm 1\sigma$. As highlighted by FIGURE B.3.1 (b), increasing $\sigma(I_{th})$ allows to make the network more robust to a variation of the mean threshold value, $I_{th,mean}$: F1 is less and less sensitive to a variation of the mean threshold value around the optimised threshold value with increasing $\sigma(I_{th})$ (*i.e.* curves are flatter with higher $\sigma(I_{th})$).

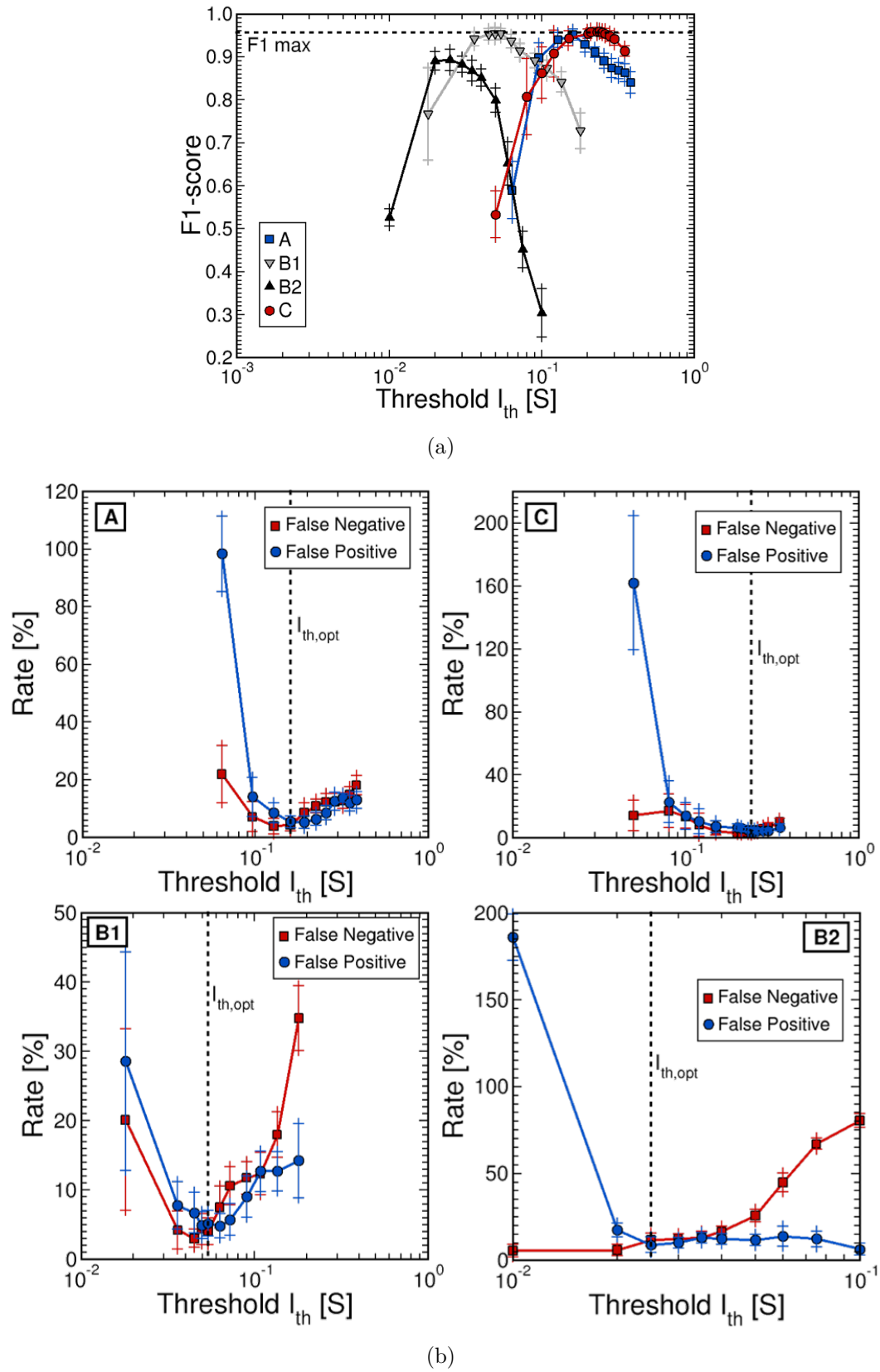


FIGURE B.1.2: (a) F1-score as a function of the firing threshold value, I_{th} , for the four studied RRAM programming conditions (A, B1, B2, and C). (b) False Negative (FN, red square) and False Positive (FP, blue circle) rates as a function of the firing threshold value, I_{th} . The optimised threshold value, $I_{th,opt}$, comes from a trade-off between FN and FP rates.

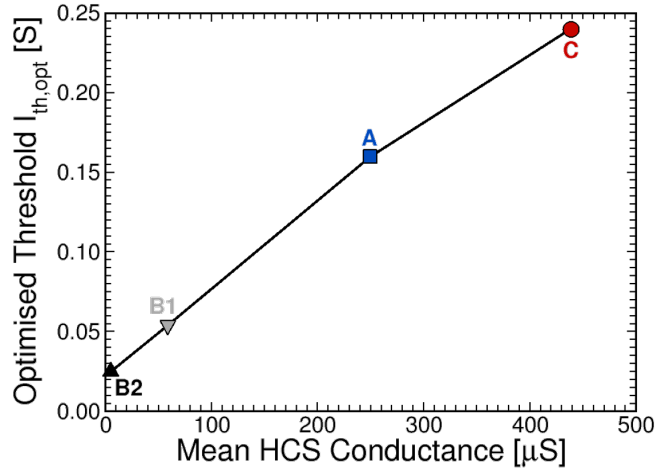
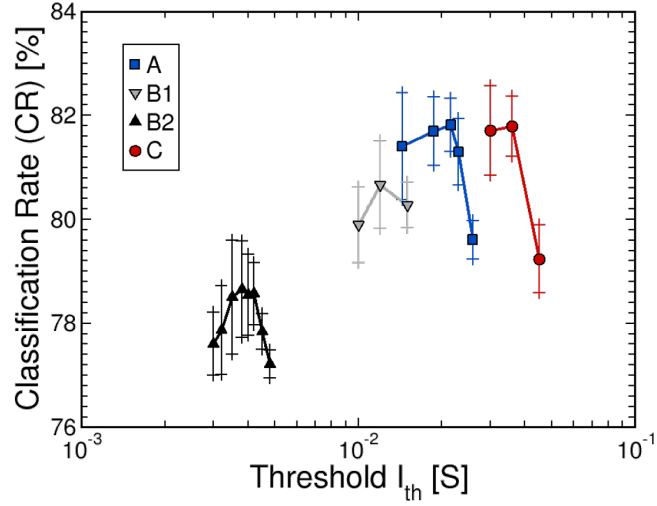
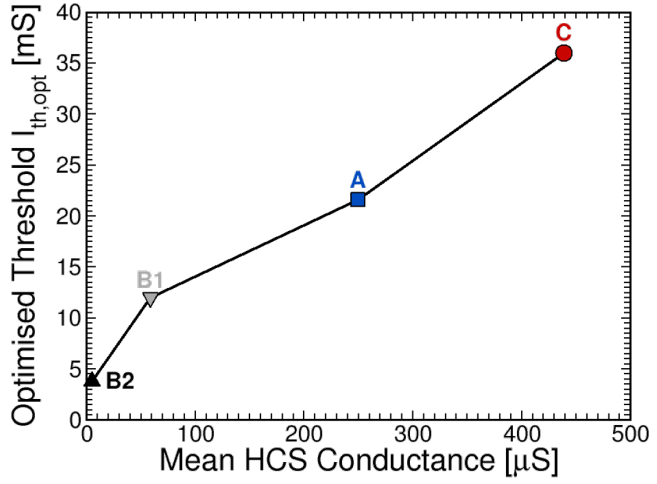


FIGURE B.1.3: Optimised firing threshold value, $I_{th,opt}$, as a function of the mean High Conductance State (HCS) conductance value of each programming condition. $I_{th,opt}$ is proportional to the mean HCS conductance value.



(a)



(b)

FIGURE B.2.1: (a) Classification Rate (CR) as a function of the firing threshold value, I_{th} , for the four studied RRAM programming conditions (A, B1, B2, and C). (b) Optimised firing threshold value, $I_{th,opt}$, as a function of the mean High Conductance State (HCS) conductance value of each programming condition. $I_{th,opt}$ is proportional to the mean HCS conductance value.

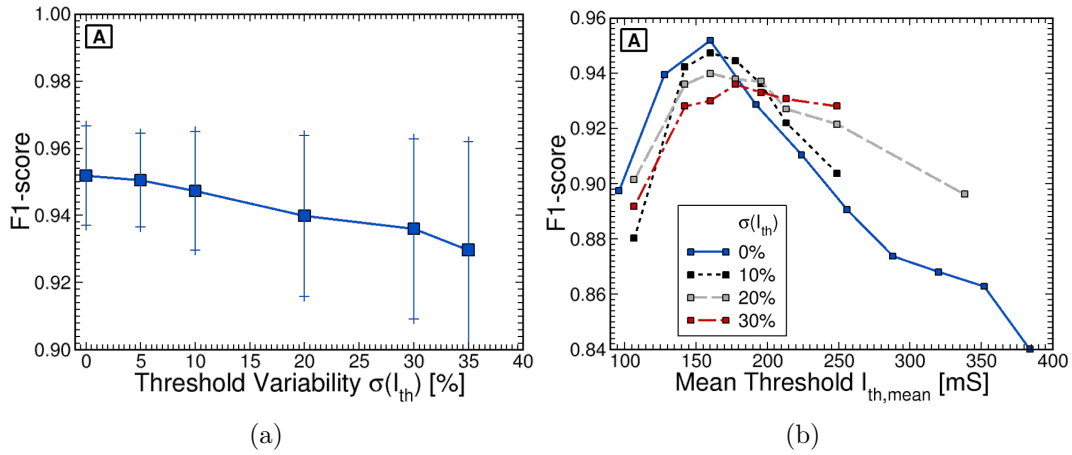


FIGURE B.3.1: (a) F1-score as a function of threshold variability values, $\sigma(I_{th})$. (b) F1-score as a function of the mean firing threshold value, $I_{th,mean}$, for different threshold variability values, $\sigma(I_{th})$. Synaptic elements are implemented with one binary RRAM device calibrated on programming conditions A.

Robustness of spiking neural networks trained with unsupervised learning to input noise

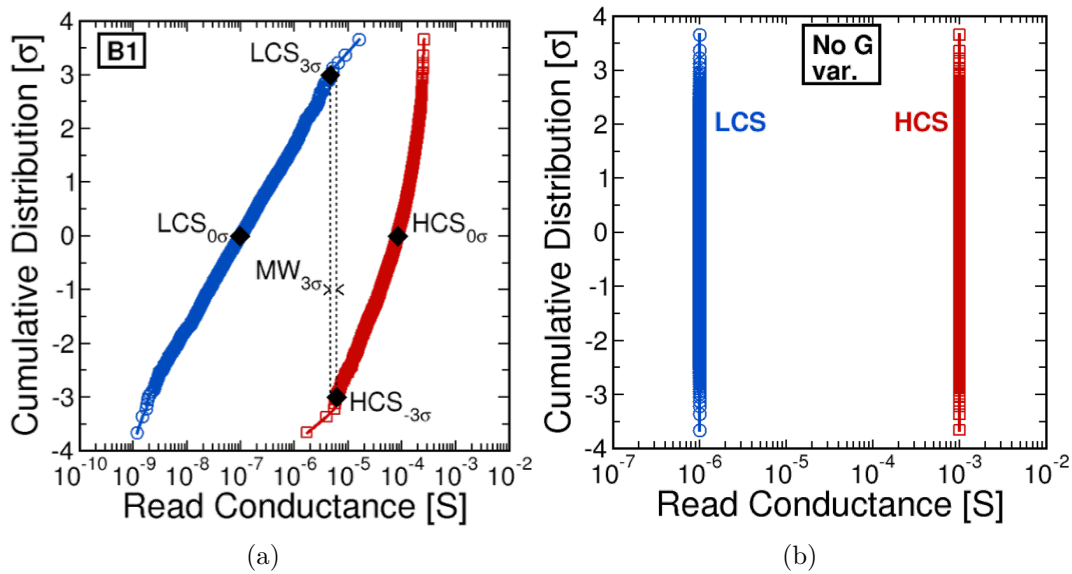


FIGURE C.0.1: Cumulative distributions of High Conductance State (HCS, red square) and Low Conductance State (LCS, blue circle) with (a) programming conditions B1, and (b) an artificial case of a synapse with zero variability.

IN this appendix, we assess the robustness of Spiking Neural Networks (SNNs) with Resistive Memory (RRAM)-based synapses trained with the unsupervised spike-timing-dependent plasticity learning paradigm to input noise. For this purpose, we simulated the detection application based on car tracking [1] (*cf* APPENDIX A - SECTION A.1.1). Synaptic elements are implemented

with n binary RRAMs operating in parallel, with $n=1$. The SNN has first been simulated with RRAMs calibrated on the experimental programming conditions B1 (FIGURE C.0.1 (a), with memory window at 3σ , $MW_{3\sigma}=1.3$, High Conductance State (HCS) variability $\sigma_{G,HCS}=0.28$, and Low Conductance State (LCS) variability $\sigma_{G,LCS}=0.58$), then with an artificial case of a synapse with zero variability (FIGURE C.0.1 (b), with $MW_{3\sigma}=1000$, $\sigma_{G,HCS}=\sigma_{G,LCS}=0$). We simulated the proposed application with a certain amount of input noise corresponding to a certain amount of random input activity. For instance, 20% of input noise means that 20% of input activity is entirely random.

FIGURE C.0.2 shows the F1-score (*cf* APPENDIX A - SECTION A.1.1) as a function of the amount of input noise for (a) the programming conditions B1, and (b) the case of synapse with zero variability. Each result has been averaged over twenty simulations, and the error bars denote the deviation at $\pm 1\sigma$. Output neuron thresholds have been optimised for each result (*cf* APPENDIX B). For each condition, up to 20% of input noise can be tolerated without noticeable degradation of performance (degradation of $\approx 1\%$), and synaptic variability has no impact on SNN robustness to input noise. A solution to improve robustness to noise is to implement synaptic short-term plasticity at the cost of increased circuit complexity [12].

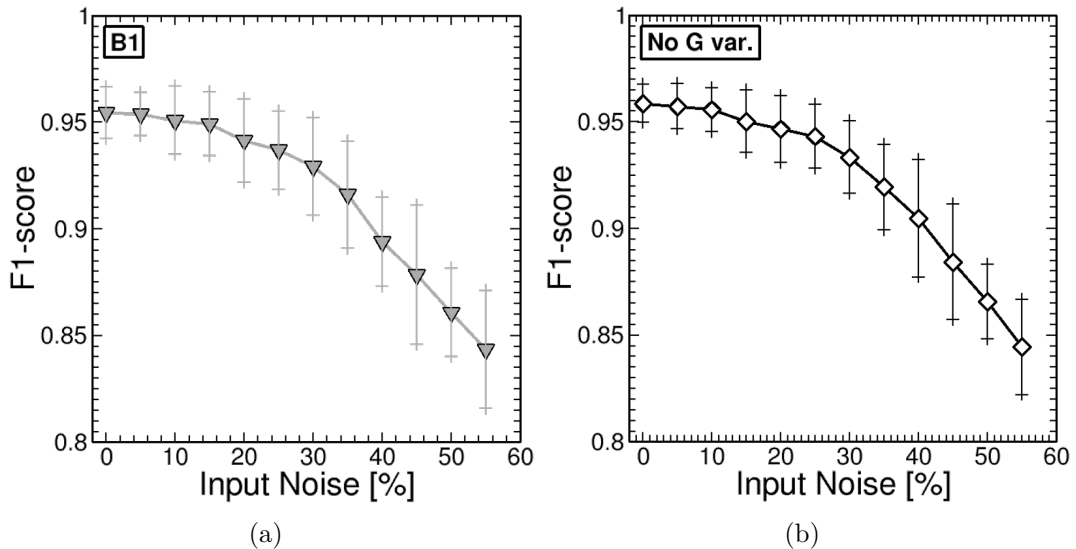


FIGURE C.0.2: F1-score as a function of input noise when synaptic elements are calibrated on (a) the experimental programming conditions B1, and (b) an artificial case of a synapse with zero variability.

References: Appendice

- [1] Tobi Delbruck. “Frame-free dynamic digital vision”. In *Intl. Symp. on Secure-Life Electronics, Advanced Electronics for Quality Life and Society*, pages 21–26, 2008. doi: <http://dx.doi.org/10.5167/uzh-17620>.
- [2] Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. “Gradient-based learning applied to document recognition”. *Proceedings of the IEEE*, 86(11):2278–2323, 1998. ISSN 00189219. doi: 10.1109/5.726791.
- [3] Steve Deiss, Rodney Douglas, Mike Fischer, Misha Mahowald, and Tony Matthews. “Address-Event Asynchronous Local Broadcast Protocol”, 1994. URL <https://www.ini.uzh.ch/~amw/scx/aeprotocol.html>.
- [4] Kwabena A. Boahen. “Point-to-point connectivity between neuromorphic chips using address events”. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 47(5):416–434, 2000. ISSN 10577130. doi: 10.1109/82.842110.
- [5] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. “A 128 x 128 120 dB 15 us latency asynchronous temporal contrast vision sensor”. *IEEE Journal of Solid-State Circuits*, 43(2):566–576, feb 2008. ISSN 00189200. doi: 10.1109/JSSC.2007.914337.
- [6] M Suri, V Sousa, L Perniola, D Vuillaume, and B DeSalvo. “Phase change memory for synaptic plasticity application in neuromorphic systems”. In *The 2011 International Joint Conference on Neural Networks*, pages 619–624, 2011. doi: 10.1109/IJCNN.2011.6033278.
- [7] M. Suri, D. Querlioz, et al. “Bio-Inspired Stochastic Computing Using Binary CBRAM Synapses”. *IEEE Transactions on Electron Devices*, 60(7):2402–2409, 2013. doi: 10.1109/TED.2013.2263000.
- [8] Daniele Garbin, Elisa Vianello, et al. “HfO₂-Based OxRAM Devices as Synapses for Convolutional Neural Networks”. *IEEE Transactions on Electron Devices*, 62(8):2494–2501, 2015. ISSN 00189383. doi: 10.1109/TED.2015.2440102.
- [9] Olivier Bichler, Damien Querlioz, Simon J. Thorpe, Jean Philippe Bourgoin, and Christian Gamrat. “Unsupervised features extraction from asynchronous silicon retina through spike-timing-dependent plasticity”. In *Proceedings of the International Joint Conference on Neural Networks*, pages 859–866. IEEE, 2011. ISBN 9781457710865. doi: 10.1109/IJCNN.2011.6033311.
- [10] Damien Querlioz, Olivier Bichler, Adrien Francis Vincent, and Christian Gamrat. “Bioinspired Programming of Memory Devices for Implementing an Inference Engine”. *Proceedings of the IEEE*, 103(8):1398–1416, aug 2015. ISSN 00189219. doi: 10.1109/JPROC.2015.2437616.
- [11] Peter U. Diehl and Matthew Cook. “Unsupervised learning of digit recognition using spike-timing-dependent plasticity”. *Frontiers in Computational Neuroscience*, 9(AUGUST), aug 2015. ISSN 16625188. doi: 10.3389/fncom.2015.00099.

- [12] T. Werner, E. Vianello, et al. “Experimental demonstration of short and long term synaptic plasticity using OxRAM multi k-bit arrays for reliable detection in highly noisy input data”. In *IEEE International Electron Devices Meeting (IEDM)*, pages 16.6.1–16.6.4, 2016. ISBN 9781509039029. doi: 10.1109/IEDM.2016.7838433.

Résumé en français

Introduction

1.1 De Von Neumann au calcul neuromorphique

EN 1945, John Von Neumann conçoit l'Electronic Discrete Variable Automatic Computer (EDVAC), un des tout premiers ordinateurs électroniques, et pose les bases de l'ordinateur moderne [1]. Initialement inspirée du fonctionnement des cerveaux biologiques [2], l'architecture de l'EDVAC diverge finalement de celle des cerveaux en raison de contraintes technologiques, et peut être résumée en trois composantes principales : une *unité centrale de traitement* (*CPU pour Central Processing Unit*), la *mémoire*, et un *élément de liaison* entre le CPU et la mémoire [3, 4]. Ce paradigme, souvent nommé *ordinateur Von Neumann* en référence à son co-inventeur, repose principalement sur des échanges séquentiels de données entre le CPU et la mémoire à travers l'élément de liaison [4–6] (*cf* FIGURE 1.1.1 (a)). Cependant, cette séparation physique entre les centres de calcul, le CPU, et la mémoire, ainsi que l'absence de parallélisme [4, 7] dégradent fortement les performances des ordinateurs Von Neumann: c'est ce qu'on appelle communément le *goulot d'étranglement de Von Neumann* ou le *mur de la mémoire* [4]. En effet, les temps de calcul du CPU sont de l'ordre de la nanoseconde, alors que les accès mémoire requièrent plusieurs millisecondes [8, 9], ce qui crée un écart de performance entre le CPU et la mémoire et limite l'efficacité énergétique des systèmes de calcul actuels [5, 9–12]. Généralement, *le processeur attend l'information*. Cet écart de performance est particulièrement visible avec l'arrivée de nouvelles applications centrées sur les données [13–15], telles que l'analytique des données massives ou l'apprentissage automatique par les machines [8], dans lesquelles la majeure partie de la puissance et du temps de calcul est perdue pour échanger les données entre le CPU et la mémoire [11, 16, 17].

À l'inverse, l'architecture des cerveaux biologiques repose sur une architecture massivement parallèle avec une co-localisation entre les centres de calcul, les *neurones*, et la mémoire, les *synapses* [5, 7] (*cf* FIGURE 1.1.1 (b)). De plus, le cerveau humain excelle aux applications cognitives grâce à sa capacité naturelle à faire de l'inférence et de l'apprentissage, pendant laquelle les neurones et synapses s'adaptent et se spécialisent à diverses tâches [15]. Ces considérations ont donné naissance à la fameuse *approche neuromorphique* [18], qui consiste à construire des systèmes de calcul inspirés du cerveau. L'objectif de l'*ingénierie neuromorphique*, ou *calcul neuromorphique*, est de bâtir de nouvelles architectures de calcul qui implémentent des modèles *bio-inspirés des réseaux de neurones*. Ce nouveau paradigme est apparu comme une solution pour résoudre les problèmes inhérents à l'architecture de Von Neumann, et plus récemment les enjeux liés à la fin de la loi de Moore [6, 19–21].

Dans le cadre de cette thèse, nous avons exploré l'*implémentation matérielle de processeurs neuromorphiques impulsionsnels et reconfigurables en exploitant de nouvelles solutions technologiques* : les *mémoires résistives (RRAM)* et les *technologies 3D monolithiques*.

1.2 Les nouvelles solutions technologiques

1.2.1 Les mémoires résistives

Les mémoires résistives (RRAM pour Resistive Random Access Memory) sont des mémoires *non volatiles* composées d'un empilage simple de deux métaux, les électrodes, qui entourent une couche d'*isolant* (*cf* FIGURE 1.2.1). Contrairement aux mémoires plus conventionnelles de type SRAM, DRAM, ou Flash, où l'information est encodée par la présence ou absence de charges électriques, les RRAMs stockent la donnée '0' ou '1' dans leur état de résistivité électrique. Le principe de fonctionnement des RRAM repose sur la formation et dissolution d'un filament conducteur qui relie les deux électrodes, et permet la conduction ou le blocage du courant électrique [23–25]. Selon la nature du filament conducteur, les RRAM peuvent être classifiées comme (i) RRAM à base d'oxydes (OxRAM pour Oxide-based RRAM) dans lesquelles le filament conducteur est composé de lacunes d'oxygène [26], et (ii) RRAM à pont conducteur (CBRAM pour Conductive-Bridge RRAM) dans lesquelles le filament conducteur est composé de cations métalliques [27]. La FIGURE 1.2.1 illustre le principe de fonctionnement d'une OxRAM. En appliquant des tensions de polarité opposée sur l'électrode du haut (pour une OxRAM bipolaire), *la mémoire commute de façon binaire entre un état de faible résistance* (LRS pour Low Resistance State, qui correspond à la donnée '1'), et un *état de haute résistance* (HRS pour High Resistance State, qui correspond à la donnée '0'). Plus précisément, en appliquant une tension positive lors d'une opération de Set, un filament conducteur composé de lacunes d'oxygène relie les deux électrodes, ce qui permet le passage du courant. En appliquant une tension négative lors d'une opération de Reset, les lacunes d'oxygène se recombinent avec des ions oxygènes, ce qui détruit le filament conducteur et bloque le passage du courant. Un cycle de commutation

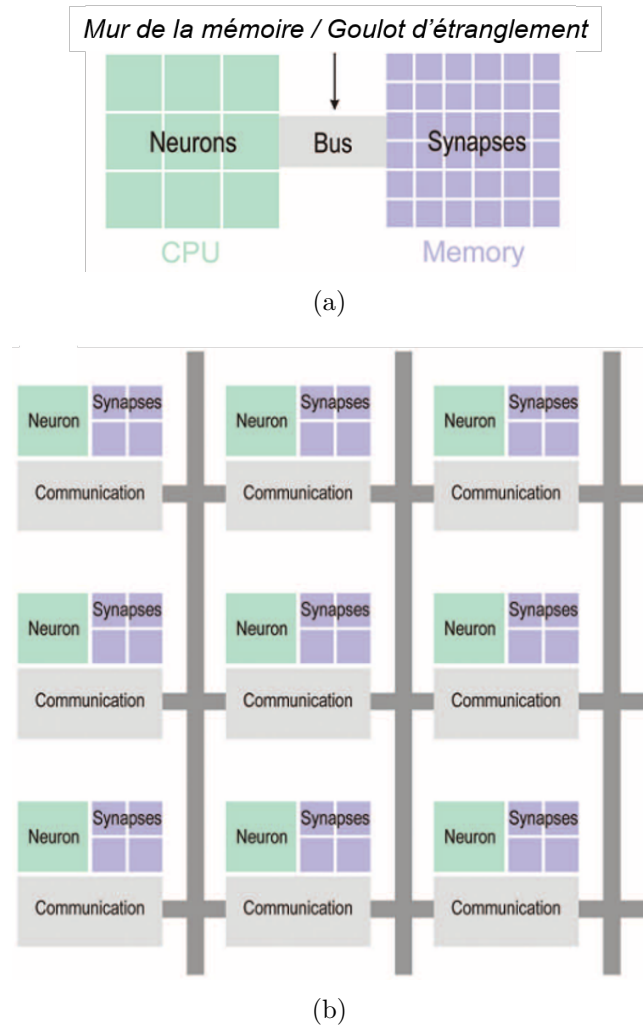


FIGURE 1.1.1: (a) Dans les architectures Von Neumann, les unités de calcul et de mémoire sont physiquement séparées par un bus, créant le fameux goulot d'étranglement de Von Neumann. (b) Schéma conceptuel d'une architecture inspirée du cerveau, où le calcul et la mémoire sont fortement co-localisés. Reproduit de [22].

correspond à une commutation entre le HRS et le LRS, et peut être répété autant de fois que la technologie le permet [28–30]. Le nombre maximal de cycles de commutation permis par une technologie définit son *endurance en programmation*.

Au cours de la dernière décennie, les RRAM ont été considérées comme des candidats potentiels pour remplacer les mémoires Flash, et plus récemment pour implémenter des processeurs neuromorphiques [32–39]. En plus de leur non volatilité, les RRAM offrent une bonne endurance en programmation ($>10^{12}$ [40, 41]), des opérations de lecture non destructrices, une commutation rapide (en-dessous de la nanoseconde [42–44]), un faible courant de programmation grâce à la nature filamentaire de la conduction de courant (de l'ordre de la dizaine de nanoampères [45–49]), et sont facilement miniaturisables (de l'ordre du nanomètre [26, 50]). De plus, leur fabrication est compatible avec le retour en fin de ligne (back-end-of-line) des procédés CMOS, ce qui permet de les fabriquer directement au-dessus des transistors CMOS. Cependant, les RRAM

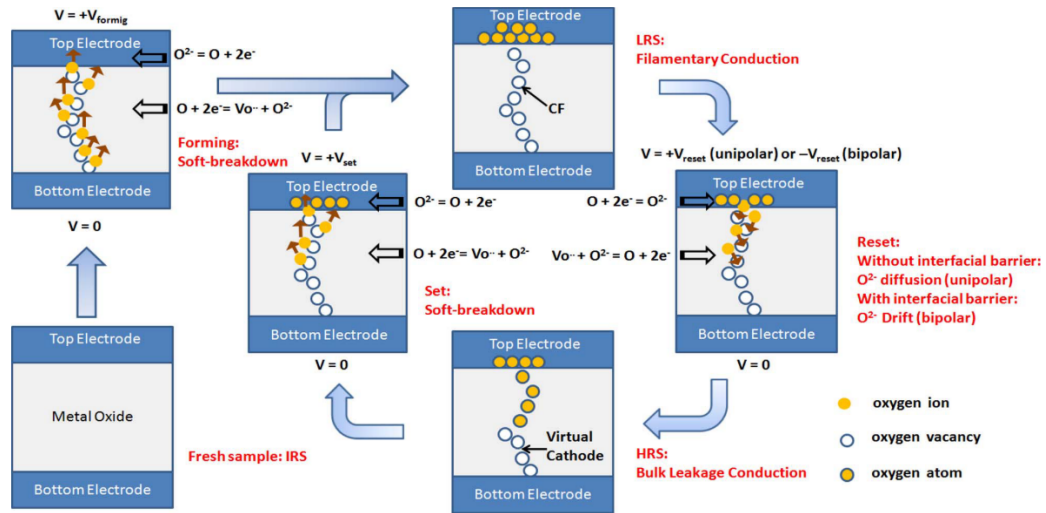


FIGURE 1.2.1: Illustration schématisée du processus de commutation des mémoires résistives à base d'oxydes (OxRAM pour Oxide-based Resistive Random Access Memory). Reproduit de [31].

souffrent de deux problèmes majeurs : (i) un rapport relativement faible entre les valeurs de résistance du HRS et LRS (≈ 10 -100), et (ii) une grande variabilité sur les états de résistance. Comme décrit précédemment, les données '0' et '1' sont stockées dans les états HRS et LRS, plus précisément dans les valeurs de résistance du HRS, R_{HRS} , et du LRS, R_{LRS} . Il est donc fondamental de garantir un rapport de résistance, R_{HRS}/R_{LRS} , suffisamment élevé pour pouvoir distinguer les deux états. Idéalement, ce rapport de résistance, R_{HRS}/R_{LRS} , généralement nommé la *fenêtre mémoire* (MW pour Memory Window), doit être maximisé afin de permettre l'intégration des RRAM dans de grandes matrices mémoires. Cependant, il a été démontré qu'un compromis existe entre la fenêtre mémoire et l'endurance en programmation : une plus grande fenêtre mémoire implique une plus faible endurance en programmation [30, 51–53]. La FIGURE 1.2.2 présente les fenêtres mémoires, MW, de différentes technologies de mémoires résistives associées à leur endurance en programmation. Afin d'assurer une endurance en programmation au moins équivalente à la technologie Flash, *i.e* 10^6 cycles de commutation [11, 54], la *fenêtre mémoire doit être relativement faible, de l'ordre de 10-100*.

Le deuxième problème majeur des RRAM est la *grande variabilité résistive sur leur états de résistance* à cause de la nature stochastique du filament conducteur. D'une part, les RRAM présentent de la variabilité résistive cycle-à-cycle, comme illustrée sur la mesure de cyclage d'une cellule RRAM de la FIGURE 1.2.3 (a). Lorsque la cellule RRAM commute successivement entre ses états HRS (bleu) et LRS (rouge), les valeurs de résistance HRS (bleu) et LRS (rouge) varient de cycle à cycle. D'autre part, les RRAM présentent également de la variabilité résistive cellule-à-cellule, comme représentée sur la distribution de valeurs de résistance d'une matrice RRAM de 4 kbit de la FIGURE 1.2.3 (b). Après une opération de Reset (bleu) et une opération de Set (rouge), les 4 kbit cellules RRAM présentent de la dispersion sur leur valeurs de résistance. Cette variabilité résistive dégrade la fenêtre mémoire des technologies RRAM si on considère le pire cas avec un rapport de résistance à 3σ . Dans le cadre de ce travail de thèse, l'impact de ces

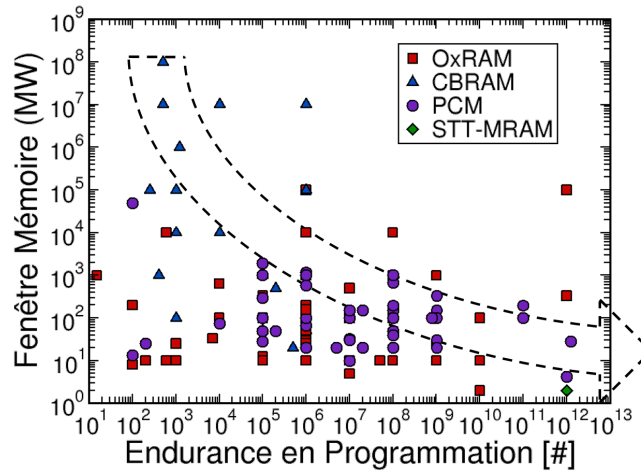


FIGURE 1.2.2: Fenêtre mémoire, MW, en fonction de l'endurance en programmation. Reproduit de [53].

deux problèmes, (i) la faible fenêtre mémoire (≈ 10 -100) et (ii) la forte variabilité résistive, sur les processeurs neuromorphiques impulsionnels sera étudié.

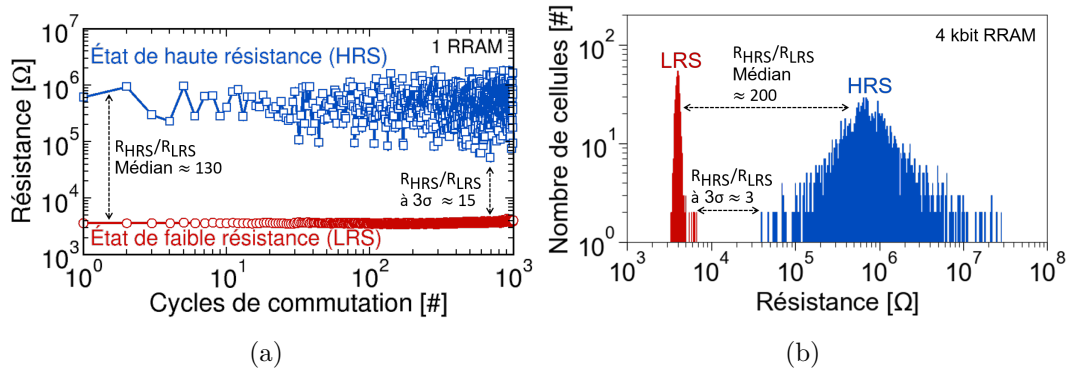


FIGURE 1.2.3: (a) Mesures de cyclage effectuées sur une cellule RRAM à base de HfO_2 . La valeur de résistance dans les états de basse (LRS, rouge) et haute (HRS, bleu) résistance varie de cycle à cycle. (b) Distributions de valeurs de résistances en LRS (rouge) et HRS (bleu) mesurées sur une matrice 4 kbit de RRAM à base de HfO_2 . La valeur de résistance en LRS et HRS varie de cellule à cellule. Reproduit de [25].

1.2.2 Les technologies 3D monolithiques

La deuxième solution technologique étudiée dans le contexte de cette thèse est la *technologie tri-dimensionnelle (3D) monolithique*, dans laquelle les composants sont fabriqués séquentiellement les uns au-dessus des autres. Deux types d'intégration 3D peuvent être distingués : (i) l'intégration parallèle, et (ii) l'intégration séquentielle, comme schématisés sur la FIGURE 1.2.4. Dans l'intégration 3D parallèle, les différents composants sont fabriqués séparément, puis sont empilés et connectés ultérieurement. Dans l'intégration 3D séquentielle, ou *technologie 3D monolithique*, les différentes couches de composants sont fabriquées directement au-dessus des couches précédentes. L'intégration 3D monolithique offre de meilleures précisions d'alignement entre les composants des

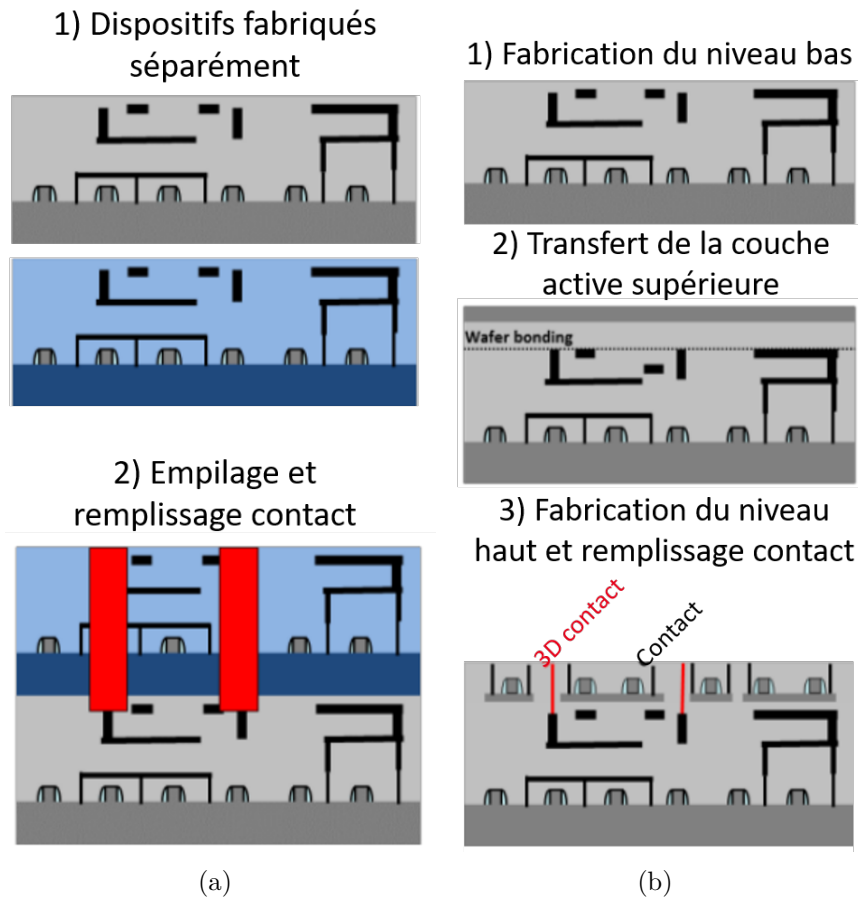


FIGURE 1.2.4: Illustration schématique de (a) l'intégration tri-dimensionnelle (3D) (a) parallèle, et (b) séquentielle. Reproduit de [55].

différentes couches, ce qui permet des densités d'interconnexions 50x supérieures à l'intégration 3D parallèle [56]. Cependant, son procédé de fabrication est bien plus complexe, car il requière des températures de fabrication suffisamment basses pour ne pas dégrader les performances des composants déjà fabriqués. Dans le cas des RRAM, leur température de fabrication est compatible avec le retour en fin de ligne des procédés CMOS, et peuvent donc être facilement fabriquées monolithiquement au-dessus des transistors CMOS. La FIGURE 1.2.5 montre une image par microscopie électronique d'une telle intégration : la RRAM est fabriquée au-dessus des contacts d'un transistor NMOS, dans la fameuse configuration *un-transistor/une-RRAM (1T1R)*.

Dans le cas d'une intégration 3D monolithique de transistors CMOS sur CMOS, la fabrication des transistors des couches du haut peut dégrader les performances des transistors des couches du bas, car les budgets thermiques impliqués sont généralement trop élevés [55–58]. La technologie CoolCube™ du CEA-Leti [59] permet la fabrication en 3D monolithique de deux couches de transistors CMOS en procédé silicium sur isolant (SOI pour Silicon On Insulator) 65 nm, sans dégradation des performances des différents transistors. De plus, cette technologie est compatible avec les requis industriels en termes de contamination. Dans le cadre de cette thèse, nous allons démontrer la *fabrication en technologie 3D monolithique de plusieurs couches de transistors CMOS hautes performances avec des mémoires résistives*.

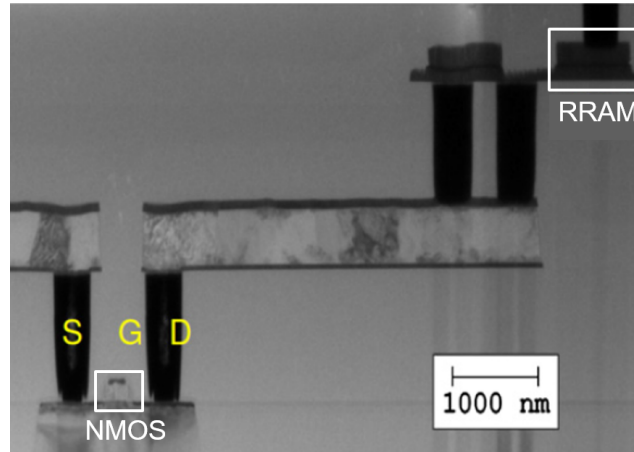


FIGURE 1.2.5: Image par microscopie électronique d'une RRAM intégrée monolithiquement au-dessus des contacts d'un transistor NMOS. Reproduit de [60].

1.3 Les réseaux de neurones impulsionsnels

Dans cette thèse, nous nous sommes intéressés à la troisième génération de réseaux de neurones : les *réseaux de neurones impulsionsnels* (SNN pour Spiking Neural Network) [61]. L'objectif de cette génération de réseaux de neurones est d'améliorer l'efficacité énergétique par rapport aux générations précédentes, en s'inspirant au plus près du fonctionnement des cerveaux biologiques [15, 62–67]. À l'image des réseaux de neurones biologiques, les SNN sont composés d'un ensemble d'unités de calcul, les *neurones*, interconnectées par des éléments mémoires, les *synapses*. Contrairement aux générations précédentes, l'information est encodée sous forme de courtes impulsions électriques qui sont transmises entre les neurones à travers les synapses. Cette représentation impulsionsnelle de l'information permet de prendre en compte le temps, qui joue un rôle primordial dans le traitement cognitif comme l'a montré la neurobiologie, lors du calcul, de la communication et de l'apprentissage du réseau.

La FIGURE 1.3.1 schématise le principe de fonctionnement de base d'un SNN. Une synapse connecte un neurone pré-synaptique à un neurone post-synaptique, et possède un *poids synaptique*. Lorsque les neurones pré-synaptiques émettent des impulsions, celles-ci sont propagées à travers les synapses et intégrées par le neurone post-synaptique. Plus le poids synaptique d'une synapse est élevé, plus les impulsions contribuent fortement à l'intégration. Lorsque l'intégration dépasse un certain seuil, le neurone post-synaptique émet une impulsion vers les neurones suivants, et sa valeur d'intégration est remise à zéro. Ainsi, l'opération de base d'un SNN est la *multiplication et accumulation* entre les données d'entrée et les poids synaptiques. Lors de la *phase d'apprentissage*, les poids synaptiques du réseau sont généralement modifiés et ajustés en fonction de l'application visée selon une règle d'apprentissage. De fait, l'application du SNN est définie par deux facteurs : d'une part les poids synaptiques, et d'autre part la topologie du réseau de neurone, c'est-à-dire l'agencement du réseau qui est défini par les différentes connexions synaptiques entre les neurones.

De ces considérations, *quatre composantes principales* sont nécessaires pour

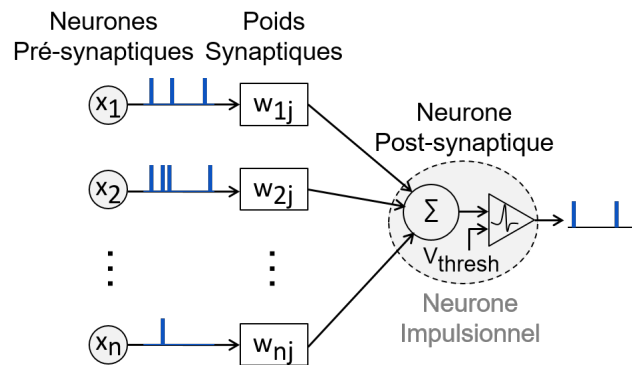


FIGURE 1.3.1: Principe de fonctionnement de base d'un réseau de neurones impulsionnel.

l'implémentation de *processeurs neuromorphiques impulsionnels*:

les circuits de neurones impulsionnels : le modèle de neurone impulsionnel le plus communément utilisé est celui d'*intégration et émission avec fuite* (LIF pour Leaky Integrate-and-Fire), un modèle canonique pour effectuer l'opération d'intégration et émission d'impulsions.

les matrices synaptiques : elles connectent les neurones impulsionnels entre eux et stockent les poids synaptiques du réseau. Un requis fondamental des matrices synaptiques est qu'elles soient *plastiques*, *i.e.* que les poids synaptiques soient ajustables.

la circuiterie d'apprentissage : elle permet d'effectuer l'apprentissage du réseau, *i.e.* de re-programmer les poids synaptiques du réseau pendant les phases d'apprentissage

les tables de routage synaptique : elles permettent de stocker la topologie du réseau de neurones, ce qui permet de modifier dynamiquement la topologie au cours de la vie du processeur en re-programmant les tables de routage. Elles sont essentielles à l'implémentation de *processeurs neuromorphiques reconfigurables*.

Dans le cadre de cette thèse, l'étude sera focalisée sur deux composantes: (i) les matrices synaptiques, et (ii) les tables de routage synaptique.

1.4 Objectif de ce travail de thèse de doctorat

Différents processeurs neuromorphiques ont déjà été démontrés dans la littérature [22, 62, 68–71]. La TABLE 1.1 synthétise les caractéristiques principales de chaque processeur. Bien que ces processeurs soient d'excellentes preuves de concept, il reste néanmoins beaucoup de marge d'amélioration possible d'un point de vue technologique, conception, et circuit. Premièrement, les matrices synaptiques sont implémentées entièrement en technologie CMOS [62, 72] ou avec des circuits numériques associés à de la mémoire SRAM [22, 68, 70, 71], ce qui n'est pas optimal en termes de surface silicium. Deuxièmement, les paramètres du réseau sont

souvent stockés dans des circuits de mémoires centralisées - généralement dans des mémoires SRAM ou DRAM [22, 68–70]. Ceci n'élimine pas véritablement le goulot d'étranglement de Von Neumann car les paramètres du réseau doivent être lus et transférés au travers d'un bus numérique pendant le calcul. Troisièmement, l'utilisation de mémoires volatiles SRAM et DRAM entraîne de la consommation statique de puissance et limite l'efficacité énergétique. Quatrièmement, les tables de routage synaptique sont implémentées soit avec des mémoires SRAM ou DRAM [22, 69–71], soit avec des mémoires adressables par contenu (CAM pour Content-Addressable Memory) [62, 68]. Il a été démontré que l'implémentation avec des CAM est plus efficace pour empêcher l'encombrement du réseau lors du routage des impulsions, et pour assouplir les contraintes sur le nombre maximal de neurones et synapses au sein de chaque cœur neuromorphique [73] grâce à la capacité des CAM à effectuer de la recherche rapide et parallèle de données. Cependant, les tables de routage synaptique à base de CAM sont habituellement implémentées avec des structures à base de SRAM [62], ce qui consomme la majeure partie de la surface silicium du processeur. Enfin, l'apprentissage en ligne n'est pas possible sur tous les processeurs.

	SpiNNaker [68]	TrueNorth [22]	Neurogrid [69]	HIAER [70]	Loihi [71]	DYNAPs [62]
Technologie	130 nm	28 nm	180 nm	130 nm	14 nm	180 nm
Neurones/cœur	~100	256	64 kbit	16 kbit	1 kbit	256
Synapses/cœur	~10 ⁵	64 kbit	N/A	~1 kbit	N/A	16 kbit
Type de Neurone	CMOS Numérique	CMOS Numérique	Mix analogue/ numérique CMOS	CMOS Numérique	CMOS Numérique	Mix analogue/ numérique CMOS
Type de Synapse	CMOS Numérique	CMOS Numérique	CMOS analogue	CMOS Numérique	CMOS Numérique	CMOS analogue
Schéma de routage	SRT (à base de CAM)	SRT (à base de SRAM)	SRT (à base de SRAM)	SRT (stockées dans une DRAM externe)	SRT (à base de SRAM)	SRT (à base de CAM intégrées avec les neurones)
Apprentissage	On-line	Off-line	Off-line	Off-line	On-line	On-line

SRT=Tables de routage synaptiques (Synaptic Routing Table)

TABLE 1.1: Synthèse des processeurs neuromorphiques impulsionnels multi-cœurs validés sur silicium rapportés dans la littérature [22, 62, 68–71].

Pour pallier ces problèmes, les nouvelles technologies présentées précédemment - les *mémoires résistives* (RRAM) et les *technologies 3D monolithiques* - sont des candidats appropriés pour améliorer l'efficacité en surface et énergie des processeurs neuromorphiques impulsionnels. Cependant, la faible fenêtre mémoire et la forte variabilité résistive des RRAM sont un frein à leur intégration dans de grandes matrices mémoires pour des applications mémoires classiques. Dans le cadre de cette étude, nous nous focaliserons sur (i) *les matrices de RRAM pour implémenter des poids synaptiques ajustables*, et (ii) *les matrices de mémoires ternaires adressables par contenu (TCAM pour Ternary Content-Addressable Memory) pour implémenter les tables de routage synaptique*. L'objectif final de ce travail de thèse de doctorat est d'évaluer rigoureusement l'impact des propriétés électriques des RRAM sur les performances et fiabilité de ces deux composants majeures, et de fournir des lignes directrices pour optimiser la programmation des RRAM au moyen de caractérisations électriques poussées et de simulations. De plus, nous ouvrirons des perspectives d'un point de vue technologique pour améliorer davantage l'efficacité en surface des processeurs neuromorphiques impulsionnels en démontrant la co-intégration 3D monolithique

de transistors CMOS hautes performances avec la technologie RRAM. Il est également intéressant de souligner que tous les résultats présentés dans ce manuscrit de thèse ne sont pas seulement limités à la technologie RRAM, et les lignes directrices pour optimiser les performances des SNN peuvent être appliquées à n'importe quelle technologie qui peut remplacer la technologie RRAM (telles que les mémoires à changement de phase, les mémoires magnétiques, ...).

Ce manuscrit de thèse est structuré de la façon suivante :

Chapitre 2: Rôle de la variabilité synaptique dans les réseaux de neurones impulsionnels à base de mémoires résistives avec apprentissage non supervisé

Dans ce chapitre, nous étudierons l'implémentation de synapses artificielles avec des RRAM dans les SNN entraînés avec l'algorithme d'apprentissage non supervisé de plasticité fonction d'occurrence des impulsions. Pour cela, deux applications simples ont été simulées: (i) une application de détection, et (ii) une application de classification. Nous présenterons d'abord des caractérisations électriques effectuées sur des matrices RRAM multi-kilobits. Nous évaluerons ensuite l'impact des propriétés électriques des RRAM sur les performances d'apprentissage des SNN grâce à des simulations niveau système calibrées sur les caractérisations électriques des RRAM. En particulier, nous clarifierons le rôle de la variabilité synaptique, qui provient de la variabilité résistive cycle-à-cycle et cellule-à-cellule des RRAM.

Chapitre 3: Reconfigurabilité du routage synaptique des réseaux de neurones impulsionnels avec des mémoires ternaires adressables par contenu à base de mémoires résistives

Dans ce chapitre, nous étudierons l'implémentation de tables de routage synaptique avec des mémoires ternaires adressables par contenu (TCAM pour Ternary Content-Addressable Memory) à base de RRAM. Nous présenterons d'abord des caractérisations électriques poussées d'un circuit TCAM à base de RRAM qui implémente la cellule unitaire TCAM la plus commune deux-transistors/deux-RRAM (2T2R). Nous présenterons ensuite une nouvelle cellule unitaire TCAM à base de RRAM dans une configuration un-transistor/deux-RRAM/un-transistor (1T2R1T) avec une surface silicium similaire à celle de la structure précédente 2T2R. La structure TCAM 1T2R1T proposée a pour objectif de résoudre les problèmes majeurs de la TCAM la plus commune 2T2R. Des caractérisations électriques poussées ont été effectuées sur un circuit TCAM 1T2R1T, et les résultats électriques obtenus sur chaque circuit TCAM ont été comparés.

Chapitre 4: Intégration tri-dimensionnelle monolithique de deux niveaux de transistors CMOS hautes performances avec un niveau de dispositifs de mémoires résistives

Dans ce chapitre, nous démontrerons la co-intégration complète de deux niveaux de transistors CMOS fabriqués en technologie tri-dimensionnelle (3D) monolithique avec un niveau de dispositifs RRAM monolithiquement fabriqué au-dessus des deux niveaux de transistors CMOS. Les dispositifs ont été fabriqués dans un procédé classique CMOS sur CMOS en technologie silicium sur isolant (SOI pour Silicon On Insulator) 65 nm. Nous présenterons d'abord le flux du processus d'intégration. Nous montrerons ensuite des caractérisations

électriques effectuées sur les dispositifs fabriqués pour démontrer la fonctionnalité de l'intégration.

Chapitre 5: Conclusion et perspectives

Ce chapitre conclut le manuscrit en synthétisant les résultats majeurs présentés dans ce travail de thèse de doctorat, et en fournissant quelques perspectives pour de possibles travaux futurs.

Rôle de la variabilité synaptique dans les réseaux de neurones impulsionnels à base de mémoires résistives avec apprentissage non supervisé

2.1 Objectif de ce chapitre

L'objectif de ce chapitre est de fournir une *étude détaillée de l'impact des propriétés électriques des mémoires résistives (RRAM pour Resistive Random Access Memory) sur les réseaux de neurones impulsionnels (SNN pour Spiking Neural Network) avec synapses à base de RRAM*. Au cours de la dernière décennie, les matrices RRAM ont été pressenties comme candidats potentiels pour implémenter les matrices synaptiques dans les réseaux de neurones électroniques [32–39]. Une des principales raisons est que les matrices RRAM implémentent naturellement des couches de réseaux de neurones. La FIGURE 2.1.1 illustre un exemple de réseau de neurones avec une topologie "connexions totales" implémenté avec une matrice de RRAM. Chaque point mémoire correspond à une synapse. Premièrement, les poids synaptiques sont encodés dans la valeur de conductance - *i.e.* l'inverse de la résistance - des RRAM. Deuxièmement, l'opération de base "multiplication et accumulation" est réalisée physiquement par la loi d'Ohm à chaque point mémoire, et par la loi de Kirchhoff par sommation des courants synaptiques à l'entrée des neurones post-synaptiques.

Dans cette partie, nous nous intéresserons aux *matrices un-transistor/une-RRAM (1T1R)*, où chaque élément mémoire est composé d'une RRAM en série avec un transistor NMOS de sélection. Cette structure matricielle est à l'heure actuelle la structure présentant la plus grande densité synaptique. Cependant, le problème majeur de cette structure matricielle 1T1R est la forte variabilité conductive des RRAM, qui empêche l'intégration des RRAM dans

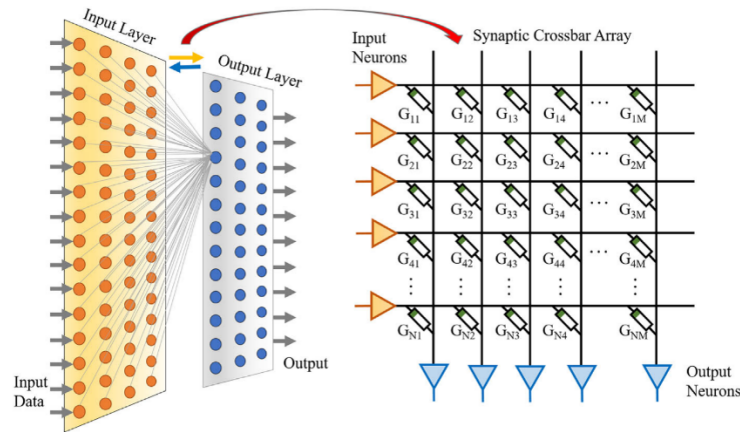


FIGURE 2.1.1: Illustration schématique d'une multiplication vecteur-matrice effectuée par une matrice de RRAM de type crossbar en un seul cycle de lecture. Reproduit de [39].

de grandes matrices mémoires pour des applications mémoires classiques. Pour des applications neuromorphiques, l'impact de cette variabilité conductive - qui crée de la *variabilité synaptique* - doit être clarifié. Il a été démontré dans de nombreux travaux que les réseaux de neurones à base de RRAM étaient intrinsèquement robustes à la variabilité synaptique [15, 74–92]. Toutefois, une étude précise expliquant l'origine de cette robustesse n'a pas encore été effectuée. En particulier, il reste à comprendre si les réseaux de neurones sont seulement robustes à la variabilité synaptique, ou s'ils peuvent tirer profit de cette variabilité. En effet, la neurobiologie a montré que de la variabilité existait dans les cerveaux biologiques, et que celle-ci pouvait être bénéfique [93–98]. Bien que de nombreuses études détaillées de l'impact des propriétés électriques des RRAM sur les réseaux de neurones avec apprentissage supervisé existent déjà [76, 79, 80, 83, 99–102], il n'en existe que très peu sur des réseaux de neurones entraînés avec des algorithmes d'apprentissage non supervisé [15, 88, 90, 103]. Dans ce chapitre, nous fournirons *une étude détaillée des requis électriques des RRAM pour implémenter des synapses dans des SNN avec apprentissage non supervisé par plasticité fonction d'occurrence des impulsions (STDP pour Spike-Timing-Dependent Plasticity)*. En particulier, le rôle de la *variabilité synaptique* dans les SNN entraînés de façon non supervisée par STDP sera clarifié.

2.2 Caractérisations électriques des RRAM

Dans ce chapitre, nous nous intéresserons à des dispositifs *RRAM à base d'oxyde HfO₂* intégrés dans le retour en fin de ligne (back-end-of-line) d'un procédé CMOS 130 nm (*cf* FIGURE 2.2.1) [25]. Les dispositifs RRAM sont composés d'un empilage TiN/HfO₂/Ti/TiN d'épaisseur 100 nm/10 nm/10 nm/100 nm. Un transistor NMOS en série avec la RRAM est utilisé en tant que sélecteur dans une configuration 1T1R. Ce transistor permet de sélectionner individuellement chaque élément mémoire d'une matrice, et de contrôler le courant de programmation, I_{prog} , pendant les opérations de programmation. Toutes les mesures ont été effectuées sur des *matrices RRAM 1T1R de 4 kbit*. En appliquant une tension

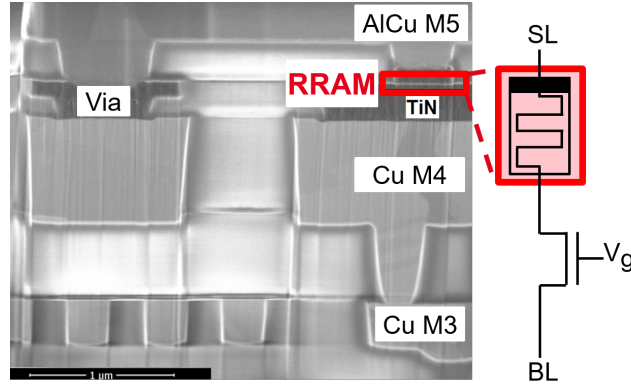


FIGURE 2.2.1: (Gauche) Photo par microscopie électronique d'une cellule RRAM TiN/HfO₂/Ti/TiN (100 nm/10 nm/10 nm/100 nm) intégrée au-dessus du quatrième niveau métallique Cu. (Droite) Vue schématique d'une configuration 1T1R.

positive sur l'électrode du haut, V_{SL} , les mémoires commutent dans un état de haute conductance avec une opération de Set (HCS pour High Conductance State). En appliquant une tension positive sur l'électrode du bas, V_{BL} , les mémoires commutent dans un état de faible conductance avec une opération de Reset (LCS pour Low Conductance State).

Pour étudier l'impact des propriétés électriques des RRAM sur les performances d'apprentissage des réseaux de neurones impulsifs, les RRAM ont été programmées avec quatre conditions de programmation (*i.e.* tension et courant de programmation) différentes. La FIGURE 2.2.2 montre les distributions cumulées de conductance en HCS et LCS associées à chaque condition de programmation. La TABLE 2.1 synthétise les paramètres de chaque condition de programmation. Pour étudier l'impact du vieillissement des RRAM sur les performances d'apprentissage des SNN, l'endurance en programmation a été mesurée en FIGURE 2.2.3 sur la matrice RRAM 4 kbit avec les conditions de programmation A. Comme le montrent les FIGURE 2.2.2 et 2.2.3, les RRAM présentent de la variabilité conductive cycle-à-cycle et cellule-à-cellule, ce qui rapproche les distributions HCS et LCS. Pour quantifier la séparation entre les distributions HCS et LCS, la *fenêtre mémoire*, $MW_{3\sigma}$, est définie comme le rapport entre la valeur de conductance HCS à -3σ , $HCS_{-3\sigma}$, et la valeur de conductance LCS à $+3\sigma$, $LCS_{+3\sigma}$:

$$MW_{3\sigma} = \frac{HCS_{-3\sigma}}{LCS_{+3\sigma}} \quad (2.2.1)$$

La *variabilité conductive* est définie comme l'écart-type en logarithme base 10 des distributions de conductance:

$$\begin{aligned} \sigma_{G,HCS} &= \text{std}[\log_{10}(G_{HCS})] \\ \sigma_{G,LCS} &= \text{std}[\log_{10}(G_{LCS})] \end{aligned} \quad (2.2.2)$$

L'endurance en programmation est définie dans ce travail comme le nombre maximal d'opérations de programmation avant que l'oxyde des RRAM ne se rompe. Les énergies de programmation en Set et Reset ont été calculées comme suit:

$$\begin{aligned} E_{\text{Set}} &= V_{\text{Set}} * I_{\text{prog,set}} * t_{\text{pulse}} \\ E_{\text{Reset}} &= V_{\text{Reset}} * I_{\text{prog,set}} * t_{\text{pulse}} \end{aligned} \quad (2.2.3)$$

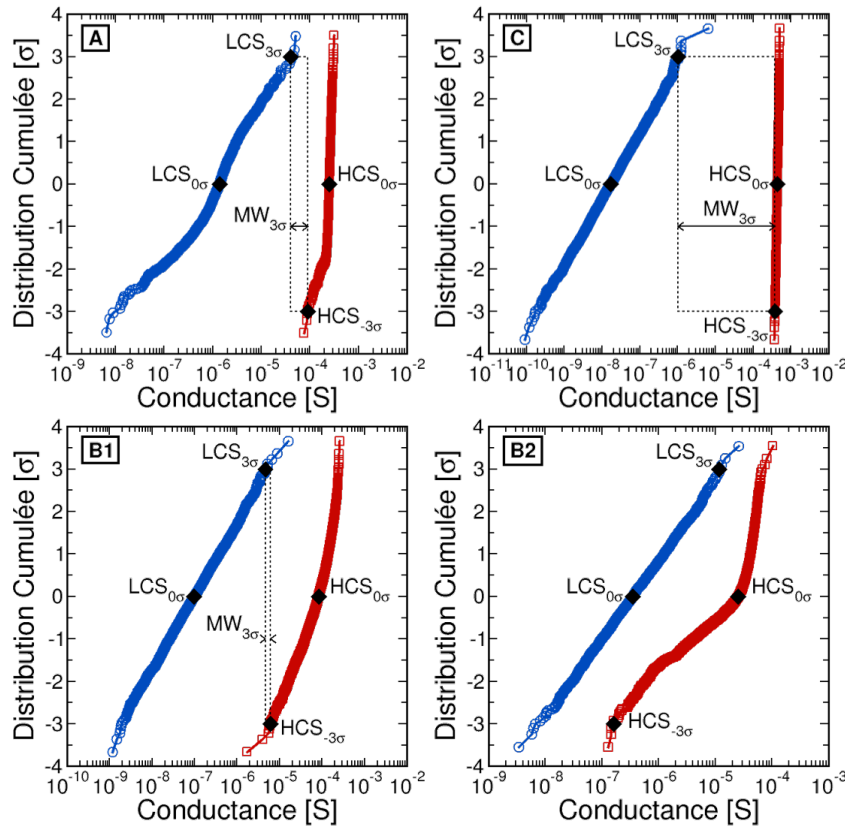


FIGURE 2.2.2: Distributions cumulées LCS et HCS mesurées sur la matrice 4 kbit avec différentes conditions de programmation.

Conditions de programmation		A	C	B1	B2
Tension [V]	V_{Set}	2	2	2	2
	V_{Reset}	2.5	2.5	2.5	2.5
$I_{\text{prog,set}}$ [μA]		250	500	57	20
$V_{\text{g,reset}}$ [μA]		3	3.5	3.5	3.5
Energie [pJ/spike]	E_{Set}	50	100	11.4	4
	E_{Reset}	62.5	125	14.25	5
$\sigma_{\text{G,HCS}}$ [$\log_{10}(\text{S})$]		0.05	0.02	0.28	0.53
$\sigma_{\text{G,LCS}}$ [$\log_{10}(\text{S})$]		0.49	0.64	0.58	0.54
$\text{MW}_{3\sigma}$ [#]		3	370	1.3	0.014
Endurance [#]		10^6	$\approx 10^2$	$\approx 10^6$	$\approx 10^8$

 TABLE 2.1: Conditions de programmation utilisées dans ce travail, avec $t_{\text{pulse}}=100$ ns.

2.3 Implémentation des éléments synaptiques et règle d'apprentissage avec les mémoires résistives

La FIGURE 2.3.1 (a) schématise l'implémentation des éléments synaptiques avec les mémoires résistives. Chaque synapse est composée de n cellules RRAM 1T1R connectées en parallèle [74, 89]. Comme les conductances parallèles se somment, le poids synaptique équivalent varie de la somme de n mémoires en LCS à

2.3. IMPLÉMENTATION DES ÉLÉMENTS SYNAPTIQUES ET RÈGLE D'APPRENTISSAGE AVEC LES MÉMOIRES RÉSISTIVES

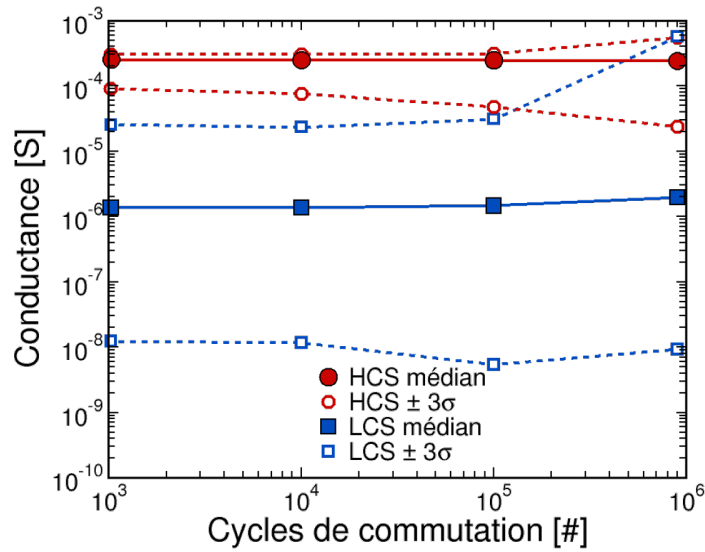
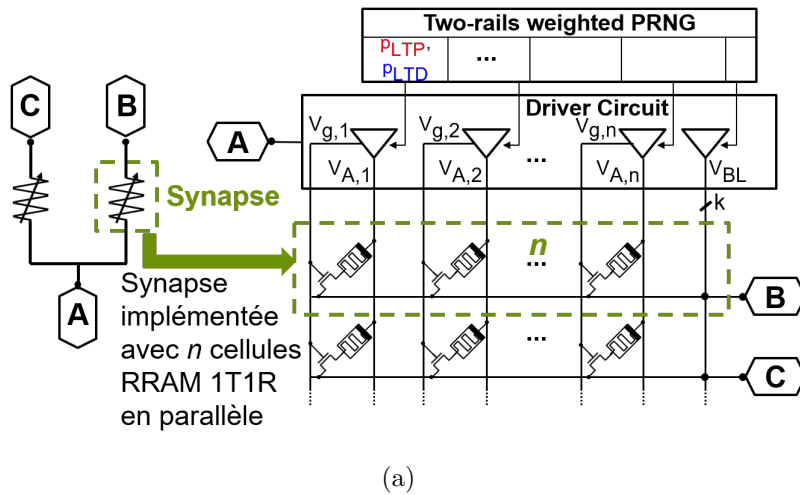
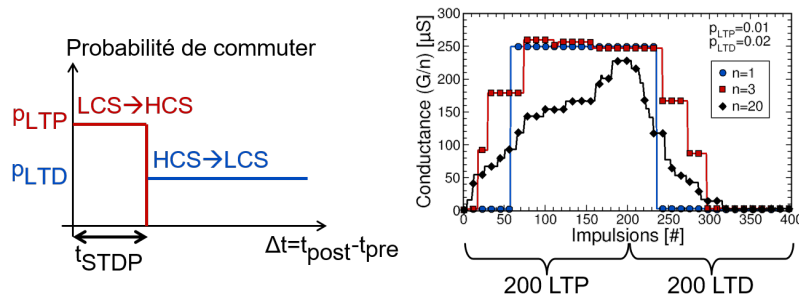


FIGURE 2.2.3: Caractérisation de l'endurance en programmation mesurée avec les conditions de programmation A de la TABLE 2.1.



(a)



(b)

FIGURE 2.3.1: (a) Schéma de l'implémentation des éléments synaptiques à base de RRAM. (b) Version stochastique de la règle d'apprentissage non-supervisée de plasticité fonction d'occurrence des impulsions (STDP).

n mémoires en HCS, avec $n+1$ niveaux de conductance intermédiaires. Cette implémentation est associée à une *version stochastique de la règle d'apprentissage bio-inspirée de plasticité fonction d'occurrence des impulsions (STDP pour Spike-Timing-Dependent Plasticity)* [104–106]. La règle d'apprentissage est représentée

sur la FIGURE 2.3.1 (b). Quand un neurone pré-synaptique émet une impulsion juste avant un neurone post-synaptique dans une fenêtre temporelle t_{STDP} , un événement de potentialisation à long-terme (LTP pour Long-Term Potentiation) se produit, et chacune des n RRAM composant l'élément synaptique a une probabilité p_{LTP} de commuter en HCS. Sinon, un événement de dépression à long-terme (LTD pour Long-Term Depression) se produit, et chacune des n RRAM composant l'élément synaptique a une probabilité p_{LTD} de commuter en LCS. Cette implémentation synaptique permet d'augmenter ou diminuer graduellement la conductance équivalente de la synapse à chaque événement de potentialisation ou dépression, respectivement.

2.4 Implications pour un système d'apprentissage: impact des caractéristiques des synapses à base de RRAM sur les performances d'un réseau

2.4.1 Topologie des réseaux de neurones impulsionsnels

Deux applications ont été simulées: (i) une application de détection, et (ii) une application de classification. L'application de détection consiste en la détection de voitures roulant sur une autoroute [107], et celle de classification en la classification de chiffres écrits à la main de la base de données MNIST [108]. Les deux applications sont basées sur un *SNN de type connexions totales avec une seule couche de synapses reliant une couche de neurones d'entrée à une couche de neurones de sortie*. Chaque élément synaptique est implémenté avec la structure de la SECTION 2.3, et calibré sur les distributions expérimentales obtenues avec les différentes conditions de programmation (*cf* FIGURE 2.2.2 : conditions A, B1, B2, et C). Les neurones de sortie sont implémentés avec le modèle d'intégration et émission avec fuite (LIF pour Leaky Integrate-and-Fire) [109]. Les SNN simulés pour chaque application sont représentés en FIGURE 2.4.1.

2.4.2 Impact de la fenêtre mémoire et variabilité conductive des RRAM

La FIGURE 2.4.2 (a) montre la performance de détection, F1, en fonction de la fenêtre mémoire, $MW_{3\sigma}$, pour le SNN simulé pour l'application de détection. Différents nombres n de RRAM par synapse ont été simulés pour étudier l'impact de la résolution synaptique, *i.e.* le nombre de niveaux de conductance, et les RRAM ont été calibrées sur les conditions de programmation A. Les différentes fenêtres mémoires, $MW_{3\sigma}$, ont été artificiellement obtenues en translatant la distribution LCS obtenue expérimentalement vers des conductances plus faibles ou plus hautes pour augmenter ou diminuer $MW_{3\sigma}$, respectivement. Comme le montre la FIGURE 2.4.2 (a), la performance n'est pas améliorée avec la résolution

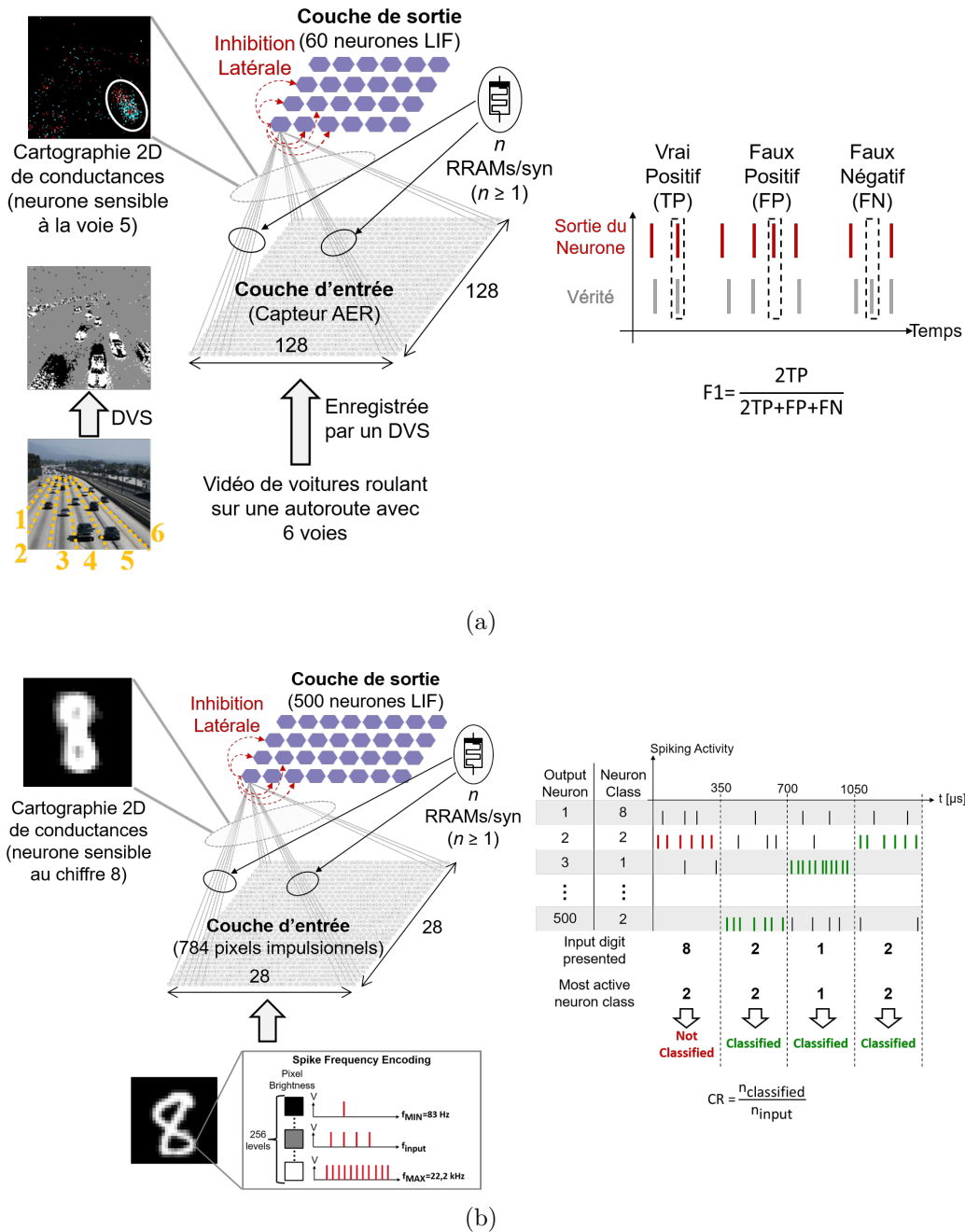


FIGURE 2.4.1: Réseaux de neurones impulsifs simulés pour les applications de (a) suivi de voitures et (b) classification de chiffres. Les scores associés pour évaluer les performances de chaque réseau sont définis sur la partie droite de chaque réseau.

synaptique, et seule la fenêtre mémoire impacte la performance. On en conclut qu'une synapse binaire ($n=1$) est suffisante, et que la performance maximale de 96% est atteinte pour des fenêtres mémoires d'au moins 10. Pour étudier l'impact de la variabilité synaptique - qui provient de la variabilité conductive des RRAM -, la FIGURE 2.4.2 (b) montre la performance de détection en fonction de la fenêtre mémoire, $MW_{3\sigma}$, lorsque le réseau de neurones est calibré avec les différentes conditions de programmation A, B1, B2, et C. Comme référence, une condition artificielle présentant zéro variabilité synaptique a été également

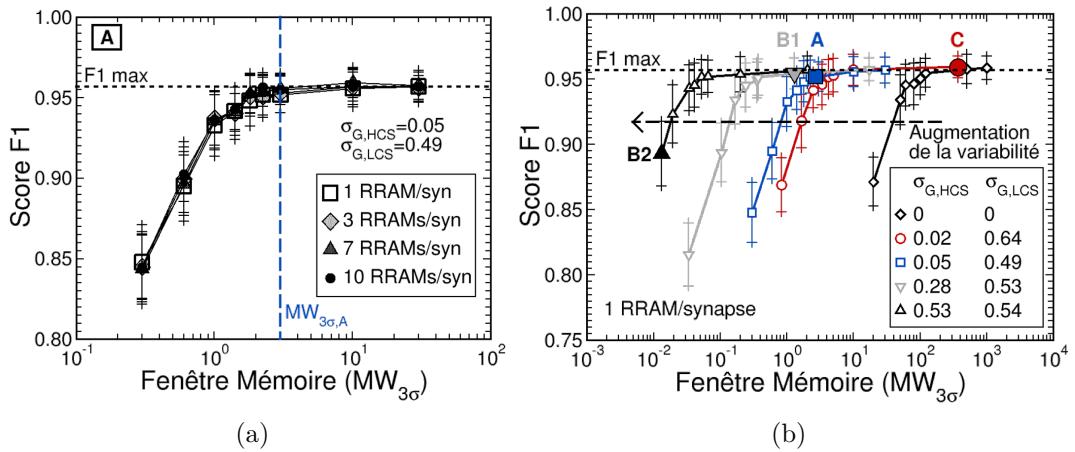


FIGURE 2.4.2: (a) Score F1 en fonction de la fenêtre mémoire à 3σ , $MW_{3\sigma}$, pour différents nombres de RRAM par synapse. Le réseau de neurones conçu pour le suivi de voitures est calibré sur les conditions de programmation A (cf TABLE 2.1). (b) Score F1 en fonction de $MW_{3\sigma}$ lorsque le réseau de neurones conçu pour le suivi de voitures est calibré avec les différentes conditions de programmation de la TABLE 2.1.

simulée (losanges noirs). Les synapses sont implémentées avec une seule RRAM. On observe que *la variabilité synaptique est bénéfique car elle permet de travailler avec des fenêtres mémoires plus petites*.

Ce protocole de simulations a été reproduit avec le SNN conçu pour l'application de classification. La FIGURE 2.4.3 (a) montre la performance de classification, CR, en fonction de la fenêtre mémoire, $MW_{3\sigma}$, pour différents nombres n de RRAM par synapse. Le réseau a été calibré avec les conditions de programmation A, B2, et la condition artificielle présentant zéro variabilité synaptique. Contrairement à l'application de détection, la performance est indépendante de la fenêtre mémoire et augmente avec la résolution synaptique. Pour cette application de classification, une résolution d'au moins 10 niveaux synaptiques est nécessaire. Pour étudier l'impact de la variabilité synaptique, le réseau a été ensuite calibré avec les différentes conditions de programmation. La FIGURE 2.4.3 (b) montre la performance de classification, CR, en fonction de la variabilité synaptique, $\sigma_{G,HCS}$. Les différentes variabilités synaptiques proviennent des différentes conditions de programmation. Les synapses sont implémentées avec 20 RRAM par synapse. On observe que la performance est maximale pour une variabilité synaptique d'environ 0.05 (CR=81.81% pour la condition A, et CR=81.78% pour la condition C). Pour comprendre les résultats obtenus avec les applications de détection et classification, la FIGURE 2.4.4 montre les distributions de poids synaptiques obtenues après apprentissage, pour chaque condition de programmation. Pour l'application de détection (FIGURE 2.4.4 (a)), il est nécessaire que la population de synapses potentialisées (rouge) soit suffisamment éloignée de la population de synapses déprimées (bleu). Ceci se traduit par la nécessité d'avoir un *rapport moyen*, au sens de *moyenne arithmétique*, suffisamment élevé pour obtenir la performance maximale de 96%. En pratique, un rapport moyen d'au moins 200 est nécessaire. Ainsi, la fenêtre mémoire et la variabilité conductive permettent d'améliorer la performance. En effet, même si les deux populations se chevauchent car la fenêtre mémoire n'est pas assez

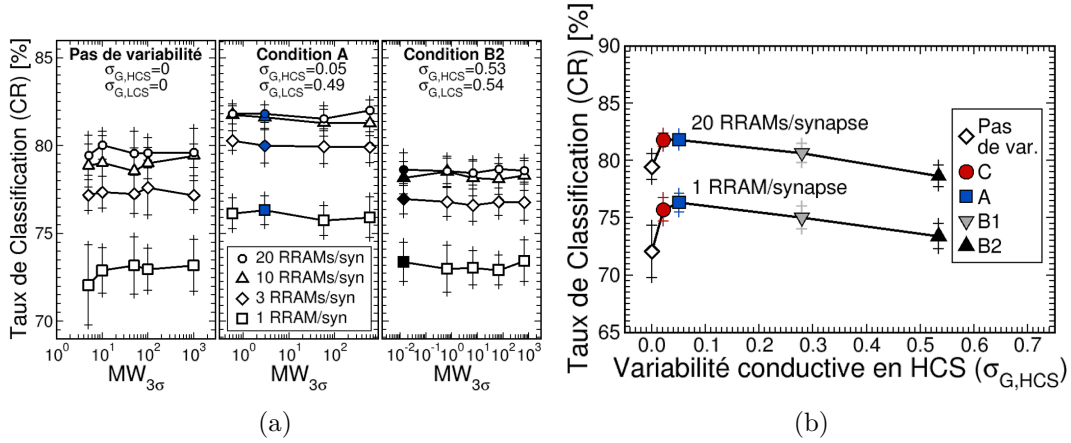


FIGURE 2.4.3: (a) Taux de classification, CR, du réseau de neurones conçu pour la classification de chiffres en fonction de la fenêtre mémoire à 3σ , $MW_{3\sigma}$, pour différents nombres de RRAM par synapse. (b) Taux de classification, CR, du réseau de neurones conçu pour la classification de chiffres en fonction de la variabilité synaptique, $\sigma_{G,HCS}$.

élevée (condition B1), la variabilité synaptique permet néanmoins d'éloigner les deux distributions par un effet de moyennage dans chaque distribution. Pour l'application de classification (FIGURE 2.4.4 (b)), il est nécessaire d'avoir une résolution synaptique suffisante pour obtenir la meilleure performance. En pratique, *au moins 10 niveaux synaptiques sont nécessaires. Comme la résolution synaptique ne dépend que de la distribution HCS, la fenêtre mémoire n'a pas d'impact sur la performance.* Ainsi, pour les conditions A, C, et la condition artificielle avec zéro variabilité synaptique, 21 niveaux synaptiques sont distinguables. Cependant, la présence d'une certaine quantité de variabilité synaptique dans le cas des conditions A et C permet un ajustement plus précis et une transition plus graduelle entre chaque niveau synaptique, ce qui améliore la performance par rapport à la condition avec zéro variabilité. Dans le cas des conditions B1 et B2 avec une plus forte variabilité synaptique, la variabilité synaptique aplattit les distributions, et seulement 9 et 7 niveaux synaptiques sont distinguables, respectivement.

2.4.3 Impact du vieillissement des RRAM

La FIGURE 2.4.5 montre la performance de détection (a) et de classification (b) en fonction du nombre d'opérations de programmation. Pour cela, le réseau a été calibré avec les distributions obtenues lors de la caractérisation de l'endurance en programmation avec la condition A de la FIGURE 2.2.3. Dans le cas de l'application de détection, la performance se dégrade avec le vieillissement des RRAM à cause des cellules détruites. En supprimant ces cellules défectueuses, il est possible de ré-obtenir la performance maximale. Pour l'application de classification, la performance n'est pas impactée par le vieillissement des RRAM, car la variabilité synaptique reste autour de la valeur de 0.05 pendant tout le cyclage.

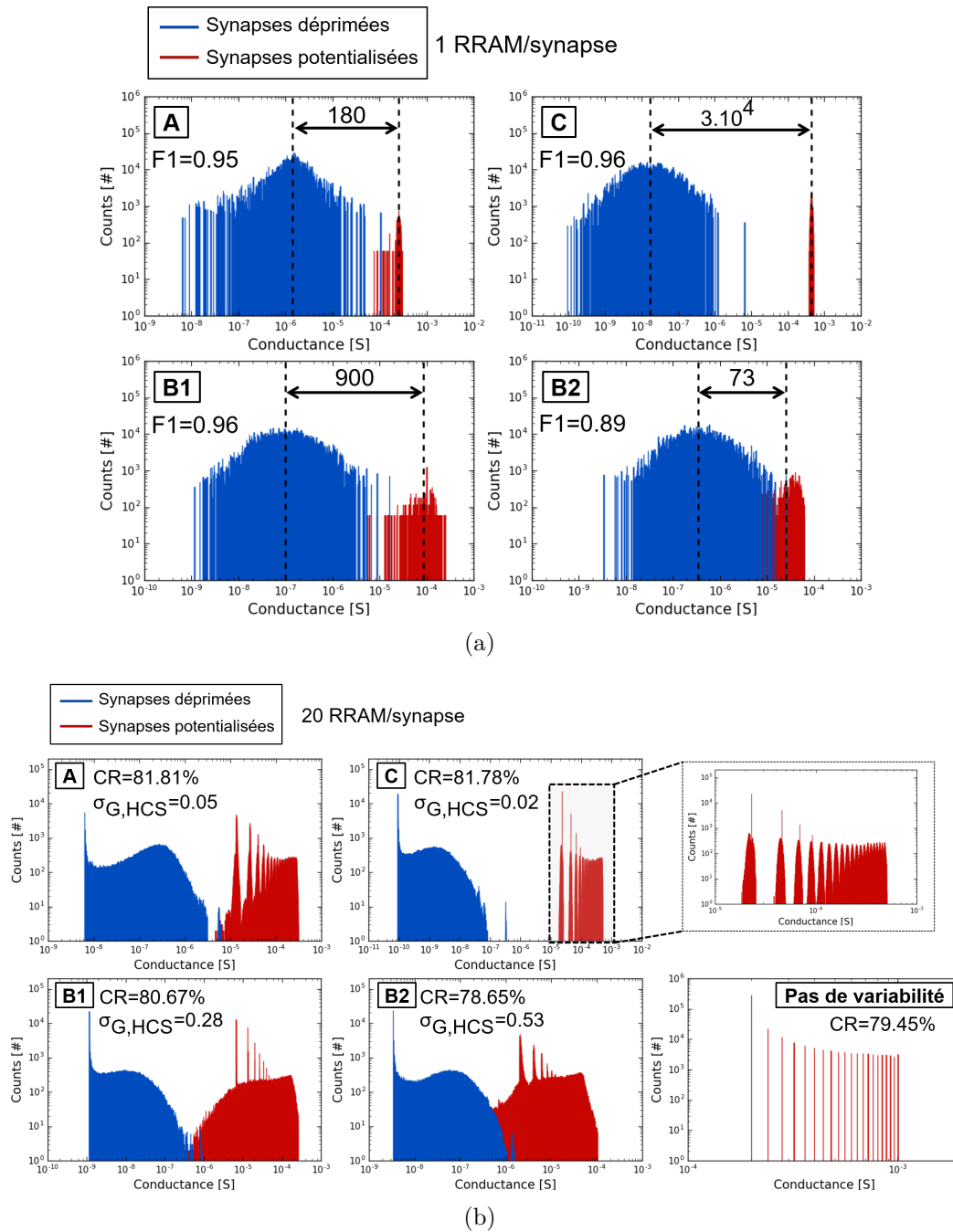


FIGURE 2.4.4: Distributions des poids synaptiques après apprentissage pour l'application de (a) détection et (b) classification.

2.5 Conclusion

Dans ce chapitre, une étude détaillée de l'impact de la variabilité conductive, puissance de programmation, et vieillissement des RRAM sur les performances d'apprentissage des SNN avec synapses à base de RRAM et entraînés de façon non supervisée par STDP a été présentée. Les données expérimentales ont été obtenues par caractérisations électriques de matrices multi-kilobits 1T1R RRAM [25], et les performances d'apprentissage ont été évaluées par simulations niveau système de SNN calibrés sur les données expérimentales, et conçus pour (i)

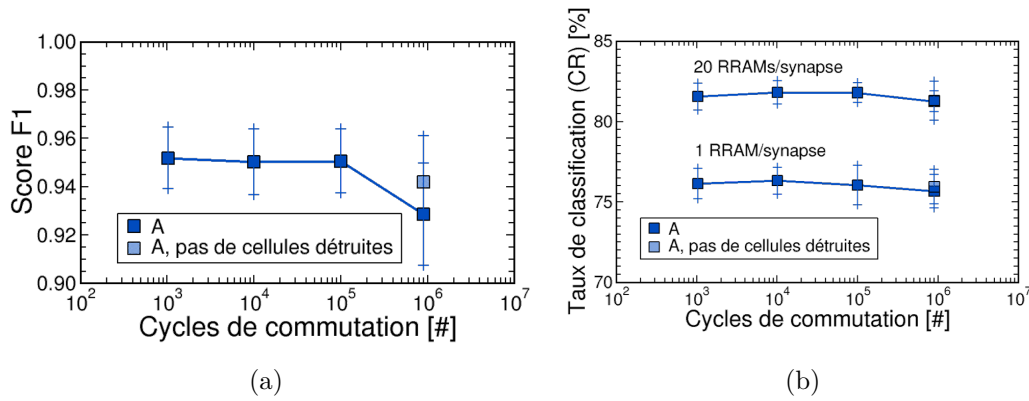


FIGURE 2.4.5: (a) Score F1 en fonction du nombre de cycles de commutation pour le réseau de neurones conçu pour le suivi de voitures. (b) Taux de classification en fonction du nombre de cycles de commutation pour le réseau de neurones conçu pour la classification de chiffres.

une application de détection [107], et (ii) une application de classification [108]. Par rapport à la littérature [15, 74–92], nous avons montré que les SNN ne sont pas seulement robustes à la variabilité synaptique, mais que celle-ci peut être bénéfique. Ceci peut s'expliquer par le fait que la variabilité synaptique permet aux synapses d'accéder à une plus grande plage de valeurs de poids synaptiques pendant l'apprentissage. De façon plus large, cette étude fournit des lignes directrices pour optimiser la programmation des RRAM dans les SNN avec synapses à base de RRAM et apprentissage non supervisé. Également, elle démontre que les dispositifs mémoires peuvent être programmés de façon plus optimale dans les applications neuromorphiques que dans les applications mémoires, et que les requis matériels des RRAM diffèrent pour les applications mémoires et neuromorphiques.

Reconfigurabilité du routage synaptique des réseaux de neurones impulsionnels avec des mémoires ternaires adressables par contenu à base de mémoires résistives

3.1 Objectif de ce chapitre

L'objectif de ce chapitre est d'évaluer la possibilité d'implémenter des tables de routage synaptique (SRT pour Synaptic Routing Table) avec des mémoires résistives (RRAM pour Resistive Random Access Memory). Le rôle des SRT est de stocker la topologie du réseau de neurones, c'est-à-dire l'agencement du réseau de neurones qui est défini par les différentes connexions synaptiques entre les neurones [110–112]. Il est alors possible de modifier dynamiquement la topologie du réseau en re-programmant les SRT, et ainsi de reconfigurer le processeur neuromorphique. Dans le cas de réseaux de neurones impulsionnels (SNN pour Spiking Neural Network), les différents processeurs neuromorphiques impulsionnels [22, 62, 68–71] utilisent la représentation d'événements par adresse (AER pour Address Event Representation) [113, 114] : chaque neurone a une adresse unique, et transmet son adresse à tous les neurones avec lesquels il est connecté lorsqu'il émet une impulsion. Dans ce contexte, les SRT correspondent à des tables de conversion (LUT pour Look-Up Table) qui stockent et associent les adresses des neurones connectés ensemble. Pour cette étude, nous nous focaliserons sur l'implémentation des SRT avec des matrices de mémoires ternaires adressables par contenu (TCAM pour Ternary Content-Addressable Memory). En effet, il a été démontré que l'implémentation des SRT avec des TCAM était une méthode efficace pour empêcher l'encombrement du réseau lors du routage des impulsions [73].

Dans cette partie, nous évaluerons l'utilisation de TCAM à base de RRAM pour

l'implémentation de SRT dans les processeurs neuromorphiques impulsionsnels. Nous présenterons tout d'abord deux circuits TCAM à base de RRAM qui ont été intégrés, fabriqués, et caractérisés électriquement. Le premier circuit TCAM à base de RRAM correspond à la structure la plus commune où chaque cellule unitaire TCAM est composée de deux transistors et deux RRAM dans une configuration deux un-transistor/une-RRAM en parallèle (2T2R) [115–119]. La structure TCAM 2T2R est à l'heure actuelle la plus petite structure TCAM [116]. Le deuxième circuit TCAM est une nouvelle structure où chaque cellule unitaire TCAM est composée également de deux transistors et deux RRAM dans une configuration un-transistor/deux-RRAM/un-transistor (1T2R1T). Cette structure présente une surface silicium similaire à celle de la TCAM 2T2R. L'objectif de cette nouvelle TCAM 1T2R1T est de résoudre les problèmes majeurs de la TCAM 2T2R en s'affranchissant des propriétés électriques des RRAM. Des caractérisations électriques poussées ont ensuite été effectuées sur chaque circuit TCAM afin de quantifier l'impact des propriétés électriques des RRAM sur les performances et fiabilité des circuits TCAM. Enfin, les résultats électriques obtenus ont été comparés.

3.2 Principes de base des mémoires adressables par contenu

Les mémoires adressables par contenu (CAM pour Content-Addressable Memory) sont des circuits spécialisés dans la recherche intensive et rapide de données [120–123]. Ils permettent de rechercher une donnée dans une table mémoire avec des données pré-enregistrées. Le principe est schématisé en FIGURE 3.2.1. Dans un système mémoire à accès aléatoire (RAM pour Random Access Memory) classique (FIGURE 3.2.1 (a)), une adresse est envoyée en entrée du système, et la donnée stockée à cette adresse est retournée en sortie. Les systèmes CAM procèdent de façon inverse : une donnée recherchée est envoyée en entrée, et chacune des données stockées est comparée à la donnée recherchée (FIGURE 3.2.1 (b)). Le résultat de la comparaison est retourné par la ligne de match. Si les

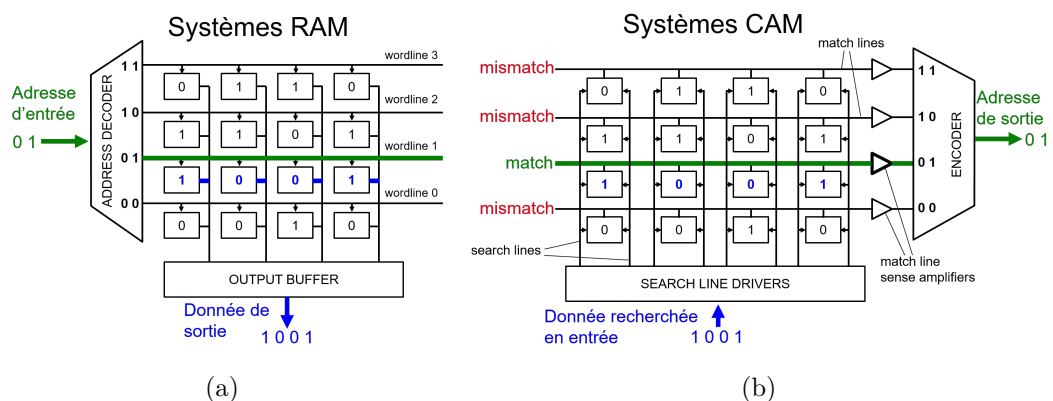


FIGURE 3.2.1: Principe de fonctionnement (a) d'un système RAM classique, et (b) un système de mémoire adressable par contenu (CAM pour Content-Addressable Memory).

données recherchées et stockées sont différentes, la ligne de match retourne un cas de *mismatch*. Si les données sont identiques, la ligne de match retourne un cas de *match*. L'avantage principal d'une CAM est que la recherche de données peut être effectuée en parallèle sur toute la matrice mémoire, en un seul cycle d'horloge, ce qui permet d'atteindre des vitesses de recherche de l'ordre de la nanoseconde [123].

Les systèmes CAM peuvent être catégorisés soit en CAM binaires (BCAM pour Binary CAM), soit en CAM ternaire (TCAM pour Ternary CAM). Alors que les BCAM ne peuvent stocker que des données '0' ou '1', une TCAM offre la possibilité de stocker un troisième type de donnée dénoté 'X'. La donnée 'X' sert de joker lors d'une opération de recherche, et retourne toujours un cas de match quelle que soit la donnée recherchée. Dans ce chapitre, nous nous intéresserons aux systèmes TCAM, en particulier ceux implémentés avec des RRAM.

3.3 Circuits de mémoires ternaires adressables par contenu à base de mémoires résistives

3.3.1 La cellule TCAM la plus commune deux-transistors/deux-RRAM (2T2R)

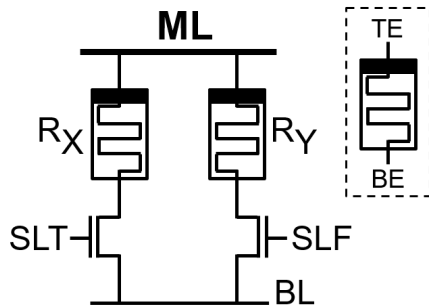


FIGURE 3.3.1: Schéma de la cellule unitaire TCAM la plus commune 2T2R.

Donnée stockée	R_X	R_Y
'0'	LRS	HRS
'1'	HRS	LRS
'X'	HRS	HRS

TABLE 3.1: Définition des états des RRAM en fonction de la donnée stockée.

Donnée recherchée	SLT	SLF
'0'	0	VDD
'1'	VDD	0

TABLE 3.2: Tensions SLT et SLF en fonction de la donnée recherchée.

La cellule TCAM la plus commune, qui est également la plus petite à l'heure actuelle [115–119], est composée de deux structures 1T1R en parallèle (2T2R). La cellule unitaire TCAM 2T2R est schématisée sur la FIGURE 3.3.1. Le stockage de donnée repose sur un encodage différentiel, en programmant les RRAM soit dans l'état faible résistance (LRS pour Low Resistance State), soit dans l'état haute résistance (HRS pour High Resistance State) en suivant les combinaisons de la TABLE 3.1. La recherche de données repose sur la décharge de la ligne de match. Lors d'une opération de recherche, la ligne de match est d'abord préchargée à une tension haute avec une tension V_{search} . Elle est ensuite laissée

flottante, et se décharge à travers les cellules TCAM. Selon la donnée recherchée, le transistor SLT ou SLF est activé et l'autre désactivé (cf TABLE 3.2). Dans un cas de match (cf FIGURE 3.3.2 (a)), la structure 1T1R activée est en HRS, la ligne de match se décharge relativement lentement et reste à l'état haut. Dans un cas de mismatch (cf FIGURE 3.3.2 (b)), la structure 1T1R activée est en LRS, la ligne de match se décharge rapidement vers l'état bas.

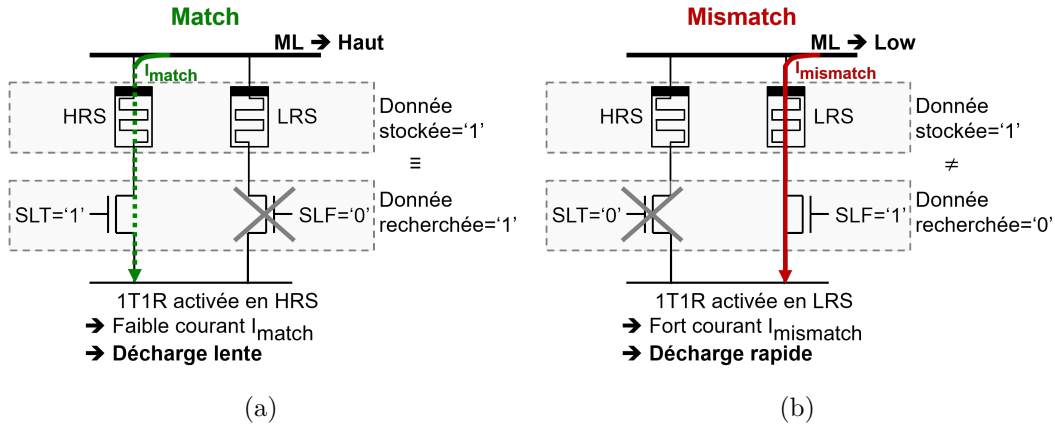


FIGURE 3.3.2: (a) Dans un cas de match, le transistor activé est en série avec une RRAM dans l'état haute résistance (HRS pour High Resistance State). (b) Dans un cas de mismatch, le transistor activé est en série avec une RRAM dans l'état faible résistance (LRS pour Low Resistance State).

3.3.2 La nouvelle cellule TCAM un-transistor/deux-RRAM/un-transistor (1T2R1T)

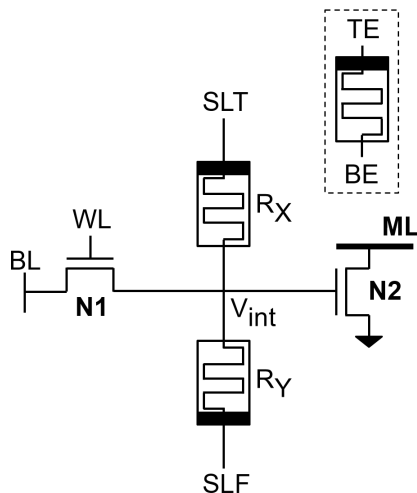


FIGURE 3.3.3: Schéma de la nouvelle cellule unitaire TCAM 1T2R1T.

Donnée stockée	R_X	R_Y
'0'	LRS	HRS
'1'	HRS	LRS
'X'	HRS	HRS

TABLE 3.3: Définition des états des RRAM en fonction de la donnée stockée.

Donnée recherchée	SLT	SLF
'0'	0	V_{search}
'1'	V_{search}	0

TABLE 3.4: Tensions SLT et SLF en fonction de la donnée recherchée.

La nouvelle cellule TCAM est composée de deux transistors et deux RRAM dans une configuration un-transistor/deux-RRAM/un-transistor (1T2R1T)

comme représentée sur la FIGURE 3.3.3. Le transistor N1 est impliqué dans les opérations de programmation des RRAM, le transistor N2 est impliqué dans les opérations de recherche. Comme pour la TCAM 2T2R, le stockage de données repose sur un encodage différentiel des deux RRAM (*cf* TABLE 3.3). Lors d'une opération de recherche, le transistor N1 est désactivé, et une tension positive V_{search} est appliquée soit sur SLT, soit sur SLF selon la TABLE 3.4. Le noeud interne entre les deux RRAM est polarisé par pont diviseur de tension entre les deux RRAM. Dans un cas de match (*cf* FIGURE 3.3.4 (a)), le noeud interne est polarisé à une tension quasiment nulle, le transistor N2 reste OFF et la ligne de match ne se décharge pas. Dans un cas de mismatch (*cf* FIGURE 3.3.4 (b)), le noeud interne est polarisé à environ V_{search} , ce qui active le transistor N2 si V_{search} est supérieure à sa tension de seuil. La ligne de match se décharge alors vers l'état bas.

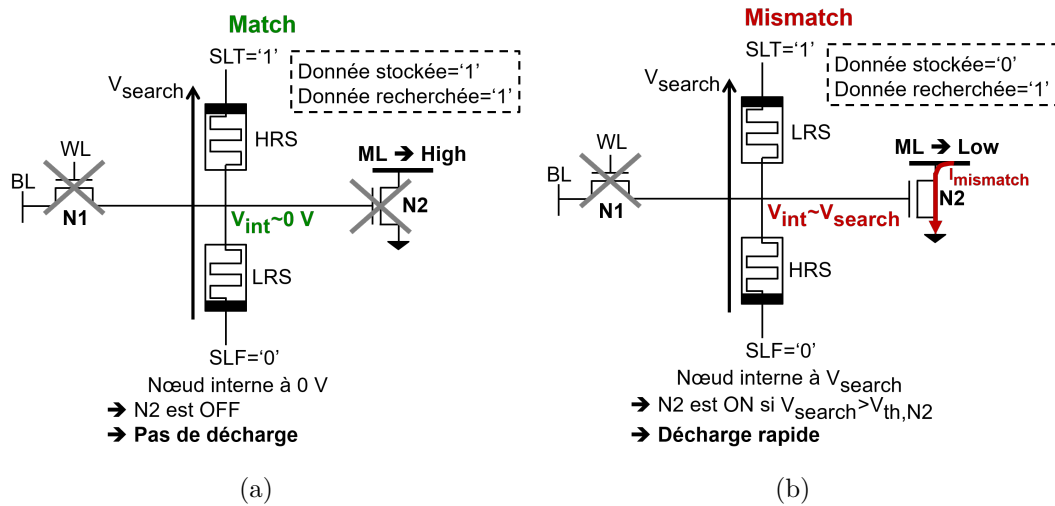


FIGURE 3.3.4: (a) Dans un cas de match, la tension du noeud interne, V_{int} , dans le pont diviseur de tension entre les deux RRAM reste à 0 V, le transistor N2 est OFF. (b) Dans un cas de mismatch, V_{int} est quasiment égale à V_{search} , le transistor N2 est ON si V_{search} est supérieure à la tension de seuil du transistor N2.

3.3.3 Comparaison des deux structures TCAM

Dans le cas de la TCAM 2T2R, la ligne de match est connectée aux électrodes du haut de chaque RRAM. Dans un cas de match (*cf* FIGURE 3.3.5 (a, haut)), la ligne de match se décharge à travers des RRAM en HRS avec des courants de fuite I_{match} . Dans un cas de mismatch (*cf* FIGURE 3.3.5 (a, bas)), la ligne de match se décharge à travers une ou des RRAM en LRS. *La marge de fiabilité du système*, c'est-à-dire sa capacité à distinguer un cas de match du pire cas de mismatch quand une seule cellule TCAM mismatch, dépend directement du rapport de courant entre le courant de mismatch, $I_{mismatch}$, et la somme des courants de fuite I_{match} dans un cas de match. La marge de fiabilité de la TCAM 2T2R dépend alors directement des valeurs de résistance en HRS et LRS des RRAM, et plus précisément de la fenêtre mémoire des RRAM (rapport entre

les valeurs de résistance en HRS et LRS) qui est de l'ordre de 10-100. Dans le cas de la TCAM 1T2R1T (*cf* FIGURE 3.3.5 (b, haut)), la ligne de match est connectée aux drains de chaque transistor N2. Dans un cas de match, la ligne de match se décharge à travers des transistors dans l'état OFF. Dans un cas de mismatch (*cf* FIGURE 3.3.5 (b, bas)), la ligne de match se décharge à travers un ou des transistors dans l'état ON. La marge de fiabilité du système dépend alors du rapport de courant I_{ON} et I_{OFF} des transistors N2, qui peut atteindre jusqu'à six ordres de grandeur (*cf* FIGURE 3.3.5 (c)).

3.3.4 Intégration et fabrication des deux circuits TCAM à base de RRAM

La FIGURE 3.3.6 (a) schématise les deux circuits TCAM fabriqués. Un registre à décalage envoie le mot recherché de 128 bits en entrée d'une matrice TCAM 3x128 bits. Seule la ligne de match de la TCAM du milieu est connectée à un circuit de lecture. Le circuit de lecture évalue la décharge de la ligne de match pendant une opération de recherche en comparant sa tension à une tension de référence, V_{REF} . Le résultat de la comparaison est fourni en sortie du comparateur, et permet d'évaluer le temps t_{search} nécessaire pour décharger la ligne de match de l'état haut vers la tension de référence, V_{REF} . Pour chaque circuit, les mesures sont effectuées sur la TCAM du milieu. Les FIGURE 3.3.6 (b) et (c) montrent des photographies des circuits fabriqués.

Les RRAM ont été intégrées dans le retour en fin de ligne (back-end-of-line) d'un procédé CMOS 130 nm [25], et sont composées d'un empilage TiN/HfO₂/Ti/TiN d'épaisseur 100 nm/10 nm/10 nm/100 nm (*cf* FIGURE 3.3.6 (d)). Pour étudier l'impact des propriétés électriques des RRAM sur les performances et fiabilité de chaque circuit TCAM, les RRAM ont été programmées avec différentes conditions de programmation. La FIGURE 3.3.7 montre les distributions cumulées de résistance mesurées directement sur les circuits TCAM. Les différentes distributions HRS ont été obtenues soit avec la condition de programmation Soft HRS, soit avec la condition de programmation Strong HRS, soit en gardant les RRAM dans l'état vierge "pristine". La fenêtre mémoire est définie ici comme le rapport de résistance entre la valeur de résistance du HRS et LRS à -2σ et $+2\sigma$ de chaque distribution, respectivement.

3.4 Caractérisations électriques des circuits TCAM à base de RRAM

3.4.1 Fonctionnalité de base des circuits : caractérisation du temps de décharge de la ligne de match

La FIGURE 3.4.1 montre les temps de décharge, t_{search} , en fonction de la tension appliquée aux bornes des RRAM pendant une opération de recherche, V_{search} , mesurés sur le circuit TCAM 2T2R (a) et 1T2R1T (b). La ligne de match se

3.4. CARACTÉRISATIONS ÉLECTRIQUES DES CIRCUITS TCAM À BASE DE RRAM

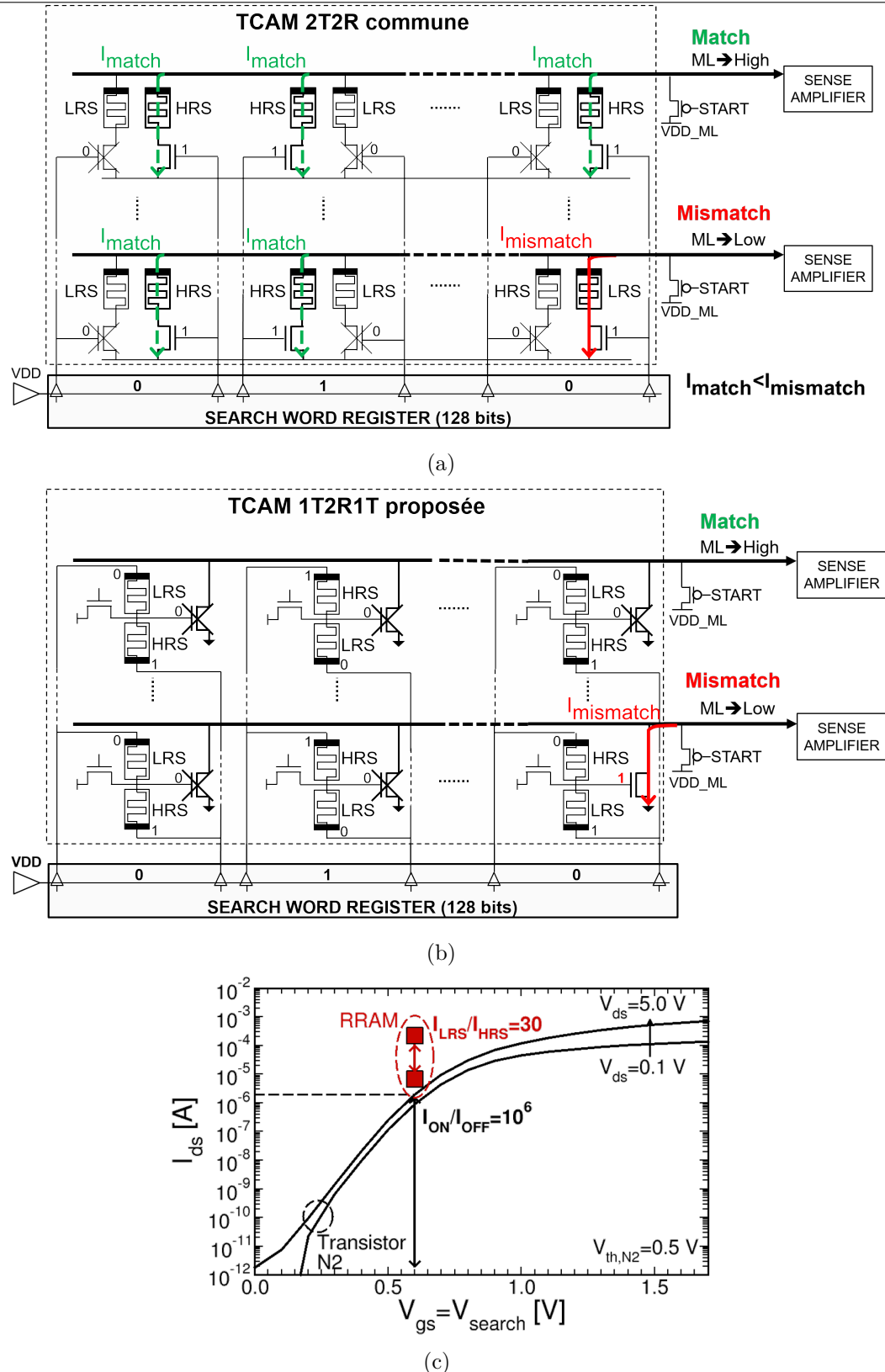


FIGURE 3.3.5: Cas de match et mismatch pour (a) la structure 2T2R commune et (b) la nouvelle structure proposée 1T2R1T. (c) Caractéristique I_{ds} - V_{gs} mesurée sur les transistors N2 de la TCAM 1T2R1T.

décharge dans un cas de mismatch pour les deux TCAM (rouge), et la décharge s'accélère avec le nombre de cellules TCAM qui mismatchent. Dans un cas de match, la ligne de match doit idéalement rester à l'état haut. Comme attendu, c'est le cas pour la TCAM 1T2R1T (vert). En revanche, la ligne de match se

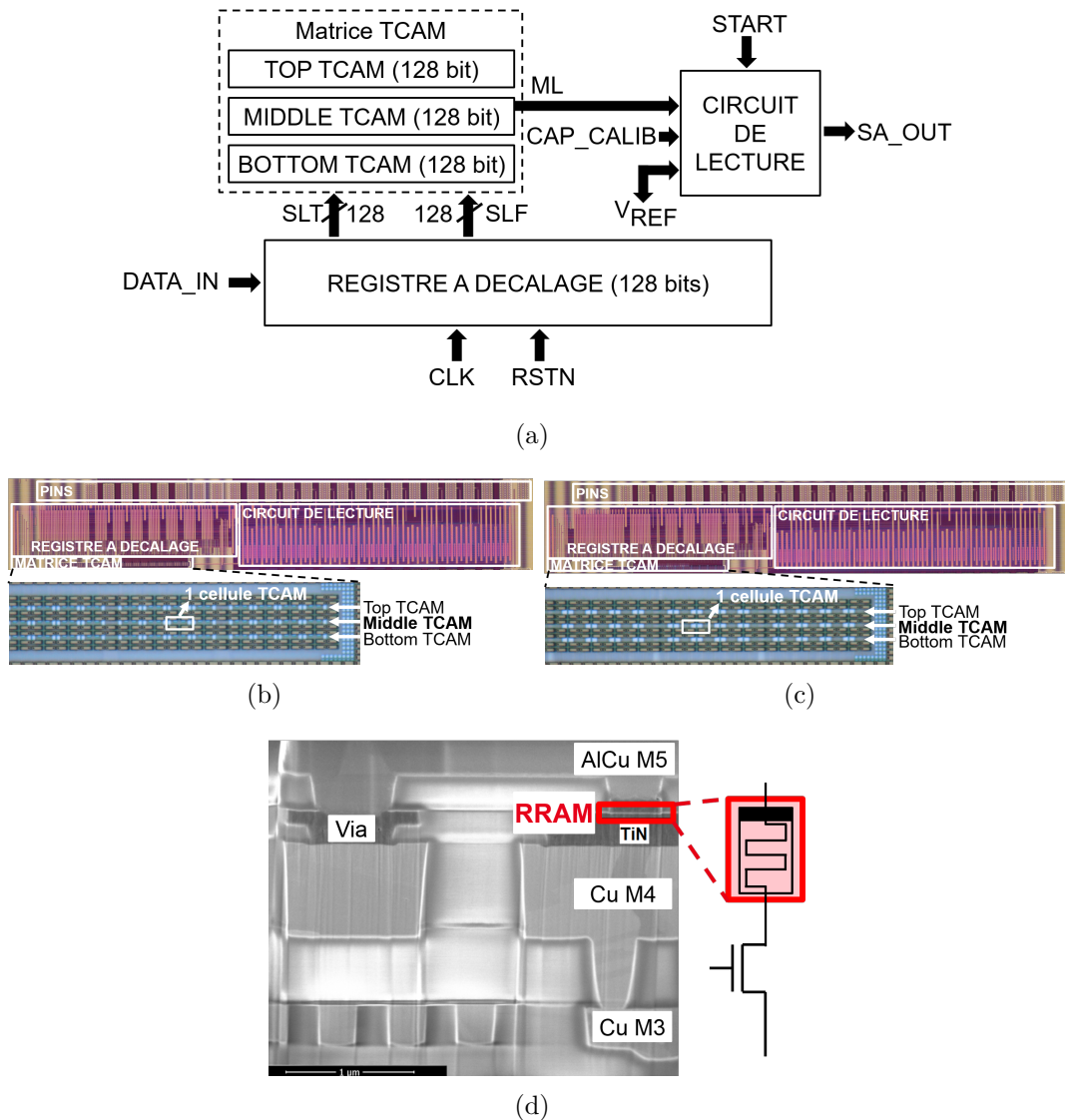


FIGURE 3.3.6: (a) Schéma bloc des circuits TCAM 2T2R et 1T2R1T fabriqués. (b) Photographies des circuits 2T2R et (c) 1T2R1T fabriqués. (d) Image par microscopie électronique des cellules RRAM à base de HfO₂ intégrées.

décharge dans un cas de match pour la TCAM 2T2R à cause de la faible fenêtre mémoire des RRAM.

3.4.2 Marge de détection et capacité de recherche

Pour quantifier l'impact des propriétés électriques des RRAM sur la fiabilité des circuits TCAM, nous avons mesuré la marge de détection lorsque les TCAM sont programmées avec les différentes conditions de programmation (Soft HRS, Strong HRS, ou état vierge "pristine"). Dans ce travail, nous avons défini la marge de détection comme le rapport de temps de décharge, t_{search} , (TR pour Time Ratio) entre un cas de match et le pire cas de mismatch quand une seule cellule TCAM mismatch. La FIGURE 3.4.2 (a) montre la marge de détection, TR, en fonction de la fenêtre mémoire, MW. Comme attendu, la marge de détection

3.4. CARACTÉRISATIONS ÉLECTRIQUES DES CIRCUITS TCAM À BASE DE RRAM

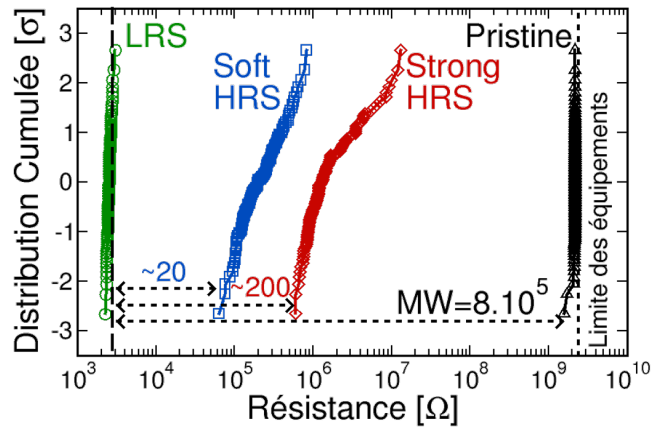


FIGURE 3.3.7: Distributions cumulées de LRS, Soft HRS, Strong HRS, et état vierge "pristine" utilisées dans ce travail, mesurées directement sur les circuits TCAM.

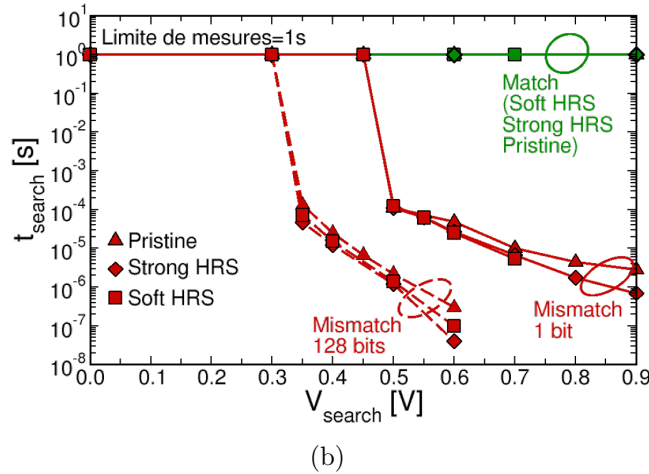
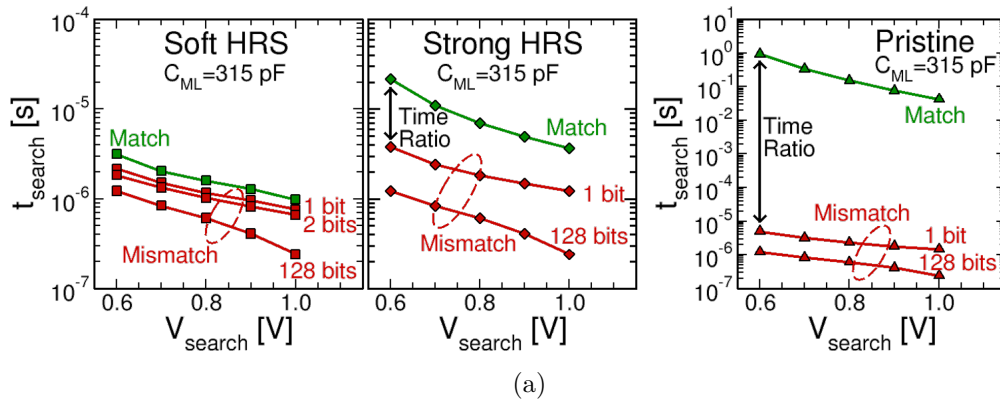
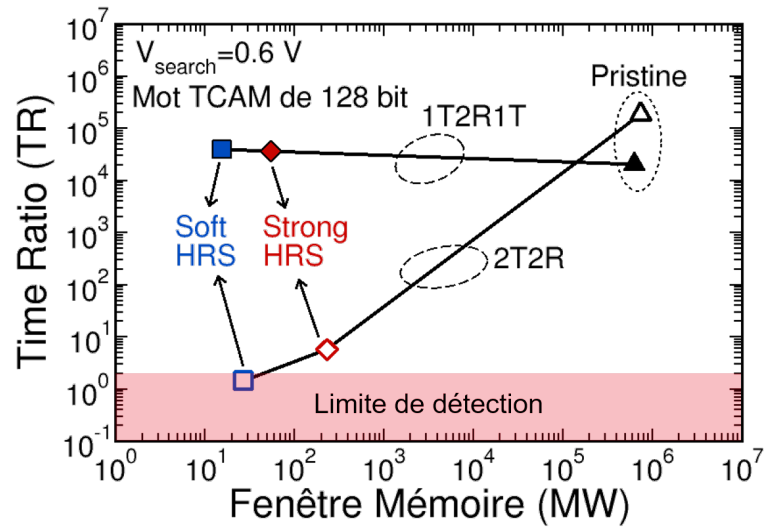
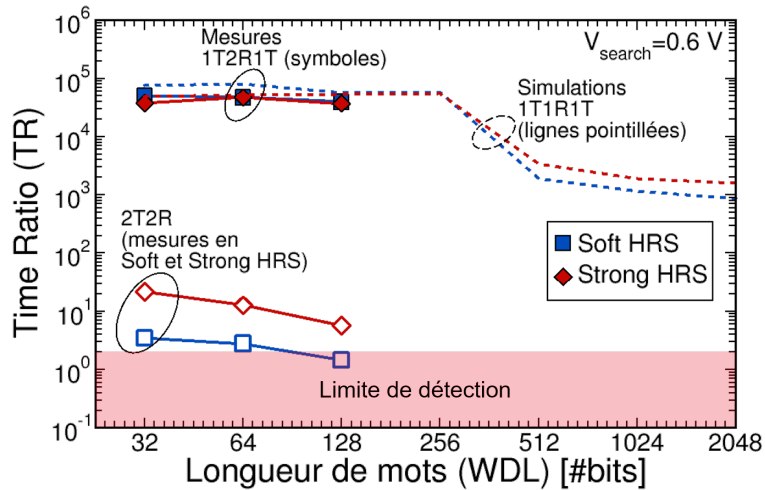


FIGURE 3.4.1: Temps de décharge, t_{search} , en fonction de la tension appliquée aux bornes des RRAM, V_{search} , mesurés sur (a) le circuit TCAM 2T2R et (b) le circuit TCAM 1T2R1T. Les RRAM ont été programmées avec les différentes conditions de programmation de la FIGURE 3.3.7.

de la TCAM 2T2R augmente avec la fenêtre mémoire car les courants de fuite, I_{match} , sont diminués. Dans le cas de la TCAM 1T2R1T, la marge de détection est indépendante de la fenêtre mémoire car elle ne dépend que des courants I_{ON} et I_{OFF} des transistors N2. De plus, on observe une augmentation drastique de



(a)



(b)

FIGURE 3.4.2: Marges de détection, TR (pour Time Ratio), en fonction de (a) la fenêtre mémoire, MW, et (b) la longueur de mots TCAM, WDL, mesurées sur les circuits TCAM 2T2R (symboles ouverts) et 1T2R1T (symboles pleins).

la marge de détection de la TCAM 1T2R1T par rapport à la TCAM 2T2R. Si on considère une marge de détection minimale de 2 pour garantir la fiabilité des circuits TCAM, une fenêtre mémoire d'au moins 50 est nécessaire pour la TCAM 2T2R, et une TCAM 2T2R de 128 bit ne peut pas être programmée en Soft HRS. Il est alors nécessaire d'augmenter la fenêtre mémoire pour travailler avec des mots TCAM plus longs. Ceci amène au deuxième problème majeur de la TCAM 2T2R : la faible fenêtre mémoire des RRAM limite les longueurs de mots TCAM. Pour quantifier ce deuxième problème, nous avons tracé en FIGURE 3.4.2 (b) la marge de détection, TR, en fonction de la longueur de mots TCAM, WDL (pour Word Length). La marge de détection de la TCAM 2T2R diminue avec la longueur de mot TCAM, ce qui limite les mots TCAM à environ 100 bit et 256 bit lorsque la TCAM 2T2R est programmée en Soft HRS et Strong HRS, respectivement. Dans le cas de la TCAM 1T2R1T, la marge

de détection est peu sensible à la longueur de mots TCAM, ce qui permet de travailler avec des mots TCAM bien plus longs jusqu'à plus de 2 kbit.

3.4.3 Caractérisation de l'endurance en recherche

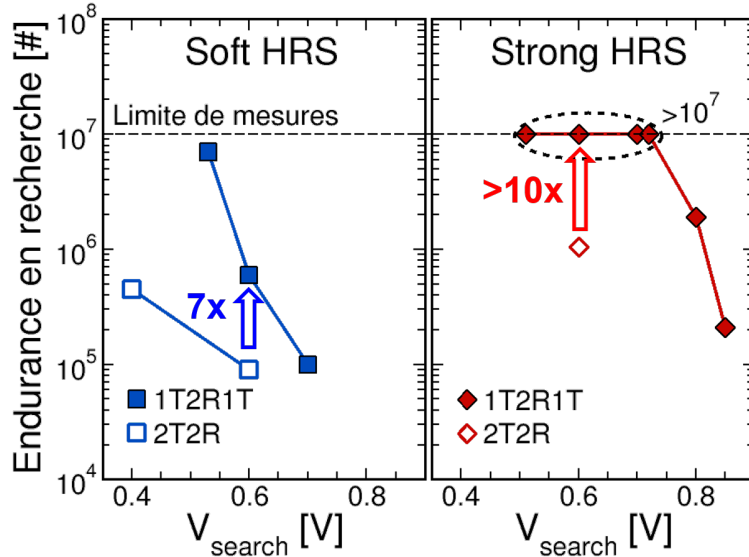


FIGURE 3.4.3: Endurances en recherche en fonction de la tension appliquée aux bornes des RRAM, V_{search} , mesurées sur les circuits TCAM 2T2R (symboles ouverts) et 1T2R1T (symboles pleins).

Pendant une opération de recherche, la tension positive V_{search} est appliquée sur l'électrode du haut d'une des deux RRAM (R_X pour la recherche de '1', R_Y pour la recherche de '0') dans la même polarité qu'une opération de Set. Dans le cas d'un match, V_{search} est appliquée aux bornes de RRAM en HRS, il y a alors risque de commuter les RRAM en LRS pendant les opérations de recherche, et de perdre l'état de match. Nous avons caractérisé l'endurance en recherche de chaque circuit TCAM en appliquant une série d'opérations de recherche lorsque les TCAM sont initialement dans une configuration de match. L'endurance en recherche a été définie comme le nombre maximal d'opérations de recherche avant qu'au moins une RRAM commute du HRS vers le LRS. La FIGURE 3.4.3 compare les endurances en recherche de chaque circuit TCAM en fonction de la tension V_{search} . On observe une amélioration de l'endurance en recherche avec la nouvelle cellule TCAM 1T2R1T par rapport à la TCAM 2T2R.

3.5 Conclusion

Dans ce chapitre, nous avons expérimentalement caractérisé deux circuits TCAM à base de RRAM : (i) la structure TCAM la plus commune 2T2R [115–119], et (ii) une nouvelle cellule TCAM dans une configuration un-transistor/deux-RRAM/un-transistor (1T2R1T). Ces travaux avancent l'état de l'art en proposant pour la première fois des caractérisations électriques complètes de circuits TCAM à base de RRAM. Comme attendu, les résultats obtenus montrent une

amélioration drastique de la marge de détection de la TCAM 1T2R1T par rapport à la TCAM 2T2R, ce qui permet de travailler avec des mots TCAM de plus de 2 kbit. De plus, l'endurance en recherche est également améliorée avec la TCAM 1T2R1T. En termes d'endurance en programmation, la technologie de RRAM intégrée dans ce travail peut aller jusqu'à 10^6 et 10^4 opérations de programmation lorsque les TCAM sont programmées en Soft HRS et Strong HRS, respectivement [25], ce qui est suffisant pour la plupart des applications TCAM [118]. Pour des applications TCAM classiques, par exemple le routage de paquets internet [123], il est généralement nécessaire de travailler avec des mots TCAM de plus de 128 bit, ce qui rend la structure 2T2R non adaptée à ce type d'applications. Pour des applications neuromorphiques, plus précisément pour l'implémentation de tables de routage synaptique, les adresses des neurones ne dépassent généralement pas 32 bit [62]. Les deux structures TCAM sont alors adaptées en termes de longueur de mot. Cependant, l'endurance en recherche n'est probablement pas assez élevée pour permettre aux processeurs neuromorphiques d'opérer continuellement.

Intégration tri-dimensionnelle monolithique de deux niveaux de transistors CMOS hautes performances avec un niveau de dispositifs de mémoires résistives

4.1 Objectif de ce chapitre

CE chapitre conclut les travaux présentés dans ce manuscrit de thèse de doctorat en ouvrant des perspectives technologiques pour améliorer l'efficacité en surface des processeurs neuromorphiques impulsionnels. Pour cela, nous démontrerons l'*intégration hétérogène 3D monolithique de deux niveaux de transistors CMOS et un niveau de mémoires résistives (RRAM pour Resistive Random Access Memory) en procédé CMOS SOI 65 nm*. Les dispositifs ont été intégrés, fabriqués, et sont électriquement fonctionnels. En particulier, les dispositifs présentés dans ce travail implémentent les mêmes structures que celles étudiées dans les CHAPITRE 2 et CHAPITRE 3.

4.2 Intégration tri-dimensionnelle monolithique de mémoires résistives et transistors CMOS

4.2.1 La technologie CoolCubeTM

Les deux niveaux de transistors ont été intégrés avec la technologie CoolCubeTM du CEA-Leti [59]. Cette technologie permet l'intégration de deux niveaux de

transistors CMOS hautes performances en CMOS SOI 65 nm. La FIGURE 4.2.1 synthétise le flux du processus d'intégration.

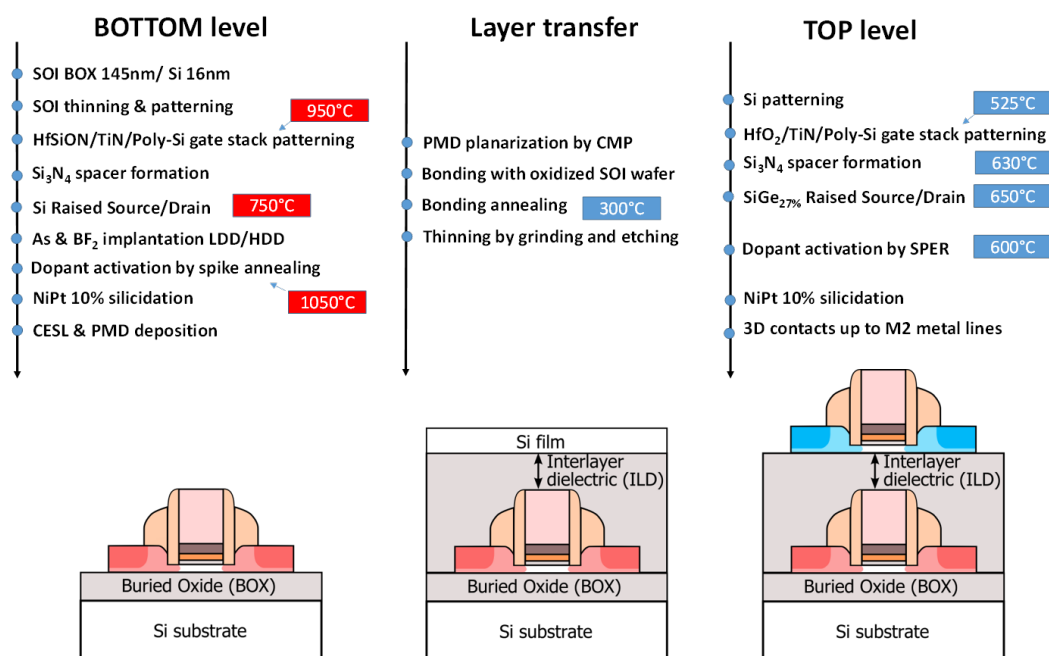


FIGURE 4.2.1: Flux du processus d'intégration de la technologie CoolCube™ du CEA-Leti [59].

4.2.2 Intégration des mémoires résistives

Les RRAM ont été intégrées dans le retour en fin de ligne (back-end-of-line) des transistors CMOS fabriqués avec la technologie CoolCube™. La FIGURE 4.2.2 schématise le flux du processus d'intégration. Nous avons tout d'abord récupéré des plaques CoolCube™ juste avant le premier niveau de lignes métalliques (a). Par une première photo-lithographie par faisceau électronique, les dispositifs RRAM ont été intégrés au-dessus des contacts des transistors (b). Les RRAM sont composées d'un empilage TiN/HfO₂/Ti/TiN d'épaisseur 10 nm/5 nm/5 nm/30 nm, similaires aux RRAM étudiées dans les CHAPITRE 2 et CHAPITRE 3 [25]. Après remplissage espaceur et oxide, les contacts sont récupérés au-dessus des RRAM par une deuxième photo-lithographie par faisceau électronique (c). Enfin, l'intégration est achevée par un niveau de retour en fin de ligne standard (d). La FIGURE 4.2.3 montre une image obtenue par microscopie électronique des dispositifs fabriqués.

4.3. CARACTÉRISATIONS ÉLECTRIQUES DE L'INTÉGRATION TRI-DIMENSIONNELLE MONOLITHIQUE DE DEUX NIVEAUX DE TRANSISTORS NMOS ET UN NIVEAU DE MÉMOIRES RÉSTIVES

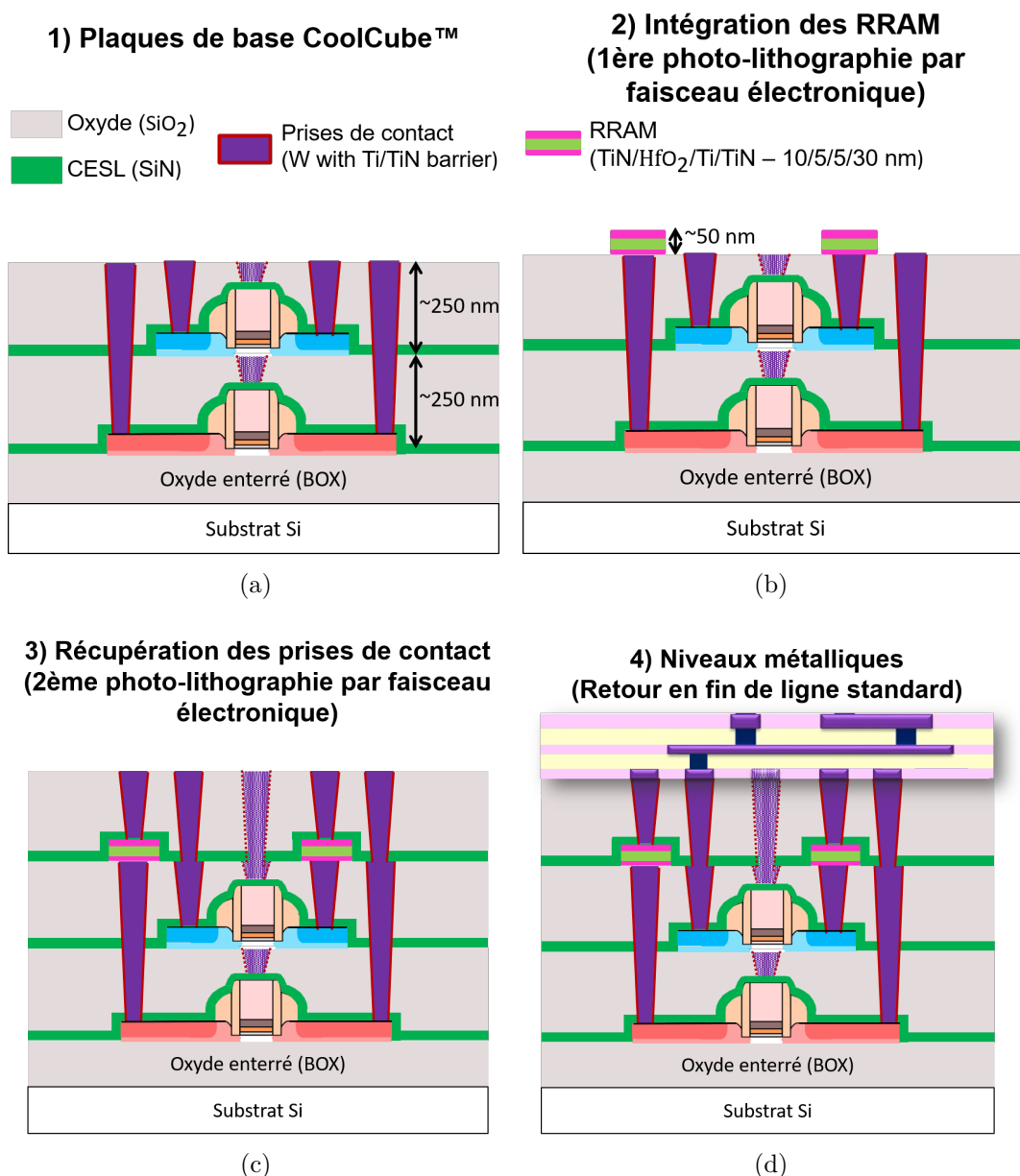


FIGURE 4.2.2: (a) Illustration schématique des plaques CoolCube™ de base, avant intégration des RRAM. (b) Intégration des dispositifs RRAM par une première photo-lithographie par faisceau électronique. (c) Récupération des prises de contact par une seconde photo-lithographie par faisceau électronique. (d) Ajout des niveaux métalliques par un processus de retour en fin de ligne standard.

4.3 Caractérisations électriques de l'intégration tri-dimensionnelle monolithique de deux niveaux de transistors NMOS et un niveau de mémoires résistives

Les structures obtenues sont équivalentes à deux structures un-transistor/une-RRAM (1T1R) en parallèle. Une RRAM est connectée au transistor NMOS du

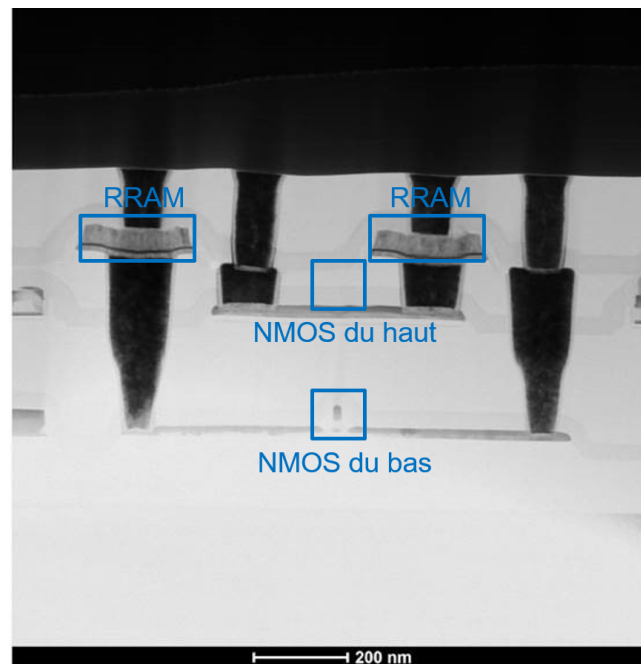


FIGURE 4.2.3: Image par microscopie électronique des dispositifs fabriqués par intégration 3D monolithique, avec deux niveaux de transistors et un niveau de RRAM.

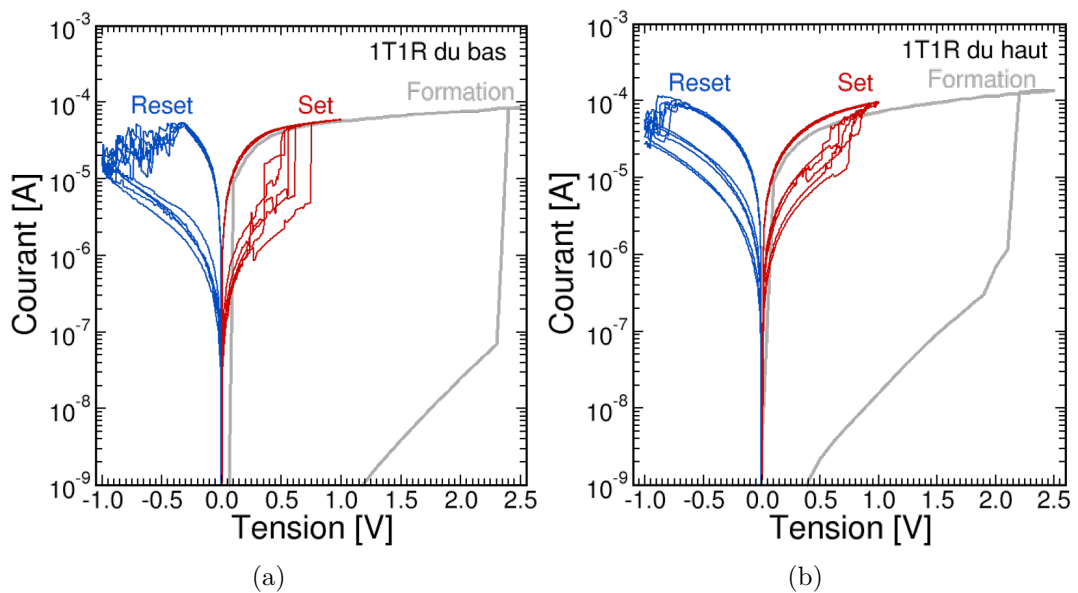


FIGURE 4.3.1: Caractérisations électriques I-V en quasi-statique de (a) la 1T1R du bas et (b) la 1T1R du haut.

bas - la 1T1R du bas -, et une RRAM est connectée au transistor NMOS du haut - la 1T1R du haut. Nous avons caractérisé électriquement chaque structure 1T1R indépendamment. La FIGURE 4.3.1 montre la caractérisation électrique I-V en quasi-statique de la 1T1R du bas (a) et de la 1T1R du haut (b). En appliquant une tension positive sur l'électrode du haut des RRAM, il est possible de former et commuter vers l'état de basse résistance (LRS pour Low Resistance State) les mémoires. En appliquant une tension négative sur l'électrode du haut des RRAM, il est possible de commuter vers l'état de haute résistance (HRS

4.3. CARACTÉRISATIONS ÉLECTRIQUES

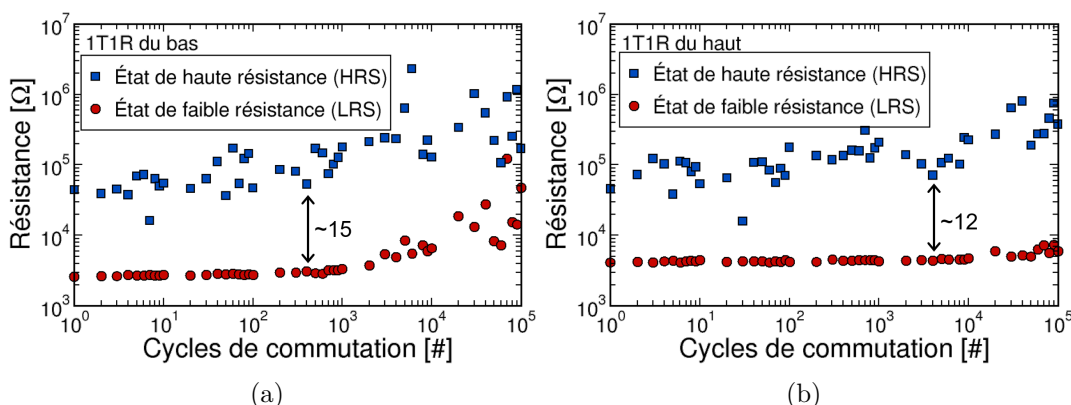


FIGURE 4.3.2: Caractérisations électriques en mesures pulsées de (a) la 1T1R du bas et (b) la 1T1R du haut.

pour High Resistance State). Nous avons ensuite caractérisé les structures 1T1R en mesures pulsées, tracées sur la FIGURE 4.3.2 pour la 1T1R du bas (a) et la 1T1R du haut (b). Il est possible de commuter les mémoires entre le LRS et HRS pendant plus de 10^5 cycles en maintenant une fenêtre mémoire d'environ 10 avec chaque structure. Enfin, nous avons vérifié qu'il était possible de contrôler la valeur de résistance du LRS avec le courant de programmation, I_{prog} , pendant les opérations de Set. La FIGURE 4.3.3 trace la valeur de résistance du LRS en fonction du courant de programmation, I_{prog} . Les données ont été obtenues à partir de différentes technologies de mémoires résistives de la littérature [124] (symboles gris). Une des caractéristiques universelles des RRAM est qu'il existe une loi de puissance entre la valeur de résistance du LRS et le courant de programmation. Nous avons ajouté à cette courbe les données issues des mesures sur la 1T1R du bas (losanges rouges) et la 1T1R du haut (triangles bleus). Nos données concordent avec les données de la littérature, ce qui valide la fonctionnalité électrique basique complète de notre intégration.

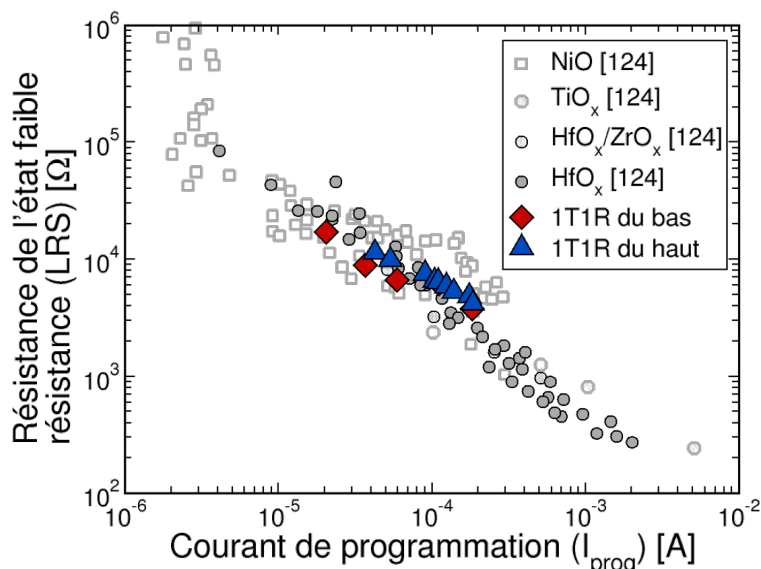


FIGURE 4.3.3: Valeurs de résistance de l'état faible résistance (LRS pour Low Resistance State) en fonction du courant de programmation, I_{prog} .

4.4 Conclusion

Dans ce chapitre, nous avons démontré la première intégration hétérogène 3D monolithique de deux niveaux de transistors CMOS et un niveau de RRAM, jusqu'alors inexistante dans la littérature. Des structures classiques 1T1R ont été intégrées, fabriquées, et sont électriquement fonctionnelles. L'avantage principal de cette intégration est le gain en surface silicium grâce à l'exploitation de la dimension verticale. En particulier, les structures 1T1R intégrées sont les mêmes que celles utilisées pour implémenter (i) des synapses artificielles dans le CHAPITRE 2, et (ii) des tables de routage synaptique dans le CHAPITRE 3. En première approximation, cette intégration 3D monolithique permettrait d'améliorer d'un facteur 2 la surface silicium de ces deux composantes des processeurs neuromorphiques impulsionsnels.

Conclusion et perspectives

L'objectif de ce travail de thèse de doctorat était d'évaluer l'utilisation des mémoires résistives et technologies 3D monolithiques pour permettre l'implémentation matérielle compacte et efficace énergétiquement de processeurs neuromorphiques impulsionsnels et reconfigurables. Pour cela, l'étude était focalisée sur deux composantes de base des processeurs neuromorphiques impulsionsnels: (i) les matrices synaptiques avec poids synaptiques ajustables, et (ii) les tables de routage synaptique, toutes deux implémentées avec des mémoires résistives (RRAM pour Resistive Random Access Memory). Dans le CHAPITRE 2, nous avons étudié de façon détaillée l'impact des propriétés électriques des RRAM (fenêtre mémoire, variabilité conductive, et endurance en programmation) sur les performances d'apprentissage de réseaux de neurones impulsionsnels (SNN pour Spiking Neural Network) avec synapses à base de RRAM et entraînés de façon non supervisée par STDP. Nous avons notamment clarifié le rôle de la variabilité synaptique, qui provient de la variabilité conductive des RRAM, sur les performances d'apprentissage, et avons démontré que la variabilité synaptique pouvait être bénéfique. L'une des raisons est qu'elle permet aux synapses d'accéder à une plus grande plage de valeurs de poids synaptiques pendant l'apprentissage. Dans le CHAPITRE 3, nous avons évalué l'utilisation de mémoires ternaires adressables par contenu (TCAM pour Ternary Content-Addressable Memory) à base de RRAM pour implémenter des tables de routage synaptique. Pour cela, deux circuits TCAM à base de RRAM ont été intégrés, fabriqués, et des caractérisations électriques poussées ont été effectuées sur chaque circuit. Notamment, nous avons expérimentalement démontré que la structure TCAM la plus commune, qui est également la plus petite, faite de deux structures un-transistor/une-RRAM (2T2R), ainsi que la nouvelle structure proposée dans ce travail un-transistor/deux-RRAM/un-transistor (1T2R1T), de surface silicium similaire à celle de la 2T2R, sont toutes deux adaptées en termes de performances et fiabilité pour implémenter les tables de routage synaptique. Cependant, l'endurance en recherche doit encore être améliorée pour

permettre aux processeurs neuromorphiques impulsionsnels d'opérer continuellement sans défaillance. Enfin, dans le CHAPITRE 4, nous avons démontré la première intégration hétérogène 3D monolithique de deux niveaux de transistors CMOS hautes performances avec un niveau de RRAM en technologie CMOS SOI 65 nm. Notamment, des structures classiques 1T1R ont été intégrées, fabriquées, et sont électriquement fonctionnelles. Cette démonstration technologique ouvre des perspectives pour améliorer l'efficacité en surface des deux composants des processeurs neuromorphiques impulsionsnels étudiées dans les deux chapitres précédents. En première approximation, la surface silicium pourrait être améliorée d'un facteur 2.

En termes de perspectives, les travaux futurs devraient couvrir les sujets suivants:

- Une meilleure compréhension de la physique des RRAM pour bénéficier autant que possible des propriétés électriques des RRAM à la fois pour des applications mémoires et des cœurs neuromorphiques.
- Identification de modèles de matrices synaptiques associées à une circuiterie d'apprentissage appropriée pour permettre une implémentation matérielle plus efficace et adaptée, ainsi que l'apprentissage en ligne et en temps réel.
- Développement d'un schéma et circuiterie de programmation adaptés des RRAM pour la TCAM 1T2R1T.
- Évaluation minutieuse des opportunités offertes par l'intégration 3D monolithique, ainsi que des coûts associés pour bénéficier au mieux de cette intégration au niveau circuit.
- Identification de modèles de circuits de neurones efficaces : les modèles de neurones actuels à base de CMOS consomment une grande partie de la surface silicium dans les cœurs neuromorphiques.

Références : Résumé en français

- [1] John Von Neumann and Michael D. Godfrey. “First Draft of a Report on the EDVAC”. *IEEE Annals of the History of Computing*, 15(4):27–75, aug 1993. ISSN 10586180. doi: 10.1109/85.238389.
- [2] W. McCulloch and W. Pitts. “A logical calculus of the ideas immanent in nervous activity (reprinted from 1943)”. *Bulletin of Mathematical Biology*, 52(1/2):99–115, 1990.
- [3] M D Godfrey and D F Hendry. “The Computer as von Neumann Planned It”. *IEEE Annals of the History of Computing*, 15(1):11–21, 1993. ISSN 10586180. doi: 10.1109/85.194088.
- [4] J. Backus. “Can Programming Be Liberated from the von Neumann Style? A Functional Style and Its Algebra of Programs”. *Communications of the ACM*, 21(8):613–641, 1978. doi: 10.1145/359576.359579.
- [5] Carver Mead. “Neuromorphic Electronic Systems”. *Proceedings of the IEEE*, 78(10):1629–1636, 1990.
- [6] Thomas N Theis and H.-S Philip Wong. “The End of Moore’s Law: A New Beginning for Information Technology”. *Computing in Science Engineering*, 19(2):41–50, 2017. doi: 10.1109/MCSE.2017.29.
- [7] Massimiliano Di Ventra and Yuriy V. Pershin. “The parallel approach”. *Nature Physics*, 9(4):200–202, apr 2013. ISSN 17452481. doi: 10.1038/nphys2566.
- [8] H. S. Philip Wong and Sayeef Salahuddin. “Memory leads the way to better computing”. *Nature Nanotechnology*, 10(3):191–194, 2015. ISSN 17483395. doi: 10.1038/nnano.2015.29.
- [9] M. T. Bohr. “Interconnect scaling—the real limiter to high performance ULSI”. In *Proceedings of International Electron Devices Meeting*, pages 241–244. Institute of Electrical and Electronics Engineers (IEEE), 1995. doi: 10.1109/iedm.1995.499187.
- [10] Saber Moradi and Rajit Manohar. “The impact of on-chip communication on memory technologies for neuromorphic systems”. *Journal of Physics D: Applied Physics*, 52(1):1–25, 2019. ISSN 13616463. doi: 10.1088/1361-6463/aae641.
- [11] Hongsik Jeong and Luping Shi. “Memristor devices for neural networks”. *Journal of Physics D: Applied Physics*, 52(2):023003, 2019. ISSN 13616463. doi: 10.1088/1361-6463/aae223.
- [12] Yuan Taur, Douglas A. Buchanan, et al. “CMOS scaling into the nanometer regime”. *Proceedings of the IEEE*, 85(4):486–503, 1997. ISSN 00189219. doi: 10.1109/5.573737.
- [13] Yen Kuang Chen, Jatin Chhugani, et al. “Convergence of recognition, mining, and synthesis workloads and its implications”. *Proceedings of the IEEE*, 96(5):790–807, 2008. ISSN 00189219. doi: 10.1109/JPROC.2008.917729.

RÉFÉRENCES : RÉSUMÉ EN FRANÇAIS

- [14] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. *Nature*, 521(7553): 436–444, 2015. ISSN 14764687. doi: 10.1038/nature14539.
- [15] Damien Querlioz, Olivier Bichler, Adrien Francis Vincent, and Christian Gamrat. “Bioinspired Programming of Memory Devices for Implementing an Inference Engine”. *Proceedings of the IEEE*, 103(8):1398–1416, aug 2015. ISSN 00189219. doi: 10.1109/JPROC.2015.2437616.
- [16] Igor L Markov. “Limits on fundamental limits to computation”. *Nature*, 512(7513): 147–154, 2014. ISSN 14764687. doi: 10.1038/nature13570.
- [17] Chengning Wang, Dan Feng, et al. “Cross-point Resistive Memory: Nonideal properties and solutions”. *ACM Transactions on Design Automation of Electronic Systems*, 24(4): 1–37, jun 2019. ISSN 15577309. doi: 10.1145/3325067.
- [18] Carver Mead. *Analog VLSI and Neural Systems*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989. ISBN 0-201-05992-4.
- [19] M. M. Sabry Aly, M. Gao, et al. “Energy-Efficient Abundant-Data Computing: The N3XT 1,000x”. *Computer*, 48(12):24–33, 2015. doi: 10.1109/MC.2015.376.
- [20] T.J. Biggerstaff. “Moore’s law: Change or Die”. *IEEE Software*, 13(1):4, 1996. doi: 10.1109/MS.1996.476277.
- [21] Wolfgang Arden, Michel Brillouet, et al. “”More-than-Moore” (ITRS Whitepaper)”. Technical report, International Technology Roadmap for Semiconductors (ITRS). URL http://www.itrs2.net/uploads/4/9/7/7/49775221/irc-itrs-mtm-v2{_}3.pdf.
- [22] Paul A. Merolla, John V. Arthur, et al. “A million spiking-neuron integrated circuit with a scalable communication network and interface”. *Science*, 345(6197):668–673, 2014. ISSN 10959203. doi: 10.1126/science.1254642.
- [23] Ximeng Guan, Shimeng Yu, and H. S.Philip Wong. “On the switching parameter variation of metal-oxide RRAM - Part I: Physical modeling and simulation methodology”. *IEEE Transactions on Electron Devices*, 59(4):1172–1182, 2012. ISSN 00189383. doi: 10.1109/TED.2012.2184545.
- [24] P. Gonon, C. Vallee, C. Mannequin, M. Saadi, and .F Jomni. “Mechanisms of resistance switching in nanometric metal oxides and their dependence on electrodes”. In *Proceedings of the IEEE International Conference on Properties and Applications of Dielectric Materials*, volume 2015-October, pages 56–59, 2015. ISBN 9781479989034. doi: 10.1109/ICPADM.2015.7295207.
- [25] A. Grossi, E. Nowak, et al. “Fundamental variability limits of filament-based RRAM”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 4.7.1–4.7.4, 2016. ISBN 9781509039012. doi: 10.1109/IEDM.2016.7838348.
- [26] Deok Hwang Kwon, Kyung Min Kim, et al. “Atomic structure of conducting nanofilaments in TiO₂ resistive switching memory”. *Nature Nanotechnology*, 5(2):148–153, 2010. ISSN 17483395. doi: 10.1038/nnano.2009.456.
- [27] Yuchao Yang, Peng Gao, et al. “Observation of conducting filament growth in nanoscale resistive memories”. *Nature Communications*, 3(1):732, 2012. ISSN 20411723. doi: 10.1038/ncomms1737.
- [28] Y. Y. Chen, R. Roelofs, et al. “Tailoring switching and endurance / retention reliability characteristics of HfO₂ / Hf RRAM with Ti, Al, Si dopants”. In *Digest of Technical Papers - Symposium on VLSI Technology*, pages 1–2. IEEE, 2014. ISBN 9781479933310. doi: 10.1109/VLSIT.2014.6894403.

- [29] R. Degraeve, A. Fantini, et al. “Quantitative endurance failure model for filamentary RRAM”. In *Digest of Technical Papers - Symposium on VLSI Technology*, pages T188–T189. JSAP, 2015. ISBN 9784863485013. doi: 10.1109/VLSIT.2015.7223673.
- [30] C. Nail, G. Molas, et al. “Understanding RRAM endurance, retention and window margin trade-off using experimental results and simulations”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 4.5.1–4.5.4, 2016. ISBN 9781509039012. doi: 10.1109/IEDM.2016.7838346.
- [31] H. S.Philip Wong, Heng Yuan Lee, et al. “Metal-oxide RRAM”. *Proceedings of the IEEE*, 100(6):1951–1970, 2012. ISSN 00189219. doi: 10.1109/JPROC.2012.2190369.
- [32] Geoffrey W Burr, Robert M Shelby, et al. “Neuromorphic computing using non-volatile memory”. *Advances in Physics: X*, 2(1):89–124, 2017. ISSN 23746149. doi: 10.1080/23746149.2016.1259585.
- [33] Changhyuck Sung, Hyunsang Hwang, and In Kyeong Yoo. “Perspective: A review on memristive hardware for neuromorphic computation”. *Journal of Applied Physics*, 124(15):151903, oct 2018. ISSN 10897550. doi: 10.1063/1.5037835.
- [34] Hsinyu Tsai, Stefano Ambrogio, Pritish Narayanan, Robert M Shelby, and Geoffrey W Burr. “Recent progress in analog memory-based accelerators for deep learning”. *Journal of Physics D: Applied Physics*, 51(28):283001, 2018. doi: 10.1088/1361-6463/aac8a5.
- [35] A Fumarola, P Narayanan, et al. “Accelerating machine learning with Non-Volatile Memory: Exploring device and circuit tradeoffs”. In *2016 IEEE International Conference on Rebooting Computing (ICRC)*, pages 1–8, oct 2016. doi: 10.1109/ICRC.2016.7738684.
- [36] Robert Legenstein. “Nanoscale connections for brain-like circuits”. *Nature*, 521(7550): 37–38, 2015. doi: 10.1038/521037a.
- [37] Yue Zha, Etienne Nowak, and Jing Li. “Liquid Silicon: A Nonvolatile Fully Programmable Processing-In-Memory Processor with Monolithically Integrated ReRAM for Big Data/Machine Learning Applications”. In *IEEE Symposium on VLSI Circuits, Digest of Technical Papers*, pages C206–C207, 2019. ISBN 9784863487185. doi: 10.23919/VLSIC.2019.8778064.
- [38] Reiji Mochida, Kazuyuki Kouno, et al. “A 4M synapses integrated analog ReRAM based 66.5 TOPS/W neural-network processor with cell current controlled writing and flexible network architecture”. *Digest of Technical Papers - Symposium on VLSI Technology*, 2018-June:175–176, 2018. ISSN 07431562. doi: 10.1109/VLSIT.2018.8510676.
- [39] Teng Zhang, Ke Yang, et al. “Memristive Devices and Networks for Brain-Inspired Computing”. *Physica Status Solidi - Rapid Research Letters*, 13(8):1900029, 2019. ISSN 18626270. doi: 10.1002/pssr.201900029.
- [40] Young Bae Kim, Seung Ryul Lee, et al. “Bi-layered RRAM with unlimited endurance and extremely uniform switching”. In *Digest of Technical Papers - Symposium on VLSI Technology*, pages 52–53. IEEE, 2011. ISBN 9784863481640.
- [41] Chung-wei Hsu, I-ting Wang, Chun-li Lo, Ming-chung Chiang, and Wen-yueh Jang. “Self-rectifying bipolar TaOx/TiO2 RRAM with superior endurance over 1012 cycles for 3D high-density storage-class memory”. In *Digest of Technical Papers - Symposium on VLSI Technology*, pages T166–T167, 2013. ISBN 9784863483477.
- [42] K. Tsunoda, K. Kinoshita, et al. “Low power and high speed switching of Ti-doped NiO ReRAM under the unipolar voltage source of less than 3 V”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 767–770, 2007. doi: 10.1109/IEDM.2007.4419060.

- [43] H. Y. Lee, Y. S. Chen, et al. “Evidence and solution of over-RESET problem for HfOx based resistive memory with sub-ns switching speed and high endurance”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 19.7.1–19.7.4, 2010. ISBN 9781424474196. doi: 10.1109/IEDM.2010.5703395.
- [44] L. Goux, A. Fantini, et al. “Role of the Ta scavenger electrode in the excellent switching control and reliability of a scalable low-current operated TiN/Ta2O5/Ta RRAM device”. In *Digest of Technical Papers - Symposium on VLSI Technology*, pages 1–2. IEEE, 2014. ISBN 9781479933310. doi: 10.1109/VLSIT.2014.6894401.
- [45] Qing Luo, Xiaoxin Xu, et al. “Demonstration of 3D vertical RRAM with ultra low-leakage, high-selectivity and self-compliance memory cells”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 10.2.1–10.2.4, 2015. ISBN 9781467398930. doi: 10.1109/IEDM.2015.7409667.
- [46] Chung Wei Hsu, Chia Chen Wan, et al. “3D vertical TaOx/TiO2 RRAM with over 103 self-rectifying ratio and sub-uA operating current”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 10.4.1–10.4.4, 2013. ISBN 9781479923076. doi: 10.1109/IEDM.2013.6724601.
- [47] X. P. Wang, Z. Fang, et al. “Highly compact 1T-1R architecture (4F2footprint) involving fully CMOS compatible vertical GAA nano-pillar transistors and oxide-based RRAM cells exhibiting excellent NVM properties and ultra-low power operation”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 20.6.1–20.6.4, 2012. ISBN 9781467348706. doi: 10.1109/IEDM.2012.6479082.
- [48] Wanki Kim, Sung Il Park, et al. “Forming-free nitrogen-doped AlOX RRAM with sub-uA programming current”. In *Digest of Technical Papers - Symposium on VLSI Technology*, pages 22–23. IEEE, 2011. ISBN 9784863481640.
- [49] L. Goux, A. Fantini, et al. “Ultralow sub-500nA operating current high-performance TiN/Al2O3/HfO2/Hf/TiN bipolar RRAM achieved through understanding-based stack-engineering”. In *Digest of Technical Papers - Symposium on VLSI Technology*, pages 159–160, 2012. ISBN 9781467308458. doi: 10.1109/VLSIT.2012.6242510.
- [50] B. Govoreanu, G. S. Kar, et al. “10×10nm2 Hf/HfOx crossbar resistive RAM with excellent performance, reliability and low-energy operation”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 31.6.1–31.6.4. IEEE, 2011. ISBN 9781457705052. doi: 10.1109/IEDM.2011.6131652.
- [51] E. Vianello, O. Thomas, et al. “Resistive Memories for Ultra-Low-Power embedded computing design”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 6.3.1–6.3.4, 2014. ISBN 9781479980017. doi: 10.1109/IEDM.2014.7046995.
- [52] L Perniola, G Molas, et al. “Universal Signatures from Non-Universal Memories: Clues for the Future.”. In *2016 IEEE 8th International Memory Workshop, IMW 2016*, pages 1–3, 2016. ISBN 9781467388313. doi: 10.1109/IMW.2016.7495295.
- [53] H.-S. P. Wong, C. Ahn, et al. “Stanford Memory Trends”, 2018. URL <https://nano.stanford.edu/stanford-memory-trends/>.
- [54] A A Chien and V Karamcheti. “Moore’s Law: The First Ending and a New Beginning”. *Computer*, 46(12):48–53, 2013. ISSN 1558-0814. doi: 10.1109/MC.2013.431.
- [55] Cao-Minh Lu. *Fabrication de CMOS à basse température pour l’intégration 3D séquentielle*. PhD thesis, Université Grenoble Alpes, 2017.
- [56] Perrine Batude. *Intégration à trois dimensions séquentielle : Etude, fabrication et caractérisation*. PhD thesis, Institut National Polytechnique de Grenoble - INPG, 2009.

- [57] Jessy Micout. *Fabrication et caractérisation de transistor réalisée à basse température pour l'intégration 3D séquentielle*. PhD thesis, Université Grenoble Alpes, 2019.
- [58] C. Fenouillet-Beranger, B. Previtali, et al. “FDSOI bottom MOSFETs stability versus top transistor thermal budget featuring 3D monolithic integration”. In *European Solid-State Device Research Conference*, pages 110–113, 2014. ISBN 9781479943784. doi: 10.1109/ESSDERC.2014.6948770.
- [59] L. Brunet, P. Batude, et al. “First demonstration of a CMOS over CMOS 3D VLSI CoolCube™ integration on 300mm wafers”. In *Digest of Technical Papers - Symposium on VLSI Technology*, pages 1–2, 2016. ISBN 9781509006373. doi: 10.1109/VLSIT.2016.7573428.
- [60] D. Walczyk, Ch Walczyk, et al. “Resistive switching characteristics of CMOS embedded HfO₂-based 1T1R cells”. *Microelectronic Engineering*, 88(7):1133–1135, 2011. ISSN 01679317. doi: 10.1016/j.mee.2011.03.123.
- [61] Wolfgang Maass. “Networks of spiking neurons: The third generation of neural network models”. *Neural Networks*, 10(9):1659–1671, 1997. ISSN 08936080. doi: 10.1016/S0893-6080(97)00011-7.
- [62] Saber Moradi, Ning Qiao, Fabio Stefanini, and Giacomo Indiveri. “A Scalable Multi-core Architecture with Heterogeneous Memory Structures for Dynamic Neuromorphic Asynchronous Processors (DYNAPs)”. *IEEE Transactions on Biomedical Circuits and Systems*, 12(1):106–122, feb 2018. ISSN 19324545. doi: 10.1109/TBCAS.2017.2759700.
- [63] Alice Mizrahi, Tifenn Hirtzlin, et al. “Neural-like computing with populations of superparamagnetic basis functions”. *Nature Communications*, 9(1), dec 2018. ISSN 20411723. doi: 10.1038/s41467-018-03963-w.
- [64] G. Indiveri, F. Corradi, and N. Qiao. “Neuromorphic architectures for spiking deep neural networks”. In *2015 IEEE International Electron Devices Meeting (IEDM)*, pages 4.2.1–4.2.4. IEEE, 2015. doi: 10.1109/IEDM.2015.7409623.
- [65] Giacomo Indiveri, Bernabe Linares-Barranco, et al. “Neuromorphic silicon neuron circuits”. *Frontiers in Neuroscience*, 5(73):1–23, 2011. ISSN 16624548. doi: 10.3389/fnins.2011.00073.
- [66] Sebastian Schmitt, Johann Klahn, et al. “Neuromorphic hardware in the loop: Training a deep spiking network on the BrainScaleS wafer-scale system”. *Proceedings of the International Joint Conference on Neural Networks*, 2017-May:2227–2234, 2017. doi: 10.1109/IJCNN.2017.7966125.
- [67] Kaushik Roy, Akhilesh Jaiswal, and Priyadarshini Panda. “Towards spike-based machine intelligence with neuromorphic computing”. *Nature*, 575(7784):607–617, 2019. ISSN 14764687. doi: 10.1038/s41586-019-1677-2. URL <http://dx.doi.org/10.1038/s41586-019-1677-2>.
- [68] Steve B. Furber, Francesco Galluppi, Steve Temple, and Luis A. Plana. “The SpiNNaker project”. *Proceedings of the IEEE*, 102(5):652–665, 2014. ISSN 00189219. doi: 10.1109/JPROC.2014.2304638.
- [69] Ben Varkey Benjamin, Peiran Gao, et al. “Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations”. *Proceedings of the IEEE*, 102(5):699–716, 2014. ISSN 00189219. doi: 10.1109/JPROC.2014.2313565.
- [70] Jongkil Park, Theodore Yu, Siddharth Joshi, Christoph Maier, and Gert Cauwenberghs. “Hierarchical Address Event Routing for Reconfigurable Large-Scale Neuromorphic Systems”. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10):2408–2422, oct 2017. ISSN 21622388. doi: 10.1109/TNNLS.2016.2572164.

RÉFÉRENCES : RÉSUMÉ EN FRANÇAIS

- [71] Mike Davies, Narayan Srinivasa, et al. “Loihi: A Neuromorphic Manycore Processor with On-Chip Learning”. *IEEE Micro*, 38(1):82–99, 2018. ISSN 02721732. doi: 10.1109/MM.2018.112130359.
- [72] Chiara Bartolozzi and Giacomo Indiveri. “Synaptic dynamics in analog VLSI”. *Neural Computation*, 19(10):2581–2603, 2007. ISSN 08997667. doi: 10.1162/neco.2007.19.10.2581.
- [73] Vladimir Kornijcuk, Jongkil Park, et al. “Reconfigurable Spike Routing Architectures for On-Chip Local Learning in Neuromorphic Systems”. *Advanced Materials Technologies*, 4(1):1800345, jan 2019. ISSN 2365709X. doi: 10.1002/admt.201800345.
- [74] D. Garbin, E. Vianello, et al. “On the impact of OxRAM-based synapses variability on convolutional neural networks performance”. In *Proceedings of the 2015 IEEE/ACM International Symposium on Nanoscale Architectures, NANOARCH 2015*, pages 193–198, 2015. ISBN 9781467378482. doi: 10.1109/NANOARCH.2015.7180611.
- [75] M. Suri, V. Parmar, A. Kumar, D. Querlioz, and F. Alibart. “Neuromorphic hybrid RRAM-CMOS RBM architecture”. In *15th Non-Volatile Memory Technology Symposium (NVMTS)*, pages 1–6, 2015. ISBN 9781509021260. doi: 10.1109/NVMTS.2015.7457484.
- [76] Geoffrey W. Burr, Robert M. Shelby, et al. “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, jul 2015. ISSN 00189383. doi: 10.1109/TED.2015.2439635.
- [77] Sheng Sung Yang, Chia Lu Ho, and Sammy Siu. “Computing and analyzing the sensitivity of MLP due to the errors of the i.i.d. inputs and weights based on CLT”. *IEEE Transactions on Neural Networks*, 21(12):1882–1891, dec 2010. ISSN 10459227. doi: 10.1109/TNN.2010.2077681.
- [78] Mohammad Mahvash and Alice C. Parker. “Synaptic variability in a cortical neuromorphic circuit”. *IEEE Transactions on Neural Networks and Learning Systems*, 24(3):397–409, 2013. ISSN 2162237X. doi: 10.1109/TNNLS.2012.2231879.
- [79] Robert M. Shelby, Geoffrey W. Burr, Irem Boybat, and Carmelo Di Nolfo. “Non-volatile memory as hardware synapse in neuromorphic computing: A first look at reliability issues”. In *IEEE International Reliability Physics Symposium Proceedings*, volume 2015-May, pages 6A11–6A16, 2015. ISBN 9781467373623. doi: 10.1109/IRPS.2015.7112755.
- [80] Suhwan Lim, Jong Ho Bae, et al. “Hardware-based Neural Networks using a Gated Schottky Diode as a Synapse Device”. In *Proceedings - IEEE International Symposium on Circuits and Systems*, volume 2018-May, 2018. ISBN 9781538648810. doi: 10.1109/ISCAS.2018.8351152.
- [81] Sungmin Hwang, Hyungjin Kim, et al. “System-level simulation of hardware spiking neural network based on synaptic transistors and if neuron circuits”. *IEEE Electron Device Letters*, 39(9):1441–1444, sep 2018. ISSN 07413106. doi: 10.1109/LED.2018.2853635.
- [82] Anakha V. Babu, Sandip Lashkare, Udayan Ganguly, and Bipin Rajendran. “Stochastic learning in deep neural networks based on nanoscale PCMO device characteristics”. *Neurocomputing*, 321:227–236, dec 2018. ISSN 18728286. doi: 10.1016/j.neucom.2018.09.019.
- [83] Giorgio Cristiano, Massimo Giordano, et al. “Perspective on training fully connected networks with resistive memories: Device requirements for multiple conductances of varying significance”. *Journal of Applied Physics*, 124(15), oct 2018. ISSN 10897550. doi: 10.1063/1.5042462.

- [84] T. Werner, D. Garbin, et al. “Real-time decoding of brain activity by embedded Spiking Neural Networks using OxRAM synapses”. In *IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 2318–2321, 2016. ISBN 9781479953417. doi: 10.1109/ISCAS.2016.7539048.
- [85] Alexander Serb, Johannes Bill, et al. “Unsupervised learning in probabilistic neural networks with multi-state metal-oxide memristive synapses”. *Nature Communications*, 7, sep 2016. ISSN 20411723. doi: 10.1038/ncomms12611.
- [86] Erika Covi, Stefano Brivio, et al. “Analog memristive synapse in spiking networks implementing unsupervised learning”. *Frontiers in Neuroscience*, 10(OCT), oct 2016. ISSN 1662453X. doi: 10.3389/fnins.2016.00482.
- [87] Manan Suri, Olivier Bichler, et al. “Phase change memory as synapse for ultra-dense neuromorphic systems: Application to complex visual pattern extraction”. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 4.4.1–4.4.4. IEEE, 2011. ISBN 9781457705052. doi: 10.1109/IEDM.2011.6131488.
- [88] Damien Querlioz, Olivier Bichler, Philippe Dollfus, and Christian Gamrat. “Immunity to device variations in a spiking neural network with memristive nanodevices”. *IEEE Transactions on Nanotechnology*, 12(3):288–295, 2013. ISSN 1536125X. doi: 10.1109/TNANO.2013.2250995.
- [89] D. Garbin, O. Bichler, et al. “Variability-tolerant Convolutional Neural Network for Pattern Recognition applications based on OxRAM synapses”. In *Technical Digest - International Electron Devices Meeting, IEDM*, volume 2015-Febru, pages 28.4.1–28.4.4, 2015. ISBN 9781479980017. doi: 10.1109/IEDM.2014.7047126.
- [90] Johannes Bill and Robert Legenstein. “A compound memristive synapse model for statistical learning through STDP in spiking neural networks”. *Frontiers in Neuroscience*, 8(DEC), 2014. ISSN 1662453X. doi: 10.3389/fnins.2014.00412.
- [91] T. Werner, E. Vianello, et al. “Experimental demonstration of short and long term synaptic plasticity using OxRAM multi k-bit arrays for reliable detection in highly noisy input data”. In *IEEE International Electron Devices Meeting (IEDM)*, pages 16.6.1–16.6.4, 2016. ISBN 9781509039029. doi: 10.1109/IEDM.2016.7838433.
- [92] Giacomo Pedretti, Valerio Milo, et al. “Stochastic Learning in Neuromorphic Hardware via Spike Timing Dependent Plasticity with RRAM Synapses”. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 8(1):77–85, mar 2018. ISSN 21563357. doi: 10.1109/JETCAS.2017.2773124.
- [93] John J Hopfield and D. W. Tank. “”Neural” computation of decisions in optimization problems”. *Biological Cybernetics*, 52(3):141–152, 1985. ISSN 03401200. doi: 10.1007/BF00339943.
- [94] Mark D. McDonnell and Derek Abbott. “What is stochastic resonance? Definitions, misconceptions, debates, and its relevance to biology”. *PLoS Computational Biology*, 5(5), may 2009. ISSN 1553734X. doi: 10.1371/journal.pcbi.1000348.
- [95] G. Bard Ermentrout, Roberto F. Galán, and Nathaniel N. Urban. “Reliability, synchrony and noise”. *Trends in Neurosciences*, 31(8):428–434, aug 2008. ISSN 01662236. doi: 10.1016/j.tins.2008.06.002.
- [96] David C. Knill and Alexandre Pouget. “The Bayesian brain: The role of uncertainty in neural coding and computation”. *Trends in Neurosciences*, 27(12):712–719, 2004. ISSN 01662236. doi: 10.1016/j.tins.2004.10.007.
- [97] Edwin R Lewis, Kenneth R Henry, and Walter M Yamada. “Essential roles of noise in neural coding and in studies of neural coding”. *BioSystems*, 58(1):109–115, 2000. ISSN 03032647. doi: 10.1016/S0303-2647(00)00113-1.

RÉFÉRENCES : RÉSUMÉ EN FRANÇAIS

- [98] Anthony Randal McIntosh, Natasa Kovacevic, and Roxane J. Itier. “Increased brain signal variability accompanies lower behavioral variability in development”. *PLoS Computational Biology*, 4(7), jul 2008. ISSN 1553734X. doi: 10.1371/journal.pcbi.1000106.
- [99] Pritish Narayanan, Geoffrey W. Burr, Stefano Ambrogio, and Robert M. Shelby. “Neuromorphic technologies for next-generation cognitive computing”. In *2017 IEEE 9th International Memory Workshop, IMW 2017*, pages 1–4, 2017. ISBN 9781509032723. doi: 10.1109/IMW.2017.7939095.
- [100] S. R. Nandakumar, Manuel Le Gallo, et al. “Mixed-precision architecture based on computational memory for training deep neural networks”. In *Proceedings - IEEE International Symposium on Circuits and Systems*, volume 2018-May, pages 1–5, 2018. ISBN 9781538648810. doi: 10.1109/ISCAS.2018.8351656.
- [101] Alessandro Fumarola, Severin Sidler, et al. “Bidirectional Non-Filamentary RRAM as an Analog Neuromorphic Synapse, Part II: Impact of Al/Mo/Pr 0.7 Ca 0.3 MnO 3 Device Characteristics on Neural Network Training Accuracy”. *IEEE Journal of the Electron Devices Society*, 6(1):169–178, 2018. ISSN 21686734. doi: 10.1109/JEDS.2017.2782184.
- [102] Sang Gyun Gi, Injune Yeo, et al. “Modeling and System-Level Simulation for Nonideal Conductance Response of Synaptic Devices”. *IEEE Transactions on Electron Devices*, 65(9):3996–4003, sep 2018. ISSN 00189383. doi: 10.1109/TED.2018.2858762.
- [103] Cheol Kim, Rak Joo Sung, Sung Gi Ahn, Jisu Min, and Kee Won Kwon. “Low Power Search Engine using Non-volatile Memory based TCAM with Priority Encoding and Selective Activation of Search Line and Match Line”. In *Proceedings - IEEE International Symposium on Circuits and Systems*, volume 2018-May. Institute of Electrical and Electronics Engineers Inc., apr 2018. ISBN 9781538648810. doi: 10.1109/ISCAS.2018.8351237.
- [104] Daniele Garbin, Quentin Raffay, et al. “Modeling of OxRAM variability from low to high resistance state using a stochastic trap assisted tunneling-based resistor network”. In *EUROSIOI-ULIS 2015 - 2015 Joint International EUROSIOI Workshop and International Conference on Ultimate Integration on Silicon*, pages 125–128, 2015. ISBN 9781479969111. doi: 10.1109/ULIS.2015.7063789.
- [105] Henry Markram, Joachim Lübke, Michael Frotscher, and Bert Sakmann. “Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs”. *Science*, 275(5297): 213–215, 1997. ISSN 00368075. doi: 10.1126/science.275.5297.213.
- [106] Guo Qiang Bi and Mu Ming Poo. “Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type”. *Journal of Neuroscience*, 18(24):10464–10472, 1998. ISSN 02706474. doi: 10.1523/jneurosci.18-24-10464.1998.
- [107] Tobi Delbruck. “Frame-free dynamic digital vision”. In *Intl. Symp. on Secure-Life Electronics, Advanced Electronics for Quality Life and Society*, pages 21–26, 2008. doi: <http://dx.doi.org/10.5167/uzh-17620>.
- [108] Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. “Gradient-based learning applied to document recognition”. *Proceedings of the IEEE*, 86(11):2278–2323, 1998. ISSN 00189219. doi: 10.1109/5.726791.
- [109] Olivier Bichler, Damien Querlioz, Simon J. Thorpe, Jean Philippe Bourgoin, and Christian Gamrat. “Unsupervised features extraction from asynchronous silicon retina through spike-timing-dependent plasticity”. In *Proceedings of the International Joint Conference on Neural Networks*, pages 859–866. IEEE, 2011. ISBN 9781457710865. doi: 10.1109/IJCNN.2011.6033311.

- [110] Marcus Kaiser. “A tutorial in connectome analysis : Topological and spatial features of brain networks”. *NeuroImage*, 57(3):892–907, 2011. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2011.05.025. URL <http://dx.doi.org/10.1016/j.neuroimage.2011.05.025>.
- [111] Juan Wang, Reza Khosrowabadi, et al. “Alterations in Brain Network Topology and Structural-Functional Connectome Coupling Relate to Cognitive Impairment”. *Frontiers in Aging Neuroscience*, 10(December):1–15, 2018. ISSN 1663-4365. doi: 10.3389/fnagi.2018.00404.
- [112] Fabrizio Damicelli, Claus C. Hilgetag, Marc-Thorsten Hutt, and Arnaud Messe. “Topological reinforcement as a principle of modularity emergence in brain networks”. *Network Neuroscience*, 3(2):589–605, 2019. doi: 10.1162/netn.a.00085.
- [113] Steve Deiss, Rodney Douglas, Mike Fischer, Misha Mahowald, and Tony Matthews. “Address-Event Asynchronous Local Broadcast Protocol”, 1994. URL <https://www.ini.uzh.ch/~jamw/scx/aeprotocol.html>.
- [114] Kwabena A. Boahen. “Point-to-point connectivity between neuromorphic chips using address events”. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 47(5):416–434, 2000. ISSN 10577130. doi: 10.1109/82.842110.
- [115] Bipin Rajendran, Roger W. Cheek, et al. “Demonstration of CAM and TCAM using phase change devices”. In *2011 3rd IEEE International Memory Workshop, IMW 2011*, 2011. ISBN 9781457702266. doi: 10.1109/IMW.2011.5873229.
- [116] Jing Li, Robert K. Montoye, Masatoshi Ishii, and Leland Chang. “1 Mb 0.41 μm^2 2T-2R cell nonvolatile TCAM with two-bit encoding and clocked self-referenced sensing”. *IEEE Journal of Solid-State Circuits*, 49(4):896–907, 2014. ISSN 00189200. doi: 10.1109/JSSC.2013.2292055.
- [117] J. Li, R. Montoye, et al. “1Mb 0.41 μm^2 2T-2R cell nonvolatile TCAM with two-bit encoding and clocked self-referenced sensing”. In *Digest of Technical Papers - Symposium on VLSI Technology*, pages 104–105. IEEE, 2013. ISBN 978-1-4673-5226-0.
- [118] Alessandro Grossi, Elisa Vianello, et al. “Experimental Investigation of 4-kb RRAM Arrays Programming Conditions Suitable for TCAM”. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 26(12):2599–2607, dec 2018. ISSN 10638210. doi: 10.1109/TVLSI.2018.2805470.
- [119] Rui Yang, Haitong Li, et al. “Ternary content-addressable memory with MoS₂ transistors for massively parallel data search”. *Nature Electronics*, 2(3):108–114, mar 2019. ISSN 25201131. doi: 10.1038/s41928-019-0220-7.
- [120] A. V. Campi, R. M. Dunn, and B. H. Gray. “Content Addressable Memory System Concepts”. *IEEE Transactions on Aerospace and Electronic Systems*, AES-1(2):168–173, 1965. doi: 10.1109/TAES.1965.4501678.
- [121] Anthony J. McAuley and Paul Francis. “Fast routing table lookup using CAMs”. In *Proceedings - IEEE INFOCOM*, volume 3, pages 1382–1891. Institute of Electrical and Electronics Engineers (IEEE), mar 1993. ISBN 0818635800. doi: 10.1109/incom.1993.253403.
- [122] Kostas Pagiamtzis and Ali Sheikholeslami. “Content-addressable memory (CAM) circuits and architectures: A tutorial and survey”. *IEEE Journal of Solid-State Circuits*, 41(3): 712–727, mar 2006. ISSN 00189200. doi: 10.1109/JSSC.2005.864128.
- [123] Dennis Dudeck and Lisa Minwell. “Why is TCAM Essential for the Cloud?”. Technical report, 2014. URL www.esilicon.com.
- [124] Daniele Ielmini. “Resistive switching memories based on metal oxides: Mechanisms, reliability and scaling”. *Semiconductor Science and Technology*, 31(6):063002, 2016. ISSN 13616641. doi: 10.1088/0268-1242/31/6/063002.

List of Figures

- 1.1.1 (a) Power density as a function of clock frequency. Current Von Neumann-based architectures are inefficient for representing massively interconnected neural networks. Brains differ from today’s computers by their architecture: they feature a parallel, distributed architecture, whereas Von Neumann systems exhibit sequential, centralised architectures. (b) In Von Neumann-based architectures computation and memory units are physically separated by a bus leading to the so-called Von Neumann bottleneck. (c) Conceptual blueprint of a brain-like architecture wherein computation and memory are tightly co-localised. Reproduced from [22]. 3
- 1.2.1 (a) Overview of established charge-based memories (blue) and new resistance-based non-volatile memories (green). (b) Memory hierarchy of today’s computers. Speed, number of processors cycles (CPU cycles), and typical capacity (size) of the different memories are shown in the lower panel. The closer to processing cores (CPUs), the faster the memory (cache memory). Reproduced from [9]. 5
- 1.2.2 Schematic illustration of the switching process in Oxide-based Resistive Memories (OxRAMs). An initial forming process generates oxygen vacancies in the oxide layer by soft dielectric breakdown. Subsequent Set and Reset operations lead to the formation and dissolution of a Conductive Filament (CF) made of oxygen vacancies, respectively. The interface between the oxide layer and the top electrode acts like an oxygen reservoir. Reproduced from [44]. 7
- 1.2.3 (a) Reported storage capacity over time of different non-volatile memory technologies. Multi-gigabits prototypes with the new non-volatile memories have been reported [75–78]. Data on Flash NAND technology are reported for comparison [79, 80]. (b) Programming energy as a function of cell area. Programming energy of Resistive Memories (RRAMs, encompassing OxRAM and CBRAM technologies) does not depend on cell area due to the filament conduction nature of RRAMs. Reproduced from [54]. 9

1.2.4	<p>Low Resistance State (LRS) resistance value of different Resistive Memory (RRAM) technologies as a function of compliance current, I_{cc}, during a Set operation. A power law relationship exists between LRS resistance values and I_{cc}. Reproduced from [46]. (b) High Resistance State (HRS) resistance value as a function of the voltage applied during Reset operations, V_{reset}. Measurements have been performed on a TiN/HfO₂/Ti/TiN RRAM device. The mean HRS resistance value over 1000 Reset operations is shown (solid line) as well as the spread at two standard deviations (shaded area). Reproduced from [91]. . . .</p>	11
1.2.5	<p>Typical endurance characterisations performed on (a) a GeS₂/Ag and (b) a HfO₂/GeS₂/Ag Resistive Memory (RRAM) stack. While it is possible to sustain a low resistance ratio R_{off}/R_{on} of 10 during 10^8 switching cycles, only 10^3 switching cycles can be performed with a large resistance ratio of 10^6. Reproduced from [90]. (c) Reported Memory Window (MW) as a function of programming endurance for different RRAM technologies. Data on Phase-Change Memory (PCM) and Spin-Torque-Transfer Magnetic Memory (STT-MRAM) are reported for comparison. A general trend of lower MWs with higher endurance performance is observed. Reproduced from [54].</p>	12
1.2.6	<p>High Resistance State (HRS, red) and Low Resistance State (LRS, black) resistance distributions measured on a 4-kbit TiN/HfO₂/Ti/TiN Resistive Memory (RRAM) array, after one Reset/Set cycle, respectively. While a resistance ratio of 2500 is measured between the median HRS and LRS resistance values, it is reduced to 600 at three standard deviations, 3σ, due to device-to-device resistance variability. Reproduced from [43]. .</p>	12
1.2.7	<p>Transmission electron microscopy of a TiN/HfO₂/Ti/TiN RRAM fabricated on top of a NMOS transistor. RRAMs have been integrated in the back-end-of-line. Reproduced from [103].</p>	13
1.2.8	<p>(a) Schematic drawing of a three-dimensional (3D) cross-point Resistive Memory (RRAM) structure. RRAM cells are located in between densely stacked word-lines and bit-lines. Reproduced from [109]. (b) (Left) Schematic drawing of Vertical RRAM (VRRAM) arrays (reproduced from [109]), and (Right) transmission electron microscopy of a four-layers TiN/Ti/HfO_x/TiN VRRAM (reproduced from [110]).</p>	14

1.2.9	Schematic illustrations of three-dimensional (3D) (a) parallel integration, and (b) sequential integration. In the parallel integration both layers are fabricated separately, then vertically stacked and connected. In the sequential integration the top layer is fabricated directly on top of the bottom layer. Reproduced from [111]. (c) Alignment accuracy as a function of 3D contact width. 3D sequential integration allows for higher alignment accuracy than parallel integration since it only depends on lithographic alignment on the stepper. Reproduced from [112].	15
1.2.10	(a) (Left) NMOS and (Right) PMOS $I_{\text{off}}/I_{\text{on}}$ performance for different thermal annealings. Transistor performance can be ensured for annealing up to 500°C for 5 hours. Reproduced from [115]. (b) Transmission electron microscopy of two tiers of NMOS transistors fabricated in a 3D monolithic 65-nm SOI process with CoolCube™ technology. Reproduced from [112].	17
1.3.1	Drawing of two connected neurons. A neuron is mainly composed of a soma, several dendrites, and an axon. Neurons transmit electrical signal events (action potentials) along their axon connected to other neuron dendrites. Axon terminals connect to other neuron dendrites through terminal buttons forming synapses.	19
1.3.2	The three main neural network architectures: Fully-Connected Neural Network (FCNN), Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN). Adapted from [12].	22
1.3.3	(a) Example of a bio-inspired silicon-based Leaky Integrate-and-Fire (LIF) neuron circuit from [181]. (b) Evolution of the membrane capacitance potential, V_{mem} , during the generation of an action potential. Reproduced from [148].	24
1.3.4	Experimental Spike-Timing Dependent Plasticity (STDP) observed by Bi and Poo [133]. If a post-synaptic neuron spikes shortly after a pre-synaptic neuron within a time window of about 100 ms (right-hand side), the synaptic weight increases. Otherwise, the synaptic weight decreases (left-hand side). Adapted from [133].	26
1.3.5	Differential-Pair Integrator (DPI) synapse. Reproduced from [203].	26
1.3.6	Schematic illustration of a vector-matrix multiplication performed by a memristor crossbar array in a single read cycle. Multiplication operations are performed on each memory element by Ohm's law, while accumulate operations are performed on every column or row by Kirchhoff's law. Reproduced from [207].	27

1.3.7 (a) Schematic illustration of a Resistive Memory (RRAM) used as a synapse between a pre- and post-synaptic neuron. (b) Conductance response (represented by the read current at 1 V) of a Ag/Si-based active layer RRAM after a series of 100 identical potentiation pulses (3.2 V for 300 μ s) followed by 100 identical depression pulses (-2.8 V for 300 μ s). (c) Experimental demonstration of Spike-Timing-Dependent Plasticity (STDP) measured on the Ag/Si-based RRAM. Timing difference between the pre- and post-synaptic neuron, Δ Spike Timing, was captured and mapped with a time-division multiplexing approach. Reproduced from [219]. 28

1.3.8 (a) Pre- and post-synaptic spike sequences to enable Spike-Timing-Dependent Plasticity (STDP) with pulse amplitude modulation. Pulse amplitudes are: -1.4 V, 1 V, 0.9 V, 0.8 V, 0.7 V, and 0.6 V (pre-synaptic spikes); -1 V, 1.4 V, 1.3 V, 1.2 V, 1.1 V, and 1 V (post-synaptic spikes). (b) Experimental demonstration of STDP measured on a TiN/HfO_x/AlO_x/Pt RRAM stack using the previous STDP scheme. Reproduced from [220]. 28

1.3.9 (a) Example of a one-transistor/one-RRAM (1T1R) synapse connecting a pre- and post-synaptic neuron. The 1T1R synapse is activated via the transistor gate only at pre-synaptic pulses. (b) Pre- (top) and post-synaptic (bottom) pulses to enable Spike-Timing-Dependent Plasticity (STDP) learning in the 1T1R synapse. Only the overlap between pre- and post-synaptic pulses induces an increase ($\Delta t > 0$) or decrease ($\Delta t < 0$) in conductance of the RRAM. Reproduced from [222]. 29

2.2.1 (Left) Scanning electron microscope cross-section of the TiN/HfO₂/Ti/TiN (100 nm/10 nm/10 nm/100 nm) RRAM cell integrated on top of the fourth Cu metal layer. (Right) Schematic view of the 1T1R cell configuration. The NMOS transistor is used as a selector device. 58

2.2.2 Cumulative distributions of the LCS and HCS distributions measured on the 4-kbit array (Top left) after 1000 switching cycles with condition A, (Top right) with condition C, (Bottom left) with condition B1, and (Bottom right) with condition B2. TABLE 2.1 summarises the parameters of each programming condition. These distributions represent the device-to-device variability. 60

2.2.3 (a) Programming endurance characterisation with programming conditions A in TABLE 2.1. (b) Evolution of HCS and LCS conductance variability with programming conditions A during RRAM aging. Conductance variability is defined in EQUATION 2.2.2. 62

2.2.4	Conductance variability as a function of the median conductance value for different programming conditions. Conductance variability is defined in EQUATION 2.2.2.	62
2.2.5	Conductance evolution during the application of a series of 20 identical Set pulses and Reset pulses with (Left) programming conditions B2 and (Right) programming conditions A of TABLE 2.1. Grey lines are representative of ten single cells behaving as analog devices (gradual increase or decrease of the conductance value). Black lines are representative of ten single cells behaving as binary devices (abrupt switching between the HCS and LCS). Red circles and blue squares correspond to the median conductance value calculated on 4 kbit cells during potentiation and depression, respectively. The pulse 0 is the conductance value before the first Set pulse.	63
2.2.6	(a) RRAM-based synapse implementation. The Pseudo-Random Number Generator (PRNG) is used to tune the switching probabilities. (b) Stochastic STDP rule and conductance evolution of a RRAM-based synapse composed of 1, 3, and 20 RRAM cells in parallel. 200 potentiation pulses followed by 200 depression pulses are applied. Condition A in TABLE 2.1 is used. (c) Conductance evolution of 100 RRAM-based synapses composed of 20 RRAM cells when 200 potentiation pulses and 200 depression pulses are applied. Condition A (Top left), C (Top right), B1 (Bottom left), and B2 (Bottom right) are used. Red and blue symbols represent the median conductance evolution; grey lines represent the evolution of each synapse.	65
2.2.7	Simulated spiking neural networks used for (a) the car tracking and (b) the digit classification applications. The associated score definition to assess network performance is shown on the right-hand side of each network. See APPENDIX A for more details.	67
2.2.8	(a) F1-score as a function of the memory window at 3σ , $MW_{3\sigma}$ defined in EQUATION 2.2.1, for different numbers of RRAMs per synapse. The HCS and LCS distributions measured under the programming conditions A on the 4-kbit array are used (<i>cf</i> FIGURE 2.2.2 (a)). (b) F1-score as a function of the $MW_{3\sigma}$. One RRAM device per synapse is used. The HCS and LCS distributions measured on the 4-kbit array for the four conditions of TABLE 2.1 and an artificial case with zero variability are used. The $MW_{3\sigma}$ is varied by a translation of the LCS distributions to lower or higher conductance values.	68

2.2.9	(a) Cumulative distributions of the nine artificial log-normal distributions used to quantify the impact of synaptic variability. (b) Minimal memory window at 3σ , $MW_{3\sigma,\min}$ (z-axis), required to reach the maximal F1-score of 0.96 as a function of the HCS (x-axis) and LCS (y-axis) conductance variability values. Higher conductance variability values allow to relax the constraints on $MW_{3\sigma,\min}$	70
2.2.10	(a) Example of the two-dimensional conductance mapping of one arbitrary output neuron after learning. Potentiated synapses (RRAMs in HCS) are represented by coloured dots (red, blue, and grey); depressed synapses (RRAMs in LCS) are represented by black dots. (b) F1-score as a function of the synaptic window, SW, defined in EQUATION 2.2.4. (c) Synaptic weight distributions after the learning phase for the four programming conditions of TABLE 2.1. High performance after learning (F1=0.96) is reached when the ratio between the peaks of the HCS and LCS distributions is larger than 200. (d) Learning time as a function of the SW.	72
2.2.11	(a) Impact of the RRAM aging on the F1-score. Simulations have been calibrated using the data of FIGURE 2.2.3 (a) (condition A). Both device-to-device and cycle-to-cycle variability are taken into account. (b) Average number of Set (red circle) and Reset (blue square) operations during the learning phase for each programming condition. Red and blue shaded areas represent the evolution at $\pm 3\sigma$, respectively.	73
2.2.12	(a) Classification rate, CR, as a function of the memory window at 3σ , $MW_{3\sigma}$ defined in EQUATION 2.2.1, for different numbers of RRAMs per synapse. The HCS and LCS distributions measured on the 4-kbit array for the conditions A and B2 of TABLE 2.1, and an artificial case with zero variability were used. The $MW_{3\sigma}$ was varied by a translation of the LCS distribution to lower or higher conductance values. (b) CR as a function of the conductance variability in HCS, $\sigma_{G,\text{HCS}}$, for 1 and 20 RRAMs per synapse. The four HCS and LCS distributions measured on the 4-kbit array for the conditions of TABLE 2.1 and an artificial case with zero variability were used. (c) Synaptic weight distributions after the learning phase for the four programming conditions of TABLE 2.1 and the synapse with zero variability. Higher performance after learning is reached with a small amount of HCS conductance variability (conditions A and C).	75
2.2.13	Impact of the RRAM aging on the classification rate, CR. Simulations have been calibrated using the data of FIGURE 2.2.3 (a) (condition A). Both device-to-device and cycle-to-cycle variability are taken into account. (b) Average number of Set (red circle) and Reset (blue square) operations during the learning phase for each programming condition. Red and blue shaded areas represent the evolution at $\pm 3\sigma$, respectively.	76

2.2.14	Classification rate, CR, as a function of the HCS conductance variability, $\sigma_{G,HCS}$, for the four programming conditions of TABLE 2.1. The CR has been calculated assuming the experimental LCS conductance variability (black line) and no LCS conductance variability (dotted grey line).	77
2.3.1	(a) I-V characteristics of a TiN/HfO _x /AlO _x /Pt RRAM device with increasing compliance current for each Set operation. The device conductance value (represented by the read current) increases with the compliance current. Reproduced from [97]. (b) 2-PCM synapse implementation. One synapse is implemented with two PCM devices in parallel (LTP and LTD PCMs) with opposite current contributions. Analog conductance modulation is obtained by exploiting the gradual crystallisation of both PCMs. Reproduced from [37]. (c) Programming strategy proposed in [36] to obtain analog conductance modulation with a single PCM device. In addition to the gradual crystallisation of PCM (potentiation, green area), initialising the PCM at an intermediate resistance value ($R_1 \approx 30 \text{ k}\Omega$) enables gradual amorphisation (depression, blue area) with short Reset pulses (<50 ns). Reproduced from [36].	80
2.3.2	(a) Simplified STDP rule and synaptic weight increment and decrement equations. (b) Conductance response with the model described in EQUATION 2.3.1 for different linearity factors, β_+ and β_- . Potentiation and depression levels are fixed at 200 ($n_{pot}=n_{dep}=200$). β_+ and β_- control the linearity of the conductance response. (c) Conductance response with the model described in EQUATION 2.3.1 for different numbers of potentiation and depression levels. Linearity factors are fixed at 3 ($\beta_+=\beta_-=3$). (d) Conductance response of the PCM technology presented in [36] (filled symbol) and fitting with EQUATION 2.3.1 (grey line).	83
2.3.3	Simulated spiking neural networks used for the digit classification task with analog devices as synaptic elements. The associated score definition to assess network performance is shown on the right-hand side. See APPENDIX A for more details.	84
2.3.4	(a) Classification rate, CR, as a function of the number of output neurons, $n_{neurons}$. The conductance responses in FIGURE 2.3.2 (c) have been used for the simulations. (b) Conductance responses with EQUATION 2.3.1 for a fixed number of potentiation levels $n_{pot}=200$ and different numbers of depression levels, n_{dep} . (c) CR as a function of the number of depression levels, n_{dep} , for a fixed $n_{pot}=200$ (<i>cf</i> (b)).	85

2.3.5	(Top) Classification Rate, CR, as a function of the linearity factor in potentiation, β_+ , for 200 potentiation levels and 30 depression levels. Open symbols correspond to symmetric responses in linearity ($\beta_+=\beta_-$). The filled symbol corresponds to an asymmetric case ($\beta_+=3$ and $\beta_-=1$) fitted with the PCM technology presented in [36] (<i>cf</i> FIGURE 2.3.2 (d)). (Bottom) Synaptic weight evolution of 100 synapses during the training phase for (Left) a linear conductance response and (Right) a non-linear conductance response.	87
3.1.1	Block diagram of (a) a Random Access Memory (RAM) system and (b) a Content-Addressable Memory (CAM) system. In RAM systems, stored data are accessed by their physical address location, and the system outputs the stored content. In CAM systems, stored data are accessed by their content rather than by their address, and the system outputs the address of the searched data.	100
3.1.2	Block diagram of a Ternary Content-Addressable Memory (TCAM) system. The use of the <i>don't care</i> state, 'X', allows to perform local masking and store data ranges. As more than one word may match, a priority encoder is used instead of an encoder. A single address is output based on the highest priority matching location (<i>e.g.</i> lowest address location, most matching bits that are not in the 'X' states, ...).	101
3.1.3	(a) Conventional sixteen-transistors (16T) SRAM-based TCAM. (b) Common two-transistors/two-RRAMs (2T2R) RRAM-based TCAM.	102
3.1.4	Reported SRAM-based (black circle) [3–5, 12, 13, 33, 35–48] and RRAM-based (blue diamond) [14, 15, 18, 19, 21, 26, 27, 29, 49] (a) TCAM bitcell size, (b) TCAM search time, and (c) TCAM search energy as a function of technology node. Search times in (b) have been normalised by the number of bits per TCAM word to provide a fair comparison. TCAM bitcell size with Magnetic Memories (MRAMs, grey triangle) [50, 51] have been plotted for comparison.	104
3.1.5	TCAM-based implementation of a network router address lookup table. Reproduced from [2].	105

3.1.6	(a) Simplified block diagram of one computing node in DYNAPs [59] composed of a 64x10 bits BCAM table, 64 Pulse Generators (PGs), a Differential-Pair Integrator (DPI) synapse circuit [70], and a CMOS-based leaky Integrate-and-Fire (IF) neuron. (b) Working principle of DYNAPs. When the neuron 6 spikes, its address, 6, is broadcast to every other neuron (including itself). As its address is stored in the CAM table of the computing nodes 2 and 6 (green rows), a pulse is locally generated and transmitted to the corresponding leaky IF neuron circuits. The use of TCAMs instead of BCAMs allows to increase the fan-in of each neuron. Adapted from [59].	107
3.2.1	(a) Common block diagram of the fabricated 2T2R and 1T2R1T TCAM circuits. Only the TCAM array is different between both circuits. (b) Die picture of the fabricated 2T2R and (c) 1T2R1T circuits. (d) Scanning electron microscope cross-section of the integrated HfO ₂ -based RRAMs.	109
3.2.2	Example of the search operation principle. The Match Line (ML) is first pre-charged at VDD_ML, then it is left floating. (Top) In a match case, the ML stays high. (Bottom) In the mismatch case, the ML is pulled down to a low level.	110
3.2.3	Common 2T2R TCAM bitcell schematic. Top (TE) and Bottom (BE) Electrodes are indicated with the black rectangle.	110
3.2.4	(a) In the Match Line (ML) pre-charge phase, the ML is pre-charged high at a voltage V_{search} . (b) In the ML sensing phase, the ML is left floating and discharges through each TCAM cell. The discharge follows that of a RC circuit.	111
3.2.5	(a) In a match case, the activated transistor is in series with a RRAM in the High Resistance State (HRS). (b) In a mismatch case, the activated transistor is in series with a RRAM in the Low Resistance State (LRS).	112
3.2.6	(Top) Example of the Match Line (ML) voltage evolution during a search operation in the case of match (green) and 1-bit mismatch (red). (Bottom) Corresponding measured waveforms output by the sense amplifier, SA_OUT. The duration for which SA_OUT stays at '0' defines the ML discharge time, t_{search} . . .	112
3.2.7	Low Resistance State (LRS), High Resistance State (HRS), and pristine resistance cumulative distributions directly measured on the TCAM cells. HRS resistance distribution can be obtained using the Soft HRS or Strong HRS programming conditions. .	114
3.2.8	Discharge time, t_{search} , as a function of (a) Match Line (ML) capacitance (the search voltage, V_{search} , is fixed at 0.6 V), and (b) V_{search} (ML capacitance is fixed at 315 pF).	115

LIST OF FIGURES

3.2.9 Measured Time Ratio (TR) as a function of (a) match line capacitance, (b) search voltage V_{search} , (c) memory window, and (d) TCAM word length. 116

3.2.10 (a) During a search operation, a positive voltage is applied on RRAM top electrodes in the same configuration as a Set operation. (b) Characterisation of the search endurance. Measured discharge times, t_{search} , as a function of the number of search operations are reported. 118

3.2.11 (Top) Soft HRS programming endurance characterisation. (Bottom) Probabilities of match (square) and mismatch (circle) failures as a function of the number of Set/Reset switching cycles. 119

3.2.12 (a) Extrapolated t_{search} as a function of the match line capacitance. (b) Extrapolated search endurance as a function of the match line capacitance. 120

3.2.13 Two-bits encoding principle. (Left) When no encoding is used, two bits are encoded with two distinct TCAM bitcells. During a search operation, two out of the four transistors are turned ON. In the case of match, leakage currents, I_{match} , flow through two 1T1R structures in HRS. (Right) When the two-bits encoding scheme is used, two bits are encoded with an association of two TCAM bitcells. During a search operation, only one out of four transistors is turned ON. In the case of match, leakage currents are halved with respect to the case of no encoding as leakage currents only flow in one 1T1R structure in HRS. Reproduced from [30]. 123

3.2.14 Measured Time Ratio (TR) as a function of the match line capacitance using the two-bits encoding scheme of [29]. The TR improves by 3.8x with the two-bits encoding scheme. 124

3.2.15 Proposed 1T2R1T TCAM bitcell schematic. Top (TE) and Bottom (BE) Electrodes are indicated with the black rectangle. 124

3.2.16 (a) In a match case, the internal node voltage, V_{int} , in the RRAM voltage divider is kept at 0 V, the transistor N2 is OFF. (b) In a mismatch case, V_{int} is almost equal to V_{search} , the transistor N2 turns ON if V_{search} is higher than the threshold voltage of transistor N2, $V_{\text{th},N2}$ 125

3.2.17 Match and mismatch cases for (a) the common 2T2R and (b) the proposed 1T2R1T structures. The sensing margin ($\approx I_{\text{mismatch}}/\sum I_{\text{match}}$) of the common 2T2R structure depends on the memory window, whereas for the proposed 1T2R1T structure, it depends on the transistor N2 characteristic. (c) Measured $I_{\text{ds}}-V_{\text{gs}}$ characteristic of transistors N2. In the case of match, $V_{\text{gs}}=0$ V. In the case of mismatch, $V_{\text{gs}}=V_{\text{search}}$ 126

3.2.18 (Top) Example of the Match Line (ML) voltage evolution during a search operation in the case of match (green) and 1-bit mismatch (red). The ML does not discharge in the match case. (Bottom) Corresponding measured waveforms output by the sense amplifier, SA_OUT. The duration for which SA_OUT stays at '0' defines the ML discharge time, t_{search} . t_{search} is longer than the measurement limit (one second) in the match case. 127

3.2.19 Low Resistance State (LRS), High Resistance State (HRS), and pristine resistance cumulative distributions directly measured on the TCAM cells. HRS resistance distribution can be obtained using the Soft HRS or Strong HRS programming conditions. . . 128

3.2.20 Discharge time, t_{search} , as a function of the search voltage, V_{search} , in the case of match (green) and mismatch (red) of 1 bit and 128 bits. t_{search} is almost independent of RRAM programming conditions. 129

3.2.21 Measured time ratio for the proposed 1T2R1T structure (filled symbol) and the common 2T2R structure measured in the previous section (open symbol) as a function of (a) the search voltage V_{search} , (b) the memory window, and (c) the TCAM word length. The measurements performed on the 2T2R structure using the two-bits encoding scheme of Li et al. [29] is represented by the shaded red diamond. 130

3.2.22 (a) During a search operation, a positive voltage is applied on the top electrode of one RRAM cell (here R_x for a search '1') in the same configuration as a Set operation. (b) Search endurance characterisation for the proposed 1T2R1T structure. (c) Comparison between the search endurance of the common 2T2R TCAM (open symbol) and the 1T2R1T TCAM (filled symbol) as a function of V_{search} for Soft and Strong HRS. . . . 131

3.2.23 (a) With the proposed 1T2R1T structure, the sensing circuit can be implemented either with (Top) a comparator circuit (low swing) or (Bottom) a digital inverter (full swing). (b) Measured (symbol) and simulated (line) discharge times, t_{search} , as a function of the search voltage, V_{search} , in the (Left) 128-bits and (Right) 1-bit mismatch states. (c) Simulated search energy consumption as a function of V_{search} in the 1-bit mismatch state when transistors N2 are implemented with thick oxide MOS (solid line) and thin oxide MOS (dotted line). 133

3.2.24 Comparison in terms of (a) TCAM bitcell size, (b) search time, and (c) search energy as a function of technology node with reported silicon-proven SRAM- (black circle) [3, 5, 12, 13, 33, 36–39, 41, 43–47] and RRAM-based (blue diamond) [19, 21, 26, 27, 29] TCAM circuits. 136

4.2.1	Process flow scheme of 3D CoolCube™ integration. (Left) The bottom level is fabricated at high temperature in a conventional 65-nm SOI CMOS process. (Middle) A new SOI wafer is transferred on top of the bottom level by oxide bonding, and it represents the top active area. (Right) The top level is fabricated at low thermal budget directly on top of the bottom level. Reproduced from [1].	145
4.2.2	(a) Schematic illustration of CoolCube™ wafers before RRAM integration. Gate contact plugs (shaded purple area) are deported. (b) TiN/HfO ₂ /Ti/TiN (10 nm/5 nm/5 nm/30 nm) RRAM devices are fabricated directly on top of contact plugs by a first e-beam photo-lithography. (c) Oxide and Contact Etch Stop Layer (CESL) are deposited on top of RRAM devices. Then, contact plugs are recovered by a second e-beam photo-lithography. (d) Integration is finished by standard CoolCube™ back-end-of-line process.	146
4.2.3	Transmission electron microscopy of the co-integration of HfO ₂ -based RRAMs on top of two NMOS transistors fabricated with CoolCube™ technology.	147
4.3.1	I _{ds} -V _{gs} characteristics measured on (a) a bottom NMOS transistor with W=L=10 μm, (b) a top NMOS transistor with W=L=10 μm, and (c) a top NMOS transistor with W=60 nm and L=50 nm. Characteristics have been measured at V _{ds} =50 mV (dotted line) and V _{ds} =1 V (solid line).	148
4.3.2	(a) Butterfly I-V curves measured on the bottom 1T1R and (b) the top 1T1R. Forming, then five Reset-Set cycles have been performed with the programming conditions in TABLE 4.1. (c) Read resistance values measured after each switching operation on the bottom 1T1R and (d) on the top 1T1R.	149
4.3.3	Demonstration of the absence of crosstalk between the bottom and top 1T1R. (a) Set and Reset operations have been performed on the top 1T1R, and the resistance values of the bottom and top 1T1R have been read after each switching operation. Bottom 1T1R resistance states remain unaltered after each switching operation on the top 1T1R. (b) Similarly, Set and Reset operations have been performed on the bottom 1T1R. Top 1T1R resistance states remain unaltered after each switching operation on the bottom 1T1R. Programming conditions in TABLE 4.1 have been used.	152
4.3.4	Endurance characterisations performed on (a) the bottom 1T1R and (b) the top 1T1R for 10 ⁵ switching cycles. Measurements have been performed with the programming conditions in TABLE 4.2.	152

4.3.5 (a) Butterfly I-V curves measured on the bottom 1T1R and (b) the top 1T1R with different programming currents, I_{prog} , for each Set operation. Increasing I_{cc} allows to decrease resistance values after each Set operation. (c) Low Resistance State (LRS) resistance values as a function of programming current, I_{prog} , for the bottom 1T1R (red diamond) and the top 1T1R (blue triangle). Data on different RRAM technologies reproduced from [9] are reported for comparison. Measured data of the bottom and top 1T1R are in good agreement with previous works. 153

A.1.1 (a) Simulated spiking neural network for the car tracking application with binary devices, trained with an unsupervised stochastic Spike-Timing-Dependent Plasticity (STDP) learning rule and lateral inhibition. (b) Example of spiking activity of one output neuron (red) and the actual traffic (a grey spike corresponds to a car passing on the lane). True Positive (TP) events, False Positive (FP) events, and False Negative (FN) events are put in evidence. The F1-score is used to assess network performance. 174

A.1.2 Stochastic Spike-Timing-Dependent Plasticity (STDP) learning rule. If the post-synaptic neuron spikes after the pre-synaptic neuron within a time window t_{STDP} (the STDP time window), the synapse undergoes a potentiation event. Otherwise, it undergoes a depression event. At each potentiation (resp. depression) event, each RRAM device has a probability p_{LTP} (resp. p_{LTD}) to switch to the High Conductance State, HCS (resp. Low Conductance State, LCS). 176

A.1.3 (a) Simulated spiking neural network for the digit classification application with binary devices, trained with a stochastic STDP learning rule and lateral inhibition. (b) Example of spiking activity of four output neurons when four different input digits are presented. If the class of the most active neuron corresponds to the input digit, the digit is successfully classified (green), otherwise the digit is not classified (red). 178

A.2.1 (a) Simulated spiking neural network for the digit classification application with analog devices, trained with a simplified Spike-Timing-Dependent Plasticity (STDP) learning rule and lateral inhibition. (b) Simplified STDP learning rule. If the post-synaptic neuron spikes after the pre-synaptic neuron within a time window t_{STDP} (the STDP time window), the synapse undergoes a potentiation event. Otherwise, it undergoes a depression event. At each potentiation (resp. depression) event, the synaptic weight increases by a quantity δw_+ (resp. δw_-). α_+ , α_- , β_+ , β_- , W_{MIN} , and W_{MAX} are fitting parameters of the conductance response of synaptic elements. 179

LIST OF FIGURES

B.1.1 Cumulative distributions of the Low Conductance State (LCS) and High Conductance State (HCS) distributions measured on 4-kbit RRAM arrays (Top left) after 1000 switching cycles with condition A, (Top right) with condition C, (Bottom left) with condition B1, and (Bottom right) with condition B2 (see CHAPTER 2 - SECTION 2.2). 182

B.1.2 (a) F1-score as a function of the firing threshold value, I_{th} , for the four studied RRAM programming conditions (A, B1, B2, and C). (b) False Negative (FN, red square) and False Positive (FP, blue circle) rates as a function of the firing threshold value, I_{th} . The optimised threshold value, $I_{th,opt}$, comes from a trade-off between FN and FP rates. 185

B.1.3 Optimised firing threshold value, $I_{th,opt}$, as a function of the mean High Conductance State (HCS) conductance value of each programming condition. $I_{th,opt}$ is proportional to the mean HCS conductance value. 186

B.2.1 (a) Classification Rate (CR) as a function of the firing threshold value, I_{th} , for the four studied RRAM programming conditions (A, B1, B2, and C). (b) Optimised firing threshold value, $I_{th,opt}$, as a function of the mean High Conductance State (HCS) conductance value of each programming condition. $I_{th,opt}$ is proportional to the mean HCS conductance value. 186

B.3.1 (a) F1-score as a function of threshold variability values, $\sigma(I_{th})$. (b) F1-score as a function of the mean firing threshold value, $I_{th,mean}$, for different threshold variability values, $\sigma(I_{th})$. Synaptic elements are implemented with one binary RRAM device calibrated on programming conditions A. 187

C.0.1 Cumulative distributions of High Conductance State (HCS, red square) and Low Conductance State (LCS, blue circle) with (a) programming conditions B1, and (b) an artificial case of a synapse with zero variability. 189

C.0.2 F1-score as a function of input noise when synaptic elements are calibrated on (a) the experimental programming conditions B1, and (b) an artificial case of a synapse with zero variability. 190

1.1.1 (a) Dans les architectures Von Neumann, les unités de calcul de de mémoire sont physiquement séparées par un bus, créant le fameux goulot d'étranglement de Von Neumann. (b) Schéma conceptuel d'une architecture inspirée du cerveau, où le calcul et la mémoire sont fortement co-localisés. Reproduit de [22]. . 197

1.2.1 Illustration schématisée du processus de commutation des mémoires résistives à base d'oxydes (OxRAM pour Oxide-based Resistive Random Access Memory). Reproduit de [31]. 198

1.2.2	Fenêtre mémoire, MW, en fonction de l'endurance en programmation. Reproduit de [53].	199
1.2.3	(a) Mesures de cyclage effectuées sur une cellule RRAM à base de HfO ₂ . La valeur de résistance dans les états de basse (LRS, rouge) et haute (HRS, bleu) résistance varie de cycle à cycle. (b) Distributions de valeurs de résistances en LRS (rouge) et HRS (bleu) mesurées sur une matrice 4 kbit de RRAM à base de HfO ₂ . La valeur de résistance en LRS et HRS varie de cellule à cellule. Reproduit de [25].	199
1.2.4	Illustration schématique de (a) l'intégration tri-dimensionnelle (3D) (a) parallèle, et (b) séquentielle. Reproduit de [55].	200
1.2.5	Image par microscopie électronique d'une RRAM intégrée monolithiquement au-dessus des contacts d'un transistor NMOS. Reproduit de [60].	201
1.3.1	Principe de fonctionnement de base d'un réseau de neurones impulsif.	202
2.1.1	Illustration schématique d'une multiplication vecteur-matrice effectuée par une matrice de RRAM de type crossbar en un seul cycle de lecture. Reproduit de [39].	208
2.2.1	(Gauche) Photo par microscopie électronique d'une cellule RRAM TiN/HfO ₂ /Ti/TiN (100 nm/10 nm/10 nm/100 nm) intégrée au-dessus du quatrième niveau métallique Cu. (Droite) Vue schématique d'une configuration 1T1R.	209
2.2.2	Distributions cumulées LCS et HCS mesurées sur la matrice 4 kbit avec différentes conditions de programmation.	210
2.2.3	Caractérisation de l'endurance en programmation mesurée avec les conditions de programmation A de la TABLE 2.1.	211
2.3.1	(a) Schéma de l'implémentation des éléments synaptiques à base de RRAM. (b) Version stochastique de la règle d'apprentissage non-supervisée de plasticité fonction d'occurrence des impulsions (STDP).	211
2.4.1	Réseaux de neurones impulsifs simulés pour les applications de (a) suivi de voitures et (b) classification de chiffres. Les scores associés pour évaluer les performances de chaque réseau sont définis sur la partie droite de chaque réseau.	213
2.4.2	(a) Score F1 en fonction de la fenêtre mémoire à 3σ , $MW_{3\sigma}$, pour différents nombres de RRAM par synapse. Le réseau de neurones conçu pour le suivi de voitures est calibré sur les conditions de programmation A (<i>cf</i> TABLE 2.1). (b) Score F1 en fonction de $MW_{3\sigma}$ lorsque le réseau de neurones conçu pour le suivi de voitures est calibré avec les différentes conditions de programmation de la TABLE 2.1.	214

LIST OF FIGURES

2.4.3	(a) Taux de classification, CR, du réseau de neurones conçu pour la classification de chiffres en fonction de la fenêtre mémoire à 3σ , $MW_{3\sigma}$, pour différents nombres de RRAM par synapse. (b) Taux de classification, CR, du réseau de neurones conçu pour la classification de chiffres en fonction de la variabilité synaptique, $\sigma_{G,HCS}$	215
2.4.4	Distributions des poids synaptiques après apprentissage pour l'application de (a) détection et (b) classification.	216
2.4.5	(a) Score F1 en fonction du nombre de cycles de commutation pour le réseau de neurones conçu pour le suivi de voitures. (b) Taux de classification en fonction du nombre de cycles de commutation pour le réseau de neurones conçu pour la classification de chiffres.	217
3.2.1	Principe de fonctionnement (a) d'un système RAM classique, et (b) un système de mémoire adressable par contenu (CAM pour Content-Addressable Memory).	220
3.3.1	Schéma de la cellule unitaire TCAM la plus commune 2T2R.	221
3.3.2	(a) Dans un cas de match, le transistor activé est en série avec une RRAM dans l'état haute résistance (HRS pour High Resistance State). (b) Dans un cas de mismatch, le transistor activé est en série avec une RRAM dans l'état faible résistance (LRS pour Low Resistance State).	222
3.3.3	Schéma de la nouvelle cellule unitaire TCAM 1T2R1T.	222
3.3.4	(a) Dans un cas de match, la tension du noeud interne, V_{int} , dans le pont diviseur de tension entre les deux RRAM reste à 0 V, le transistor N2 est OFF. (b) Dans un cas de mismatch, V_{int} est quasiment égale à V_{search} , le transistor N2 est ON is V_{search} est supérieure à la tension de seuil du transistor N2.	223
3.3.5	Cas de match et mismatch pour (a) la structure 2T2R commune et (b) la nouvelle structure proposée 1T2R1T. (c) Caractéristique I_{ds} - V_{gs} mesurée sur les transistors N2 de la TCAM 1T2R1T.	225
3.3.6	(a) Schéma bloc des circuits TCAM 2T2R et 1T2R1T fabriqués. (b) Photographies des circuits 2T2R et (c) 1T2R1T fabriqués. (d) Image par microscopie électronique des cellules RRAM à base de HfO_2 intégrées.	226
3.3.7	Distributions cumulées de LRS, Soft HRS, Strong HRS, et état vierge "pristine" utilisées dans ce travail, mesurées directement sur les circuits TCAM.	227
3.4.1	Temps de décharge, t_{search} , en fonction de la tension appliquée aux bornes des RRAM, V_{search} , mesurés sur (a) le circuit TCAM 2T2R et (b) le circuit TCAM 1T2R1T. Les RRAM ont été programmées avec les différentes conditions de programmation de la FIGURE 3.3.7.	227

3.4.2	Marges de détection, TR (pour Time Ratio), en fonction de (a) la fenêtre mémoire, MW, et (b) la longueur de mots TCAM, WDL, mesurées sur les circuits TCAM 2T2R (symboles ouverts) et 1T2R1T (symboles pleins).	228
3.4.3	Endurances en recherche en fonction de la tension appliquée aux bornes des RRAM, V_{search} , mesurées sur les circuits TCAM 2T2R (symboles ouverts) et 1T2R1T (symboles pleins).	229
4.2.1	Flux du processus d'intégration de la technologie CoolCube™ du CEA-Leti [59].	232
4.2.2	(a) Illustration schématique des plaques CoolCube™ de base, avant intégration des RRAM. (b) Intégration des dispositifs RRAM par une première photo-lithographie par faisceau électronique. (c) Récupération des prises de contact par une seconde photo-lithographie par faisceau électronique. (d) Ajout des niveaux métalliques par un processus de retour en fin de ligne standard.	233
4.2.3	Image par microscopie électronique des dispositifs fabriqués par intégration 3D monolithique, avec deux niveaux de transistors et un niveau de RRAM.	234
4.3.1	Caractérisations électriques I-V en quasi-statique de (a) la 1T1R du bas et (b) la 1T1R du haut.	234
4.3.2	Caractérisations électriques en mesures pulsées de (a) la 1T1R du bas et (b) la 1T1R du haut.	235
4.3.3	Valeurs de résistance de l'état faible résistance (LRS pour Low Resistance State) en fonction du courant de programmation, I_{prog}	235

LIST OF FIGURES

List of Tables

1.1	Summary of reported silicon-proven multi-core spiking neuromorphic processors [22, 129, 130, 145, 161, 162].	31
2.1	Programming conditions used in this work, with $t_{\text{pulse}}=100$ ns. The programming energy is defined in EQUATION 2.2.3.	60
2.2	Performance and power required for learning with each programming condition of TABLE 2.1, for the detection and classification tasks. The learning power has been calculated with EQUATION 2.2.5.	78
3.1	RRAM state definition as a function of the stored data.	110
3.2	SLT and SLF voltages as a function of the searched data.	110
3.3	Performance and reliability metric definition used in this work.	113
3.4	Programming conditions used for the characterisation of the 2T2R structure.	114
3.5	Summary of the characterisation performed in this section on the common 2T2R TCAM bitcell.	121
3.6	RRAM state definition as a function of the stored data for the two-bits encoding scheme. Reproduced from [30].	123
3.7	SLT and SLF voltages as a function of the searched data for the two-bits encoding scheme. Reproduced from [30].	123
3.8	RRAM state definition as a function of the stored data.	124
3.9	SLT and SLF voltages as a function of the searched data.	124
3.10	Programming scheme for the proposed 1T2R1T TCAM.	125
3.11	Programming conditions used for the characterisation of the 1T2R1R structure.	128
3.12	Duration of each search operation as a function of the search voltage, V_{search} , for the search endurance characterisation. Note that t_{search} in the 1-bit mismatch state only depends on V_{search} , and it is similar whether RRAMs are programmed in Soft or Strong HRS.	132

LIST OF TABLES

3.13	Comparison of the characterisation results obtained with the common 2T2R and the proposed 1T2R1T structure.	135
3.14	Comparison with silicon-proven RRAM-based TCAM circuits presented in the literature [19, 21, 26, 27, 29]. Search times have been normalised by the TCAM word length.	135
4.1	Programming conditions used for the bottom and top 1T1R measurements in quasi-static mode (butterfly I-V curves). For forming and Set operations, the switching voltage $V_{\text{switching}}$ corresponds to the voltage required to abruptly switch from the High Resistance State (HRS) to the Low Resistance State (LRS). For Reset operations, it corresponds to the voltage at which the resistance value of RRAM devices starts to decrease (onset of the switching from LRS to HRS). $V_{\text{switching}}$ and read resistance values have been averaged over five Set or Reset operations.	151
4.2	Programming conditions used for the bottom and top 1T1R measurements in pulsed mode.	151
A.1	Spiking neural network parameters used for the simulations of the car tracking and digit classification applications with binary devices. All the parameters have been obtained with a genetic algorithm.	177
1.1	Synthèse des processeurs neuromorphiques impulsionsnels multi-cœurs validés sur silicium rapportés dans la littérature [22, 62, 68–71].	203
2.1	Conditions de programmation utilisées dans ce travail, avec $t_{\text{pulse}}=100$ ns.	210
3.1	Définition des états des RRAM en fonction de la donnée stockée.	221
3.2	Tensions SLT et SLF en fonction de la donnée recherchée. . . .	221
3.3	Définition des états des RRAM en fonction de la donnée stockée.	222
3.4	Tensions SLT et SLF en fonction de la donnée recherchée. . . .	222