Analog resistive switching devices optimized for deep learning applications

Von der Fakultät für Elektrotechnik und Informationstechnik der Rheinisch-Westfälischen Technischen Hochschule Aachen zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

genehmigte Dissertation vorgelegt von

Tommaso Stecconi, M. Sc.

aus Ancona, Italien

Berichter*innen: Universitätsprofessor Dr. Regina Dittmann
Universitätsprofessor Dr. Bert Jan Offrein

Tag der mündlichen Prüfung: 11.11.2024

Diese Dissertation ist auf den Internetseiten der Universitätsbibliothek online verfügbar.

Curriculum Vitae Tommaso Stecconi

Date	Event
11.11.2024	Examination for Dr.-Ing.
2020 - 2024	PhD student at IBM Research Europe - Zürich
06.02.2020	Diplôme d'ingenieur de l'École Centrale des Arts et des Manufactures (M. Sc. equivalent)
18.12.2019	Laurea Magistrale in Electronics Engineering, Politecnico di Milano (M. Sc. equivalent)
24.02.2017	Laurea in Ingegneria Fisica, Politecnico di Milano (B. Sc. equivalent)

Acknowledgements

The research work in this dissertation is the result of a tight collaboration among all the members of the Neuromorphic Devices and Systems (NDS) team at IBM Research Europe - Zürich.
Therefore, I would like to first thank Prof. Dr. Bert J. Offrein, our group manager, for having always nurtured a culture of sharing and teamwork among us. Thank you also for the trust you always gave to my work and for teaching me many valuable technical insights and communicational skills.
Another very important person for my PhD is Dr. Valeria Bragaglia, my direct supervisor, who has always supported my ideas and helped me to shape my research work into concrete and valuable results. Thank you for your daily effort of technical supervision and for helping me growing both at scientific level and in the soft-skills/relational aspect.
Thank you to my academic supervisor, Prof. Dr. Regina Dittmann, for the valuable and pleasant interactions we had. Your support on the TaO_x material development was critical for my research. Thank you also for welcoming so warmly during my stay in Jülich.
I would like to address special thanks to the Master students, PhDs, PostDocs and staff members that shared with me their valuable knowledge and introduced me to many efficient research methods used in IBM. Youri Popoff and Mattia Halter, former PhD students from our group, warmly welcomed me in our office and gave me very useful instructions and insights about processing and programming methods. Saketh Ram Mamidala and Wooseok Choi, two new PostDocs in our group, collaborated with me in ambitious projects, strongly accelerating the positive outcomes. Donato F. Falcone, current PhD in our group, has always been very helpful to me with many processing/measuring challenges. Folkert Horst, staff member, built important and reliable setups for the electrical measurements of my research. Antonio La Porta and Daniel Jubin, staff member and engineer, respectively, collaborated with me in many of such measurements. Elger Vlieg and Felix Hermann, current PhD students, have been extremely kind, positive and supportive colleagues, favoring a pleasant work environment in the team. Thank you to the master students Sofieke Charlotte ten Kate, Roberto Guido, Ludovico Carraria Martinotti and Davide G. F. Lombardo, who collaborated with me on various projects, bringing up brilliant ideas and contributing with methodologies of excellent quality. Together with Dr. Youri Popoff and Dr. Valeria Bragaglia, I co-supervised the research work of the Master Student Roberto Guido, who successfully defended his Master Thesis in 2021. A huge thank you is also devoted to the former/current cleanroom staff, Ute Drechsler, Antonis Olzierski, Diana D. Pineda, Steffen Riedt, Daniele Caimi, Richard Stutz, who supported me in the execution of multiple fabrication tests.
This research work is funded by the European Union within the H2020 “MANIC” (grant ID: 861153), “MeM-Scales” (grant ID: 871371), “PHASTRAC” (grant ID: 101092096), and by the SNF “ALMOND” (grant ID: 198612).

Abstract of the dissertation

Motivation, Goal, and Task of the Dissertation

In today's rapidly evolving landscape, Artificial Intelligence (AI) stands at the forefront of transformative technologies, revolutionizing operations across various sectors. This evolution has been facilitated by the optimization of digital hardware accelerators, crucial for enhancing AI capabilities. However, conventional digital accelerators, following the Von-Neumann architecture, face inherent limitations due to the separation of memory and logic units, leading to energy inefficiencies and operational latencies.

In response to these challenges, crossbar arrays of resistive switching devices, or Resistive Processing Units (RPUs), have emerged as promising alternatives to traditional digital accelerators. Integrated within the Back-End-Of-the-Line (BEOL) of chips, these arrays offer high-density storage of Neural Networks' (NNs) parameters directly on the chip. Besides their non-volatile memory functionality, RPUs serve as analog signal processors, executing Vector-Matrix Multiplications (VMM) fully in parallel with O(1) time complexity, significantly enhancing the efficiency of arithmetic computations in AI applications by mitigating memory access.

While RPUs have shown competitive performance in NN inference tasks, their utilization for training NNs has been hindered by various non-idealities. To be compatible with both inference and training applications, the technology inside the RPU must meet stringent programming criteria, such as a fast, low-power, linear, and symmetric resistive-switching capability.

Filamentary Resistive-switching Random Access Memory (ReRAM) devices, particularly those based on a Conductive-Metal-Oxide (CMO)/HfO_2 material structure, have emerged as promising candidates for addressing these challenges. Recent optimizations in material properties have mitigated non-idealities, making them viable for both inference and training tasks.

This dissertation investigates ReRAM devices based on the CMO/HfO_2 material structure, focusing on key material properties essential for fast, analog resistive switching. Through material and processing optimizations, improvements in unit cell scaling and operational power reduction during programming/reading measurements were achieved. These enhanced devices were integrated into small-scale crossbar array architectures, demonstrating parallel weight updates for NN training acceleration, thus contributing to bridging the gap between research on inference and training accelerators.

Major Scientific Contributions

- **Fabrication and characterization of analog, filamentary ReRAM devices based on the CMO/HfO_2 material stack.**
  ReRAM devices based on the CMO/HfO_2 material stack exhibit enhanced analog resistive-switching properties and reduced stochasticity. The physical explaination for these improvements is still debated, hindering further optimization of these devices.
  We addressed this issue by focusing on the impact of the CMO material properties on the ReRAM electrical response. To conduct this study, we developed a DC reactive sputtering deposition method for creating conductive-TaO_x thin films and tuning their resistivity.
  Multiple variations of ReRAM devices with different TaO_x material properties were compared. Devices with more resistive TaO_x layers showed a larger operating window under identical voltages and lower currents, which we attribute to the spreading resistance role of the TaO_x material. It was also found that the thickness of the TaO_x layer strongly affects the analog properties and the on/off ratio, which we relate to the thermal enhancement role of the TaO_x material.

- **Qualification of the ReRAM devices for the Tiki-Taka algorithm.**
  Tiki-Taka is a NN training algorithm developed to mitigate the impact of resistive-switching non-idealities, such as non-linearity and asymmetry.
  The optimized ReRAM devices were tested under a programming scheme replicating the pulsing stress occurring during a Tiki-Taka training workload. Excellent values in Tiki-Taka figures of merit were achieved and validated statistically.
  Reliability tests were conducted to evaluate the retention of the conductive states and the endurance of the analog programming properties. The devices exhibit negligible relaxation in timescales shorter than 1 s, and show a minor drift ($< 5\%$) for longer timescales up to 72 hours. Robust analog properties were demonstrated up to 10^8 programming pulses.

- **1T1R crossbar arrays exhibiting parallel and disturb-free synaptic updates.**
  The integration of the ReRAM devices on Complementary-Metal-Oxide-Semiconductor (CMOS) platforms is required to minimize device-to-device variability and hence demonstrate large-scale applications.
  The optimized ReRAM devices were integrated into the BEOL of a chip with integrated circuits in 350 nm CMOS technology. Stand-alone 1-Transistor-1-ReRAM (1T1R) structures were electrically characterized, with transistors providing current compliance functionality on chip to tune filament forming currents. A crossbar array of size 5×5 was used to demonstrate the parallel and analog update functionality required for deep learning accelerators based on resistive-switching devices.

Major Publications

1. Stecconi, Tommaso, et al. "Filamentary TaOx/HfO2 ReRAM Devices for Neural Networks Training with Analog In-Memory Computing." Advanced Electronic Materials 8.10 (2022): 2200448. (IF = 6.2)

2. Stecconi, Tommaso, et al. "Analog Resistive Switching Devices for Training Deep Neural Networks with the Novel Tiki-Taka Algorithm." Nano Letters 24.3 (2024): 866-872. (IF = 10.8)

3. Choi, Wooseok*, Stecconi, Tommaso*, et al. "Write-Disturb-Tolerant, Fully Parallel Stochastic Update of 1T-1ReRAM (1T1R) Crossbar Array for Analog Deep Learning Accelerators." Manuscript in preparation. * = shared first author

Contents

Acknowledgements **iii**

Abstract of the dissertation **v**

1 Introduction **1**
1.1 Artificial Intelligence (AI) and Moore's law . . . 1
1.2 How do AI models learn? Backpropagation . . . 2
1.3 The GPU: a digital electronic design optimized for training AI models . . . 4
1.4 Silicon memory technologies and In-Memory Computing (IMC) . . . 5
1.5 Emerging memory technologies . . . 8
1.6 Deep learning with crossbar arrays of analog Non-Volatile-Memory devices . . . 9
1.7 Optimized learning algorithms for resistive-switching crossbar arrays . . . 12
1.8 ReRAM devices and materials . . . 16
1.8.1 Resistive-switching mechanisms in the filamentary MIM-ReRAM . . . 16
1.8.2 Optimized filamentary ReRAM devices based on an oxide bilayer structure . . . 18
1.9 Research status on the In-Memory-Computing (IMC) deep learning accelerators . . . 19

2 Semiconductor processing and characterization methods **21**
2.1 Thin films deposition and etching methods . . . 21
2.1.1 DC reactive sputtering . . . 21
2.1.2 Plasma-Enhanced Atomic-Layer-Deposition (PEALD) . . . 23
2.1.3 Dry etching: Reactive-Ion-Etching (RIE) and Inductive-Coupled-Plasma (ICP) 24
2.2 Materials characterization . . . 26
2.2.1 Grazing-Incidence X-Ray Diffraction (GIXRD) and X-Ray Reflectivity (XRR) 26
2.2.2 Circular Transmission Line Measurements (CTLM) . . . 27
2.3 Device characterization . . . 29
2.3.1 DC and pulsed measurements . . . 29
2.3.2 Array measurements setup . . . 31

3 Materials optimization for analog ReRAM devices **33**
3.1 TaO_x material development . . . 33
3.2 ReRAM device processing and chemical analysis . . . 35
3.3 Comparison between baseline and oxide *bilayer* ReRAM . . . 36
3.4 Speculative interpretation of the resistive-switching mechanisms . . . 37
3.5 TaO_x resistivity and ReRAM electrical properties . . . 40
3.6 TaO_x thickness variations and ReRAM electrical properties . . . 40
3.7 Generalization of the oxide *bilayer* ReRAM design . . . 43

3.8 Conclusion . . . 44

4 ReRAM device scaling and analog properties 47
4.1 Via process and optimized HfO_x layer . . . 48
4.2 DC electrical properties of the scaled ReRAM devices . . . 49
4.3 Symmetry point properties . . . 50
4.4 Analog resistive-switching endurance . . . 53
4.5 Retention of the analog conductive states . . . 55
4.6 Localized delamination of ReRAM capping layers corroborating filamentary hypothesis 57
4.7 Device response model and deep learning simulations . . . 59
4.8 Analog ReRAM benchmark . . . 62
4.9 Conclusion . . . 63

5 Back-End-Of-Line (BEOL) integration and crossbar array measurements 65
5.1 BEOL integration of ReRAM devices . . . 65
5.1.1 Motivation . . . 65
5.1.2 General overview of the chip design properties and ReRAM integration . . . 66
5.1.3 Layout for ReRAM integration on CMOS . . . 68
5.2 DC characteristics of the single 1T structures . . . 70
5.3 Tuning the operating regime of 1T1R devices by adjusting the filament forming current 71
5.3.1 High forming current - DC characteristics . . . 71
5.3.2 Low forming current - synaptic characteristics . . . 71
5.4 1T1R programming methods towards array measurements . . . 73
5.4.1 Half-bias programming scheme . . . 73
5.4.2 Stochastic conductance updates with the half-bias scheme . . . 75
5.4.3 Resistance update suppression of non-targeted RRAM cells . . . 76
5.5 Array measurements - parallel crossbar updates . . . 80
5.6 Conclusion . . . 83

6 Conclusions and outlook 85
6.1 Conclusions . . . 85
6.2 Outlook . . . 87

List of publications 89

Conference talks 91

List of abbreviations 93

List of Figures 99

Bibliography 101

Annexe - ReRAM device development overview 117

Chapter 1

Introduction

1.1 Artificial Intelligence (AI) and Moore's law

In the past decade, we've observed a remarkable surge in Artificial Intelligence (AI) capabilities across various domains, including computer vision and language processing. This progress has been fueled by the emergence of groundbreaking Deep Neural Network (DNN) models, notably Convolutional Neural Networks (CNN) [1, 2] and transformers [3].
Simultaneously, significant ethical concerns arose regarding the substantial energy and financial costs associated with training these AI models [4]. Software research has addressed efficiency concerns by introducing significant algorithmic improvements [5]. However, the prevailing trends in AI indicate the deployment of larger models, whose training requires an exponentially increasing computational effort [6–8].
In this scenario, microchips play a fundamental role. The progressive release of more efficient and powerful AI processors [9, 10] by the electronics industry has facilitated and continues to support the adoption of larger models. The performance enhancements in the latest electronic devices result from a synergy of enhanced silicon technologies, innovative architectures, and advanced packaging methods [11, 12].
In the field of silicon technology, the adoption of novel transistor designs (such as Gate-All-Around Field-Effect-Transistor (GAAFET) [13, 14]), materials (such as SiGe and 2D materials [15, 16]), and optics (such as Extreme-UltraViolet (EUV) [17] lithography) is expected to sustain Moore's law in the near future [18, 19].
Nevertheless, numerous challenges lie ahead. According to the 2022 International Roadmap for Devices and Systems (IRDS) Physics Symposium, the fundamental principle governing device miniaturization [20] is expected to slow down or even halt by 2028 [19]. Subsequently, 3D integration [21–24] of beyond Complementary-Metal-Oxide-Semiconductor (CMOS) devices [25] is projected to play a pivotal role for further enhancing computer performances [19].

In Fig. 1.1, a qualitative graph presented at the 2020 International Electron Devices Meeting (IEDM) panel for beyond-CMOS and emerging materials [26] is reported. The figure clearly illustrates the primary business drivers motivating the push to overcome the gradual pace of Moore's law, by leveraging newer silicon technologies.

During my PhD, I focused my work on the topic of beyond CMOS devices and architectures. I have conducted research on an emerging memory technology called "*Resistive-switching Random Access Memory (ReRAM)*", and an In-Memory Computing (IMC) architecture known as "*crossbar array*".

Before delving into such advanced topics, the main subject of this dissertation, it is important to first introduce a typical structure of AI models and how it is trained.

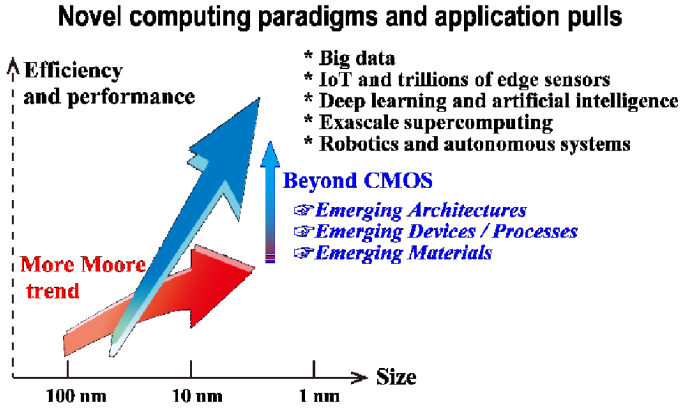


Figure 1.1: Figure presented at the IEDM 2020 by the IRDS Panel, on the need for novel elecronic devices and circuits, beyond conventional CMOS [26]. Modern applications, such as AI, Internet of Things (IoT) and Big Data, have raised a huge computational demand, which diverges from the performances offered by CMOS miniaturization alone.

1.2 How do AI models learn? Backpropagation

AI models typically consist of complex neural network (NN) structures, comprising multiple layers of neurons for efficiency and performance purposes [27]. The layers of neurons positioned between the input and the output layers are referred to as *hidden* layers. When there are multiple hidden layers, the NN is called a "*Deep Neural Network (DNN)*", and the field describing its properties and applications is called *"deep learning"* [28].
A fundamental DNN architecture is the Fully-Connected Network (FCN) [29], depicted in **Fig. 1.2 (a)**. The defining characteristic of an FCN is that each neuron is interconnected with all neurons in the preceding and succeeding layers, making it "*fully connected*". Due to this connectivity, the number of synaptic interconnections scales quadratically with the number of neurons.
While AI models consisting solely of a FCN have demonstrated lower efficiency compared to, for instance, CNNs [30], fully-connected layers remain crucial. They are frequently appended to the end of most models to facilitate final decisions [1, 3].

How do DNNs learn to make correct decisions? The algorithm conventionally adopted to train a DNN consists of two steps: *(1) Inference* and *(2) Backpropagation*. Deep learning inference is described in **Fig. 1.2 (a)**. First, the input data, for example, a 2D image, is flattened into a 1D array ($\mathbf{x}$), before being multiplied by a a 2D array of synaptic weights ($\mathbf{\Theta}$). This operation is a Vector-Matrix-Multiplication (VMM), resulting in a 1D array ($\mathbf{z}$), where each element is obtained through a Multiply-Accumulate (MAC) operation.
Subsequently, the result of the VMM is processed through a non-linear neuronal activation function, $\mathbf{a}$=f($\mathbf{z}$). The VMM + non-linear transformation task is iterated $n+1$ times, where n is the number of hidden layers, until a decision (the output $\hat{\mathbf{y}}$) is reached. This whole process is known as *"forward pass"*.

In supervised learning [31], the correct output is known ($\mathbf{y}$). The decision and the output are

compared, through a metric quantifying the decision error. The goal of the training process is to minimize this error, also called *"cost function"*. For simple linear regression applications, a common cost function is the mean square error [32].

Back-propagation is the typical process for reducing the cost function [32]. In back-propagation,

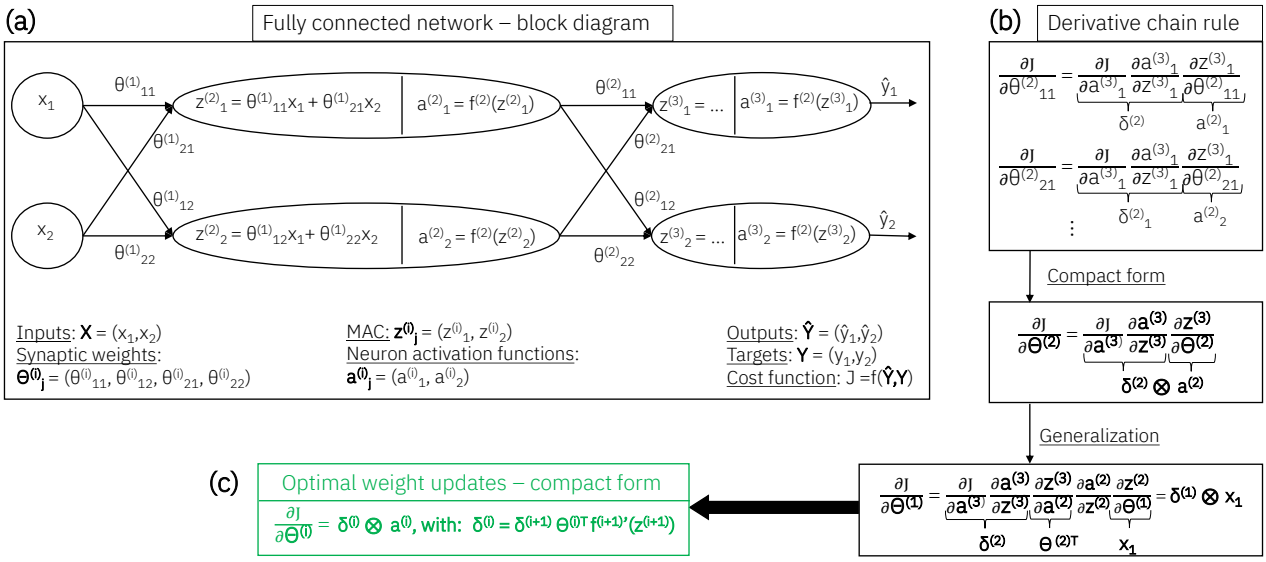


Figure 1.2: **(a)** Inference workload block diagram in a Fully-Connected-Network. **(b)** Mathematical derivation of the optimal weight updates, based on the backpropagation algorithm [32]. **(c)** Compact mathematical expression for the optimal weight upates.

the synaptic weights of the whole neural network are optimized based on the calculated error. The equations defining how the error is propagated backwards can be derived from the *chain rule*, as described in **Fig. 1.2 (b)**. At each layer *(i)*, the error, called $\boldsymbol{\delta^{(i)}}$, is a multiplication between three terms: $\boldsymbol{\delta^{(i+1)}}$, $\boldsymbol{\Theta^{(i)T}}$, $\boldsymbol{f'^{(i+1)}}$. $\boldsymbol{\delta^{(i+1)}}$ is the error at the next layer. $\boldsymbol{\Theta^{(i)T}}$ is the traspose of the 2D synaptic interconnects between layers *(i)* and *(i+1)*. $\boldsymbol{f'^{(i+1)}}$ is the derivative of the activation function of the layer *(i+1)*The backpropagation product $\boldsymbol{\delta^{(i+1)}}$ $\boldsymbol{\Theta^{(i)T}}$ is also known as *backward pass*.
This result implies that to calculate each $\boldsymbol{\delta^{(i)}}$, one must first know $\boldsymbol{\delta^{(i+1)}}$, and so on with the errors at the next layers. The chain rule is a linear property, which translates into a computational effort linearly scaling with the number of hidden layers [28]. This property makes deep learning a sustainable workload, even for very deep networks.
The expression for the optimal weight updates, also derived from the chain rule, as reported in **Fig. 1.2 (b)**, is:

$$\boldsymbol{\Delta\Theta^{(i)}} = \boldsymbol{\delta^{(i)}} \otimes \mathbf{a^{(i)}} \tag{1.1}$$

where $\otimes$ is the symbol for the outer product operation.

In deep learning, during the inference step, each input data is processed through a sequence of *n+1* VMM calculations, where *n* is the number of hidden layers. These operations are iterated for all inputs in the dataset.

Moreover, the same sequence of training data must be processed through the same model multiple times, referred to as "*epochs*", until accurate decision-making behavior is achieved. Once the model is trained, it undergoes evaluation with new data. During this process, a sequence of n+1 VMM operations is again performed for each input data.
Another significant operation discussed in this section involves the outer product, which is utilized to compute optimal weight updates. This process is iterated n/b times, with b representing the *batch size* of the input data, and it is then reiterated across numerous epochs.

To summarize, AI models involve multiple operations, such as VMMs and outer products, that entail repetitive operations on arrays of predetermined sizes. A hardware specifically designed for deep learning must account for these aspects.
In the following section, I will introduce modern processors tailored for AI, based on digital electronic systems, and pinpoint the efficiency bottlenecks.

1.3 The GPU: a digital electronic design optimized for training AI models

The Graphics Processing Unit (GPU) [33] stands out as the commercially available device specifically designed for training AI models [33]. By delving into the GPU's main features and identifying where efficiency bottlenecks occur, we can better understand the need to explore alternative computing paradigms for a more sustainable approach.
Compared to the Central Processing Unit (CPU), the conventional computer processor, a GPU reduces the area dedicated to the control logic and the cache memory, while allocating more silicon for arithmetical computations [34]. The arithmetic units of a GPU are more numerous, smaller and more uniformly distributed across the chip area [34]. Based on such properties, the GPU offers enhanced computational parallelism, compared to the CPU [34]. A representation of a recent NVIDIA GPU architecture (A100) is depicted in **Fig. 1.3** below [35].
The core processing elements of a GPU are called *Streaming Multiprocessors* (SM). The SM are designed to access the on-chip memory (the L2 cache) in parallel, as depicted in **Fig. 1.3**.

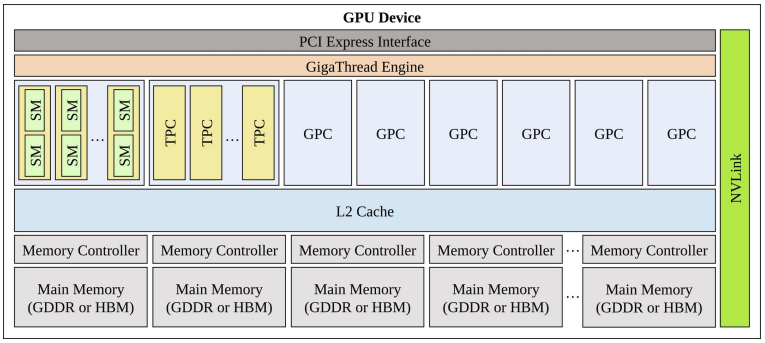


Figure 1.3: NVIDIA GPU A100 [35] layout.

Given that the L2 cache size is limited to a few megabytes [36], and modern AI models consist of several billion parameters [7, 8], the internal memory capacity of the GPU proves insufficient to accommodate such data. Additionally, certain AI models require extensive datasets to effectively learn specific functionalities. Consequently, to address these limitations, multiple off-chip memory devices with larger capacities (up to few gigabytes [37]), are integrated heterogeneously around the GPU [34, 38, 39], as

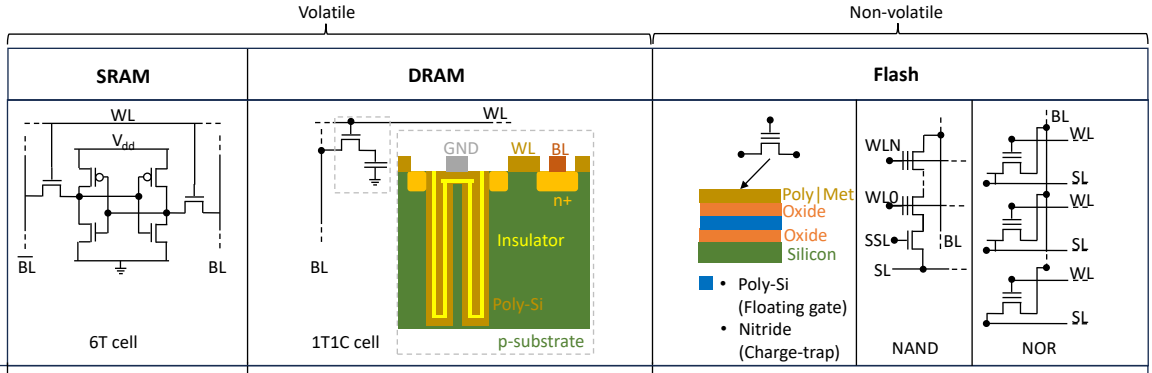
illustrated schematically in **Fig. 1.3**.
The transfer of data between these external memory devices and the processing unit is recognized as one of the primary sources of computational inefficiency [40, 41]. Two main motivations underlie this inefficiency. Firstly, the throughput of logic devices has significantly outpaced the bandwidth of main memory [42]. This creates a bottleneck for overall computer performance, commonly referred to as the *"memory wall"*. Secondly, accessing off-chip memory incurs an energy cost roughly two orders of magnitude higher than accessing internal cache. [40].
However, newer GPU designs have addressed this memory efficiency issue. Recent advancements in organic packaging technologies [43] have enabled the vertical stacking of multiple Dynamic Random-Access Memory (DRAM) devices, thereby enhancing memory bandwidth. These memory devices are known as "High Bandwidth Memory" (HBM) [11, 38].
Another important constraint for digital AI accelerators is the maximum size of the processor die, which is limited by the reticle size of state-of-the-art Extreme-Ultra-Violet (EUV) lithography, approximately ~ 800 mm^2 [38, 44]. To address this optical limitation, the latest GPU designs incorporate multiple high-end processing dies, closely packaged and operating as a single GPU (2 dies in the NVIDIA Blackwell GPU [45], 8 dies in the AMD Instinct^{TM} MI300X accelerators [46]). This approach, while not novel, involves splitting the computational functionality from a single die into multiple smaller ones, known as *"chiplets"*. It is gaining traction in the AI hardware industry [38, 47] due to the escalating cost of high-end silicon.

To summarize, I have discussed the critical aspects concerning cost and efficiency in GPU devices. The integration of large silicon dies with advanced CMOS nodes has led to escalating expenses. Additionally, the *memory wall* poses a significant bottleneck to overall hardware efficiency. To address these challenges, recent R&D efforts in both industry and academia have concentrated on 3D memory and chiplets.

As GPU design progresses, exploration into other unconventional computing paradigms is underway. The *memory wall issue*, stemming from the physical separation of memory and processor into two distant units, has led to the encouragement of novel computing architectures providing computational capabilities near or within memory. This alternative computing paradigm, known as In-Memory Computing (IMC), is rapidly gaining interest [48].
In the next paragraphs, I will introduce some popular IMC architectures, specialized in performing deep learning workloads 1.2. First, IMC circuits based on established silicon memory technologies will be presented, followed by a discussion on emerging memory technologies.

1.4 Silicon memory technologies and In-Memory Computing (IMC)

The established memory devices based on silicon technologies include Static Random-Access Memory (SRAM), Dynamic RAM (DRAM), and Flash memory. **Fig. 1.4** illustrates their fundamental cell structures and main properties such as scalability, energy consumption, programming/read speed, and endurance (specification values extracted from [49–53]). All of these memory technologies have been investigated for In-Memory Computing (IMC) applications. In the following lines, I will discuss some fundamental aspects to consider. All these memory technologies have been explored for IMC applications. In the following lines, I report some fundamental aspects to be accounted.
(1) SRAM and DRAM utilize a binary unit cell and are hence unable to perform multi-level weight representation. This limits the extent to which the MAC operation can be performed within the memory array's circuit terminals. To complete the operation, additional components must be connected to the memory circuit. Many designs incorporate mixed-signal circuits, utilizing analog components



Cell size	120-150 F^2 [49]	Embedded: 10 - 30 F^2 [49] | Stand-alone: 6 F^2 [50]	Embedded (NOR): 10 - 30 F^2 [49] | Stand-alone (NAND): 4 F^2 [51]
CMOS node compatibility	5 nm [52]	Embedded: 14 nm [52]	Embedded: 28 nm [53]
V_{write}	< 1V [49]	< 1V [49]	~ 10 V [49]
E_{write}	~ fJ [49]	10 fJ [49]	~ 1 pJ [49]
Speed	Write ~ Read ~ 1 ns [49]	Write ~ 10 ns; Read ~ 3 ns [49]	Stand-alone: Write ~ 0.1 – 1 ms; Read (random) ~ 10 µs [51] Embedded: Write ~ 0.1 – 1 ms; Read ~ 10 ns [49]
Endurance	10^{16} [49]	10^{16} [49]	10^4 – 10^6 [49]

Figure 1.4: Basic cell structure and circuit architecture of the established memory technologies based on silicon. Comparison between their main electrical properties.

like switching capacitors to adjust the bit-significance of the MAC operation [54, 55].
(2) Only SRAM technology scales with Moore's law [52]. DRAM and Flash can be embedded on a chip with processing capabilities, such as a microcontroller. However, the embedded DRAM and Flash require manufacturing processes uncompatible with very advanced logic designs (minimum 14 nm for eDRAM [52, 56] and 28 nm for eFlash [53]).
(3) Both SRAM and DRAM are volatile memory technologies, making them unsuitable for data storage in power-constrained platforms.
(4) Flash memory has a maximum endurance of 10^6 write/erase cycles [51], making its unit cell unusable for iterative updates required in DNN training tasks. For instance, considering an application with datasets larger than 10^6 inputs, and undergoing 100 iterations during training, the number of weight updates can exceed $\sim 10^8$, surpassing the endurance of Flash memory devices.
Additionally, Flash memory devices undergo a shared erase operation across a whole block, typically arrays of many kilobytes of data [51]).

Understanding the operational principles of SRAM, DRAM, and Flash designs in detail is not essential for comprehending the dissertation's subject and motivation. However, it is crucial to summarize the technology trends and what is the state-of-the-art, to competitively position our work in the IMC landscape.
The Figures of Merit (FoM) conventionally adopted for IMC chip design include *chip throughput*, measured in Tera-OPerationS per second [TOPS], and *computational efficiency*, measured in Tera-OPerationS per second per Watt [TOPS/W]).

- **SRAM:** SRAM-based IMC macros have leveraged Moore's law over the past decade, advancing from 130 nm to 4 nm technology [52, 54, 57–61].

The primary advantage offered by SRAM-based IMC is its high efficiency, reaching up to approximately $\approx$ 5000 TOPS/W [54, 61], surpassing benchmark GPU performance by almost two orders of magnitude [54, 60, 62]. Recently, SRAM-based IMC chips achieving 256 Mb memory capacity at 7 and 5 nm nodes [63, 64] were demonstrated. The continuation of Moore's law will enable further scaling of the SRAM cells and will allow to hold on chip most of neural networks weights [52]. This is especially applicable to tasks such as image classification and object detection, where the number of parameters typically range from $\sim$ tens to $\sim$ hundreds of millions [65,66].

- **eDRAM:** eDRAM provides higher memory density compared to SRAM, with the smallest conventional DRAM cell utilizing a 1-Transistor-1-Capacitor (1T1C) structure, as opposed to the minimum 6T of an SRAM cell. To shrink the 1T1C structure, the capacitor is typically fabricated using deep trenches in silicon, with a planar size similar to that of a FinField-Effect-Transistor (FinFET) channel [56], as depicted in **Fig. 1.4**.
  However, the read operation in a 1T1C cell is destructive, and leakage current discharges the memory capacitor, necessitating frequent refresh cycles ($\sim$ tens of milliseconds) to maintain bit state integrity [67].
  To address the destructive read issue, more complex structures known as *"gain cells"* [68] have been introduced. The simplest gain cell is based on a 2T1C architecture, but a 3T2C structure has been proposed [69, 70] to enable signed multiplications in-memory and provide enhanced robustness to cell leakage current.
  The size of the *gain cells* has followed a scaling trend, decreasing from 65 nm to 28 nm [55, 71–73]. At similar technology nodes (down to 16 nm), eDRAM with gain cells offers higher memory density and efficiency compared to SRAM IMC macros, with comparable throughput (refer to benchmark table in [69]).
  However, compared to the 4 nm SRAM-CIM macro presented last year by the Taiwan Semiconductor Manufacturing Company (TSMC) [61], the efficiency (Macro peak efficiency = 2316 TOPS/W/bit [69]) and throughput (=40.6 TOPS/mm^2/bit [69]) of eDRAM are lower (6163 TOPS/W/b and 4790 TOPS/mm^2, respectively, for SRAM).
  However, DRAM still provides higher memory density (Macro DRAM: 92.2 Mb/mm^2 [69] vs Macro SRAM: 3.1 Mb/mm^2 [61]).

- **eFLASH:**Flash memory serves as the established Non-Volatile Memory (NVM) in silicon, available in two architectures depending on the logic access type: NAND and NOR. **Fig. 1.4** illustrates the architecture and the main features of both Flash memory types. The NAND architecture arranges memory cells in series, akin to n-type MOSFETs in a CMOS NAND logic gate. Similarly, NOR architecture aligns memory cells in parallel, connecting the drain terminal to the same Bit Line (BL) and the source terminal to the same Source Line (SL).
  The NAND architecture excels in storage density, with capacities scaling to over 10 Gb/mm2 today >10 Gb/mm^2 [74], facilitated by 3D monolithic integration and dense 4F^2 memory reticles [51] (where F represents the minimum lithographic feature size). However, the Bit-Cost-Scalable (BiCS) process flow developed for sustainable 3D NAND production [75] is incompatible with advanced logic chip integration [52].
  Conversely, in many eFlash applications, such as automotive embedded systems, properties like low cost and high-temperature reliability outweigh storage density [53]. Therefore, eFlash is predominantly commercialized through NOR architectures [53].
  Recent advancements in IMC with Flash technologies primarily target AI inference applications. Modern NAND Flash IMC designs offer a combination of excellent efficiency (37.1 TOPS/W),

throughput (1545.6 Frames Per Second), and storage capacity (1.13 Gb) [76]. The technology's appeal lies in its high storage density and reliable, low-power read operations, facilitated by the memory cell's low current and large on/off ratios (I_{on} ∼ 1-100 nA, I_{off} > 1 pA [77,78]).
Novel 3D NOR architectures featuring independent source lines and word lines within a single block have emerged [79, 80], demonstrating high efficiency (> 3.7 TOPS/W) and very high memory density (2.7 Gb per chip), similar to 3D NAND but with ∼ 100x faster access.

In conclusion, Flash represents a promising technology for AI inference applications due to the high memory density achieved by 3D architectures. However, 3D processes are incompatible for monolithic integration on advanced logic chips. With recent developments in multibit Flash cells (recently shown up to 5 bits/cell [81]), combined with advancements in 3D integration processes (currently up to > 280 layers [82]), 3D Flash holds important prospects for AI applications, serving as a host processor where to offload AI workloads.

1.5 Emerging memory technologies

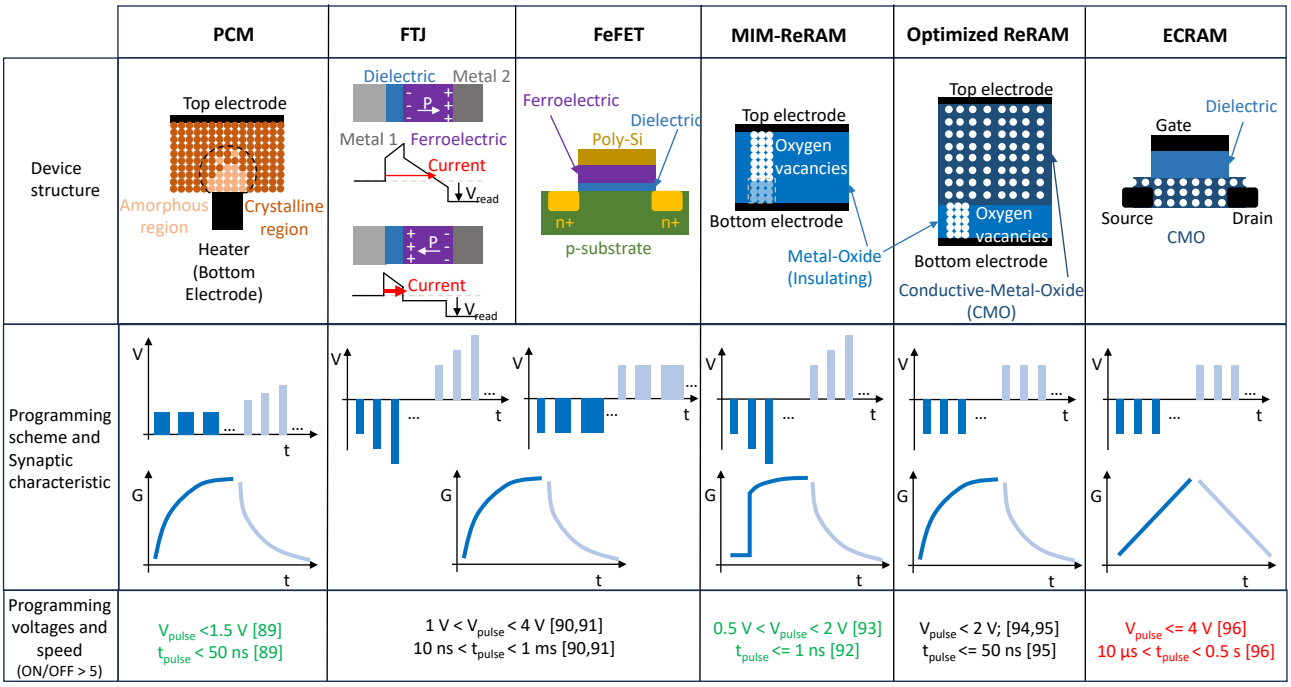


Figure 1.5: Overview of the main emerging Non-Volatile Memory (NVM) devices for deep learning applications. The structural composition, operational principles, synaptic characteristics, and key electrical parameters crucial for deep learning applications are outlined.

Over the past 20-30 years, various technologies based on materials other than silicon have emerged for IMC. These novel memory devices include those based on ferroelectric materials (such as Ferroelectric Random-Access Memory, FeRAM [83, 84], Ferroelectric Field-Effect-Transistors, FeFET [83] and Ferroelectric Tunnel Junction, FTJ [85]), phase change materials (such as the Phase Change Memory, PCM [86]), magnetic materials (such as Magnetic Tunnel Junction, MTJ [87]), and valence-change processes (such as Resistive-switching Random Access Memory, ReRAM [88]). Their material structure,

non-volatile conductance update characteristics, and some fundamental electrical properties, such as programming speed and pulse amplitude, are presented in **Fig. 1.5** (specification values extracted from [89–96]). Initially, it was believed that such devices could offer more efficient storage functionalities compared to DRAM and Flash by providing faster access times than Flash and better scalability (cost/bit) than DRAM [97]. A new class of non-volatile memory devices was thus defined under the name Storage Class Memory (SCM) [97].
During that period, expectations regarding the scalability of conventional silicon memory technologies were modest [97]. There was considerable interest in novel non-volatile memory devices with the potential for enhanced scalability [98, 99]. However, the introduction of the first BiCS 3D NAND Flash [75] changed the outlook for future storage scalability. NAND Flash has recently achieved storage densities up to 50 times higher than those of a decade ago [100]. Additionally, DRAM densities have also increased, albeit at a slightly slower pace (approximately half the increase of NAND Flash [100]), owing to various materials and cell design innovations [100, 101].
The progressive scalability advancements demonstrated by silicon memory technologies have somewhat diverted the research and development focus away from developing SCM products. In the meanwhile, AI-related businesses have emerged, gaining increasing relevance. Consequently, the emphasis on applying emerging memory technologies has shifted towards IMC.
In the following section, I will present and discuss the *crossbar array*, an IMC architecture based on NVM devices, designed for accelerating AI workloads. Subsequently, I will focus on the application of DNN training, discussing optimized algorithms and NVM devices.

1.6 Deep learning with crossbar arrays of analog Non-Volatile-Memory devices

- **Inference:** The block diagram of an FCN is shown in **Fig. 1.6 (a)**. In this network, neuronal layers are fully connected by 2D arrays of synaptic weights, denoted in the figure as $G_{i,j}^{(l)}$, where l refers to the layer depth, i and j define specific positions (row and column) in the 2D array.
  Fig. 1.6 (b) displays the circuit architecture of a crossbar array, characterized by its 2D planar design. Each memory cell, depicted as a resistor, connects to both a horizontal and vertical line. The crossbar array comprises N metallic rows and M metallic columns, facilitating parallel access to N*M memory cells.
  This architecture efficiently handles typical algebraic calculations of deep learning tasks by leveraging analog signal processing. Now, let's delve into the details. First, consider one layer of 2D synaptic interconnects, from the FCN shown in **Fig. 1.6 (a)**. For instance, let's examine the connections from the input data (x_1, x_2, ..., x_N) to the first layer of neurons, with synaptic weights labeled as: G_{11}^0, ..., G_{NM}^0.
  These synaptic weights undergo a linear transformation, where they are mapped into corresponding NVM conductance values: G_{11}, ..., G_{NM}. The elements of the crossbar array are then accurately programmed to match these values.
  The input data are mapped into a 1D array of input voltage signals for the crossbar. As illustrated in **Fig. 1.6 (b)**, a general transformation of the input data involves PWM of the input voltage signals x_1, x_2, ..., x_N. PWM is a modulation technique that addresses issues related to the non-linear current-voltage characteristics of the NVM devices [102].
  At this stage, the *forward pass* operation can be performed. By grounding the vertical lines virtually, the equation:

$$I_i = \Sigma_j G_{ij} V_j \tag{1.2}$$

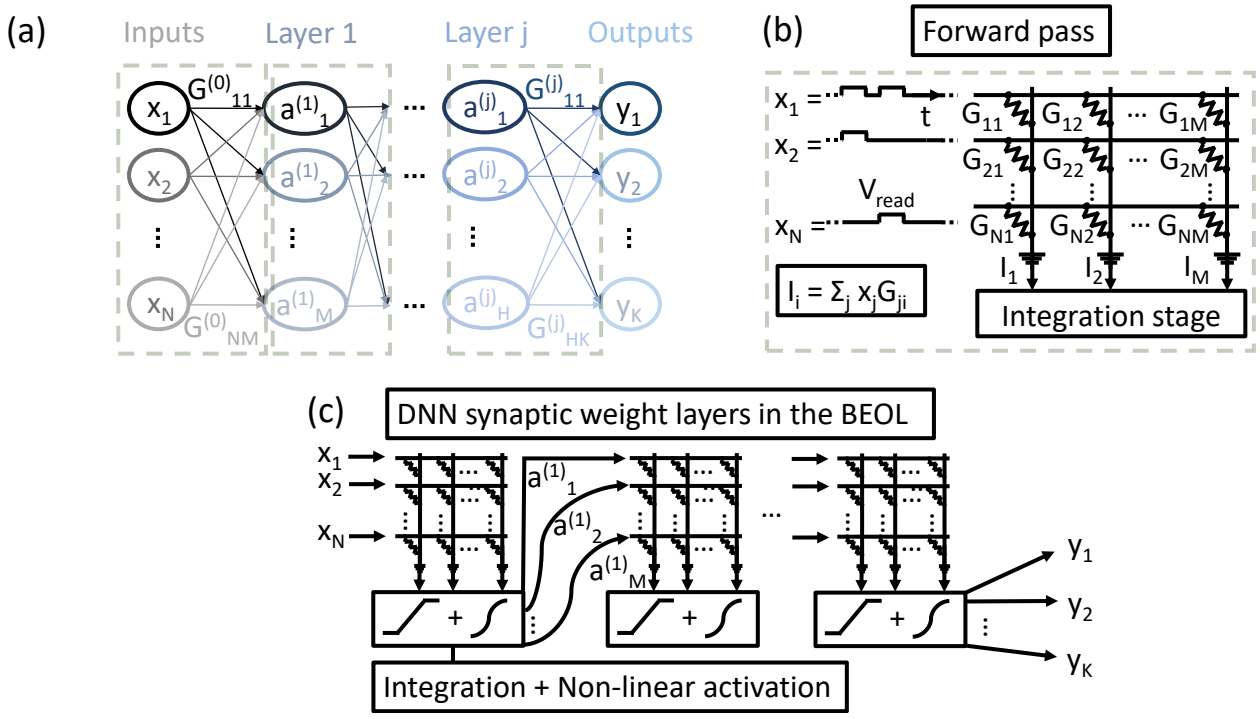


Figure 1.6: **(a)** Schematic representation of a FCN, illustrating the input layer, synaptic weights, and output layer configuration. **(b)** Implementation of the forward pass operation using a crossbar array, achieved through Pulse-Width-Modulation (PWM) of input signals and subsequent integration of output currents, resulting in a complete VMM. **(c)** Inference workload depicted through a sequence of analog tiles, integration stages, and non-linear transformation stages, cascading to execute the complete inference task.

describes each current signal I_1, I_2, ..., I_N, as shown in **Fig. 1.6 (b)**. This formula is derived from the application of Ohm's and Kirchhoff's laws, considering a generic column i and all rows (1,..., N). Since the input voltage is modulated by PWM, the output currents must be integrated before delivering the total result of the MAC operation. The resulting 1D array represents the outcome of the VMM operation between the input neuron activity $\mathbf{x}$ and the 2D array $\mathbf{G}$ of synaptic weights. Due to the linear processing based on the conductive states of the array elements, the crossbar array is referred to as a "*Resistive Processing Unit*" (RPU).

The main advantages of exploiting an RPU for deep learning inference are:
(1) Reduced time complexity: The VMM operation is performed in a $\sim$ O(1) time complexity [103], as it remains constant regardless of the array sizes (N and M). In contrast, digital circuits require $O(N^2)$ time complexity to execute the VMM operation [104]. Furthermore, if the resistive elements provide an Ohmic current-voltage (I-V) characteristic, the PWM becomes unnecessary. Instead, the input voltages can be modulated by the amplitude of the x_i signals, leading to further acceleration of the VMM operation.
(2) Large on-chip storage capacity: IMC with emerging NVM devices enables the full storage of DNN parameters on-chip. All NVM devices shown in **Fig. 1.5** are, in principle, Back-End-Of-Line (BEOL) compatible. Therefore, by designing multiple RPUs in the BEOL, all DNN parameters can be stored on-chip. Moreover, by cascading the inference task through multiple

crossbar arrays, as depicted in **Fig. 1.6 (c)**, the workload is efficiently split between BEOL and Front-End-Of-Line (FEOL). The linear algebra operations, such as the VMM operations, can be executed in the BEOL analog tiles, i.e. the RPU, as depicted in **Fig. 1.6 (c)**. Since in modern deep learning models, such as CNN and transformers, around $\approx 98\%$ of the total computational effort is related to MAC operations [105], the benefits introduced by analog processing in the BEOL is scalable and will increase the processing efficiency by more than two orders of magnitude, compared to GPU [105].
The non-linear activation functions of the neurons can be performed in the FEOL by conventional digital circuitry.

- **Training: (1) Backward pass.** The same considerations discussed for the *forward pass* also apply to the *backward pass*. In this case, the input is the backpropagated error, a 1D array labeled as $\boldsymbol{\delta}$, which must be multiplied by the transpose of the weight matrix, $\boldsymbol{\Theta}^{(\mathbf{i})\mathbf{T}}$, as discussed in the previous section 1.2. Therefore, to implement the backward pass, input and outputs are permuted compared to the schematics shown in **Fig. 1.6 (c)**. The new input $\boldsymbol{\delta}$ is applied to the vertical lines of the crossbar, while the outputs are the currents collected at the horizontal lines, which are set to virtual ground. These currents represent the result of multiplying $\boldsymbol{\delta}$ by the transpose of the weight array, G^T [102]. The backward pass operates with an approximate time complexity of $\sim$ O(1).

  (2) Weight updates. The architectural parallelism of the crossbar array offers an efficient approach to accelerate the weight updates step. In section 1.2 the optimal weight update formula was derived as $\boldsymbol{\delta}^{(i)} \otimes \mathbf{a}^{(\mathbf{i})}$ [32], representing an outer product operation resulting in a 2D array of ΔG_{ij} conductance updates. Gokmen and Vlasov [106] proposed a method to implement fully-parallel weight updates with a time complexity of O(1) [106, 107], as illustrated in **Fig. 1.7 (a)**. In this method, the inputs of the multiplication, $\boldsymbol{\delta}^{(i)}$ and $\mathbf{a}^{(\mathbf{i})}$, are transformed into two stochastic voltage pulse streams. The pulsing probability of these streams linearly depends on the values of the $\boldsymbol{\delta}^{(i)}$ and $\mathbf{a}^{(\mathbf{i})}$ arrays.
  During this process, the pulsing activity drives the $\boldsymbol{\delta_i}$ and $\mathbf{x}_i$ signals either to V_{prog} or 0 V, depending on the sign of the weight update. Whenever two coinciding pulses of opposite polarity occur at the device terminals, the full V_{prog} is applied, leading to a conductance update. Implementing signed multiplications with this approach involves carefully handling the signs of the input signals and applying appropriate voltage pulses to achieve the desired conductance updates. Let's consider a case example illustrated in **Fig. 1.7 (b)**, where the inputs are: $x_1 = +0.3$, $x_2 = -0.1$, ..., $x_N = +0.2$, and $\delta_1 = -0.2$, $\delta_2 = -0.2$, ..., $\delta_M = +0.3$. To complete the outer product operation, the task is divided in 4 steps based on the signs of the input values: (++), (+-), (-+), (–). In the first step, only signals x_1, x_N and δ_M are applied at the RPU terminals. For this step, the signs of the input signals should result in voltage pulses that lead to a positive update of the device conductances. Similarly, in the (–) step, since multiplying two negative scalars results in a positive sign, the voltage pulses must also lead to a positive update.
  Conversely, in the (+-) and (-+) steps, the polarity of the x_i and δ_j signals needs to be inverted. This ensures that the V_{prog} pulses induce a negative change in the device conductance.

  In deep learning, the equation describing the optimal weight updates 1.1 is typically adjusted by a linear factor L, known as the *learning rate*. This hyperparameter requires careful tuning to expedite (and improve) the learning process convergence. [106].
  In the implementation using RPUs, the learning rate is incorporated by dividing each update cycle (++, +- , -+, –) into L identical time intervals. Within each interval, a stochastic pulse

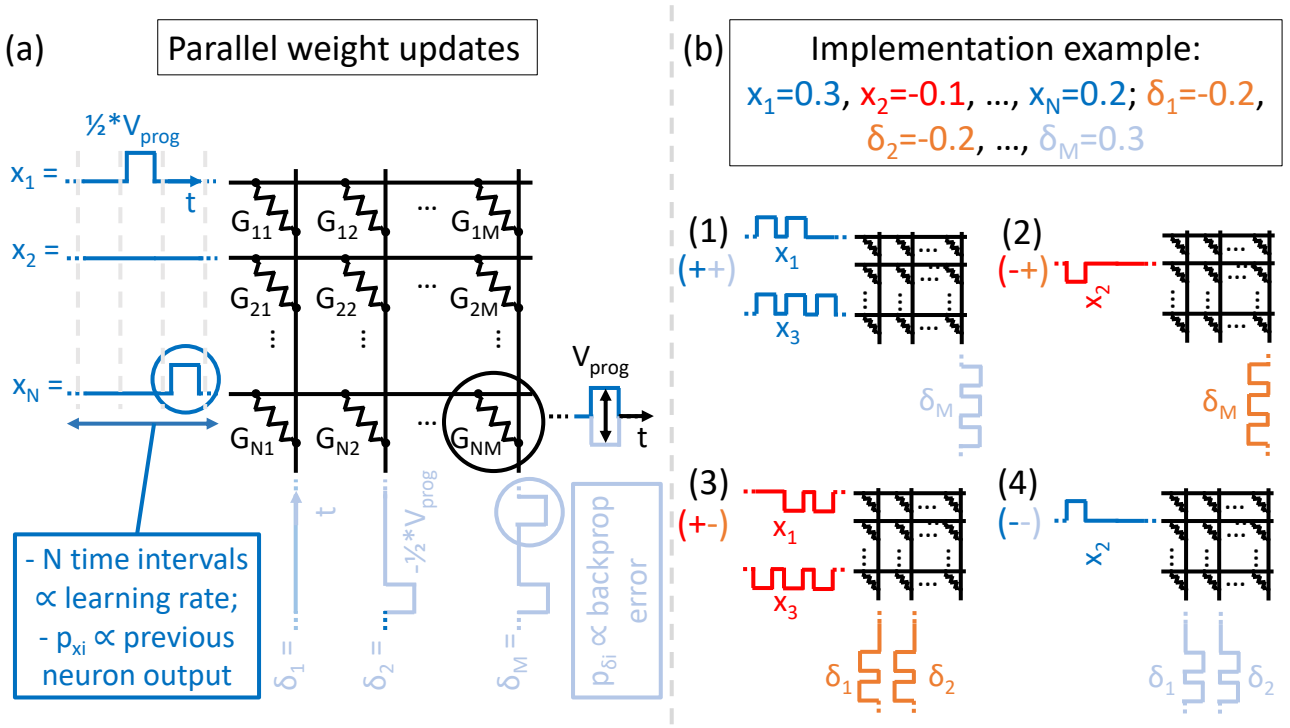


Figure 1.7: **(a)** Outer product operation executed fully in parallel with a crossbar array of NVM devices. Stochastic pulses are applied to the RPU terminals, leading to an analog conductance update whenever two coinciding pulses of opposite polarity are detected at the device terminals. **(b)** Numerical example elucidating the 4-step implementation cycle for fully-parallel weight updates.

coincidence may occur, with its probability determined by the values of the x_i and δ_j inputs.

Despite the process involving a 4-cycle algorithm, its time complexity remains O(1), as it doesn't scale with the array size. In contrast, digital hardware like GPUs require $O(N^2)$ time complexity to execute such an outer product operation [107].

1.7 Optimized learning algorithms for resistive-switching crossbar arrays

The parallel weight updates algorithm is based on the foundational assumption that the multiplication of two scalars, a and b, can be effectively replicated by a series of individual updates, repeated $a*b$ times.
To illustrate this concept, let's examine a numerical example from **Fig. 1.7 (b)**. Suppose x_N represents a scalar input of +0.2, and δ_M is another scalar input of +0.3. Their product is calculated as +0.06. Assuming a specific learning rate, say L=100, we would expect an average of 100 * 0.06 = 6 pulse coincidences instances of pulse coincidences during the (++) cycle for the G_{NM} element.
This approach accurately mirrors the theoretically optimal weight update process, under the assumption that the NVM devices are linear and symmetric. In such a scenario, each pulse coincidence would prompt a conductance update of either $+\Delta G_{min}$, or $-\Delta G_{min}$, depending on the polarity of the pulse. However, if the devices exhibit non-linear or asymmetric behavior, the accuracy of the learning process may diminish, potentially leading to a failure of the AI model to learn, as demonstrated in

simulations [106].

To summarize: **(1)** the critical requirement of the parallel weight updates method with an RPU is that the NVM device exhibits an *ideal* conductance-to-programming pulse characteristic, characterized by linearity and symmetry.
(2) Additionally, the NVM device must demonstrate a cumulative change in their conductance when subjected to identical pulse streams. If the device necessitates incremental voltage pulses, it implies that the conductance updates are influenced not only by external parameters, $\boldsymbol{\delta}$ and $\mathbf{x}$, but also by the current and the optimal G values. Determining these optimal G values requires computing the outer product outside the RPU.

The majority of emerging NVM devices necessitate incremental voltage pulses, as depicted in **Fig. 1.5**. For instance, PCM devices exhibit accumulative behavior only for positive G update directions, requiring incremental pulses for the opposite direction. Similarly, the ferroelectric-based devices and the Metal-Insulator-Metal (MIM) ReRAM demand incremental voltage pulses in both polarities.
Only Electro-Chemical Random-Access-Memory (ECRAM) devices [96], exhibit linear and symmetric conductance updates, which makes them particularly suitable for applications like parallel weight updates. However, the ECRAM technology come with critical drawbacks, as it requires long ($\gg 10$ μs) and large-amplitude (~ 5 V) voltage pulses [96] to achieve comparable on/off ratios with other emerging NVM technologies [96].

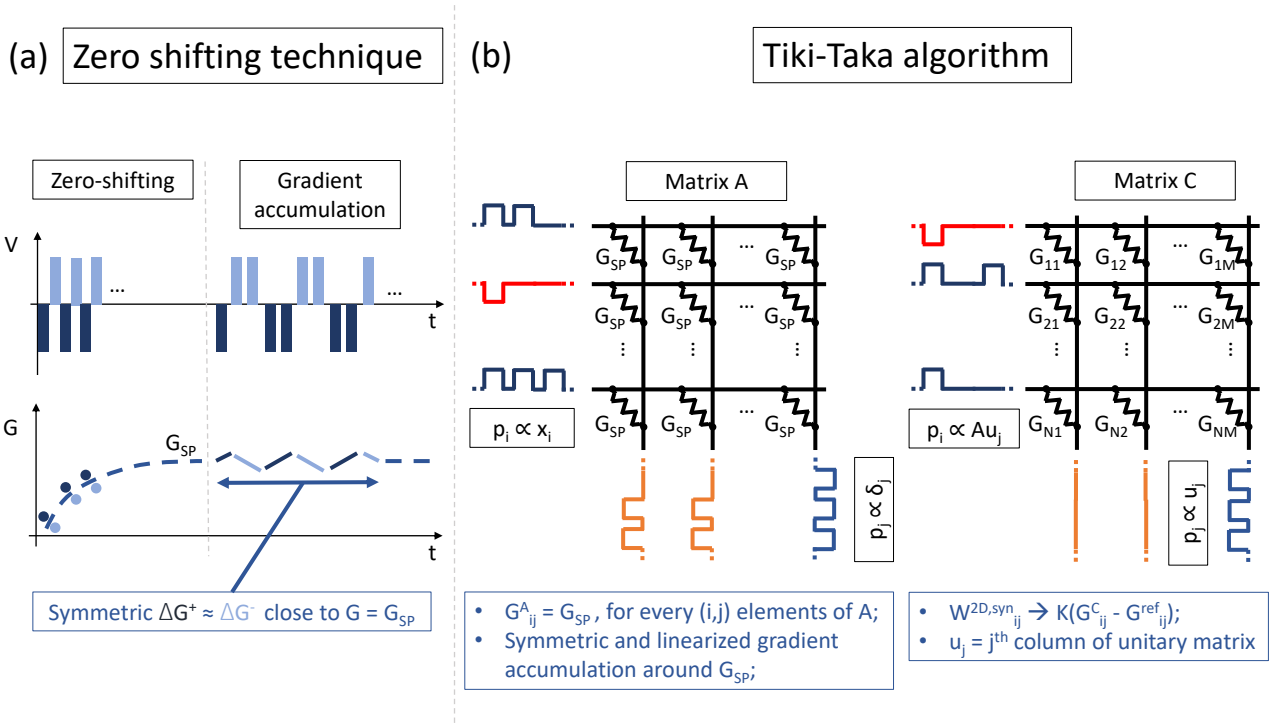


Figure 1.8: **(a)** The *zero-shifting* procedure involves applying a sequence of programming pulses with alternating polarity to the resistive-switching device. This drives the device to a state known as the *symmetry point* state, where updates become symmetric. Following this preparatory stage, the accuracy of gradient accumulation improves significantly, benefiting from enhanced symmetry and linearity. **(b)** This illustration depicts the pulsing activity surrounding matrices **A** and **C** during the operation of the Tiki-Taka algorithm.

A novel DNN training algorithm, named *Tiki-Taka*, has been recently developed to address the challenges posed by device asimmetry and non-linearity [108]. This algorithm draws inspiration from a football technique known as "Tiki-Taka", where players frequently pass the ball among themselves. In the Tiki-Taka algorithm, two crossbar arrays engage with frequent interactions, in analogy to the collaborative passing strategy seen in the football technique.
The core principle of the Tiki-Taka algorithm lies in the division of two tasks - weight updates and weight mapping - across two distinct crossbars, designated as matrix **A** and **C**, respectively. All the elements within matrix **A** undergo a preparatory step known as the *zero-shifting* procedure [108], outlined in **Fig. 1.8 (a)**. This technique involves a series of pulses with alternating polarity, driving the device's conductance state towards a convergence point termed the *Symmetry Point* (G_{SP}). At this point, conductance updates in both upward and downward directions become nearly identical ($\Delta G_{up,SP} \approx \Delta G_{dn,SP}$). Consequently, matrix **A** operates within a domain of restored symmetry and heightened linearity, thereby augmenting the accuracy of the weight update process, also referred to as *gradient accumulation*.

The Tiki-Taka algorithm consists of two phases. In the first phase, a *forward* and a *backward* pass are executed through matrices **A** and **C**, respectively. Their outcomes are combined linearly using a scalar factor γ, expressed through the equations:

$$\boldsymbol{y} = \boldsymbol{x}(\gamma \boldsymbol{A} + \boldsymbol{C}) \tag{1.3}$$

$$\boldsymbol{z} = \boldsymbol{\delta}(\gamma \boldsymbol{A} + \boldsymbol{C})^T \tag{1.4}$$

where $\boldsymbol{y}$ and $\boldsymbol{z}$ represent the results obtained from the forward and backward passes, $\boldsymbol{x}$ denotes the input 1D array, and $\boldsymbol{\delta}$ signifies the backpropagated error from the subsequent neuronal layer.
Based on the backpropagated error $\boldsymbol{\delta}$, an optimal weight update applies solely to the elements of matrix A, following equation:

$$\boldsymbol{\Delta G}^A = L(\boldsymbol{\delta} \otimes \mathbf{x}) \tag{1.5}$$

where L denotes the learning rate. This outer product is carried out using the stochastic parallel weight updates procedure already described in previous illustrations (**Fig. 1.7 (a)**). Phase 1, comprising the three steps outlined by the equations **(1)**, **(2)**, and **(3)**, is iterated a specified number of times denoted by the term: R. R is another hyperparameter, requiring meticolous fine-tuning.
Please note: If matrix A is not initially prepared with its devices at the SP, the results from cumulative updates could significantly differ from the ideal outcome due to device non-idealities. This discrepancy could potentially lead to a total update in the wrong direction [108]. By initializing the elements of matrix A to the SP, we can filter out device non-idealities occurring during the backpropagation cycles, ultimately leading to cumulative updates with the correct sign [108].

Phase 2 focuses on updating matrix **C**. This process begins by generating a hot-encoded vector u_j, which represents the j^{th} column of a unitary matrix, using the equation:

$$\boldsymbol{v}_j = \mathbf{A}\boldsymbol{u}_j \tag{1.6}$$

Then, matrix **C** undergoes an update via an outer product operation between the vector $\boldsymbol{v}_j$ and $\boldsymbol{u}_j$:

$$\Delta \boldsymbol{G}^C = \lambda(\boldsymbol{v}_j \otimes \boldsymbol{u}_j) \tag{1.7}$$

This means that matrix **C** receives only pulses at the j^{th} column, each with a pulsing probability set to 1. Along the rows of matrix **C**, pulsing activity is determined by probabilities proportional to the values from the j^{th} column of matrix **A**, adjusted by a learning rate λ.

To summarize:

- The Tiki-Taka algorithm consists of two phases. In the initial phase, matrix **A** undergoes updates through stochastic pulses, with probabilities determined by the input vector $\boldsymbol{x}$ and the backpropagated error $\boldsymbol{\delta}$. These updates are repeated R times before matrix **C** is updated.
  In the second phase, matrix **C** is updated column by column, which increases the time complexity up to a factor O(N).
- The advantage of Tiki-Taka lies in the accumulation of gradients in matrix **A**. Its elements are prepared around a symmetry point, ensuring that conductance updates are symmetric and more linear. This enhances the accuracy of the learning task.

In the newer version of the Tiki-Taka algorithm, called *"TTv2"* [109], advancements have been made to ease the requirements on the analog resistive-switching properties of the NVM devices. In TTv2, the property *number of conductance states*, a property to be elaborated on in Chapter 4 of this dissertation, has been reduced from approximately a few thousand to around a few tens [109]. Additionally, TTv2 enhances robustness by a factor of 100× against device property *programming noise* (defined in Chapter 4) and by a factor of 10× against accumulated read noise during array forward and backward passes [109]. These enhancements are achieved through the incorporation of a digital matrix, denoted as **H**, in the TTv2 algorithm. **H** acts as a filter for the gradients accumulated in matrix **A**, transferring them to matrix **C** only when the updates surpass a certain threshold, executed by single coincident

pulse updates. For further insights into the TTv2 algorithm, refer to [109].
The computational load introduced by matrix **H** entails a straightforward 1D digital sum operation, with a time complexity of O(N). Consequently, the overall time complexity of TTv2 remains consistent with that of Tiki-Taka, approximately O(N) [107].

1.8 ReRAM devices and materials

The ReRAM device stands out as a promising candidate for accelerating deep learning tasks due to its unique material and electrical properties, as illustrated in **Fig. 1.5**. Its distinctive features are:

- **Analog bipolar conductance updates:** The ReRAM devices allow for analog conductance updates in both directions, upon the application of identical pulse streams [94, 95, 110]. Despite these updates being non-linear and non-symmetric, they result in only a minor drop in training error when utilized with algorithms like Tiki-Taka or its advanced version, TTv2 [111].
- **Larger on/off ratios:** Compared to ECRAM devices, the ReRAM structures offer larger on/off ratios, indicating a more significant difference between its conductive and non-conductive states [96, 112].
- **Shorter pulses and lower amplitude:** Compared to ECRAM devices, the ReRAM structures require pulses that are significantly shorter (at least 1000×) and lower in amplitude (approximately 2× lower) [96, 112].
- **Deep scalability:** ReRAM devices, being two-terminal structures, allow for a highly compact planar layout, achieving a density of $4F^2$ [113]. Furthermore, the unit cell area can be scaled down to as small as 100 nm^2 [98].

These properties make ReRAM devices uniquely advantageous for deep learning applications. However, realizing their full potential necessitates careful material optimization. In the following subsections, I will compare the materials and electrical properties of conventional MIM-ReRAM structures with those of optimized ReRAM devices.

1.8.1 Resistive-switching mechanisms in the filamentary MIM-ReRAM

The conventional ReRAM structure is based on the metal-insulator-metal (MIM) structure. Material selection follows specific guidelines: for the insulating layer, transition metal oxides (TMOs) are commonly chosen due to their provision of ionic species crucial for resistive-switching mechanisms [114]. Regarding electrodes, it's typical to employ two metals with contrasting oxygen affinity [114]. This setup designates one electrode as chemically inert and the other as reactive, offering utility in modeling processes.
Under a sufficiently strong electric field, oxygen anions (O^{2-}) within the TMO migrate towards the anode. This ionic drift leaves behind a conductive filament comprising oxygen vacancies ($V_O^{\bullet\bullet}$ in the Kröger-Vink notation [115]) [116], a phenomenon termed "filament *forming*". Subsequent resistive-switching mechanisms can then be induced by application of bipolar voltage pulses, modulating the density profile of $V_O^{\bullet\bullet}$ in the TMO material [116].
The choice of electrode materials and the TMO significantly influence critical electrical properties such as filament forming, resistive switching on/off ratio, and programming reliability. These materials determine the energy required for the formation of oxygen vacancies [117, 118].
Additionally, the work functions (WFs) of the electrodes and the voltage polarity during the filament

forming process can also determine the switching polarity. [88]. After establishing these foundational considerations about MIM-ReRAM materials, it's crucial to discuss the relevant electrical properties to deep learning applications. Filamentary ReRAM devices typically exhibit strongly asymmetric resistive-switching processes. The transition to higher resistances, known as "*reset*", generally occurs gradually with increasing applied voltages. This gradual transition is sustained by the growth of a filament "*gap*" in the TMO material [119], which is shown in **Figure 1.9 (a)**. Conversely, the process of transitioning to lower resistances, termed "*set*", is a self-accelerated process, leading to an abrupt drop in resistance. This acceleration occurs because the applied voltage initiates a drift of oxygen vacancies in the filament *gap*, as shown in **Figure 1.9 (b)**, causing an increase in local temperature and electric field [119]. This phenomenon facilitates the continued drift of additional vacancies without the need for higher applied voltages [119].

The asymmetry between the *set* and *reset* operations in MIM-ReRAM devices renders them unsuitable for accelerating deep learning with resistive-switching-based processors. However, this asymmetry can be improved by integrating a suitable series resistance. In the following discussion, we'll explore this aspect by examining the electrodynamics of resistive-switching mechanisms. In **Figure 1.9 (e) and**

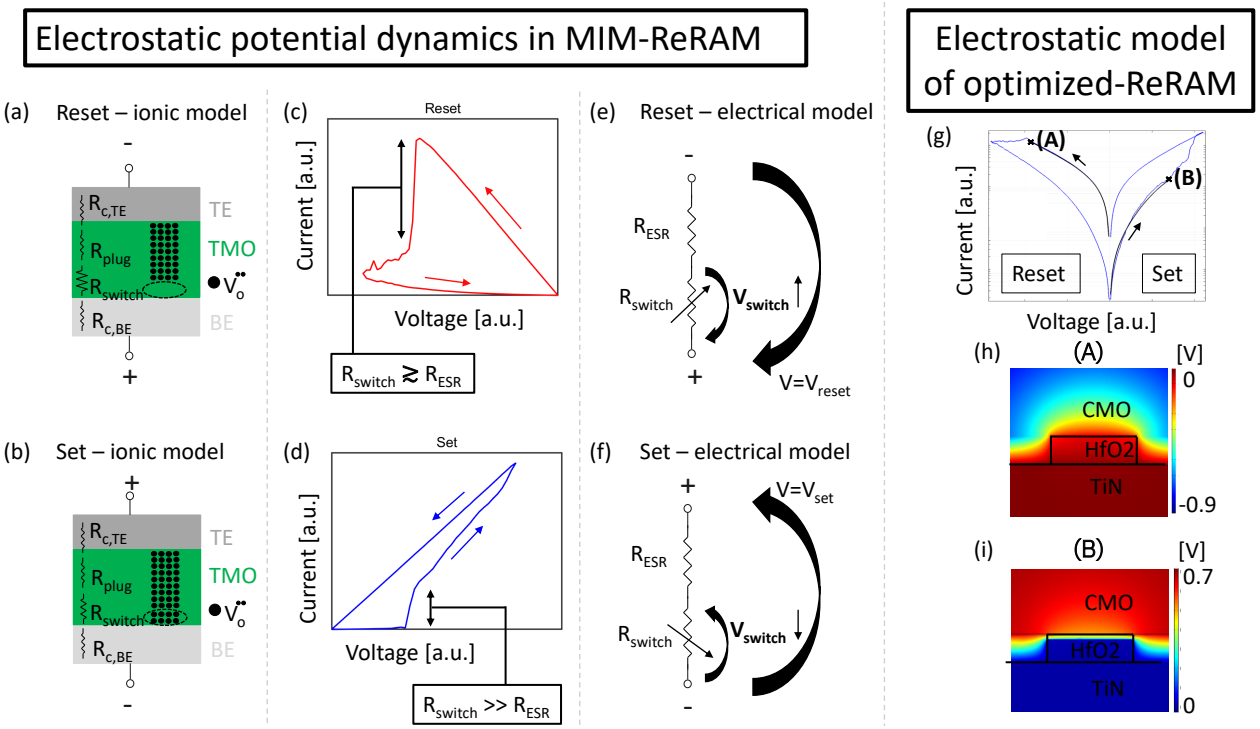


Figure 1.9: **(a)** and **(b)** Electro-ionic processes in the MIM-ReRAM device describing the *reset* and the *set* processes, respectively. **(c)** and **(d)** Current-voltage measurements during the *reset* and *set* processes, respectively. **(e)** and **(f)** Electrical equivalent model depicting the ReRAM device undergoing the *reset* and *set* processes, respectively. **(g)** Current-voltage characteristics of an optimized ReRAM device, based on the CMO/TMO *bilayer* structure. **(h)** and **(i)** Finite-element simulations demonstrating the electrostatics in the optimized ReRAM devices programmed in LRS and HRS, respectively. Modified from [94] and [120].

(f), we depict the ReRAM device using an equivalent electrical circuit consisting of two resistances in series, as done in prior literature [121]. One resistance represents the switching region of the device

(R_{switch}), while the other represents a passive series resistor, labeled as "Equivalent Series Resistor" (R_{ESR}). This resistor includes the resistance of the electrodes' contacts in series with the resistance of a portion of the filament, termed the filament [122].
The *reset* process, as mentioned earlier, is gradual and relies on increasing the applied voltage. However, when the values of R_{switch} and R_{ESR} become comparable, the *reset* process accelerates, as shown in **Figure 1.9 (c) and (e)**. This phenomenon can be explained by electrodynamics. The growth of the R_{switch} component leads to a dynamic redistribution of the electrostatic potential between R_{switch} itself and R_{ESR}. As the electrostatic potential on R_{switch} increases, a stronger electric field pushes oxygen vacancies outside towards the filament plug region, enlarging the filament *gap* in the in the TMO material. Consequently, a positive-feedback loop is initiated, continuing until the voltage drop on R_{ESR} becomes negligible compared to that on R_{switch}.
The opposite effect occurs during the *set* process. R_{switch} decreases abruptly, approaching the value of R_{ESR}. As R_{ESR} becomes significant compared to R_{switch}, a larger fraction of the external voltage is dynamically allocated to R_{ESR}. This electrical effect suppresses the self-acceleration of the *set* mechanism. To further decrease device resistance, an additional set voltage increment is necessary, as depicted in **Figure 1.9 (d)**.

1.8.2 Optimized filamentary ReRAM devices based on an oxide bilayer structure

About a decade ago, filamentary ReRAM devices with a more complex material stack than the MIM structure were already under exploration [123]. These devices incorporated a sub-stoichiometric oxide layer in the MIM-ReRAM stack, resulting in an oxide *bilayer* structure sandwiched between two electrodes. This integration led to an observed improvement in resistive-switching symmetry and even analog conductance updates when identical voltage pulses were applied [110].

However, there is no universally accepted interpretation of the physical reasons behind these improvements. One possible explanation for the enhanced symmetry is the role of the series resistor provided by the additional metal-oxide layer. In a study by Hardtdegen et al. [121], the authors compared the electrical properties of an HfO_2-MIM ReRAM with a TiO_x/HfO_2 *bilayer* ReRAM. They modeled the *bilayer* ReRAM deviced with a conductive filament extending from the HfO_2 to the TiO_x layer. The portion of the filament in the TiO_x layer was electrically modeled as a permanent series resistance, enhancing the R_{ESR} term. According to their model, the total value of R_{ESR} was significant compared to $R_{device,LRS}$, justifying a more gradual *set* and a more abrupt *reset* [121], consistently with the discussions presented earlier. Furthermore, the work by Cüppers et al. [124] demonstrated that, the same TiO_x/HfO_2 *bilayer* ReRAM, when operated at intermediate resistive states, exhibits gradual processes and analog switching under identical pulsing conditions for both directions.
Woo et al. [125] attributed the analog switching properties of a *bilayer* ReRAM to the differing mobility of the $V_O^{\bullet\bullet}$ species in the two metal-oxides.This property enables electrical control of the width of the conductive filament, allowing for a more gradual modulation of device resistance compared to conventional MIM structures.

The presence of an additional metal-oxide in the device stack, transforming the ReRAM into an oxide *bilayer*, has demonstrated improvements in various electrical properties, as discussed. However, the main drawback introduced by the oxide *bilayer* concept is the increase in *forming* voltages. Stacking two insulating materials necessitates higher voltages to induce a breakdown field.
To address this challenge while preserving the benefits of the *bilayer* concept, a highly-reduced metal-oxide material, can be utilized. The heavy reduction results in a high defect concentration, rendering the layer electrically conductive, hence its designation as Conductive-Metal-Oxide (CMO).

During the filament *forming* process, the applied voltage primarily drops across the transition metal oxide (TMO) layer, since the Conductive-Metal-Oxide (CMO) layer serves as a negligible series resistor. However, in normal set and reset operations, if the resistivity of the CMO material is designed to be orders of magnitude higher than that of a metallic electrode, it significantly contributes to the R_{ESR}, resulting value, inducing a more gradual *set* mechanism, as modelled in Zhao et al. [126].
This interpretation finds support in our Finite Element Method (FEM) simulation results as well [120, 127], illustrated in **Figure 1.9 (h) and (i)**. When the optimized ReRAM device is programmed into a Low-Resistive-State (LRS), the electrostatic potential (V) is predominantly distributed across the bulk CMO layer, which serves as a significant R_{ESR}. Conversely, in the High-Resistive-State (HRS), V is confined to a narrow region at the interface between the filament and the CMO layer, which acts as the dominant resistance R_{switch}.
Another interpretation of the analog switching property observed in CMO/TMO *bilayer* ReRAM devices is the heat confinement effect of the CMO layer. This effect could lead to the formation of multiple "weak filaments" in the TMO layer, with each contributing partially to the overall resistive-switching mechanism. Consequently, this results in gradual, analog processes [128].

1.9 Research status on the In-Memory-Computing (IMC) deep learning accelerators

Figure 1.10 (a) provides a comparison of the R&D progress between resistive-switching-based accelerators for inference and training applications. This comparison is made based on the number of cores and array size (specification values extracted from [96, 112, 129–138]). The figure indicates a significant gap between the research advancement of inference and training accelerators, primarily attributed to the non-ideal programming properties of resistive-switching devices, such as asymmetry, non-linearity, and the requirement for long/high-voltage programming pulses. **Figure 1.10 (b)** depicts the temporal evolution of the research on deep learning accelerators, based on resistive-switching devices and the parallel updates concept. The most common device technologies adopted for such application are ReRAM devices and ECRAM devices.

My PhD dissertation work is centered around the broad goal of bridging the gap between research on inference and training accelerators. This endeavor is facilitated by leveraging recent algorithmic advancements outlined in Sections 1.6 and 1.7. To achieve this objective, I have focused on engineering the material properties of optimized ReRAM devices, specifically based on the CMO/TMO *bilayer* structure. The aim is to enhance their analog response to programming pulses.

- **Chapter 2** provides an introduction to the characterization methods employed to investigate the material and electrical properties of the devices.
- **Chapter 3** delves into the material stack engineering efforts undertaken to optimize the electrical properties of CMO/TMO ReRAM.
- In **Chapter 4**, I present the processing endeavors aimed at scaling the ReRAM unit cell layout and its electrical programming parameters. Additionally, advanced electrical characterization techniques and data analysis were employed to evaluate the performance of the devices under testing conditions simulating the Tiki-Taka algorithm implementation.
- **Chapter 5** details the chip layout and processing endeavors culminating in the integration of the optimized ReRAM devices onto 350-nm CMOS technology. This led to the integration of One-Transistor-One-Resistor (1T1R) devices in RPU architectures of sizes up to 10×10. Furthermore,

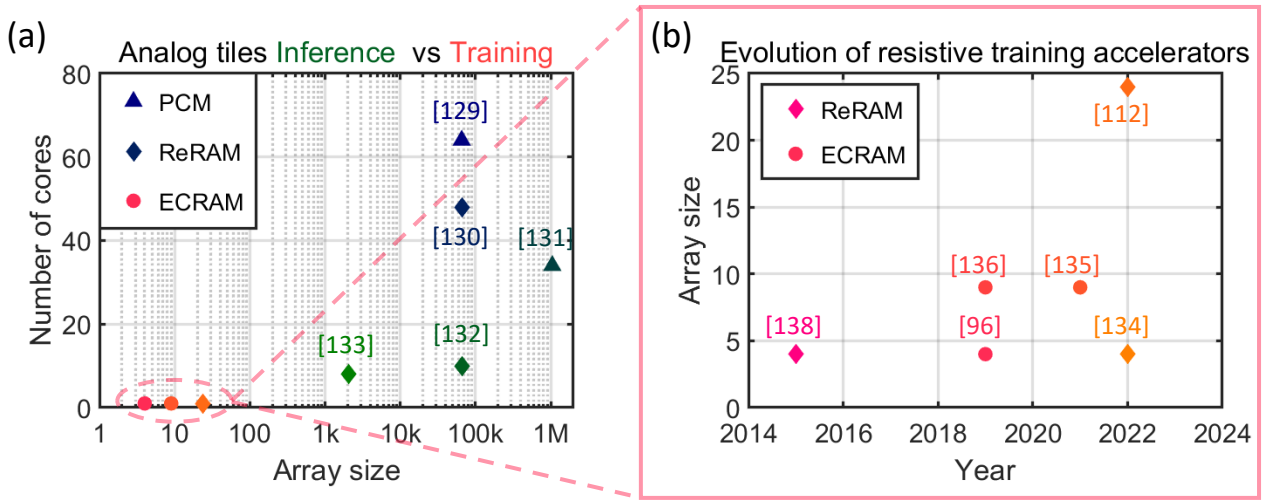


Figure 1.10: **(a)** State-Of-The-Art (SOTA) research disparity between inference and training accelerators based on resistive-switching devices. **(b)** Temporal evolution of research on deep learning accelerators exploiting the parallel weight updates concept with resistive-switching devices.

I discuss the 1T1R response to stochastic programming pulses, including an example of parallel programming for a sub-matrix of size 5×4.

- Finally, **Chapter 6** serves as the conclusion and benchmark of this dissertation work, summarizing the key findings and contributions.

Chapter 2

Semiconductor processing and characterization methods

2.1 Thin films deposition and etching methods

The experimental work of my PhD thesis began at the "Binnig and Rohrer Nanotechnology Center" at IBM Research Europe - Zürich. This facility boasts a state-of-the-art cleanroom equipped with numerous thin film deposition and etching tools. In this cleanroom, I fabricated ReRAM devices using various deposition methods including Direct Current (DC) and Radio-Frequency (RF) sputtering, Atomic-Layer-Deposition (ALD), evaporation, and etching/patterning techniques, such as lift-off, dry-etch, with Reactive-Ion-Etching (RIE) and Inductive-Coupled-Plasma (ICP) tools, as well as wet etch. Detailed descriptions of these tools' operating principles can be found in academic handbooks [139].
The first section of this chapter focuses on the engineered deposition processes developed to impart enhanced electrical and structural properties to thin films of functional TMO materials, which led to the optimization of the ReRAM device properties. Specifically, I refined deposition processes for the HfO_x and TaO_x materials. They are both TMO materials which have shown excellent electrical properties in optimized ReRAM devices [128].
Firstly, I fabricated substoichiometric TaO_x layers through DC reactive sputtering, adapting advanced recipes from previous literature [140], and fine-tuning them for our cleanroom tools. Additionally, I deposited low-density, ultrathin HfO_x films by Plasma-Enhanced ALD (PEALD) to reduce the filament *forming* voltage of ReRAM devices.

2.1.1 DC reactive sputtering

Fig. 2.1 (a) outline how a DC sputter deposition setup operates. In this sputter system, an inert gas, like argon (Ar), is introduced into a chamber at low pressure. When a DC voltage is applied across two electrodes, it generates a plasma. This plasma comprises neutral argon atoms, along with an equal mix of positive ions and free electrons, forming a conductive medium.
The top electrode, subjected to a negative DC voltage, serves as the source material for deposition. This electrode, known as the *target*, faces the bottom electrode, a grounded metal plate acting as the anode. Wafers to be coated are positioned on the anode.
Positive ions within the plasma are drawn towards the negatively charged target, which can have a voltage ranging from several hundred to several thousand volts relative to the plasma [139]. Upon striking the target, these ions dislodge or sputter atoms from its surface. These sputtered atoms then move through the plasma as a vapor and deposit onto the wafer surface, where they condense and

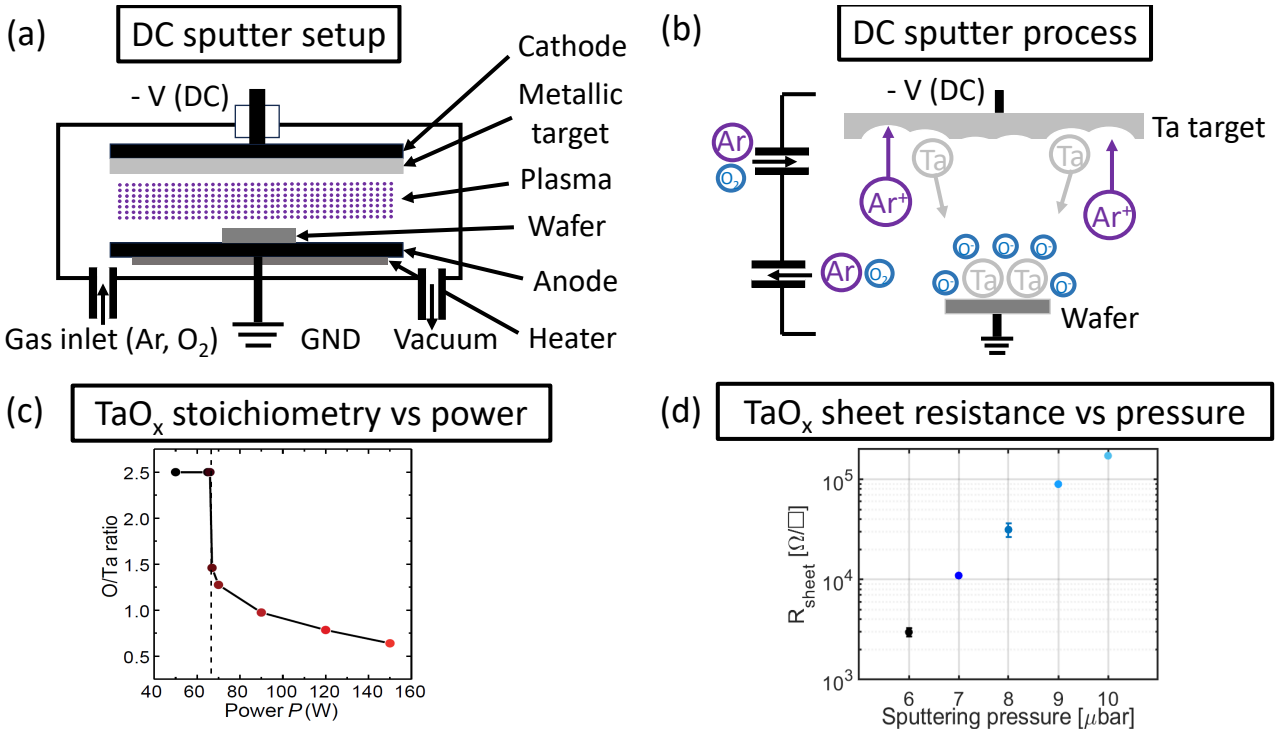


Figure 2.1: **(a)** DC sputtering setup, adapted from [139]. **(b)** Schematic depicting the DC reactive sputtering process, where a Ta target is bombarded with Ar ions, leading to the generation of oxygen radicals that react with sputtered Ta particles, resulting in the deposition of TaO_x species. **(c)** Graph demonstrating the tunable stoichiometry in TaO_x thin films deposited by RF magnetron sputtering, extracted from the PhD dissertation thesis of T. Heisig [140]. **(d)** Graph depicting the relationship between sheet resistance and sputtering pressure of our TaO_x thin films. Modified from [139] and [94].

form the desired film. This mechanism is the basis of DC sputtering.

Reactive sputter deposition involves introducing a reactive gas into the sputtering chamber alongside the Ar plasma. This gas combines with the sputtered material to form compounds, as illustrated in **Fig. 2.1 (b)**. For instance, Ta_2O_5 can be produced by sputtering Ta in the presence of oxygen. The plasma energizes the oxygen molecules, facilitating its dissociation into oxygen radicals, which then reacts with the sputtered Ta atoms. This reaction typically occurs either on the wafer surface or the target itself, not in the plasma [139].

Controlling the stoichiometry of reactively sputtered films is generally challenging [139]. However, in the PhD thesis dissertation from T. Heisig [140], the author discusses numerous recipes for the deposition of TaO_x thin films, with gradually changing stoichiometries from pure Ta, to stoichiometric Ta_2O_5. In this work, the author utilized the magnetron RF sputtering, a technique enhancing the plasma ionization efficiency [139]. By gradually tuning the RF sputtering power, from 50 to 150 W, the deposited TaO_x material exhibits linearly-controllable stoichiometry splits, which reflect into a log-scale change of the thin film's resistivity (see **Fig. 2.1 (c)**) [140].

In our work, we exploited the sputtering chamber pressure as the control parameter to adjust the material properties of our TaO_x layer, while keeping a fixed DC power of 50 W and a ratio of the O_2/Ar stream equal to 1/40 sccm. By linearly tuning the chamber pressure from 6 to 10 μbar, the sheet resistance, hence the resistivity, of the material grows exponentially [94], as illustrated in **Fig. 2.1 (d)**.

2.1.2 Plasma-Enhanced Atomic-Layer-Deposition (PEALD)

ALD is a chemical deposition process that employs reactive substances to deposit atomically thin layers through two distinct half-reactions. Initially, a reactant forms a self-limiting, monoatomic layer on the surface. Subsequently, introducing the second reactant converts this monolayer into the desired solid film, such as HfO_2. This layer-by-layer buildup ensures precise control over film thickness [141, 142].
Fig. 2.2 illustrates the PEALD deposition process using HfO_2 as an example. In this instance, tetra(ethylmethylamido)hafnium (TEMAH) and O_2 plasma serve as the two reactants [142]. TEMAH, with the chemical formula $[(CH_3)_2(H2C)N]_4Hf$, is an amide-based precursor, with the amide component denoted as "*ethylmethylamido*", often represented as *NEtMe* [143].
The reaction begins with the introduction of TEMAH precursor into the PEALD chamber, where it interacts with hydroxyl (OH) groups on the wafer surface, as depicted in **Fig. 2.2 (a)**. This interaction forms an Hf-O bond, releasing methane (CH_4) as a byproduct [141]. This sequence is self-terminating, ensuring uniform ALD deposition. Subsequently, the byproducts of this half-reaction are removed by activating the chamber pump, as illustrated in **Fig. 2.2 (b)**.
Next, the second reactant, O_2, is introduced into the chamber, and a plasma is generated by activating the chamber RF generator, as shown in **Fig. 2.2 (c)**. This plasma produces oxygen radicals that react with the amide dangling bonds and Hf, completing the second half-reaction and forming the Hf-O bonds. After deactivating the plasma, the byproducts of this reaction are purged through a final evacuation step, as depicted in **Fig. 2.2 (d)**.

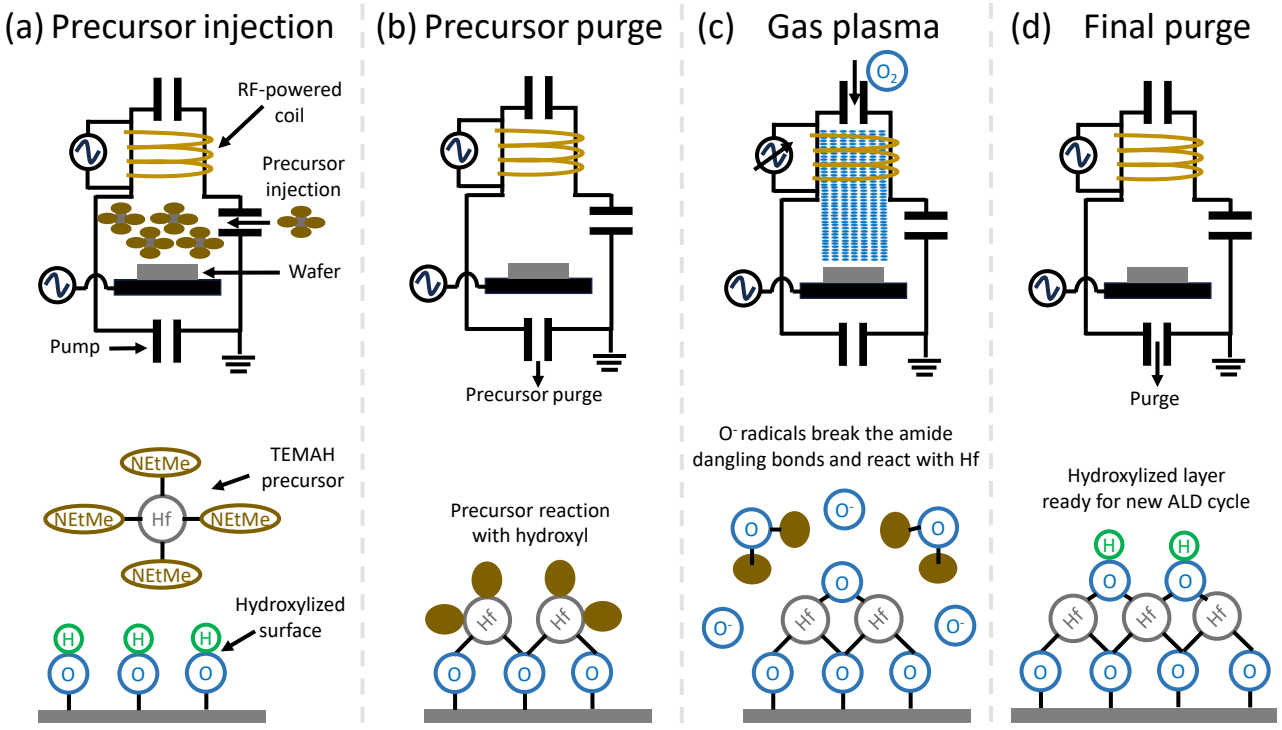


Figure 2.2: Illustration of the hafnia (HfO_2) PEALD deposition cycle (adapted from [141]): **(a)** Precursor purge, **(b)** Precursor reaction, **(c)** Injection of reacting gas and plasma reaction, and **(d)** Final purge marking the end of the cycle. Modified from [141]

In previous studies, it has been demonstrated that deposition parameters, including chamber temperature [144], Hf purge time [144, 145], and the plasma duration [146], significantly influence the quality of

the deposited layer. These parameters can introduce impurities that impact the chemical, mechanical, and structural properties of the thin film, such as stoichiometry, stress, roughness, and density [144,145]. Specifically, insufficient plasma duration may hinder oxygen from reacting with intermediate functional groups, thereby failing to effectively eliminate impurities associated with carbon and nitrogen [146]. This phenomenon may result in faster deposition rates but yields lower density thin films [146].

2.1.3 Dry etching: Reactive-Ion-Etching (RIE) and Inductive-Coupled-Plasma (ICP)

A simple RIE plasma etching setup is depicted in **Fig. 2.3 (a)**. Such etching setup utilizes a low-pressure gas (ranging from 1 millitorr to 1 torr) within the chamber and an RF generator to ignite a plasma. The parallel plate etching system can be adjusted to boost the physical aspect of etching by elevating the ion energy bombarding the wafers. This enhancement is achieved by making the electrode supporting the wafers smaller than the counterpart. Additionally, the upper electrode can be grounded and connected electrically to the chamber walls, effectively enlarging it. This configuration results in a significant voltage drop from the plasma to the wafers, ranging from 100 to 700 volts [139], intensifying ion bombardment and facilitating more directional etching.

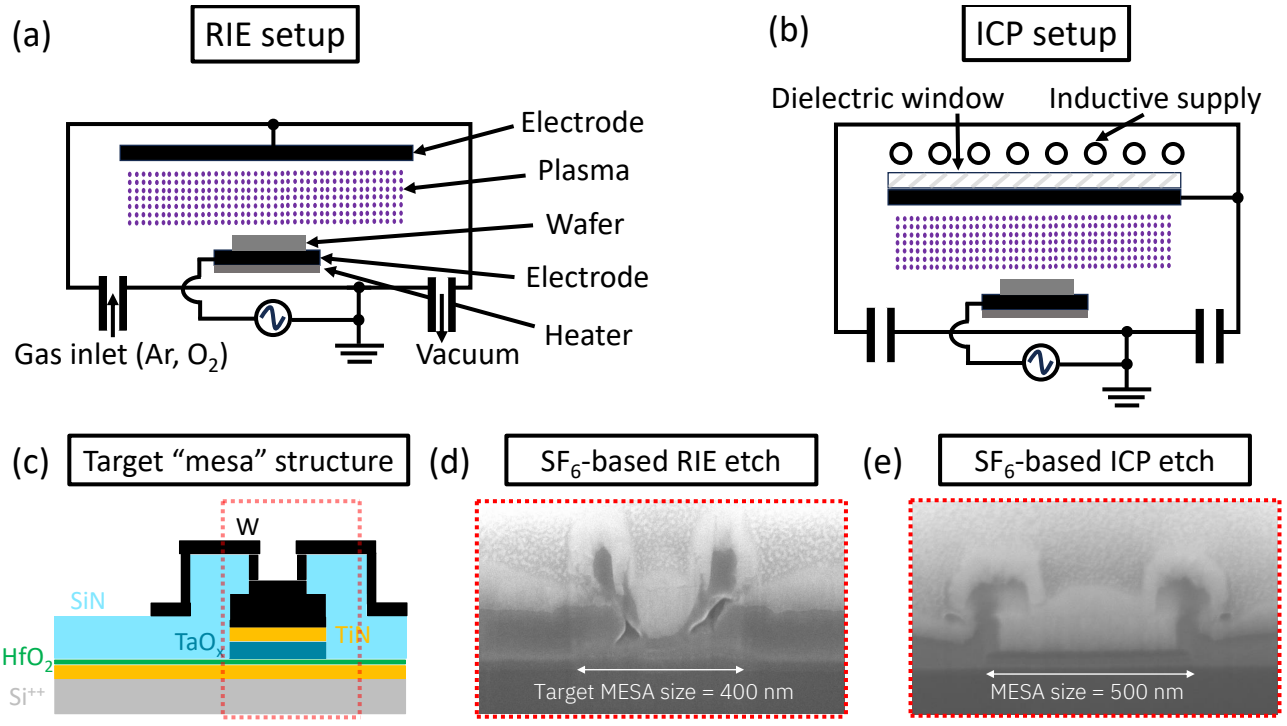


Figure 2.3: **(a)** RIE etching setup, adapted from [139]. **(b)** ICP etching setup, adapted from [139]. **(c)** Cross-sectional depiction of the targeted *mesa*-shaped ReRAM device structure, and **(d)** and **(e)** Scanning-Electron-Microscopy (SEM) cross-section of a nominally 400 nm- (500 nm-) sized mesa, patterned by RIE (ICP). Modified from [139].

Enhanced directional etching can be achieved by reducing the gas pressure, resulting in fewer ion collisions and a more focused ion flux towards the wafer [139]. An alternative plasma etching method is illustrated in **Fig. 2.3 (b)** where control over plasma density and ion energy is separated. By employing

Inductive-Coupled-Plasma (ICP) power, a high-density plasma is generated, allowing operation at lower pressures and thereby improving etching directionality [139].
In our process flows for *mesa*-shaped ReRAM devices, we transitioned from RIE to ICP etching for critical steps requiring high directionality. ICP etching enabled precise patterning of devices with smaller top electrode areas, down to approximately 500 nm×500 nm with <10% undercut, as shown in **Fig. 2.3 (e)**. A comparison is drawn between two SF_6-based etching steps: one utilizing the RIE tool (**Fig. 2.3 (d)**) and the other employing the ICP tool (**Fig. 2.3 (e)**), highlighting the significant improvements achieved through ICP. The materials sequentially etched in this step are W,

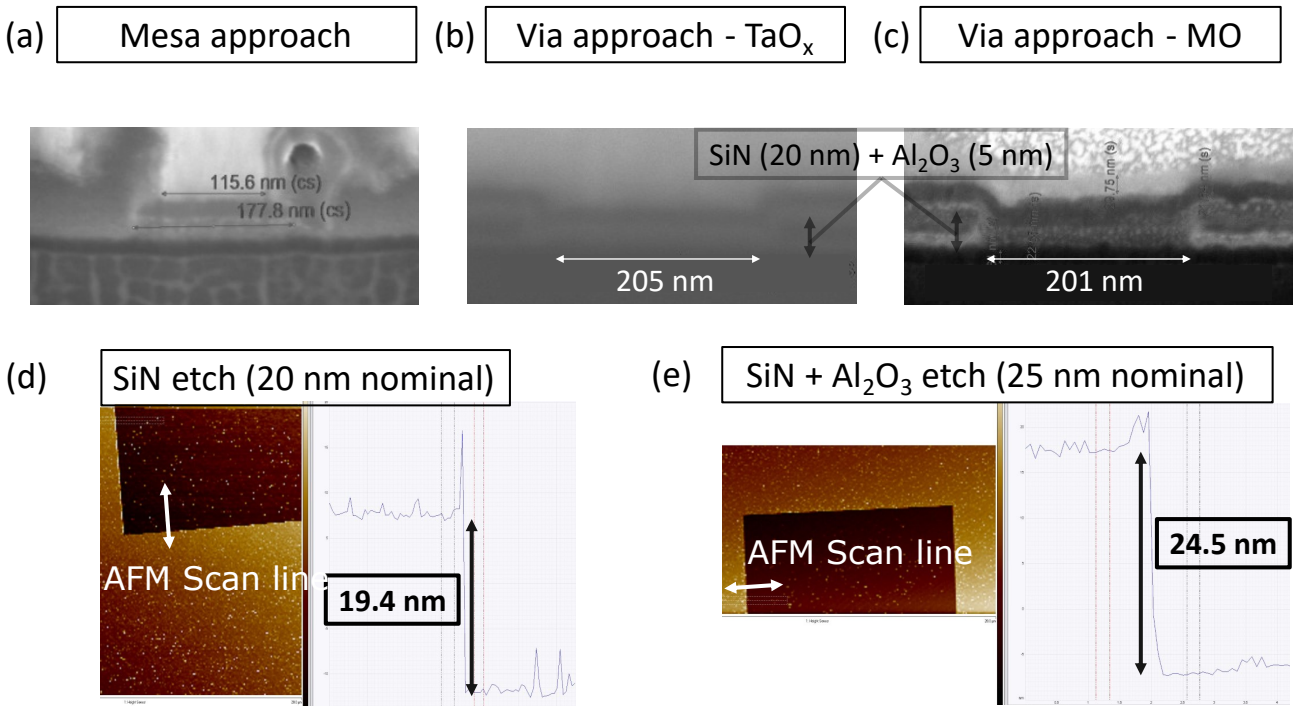


Figure 2.4: **(a)** SEM cross-section of a nominally 200 nm-sized mesa etched by ICP using an SF_6-based gas mixture. Different etching selectivities define a trapeziodal mesa shape, leading to the formation of voids in the subsequent cladding, thereby harming reproducibility. **(b)** and **(c)** SEM cross-sections of 200 nm-sized TaO_x- and MO-based ReRAM devices fabricated using the "via approach". A via is etched through 20 nm of SiN and 5 nm of Al_2O_3, first by RIE, then by wet etch, achieving perfect accuracy between the nominal and actual via sizes.. **(d)** and **(e)** Atomic-Force-Microscopy (AFM)-scans after the SiN dry etch and the Al_2O_3 wet etch, respectively. Modified from [111].

TiN and a conductive-metal-oxide (either TaO_x or a material referred to in this dissertation as "MO"). Unfortunately, these layers exhibit different etching selectivities to the SF_6-based gas mixture, leading to the formation of "*undercuts*" in the *mesa* structure, as displayed in **Figure 2.4 (a)**. Consequently, the scaling the unit cell area below 200 nm × 200 nm proved challenging, despite the highly-directional ICP etching.
To further scale the geometry of our ReRAM devices, we adopted a different approach where the unit cell area is defined by etching a "*via*" in a thin bilayer of 20 nm thick SiN and 5 nm thick Al_2O_3. A ReRAM device structure fabricated using this method, referred to as "via approach", is illustrated in **Figure 2.4 (b) and (c)**. The via in the SiN is created through a RIE process using the CHF_3 reactive gas. Subsequently, the Al_2O_3 layer is removed by wet etch using the "*AZ 726 MIF*" developer. To verify that the removal of the Al_2O_3 layer does not harm the active layer underneath (HfO_x), we deposited

the same HfO_x material onto a substrate covered by platinum (Pt). We immersed this sample in the *AZ 736 MIF* developer overnight, and then we probed electrically its surface, measuring no current. This measurement indicates that the HfO_x layer is not etched away by the *MIF* developer. Furthermore, we performed Atomic Force Microscopy (AFM) measurements to investigate this two-step etching process with nanometric precision. **Figure 2.4 (d) and (f)** depict the results from AFM scans on the sample surface after SiN and SiN+Al_2O_3 etching, respectively.

2.2 Materials characterization

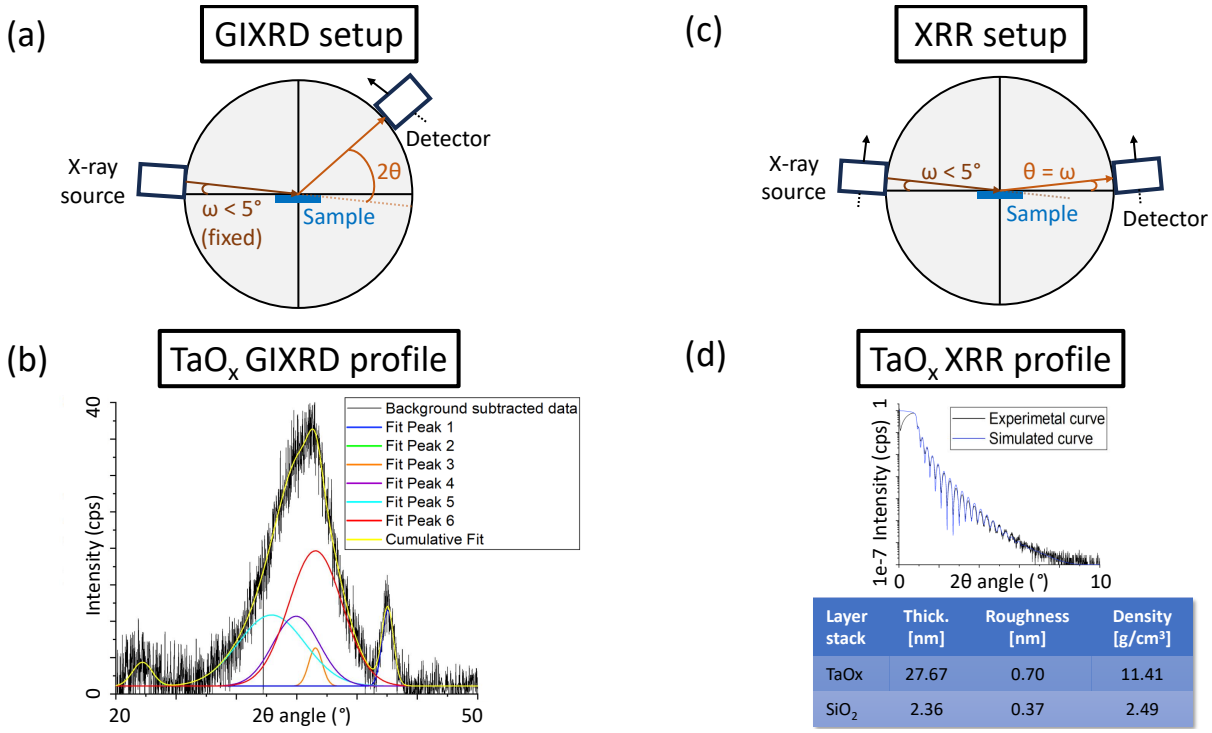
Materials characterization involves a broad range of spectroscopic, electrical and metrological techniques. techniques. These methods are essential for comprehending and assessing the material properties critical to enhancing device performance.
In this section, I will introduce the operational principles of X-Ray spectroscopy methods, such as Grazing-Incidence X-Ray Diffraction (GIXRD) and X-Ray Reflectivity (XRR), providing examples of measurements. These experiments were conducted using a Bruker D8 Discover diffractometer, equipped with a rotating anode generator. The analysis of the GIXRD and XRR scans was performed within the supervision of the master student Roberto Guido.
In this section, I will also introduce the Circular-Transmission-Line-Measurements (C-TLM), which were performed to characterize the sheet resistance of the deposited thin films and estimate their resistivity.

2.2.1 Grazing-Incidence X-Ray Diffraction (GIXRD) and X-Ray Reflectivity (XRR)

Diffraction ultimately limits the resolution capability of an imaging system. In crystalline materials, atoms are arranged in a periodic 3D structure, forming a reticle of diffraction planes for incident waves. The condition of constructive interference between waves is described by Bragg's law of diffraction:

$$n\lambda = 2dsin(\theta) \tag{2.1}$$

where λ is the wavelength of the incident wave, θ is its angle of incidence, d is the distance between two diffraction planes, and n is the diffraction order.
According to this rule, the wavelength of a probing beam must be comparable to the distance between diffraction planes. Therefore, spectroscopy techniques generally use x-rays, which enable resolutions below the nanometric scale.
In Grazing-Incidence X-Ray Diffraction (GIXRD), an incident x-ray beam strikes the sample at a fixed angle, ω, generally less than 1 degree, as illustrated in **Fig. 2.5 (a)**. The low angle of incidence avoids to collect interferences from the substrate, when very thin films ($<$ 100 nm) are scanned. An interference detector moves and collect the reflected beam, hence scanning the value of the scattering angle between the incoming and reflected beam (2θ measurement). With such notation, the angle between the surface and the reflected beam is equal to 2θ - ω, as illustrated in **Fig. 2.5 (a)**.
The result of the GIXRD scan is the intensity profile of the detected beam, as a function of the angle 2θ. Due to Bragg reflections, intensity peaks appear in the GIXRD pattern at specific angles. These peaks are compared and indexed using open source XRD databases, such as "*profile.materialsproject.org*" [147]. When measuring a polycrystalline thin film, such as our conductive-TaO_x, the resulting peak is a superposition of multiple components at different angles, leading to a broadened peak. The independent contributions of the constituent peaks can be reconstructed using Gaussian deconvolution. An example of peak deconvolution is depicted in **Fig. 2.5 (b)**.



Layer stack	Thick. [nm]	Roughness [nm]	Density [g/cm³]
TaOx	27.67	0.70	11.41
SiO_2	2.36	0.37	2.49

Figure 2.5: **(a)** Schematic of the GIXRD setup. **(b)** GIXRD profile of a TaO_x thin film deposited on amorphous-HfO_2/crystalline-TiN. The polycrystalline composition of the TaO_x is revealed through Gaussian deconvolution of its components and peak indexing, using the "*MaterialsProject*" database [147]. **(c)** Schematic of the XRR setup. **(d)** XRR profile of a TaO_x thin film deposited on SiO_2. The trace is used to fit a model of the TaO_x layer, whose properties are reported in the table below. Modified from [94].

X-Ray-Reflectivity (XRR) is a spectroscopic technique with a similar setup to GIXRD. In XRR, both the x-ray source and detector move during the scan, maintaining an identical angle $\omega = \theta$ for the incident beam and the collected reflections, as illustrated in **Fig. 2.5 (c)**. When the incident x-rays strike at an angle lower than the critical angle, known as the *total reflection angle*, θ_C, they are fully reflected, maximizing the detected signal. For higher incident angles, the reflected beam from the film's surface interferes with the beam reflected from the film-substrate interface, creating oscillations in the XRR profile, as shown in **Fig. 2.5 (d)**, known as "*fringes.*"
The shape of these fringes reveals numerous material properties. A higher frequency of reflection fringes indicates thicker films [148]. A larger total reflection angle suggests a higher film density [148]. A rapid decay in the XRR intensity indicates high surface roughness [148], and a small oscillation amplitude suggests significant interface roughness [148]. Quantitative measurements of these parameters can be obtained by fitting the experimental XRR trace with a theoretical model of the sample's layers, as displayed in **Fig. 2.5 (d)**.

2.2.2 Circular Transmission Line Measurements (CTLM)

CTLM are an electrical characterization technique used to determine the sheet resistance of semiconducting thin films [149]. Additionally, TLM provides information on the contact resistance [149]. If the

material is electrically isotropic, the material resistivity (ρ) can be inferred using the following formula:

$$\rho = R_{sheet}t \tag{2.2}$$

where R_{sheet} is the sheet resistance of the thin film and t its thickness.

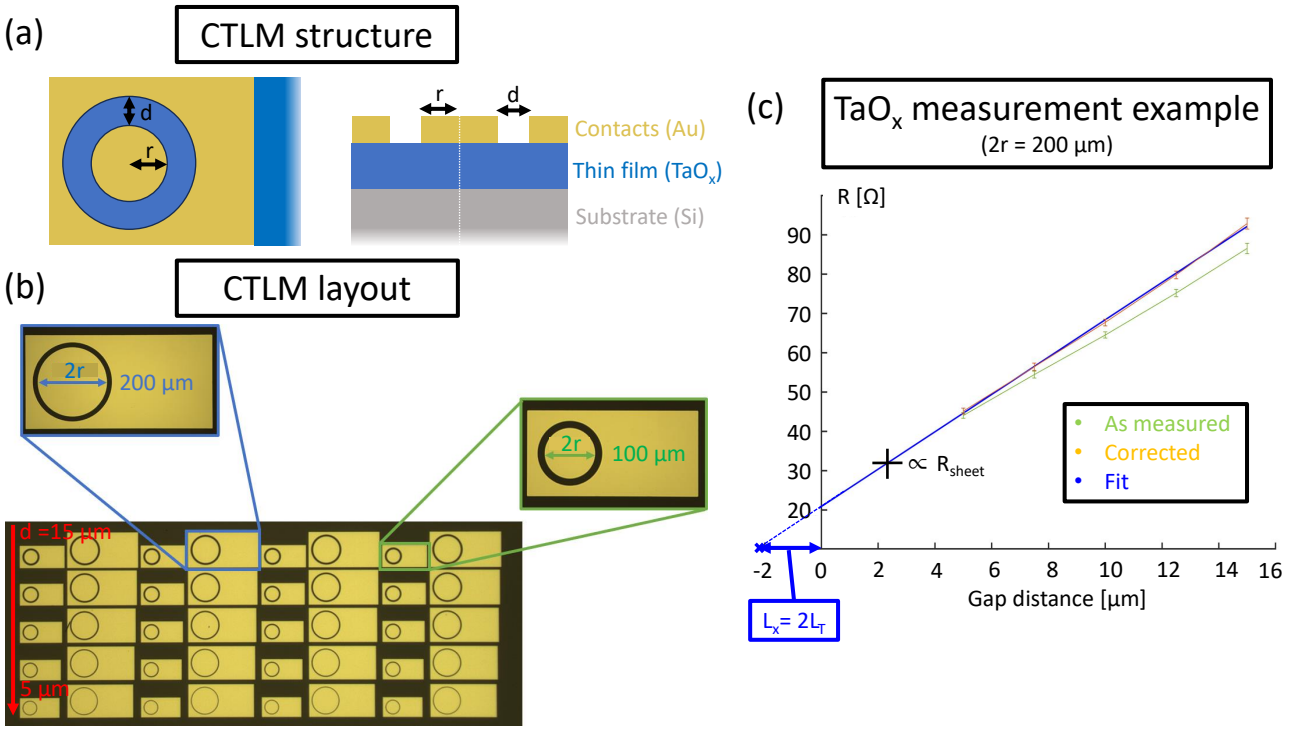


Figure 2.6: **(a)** CTLM structure and geometrical parameters (r and d), depicted in a top view and cross-section. **(b)** Layout of our CTLM structures. **(c)** Example of resistance vs. distance (R/d) transfer characteristics (in green). This example is from a 30-nm thick TaO_x thin film, measured using a CTLM structure of diameter equal to 200 μm. The results are corrected through the equations 2.3 and 2.4, yielding the orange data, which are fitted by the blue line to extract the values of contact resistance and sheet resistance of the TaO_x layer. Modified from [94].

In linear TLM structures, the electric field lines concentrate at the contact corners, leading to current crowding and inaccurate measurements [150]. To address this issue, Circular-TLM (CTLM) structures are preferred. These structures are characterized by a ring configuration, as illustrated in **Fig. 2.6 (a)**, with the inner circle radius and the spacing between contacts denoted as r and d, respectively. The CTLM method involves patterning multiple structures with varying parameters r and d. In our samples, we made rings with diameters $2r$ = 100 and 200 μm, and distances d linearly varying from 15 to 5 μm, in steps of 2.5 μm, as shown in **Fig. 2.6 (b)**.
The electrical measurement is carried out by applying a DC voltage of 200 mV using a four-probe setup to achieve more accurate results [149]. Due to the geometry, this measurement provides a non-linear output, as illustrated in **Fig. 2.6 (c)**. However, when the condition $r \gg d$ is met, a linear approximation can be used. The approximate value of the measured resistance is given by the equation:

$$R = \frac{R_{sheet}}{2\pi r}[d + 2L_T]c \tag{2.3}$$

where L_T is the "*tranfer length*", and c is the correction factor for linearization, calculated as:

$$c = \frac{r}{d} ln \frac{[r+d]}{r} \tag{2.4}$$

As depicted in **Fig. 2.6 (c)**, the tranfer length L_T can be determined by extending the corrected CTLM result until the line meets the d-axis at a value denoted as $L_x = 2L_T$. The tranfer length is a parameter describing the quality of the Ohmic contacts: the shorter the L_T, the more negligible the CTLM contact resistance [149].

2.3 Device characterization

The electrical characterization of ReRAM devices involves various types of measurements, which require flexible setups. Quasi-static measurements, often referred to as "DC", provide fundamental electrical properties such as the programming voltages and currents. However, real-world applications necessitate fast memory access in the kHz-MHz frequency range. To meet such bandwidth requirements, we built and customized an electrical setup capable of programming our devices with pulses as short as 100 ns. For the ReRAM crossbar arrays, advanced programming capabilities are required, as described in 1.6. One essential functionality for deep learning applications is the synchronization of analog programming pulses across multiple channels. To address this need, we developed a dedicated setup. My primary contribution was the development of a user-friendly Graphical User Interface (GUI) software that controls the entire system.

2.3.1 DC and pulsed measurements

- **DC measurement parameters:** The quasi-static approximation assumes that the Device Under Test (DUT) remains in an equilibrium state throughout most of the measurement time. In ReRAM devices, the timescales for electron-ionic processes that lead to resistance changes can vary exponentially, ranging from tens of *picoseconds* [92, 151], to *milliseconds* [152], depending on the applied voltages [152]. Therefore, in our DC measurements, we ensured consistent measurement conditions for at least 100 ms. For instance, in measurements involving a voltage ramp and current measurement, we applied voltage steps with an amplitude of 100 mV every 100 ms, allowing sufficient time for the current transients to settle.

  Another important dynamic process worth discussing is the filament forming. This electro-ionic phenomenon can occur on very short timescales, sometimes down to a few tens of picoseconds [151]. During filament forming, when a constant external voltage is applied, there is an abrupt increase in the ReRAM current, which can lead to a non-reversible resistance drop.
  To control the dynamics of the forming mechanism, and achieve reversible resistive-switching operations, the ReRAM device is typically integrating on CMOS technology to form "1T1R" circuits. In these circuits, the maximum current flowing through the 1T1R device is determined by the gate-source voltage of the transistor, allowing external control by the transistor's trans-characteristic (I_{ds}/V_{gs}) [153]. According to Kirchhoff's law, the current flowing through the transistor primarily originates from the ReRAM, with a minor contribution from parasitic capacitances [151, 153]. This causes a brief current overshoot for short durations, generally less than 1 ns [151]. After this period, the equation $I_{ReRAM,max} = I_{ds,sat}$ holds, meaning that the transistor limits the ReRAM current [153].

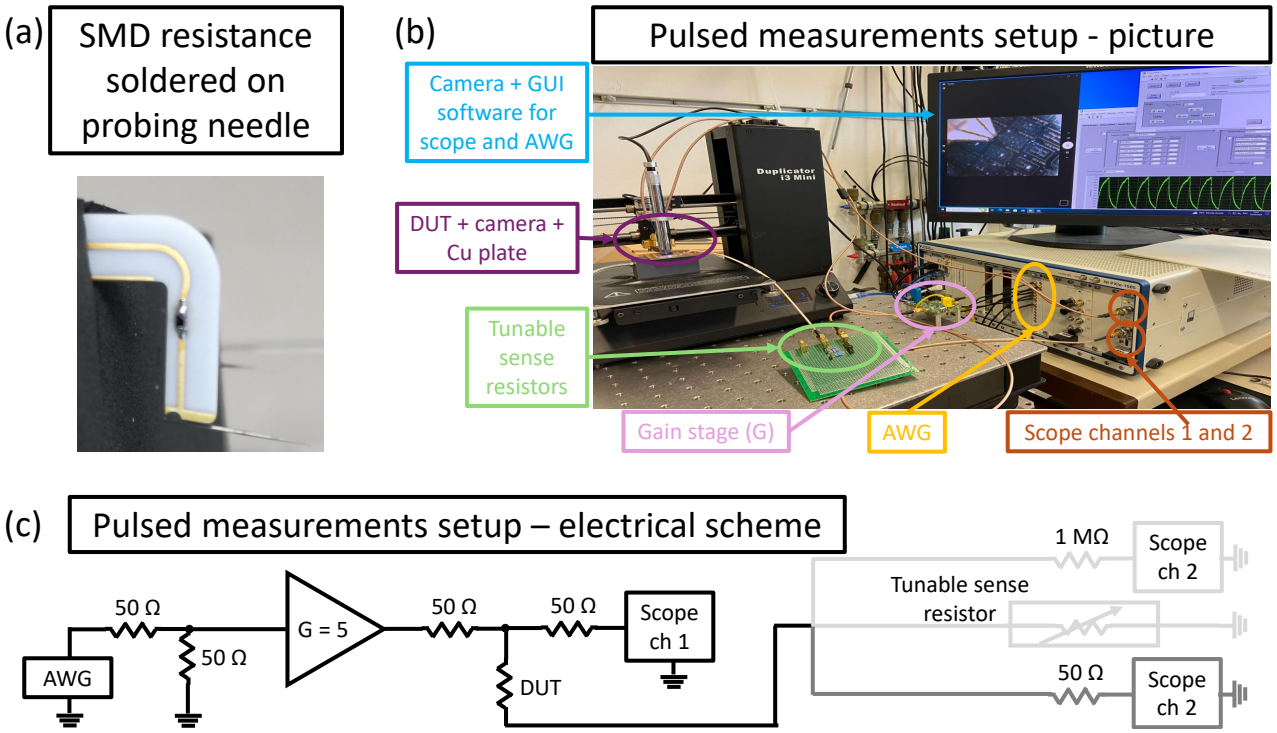


Figure 2.7: **(a)** Picture of an SMD resistor soldered onto an electrical probing tip.**(b)** Picture of our setup for pulsed measurements, which includes a 3D printer stage, allowing for precise movements in the μm range, ensuring high-precision contact with the probing needles. A camera focuses on the probing tips and the DUT, positioned on a copper plate. Two printed boards are inserted: one for amplyfing the voltage signals from the Arbitrary Waveform Generator (AWG), the other one for amplifying the electrical signal from the DUT by selecting an appropriate sense resistance value. **(c)** Circuit equivalent schematic of the pulsed measurement setup.

Without an integrated transistor, there are various, generally less effective [151], methods to limit the current in single ReRAM devices, referred to as "1R" devices. For example, when using a measurement tool called *parameter analyzer*, the system itself provides a current compliance functionality, through a Source-Measure-Unit (SMU) module. However, such modules typically operate with very low maximal slew rates, around 0.1 V/μs, which means they enforce current compliance only after a few microseconds [154].
A more effective approach involves integrating a series resistor as close as possible to the ReRAM terminals. This resistor acts as a negligible resistance before the forming process begins but functions as a voltage divider as the process evolves, limiting the ReRAM current on timescales of approximately a few nanoseconds [151, 154].
In our experiments with the 1R devices, we soldered a Surface Mount Resistor (SMD) directly onto one probing tip. A picture of the modified measurement needle is shown in **Fig. 2.7 (a)**. SMD resistors with values of 10 kΩ or 100 kΩ were used specifically for the filament forming measurement, enabling reversible ReRAM switching operations thereafter.

- **Pulsed measurements:** Measurement instruments like the *parameter analyzer* can execute fast programming pulses but have slow (and accurate) read measurements, as they switch between a Pulse Measure Unit (PMU) and an Source Measure Unit (SMU) module. The involvement of mechanical relays in this process slows down the program/read measurements, making it

impractical to collect hundreds of thousands of program/read data points. However, these measurements are essential to check the stability of the ReRAM analog response to electrical pulses.

To address this issue, we built a dedicated pulsed measurement setup. A picture of this system is illustrated in **Fig. 2.7 (b)**, while **Fig. 2.7 (c)** depicts its equivalent electrical schematics. A board with different resistances is included for sensing small electrical signals from the DUT. The corresponding circuit schematic is depicted in light grey in **Fig. 2.7 (c)**. Alternatively, for current signals sufficiently high (approximately $I_{read} > 1\ \mu A$), the sensing board is not needed, and the DUT bottom terminal can be connected to a 50 Ω input of the oscilloscope, illustrated in dark grey in **Fig. 2.7 (c)**.

The sense resistance board includes resistive elements which can be selected one at the time. Their values range from $1k\Omega$ to 3.3 $M\Omega$ in logarithmic steps. This board is connected to the 1 MΩ input resistance of the oscilloscope, defining a circuit characterized by a pole with time constant (τ) equal to: $\tau = R_{sense}C_{scope,1M\Omega}$. Since $C_{scope,1M\Omega}$ is approximately equal to 20 pF, and the sense resistors are higher than 1 kΩ, at least $3\tau = 60$ ns are needed for the capacitive transients to settle. Therefore, when measuring devices switching in the 10 kΩ - 10 MΩ range, we used waveforms with a rise/fall time of at least 100 ns. Contrarily, for devices switching in the 1 kΩ - 100 kΩ range, we could bypass the sensing resistor board and use the 50 Ω input of the oscilloscope, allowing for a total pulse duration as short as 20 ns.

Please note: to measure the resistive states of the ReRAM devices, we consistently applied voltage pulses with a peak amplitude of 0.2 V to the device terminals.

2.3.2 Array measurements setup

Fig. 2.8 (a) illustrates a picture of our crossbar array measurement setup. The setup includes three National Instruments (NI) cards: two cards for providing and sensing analog voltage signals across multiple channels, and a third card, a relay module, for connecting and disconnecting the crossbar array terminals to the probing/sensing units.

The same figure also shows the Graphic User Interface (GUI) software developed to control these NI cards. This software synchronizes the multichannel probing and sensing functionalities, enabling advanced operations such as stochastic parallel updates, as shown in **Fig. 2.8 (d)**, and parallel read operations.

Unfortunately, our analog output card (PXIe6739) operates at a maximum frequency of 1 MHz, and adjusts the applied voltages with a measured slew rate of ≈ 1 V/μs [155], which hinders fast voltage pulsing activity. As a result, we adapted the shape of the programming waveforms to triangular pulses, with nominal rise and fall times of 1 μs. **Fig. 2.8 (d)** shows oscilloscope traces from two channels with stochastic pulsing activities, with probabilities of 0.7 (red) and 0.3 (black). **Fig. 2.8 (b)** depicts a wire-bonded chip that integrates 1T1R crossbar arrays. The array holder is equipped with "comb"-shaped edges, enabling accurate insertion into a Plastic Leaded Chip Carriers (PLCC) socket with 68 pins. **Fig. 2.8 (c)** presents an optical microscopy image of a wire-bonded crossbar array.

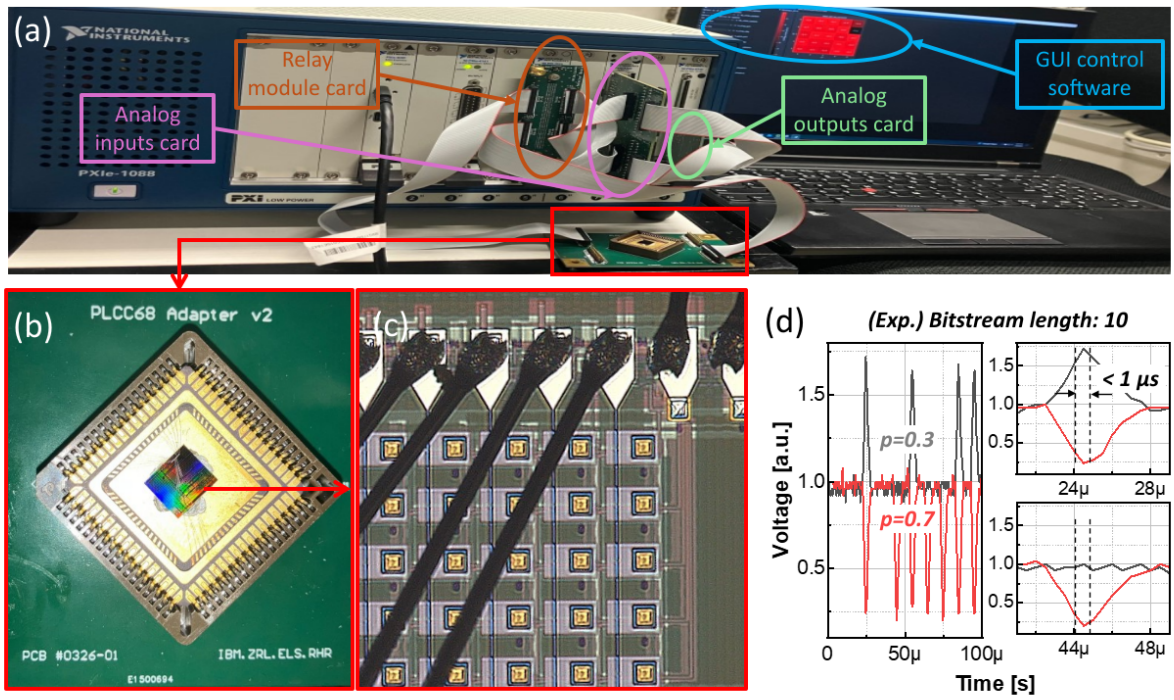


Figure 2.8: **(a)** Picture of the entire crossbar array measurement setup. **(b)** Picture of the 68-pins PLCC socket, where we glued our chip with wire-bonded arrays. **(c)** Picture of a wire-bonded 5×5 crossbar array, obtained with an optical microscope. **(d)** Synchronized stochastic pulsing activity from two analog outputs, sensed by an auxiliary oscilloscope. In the insets, the shape of the triangular waveforms is zoomed in, revealing programming pulses with a peak amplitude lasting for much less than 1 μs.

Chapter 3

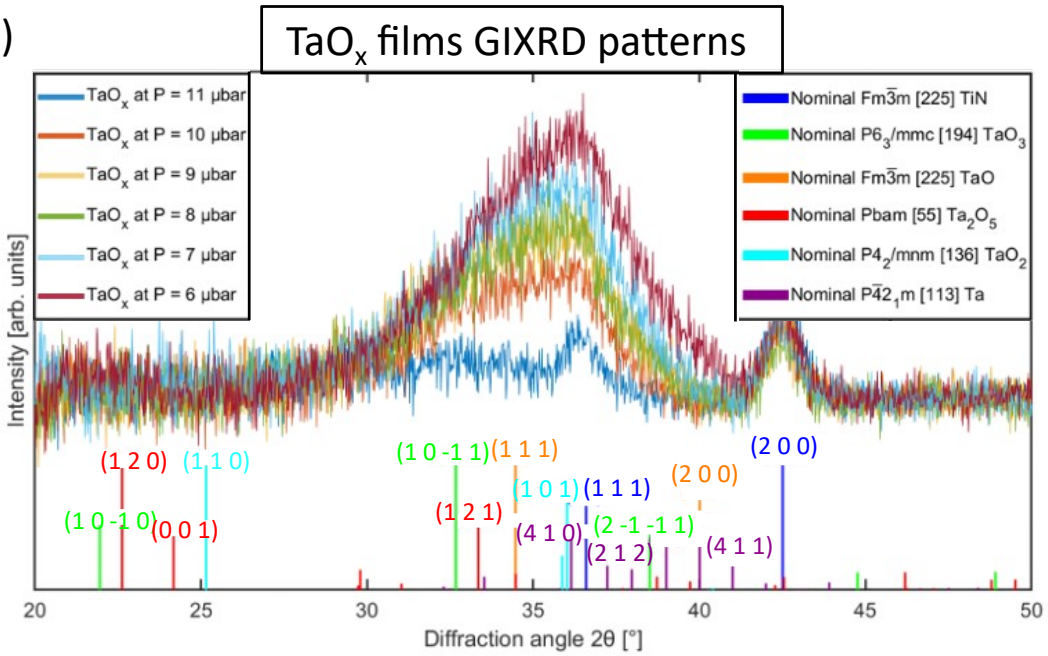
Materials optimization for analog ReRAM devices

Emerging NVM devices, compatible at the BEOL, are promising technologies for accelerating AI inference workloads, as discussed in Section 1.6. Among these novel technologies, ReRAM devices based on an oxide *bilayer* stack exhibit unique resistive-switching properties, such as bipolar, fast, and analog resistance updates driven by trains of identical voltage pulses (see Section 1.8). These properties make oxide *bilayer* ReRAM a candidate technology for deep learning acceleration (see Sections 1.6). In this chapter I will present our *bilayer* ReRAM devices, which consists of a Conductive-Metal-Oxide (CMO) layer deposited onto an insulating Transition-Metal-Oxide (TMO), such as HfO_2, all sandwiched between TiN electrodes. I will discuss the optimization of the CMO material properties, including its resistivity and thickness, to improve the electrical properties for deep learning applications.

3.1 TaO_x material development

As presented in Section 1.8, TaO_x has been explored as a CMO material in filamentary ReRAM devices, demonstrating excellent electrical properties for deep learning applications [95, 128, 133].
The deposition of conductive-TaO_x thin films can be achieved through the reactive sputtering method, as outlined in Section 2.1. I fabricated multiple layers of TaO_x for structural characterization. All these films were deposited onto HfO_2/TiN substrates to maintain the same growth conditions as in the ReRAM devices. The sputtering parameters, such as power (50 W), duration (200 s), and gas mixture stream (O_2/Ar= 1/40 sccm), were kept constant in all depositions, except for the chamber pressure, which was varied from 6 to 11 μbar in 1 μbar increments.
Grazing incidence x-ray diffraction (GIXRD) characterization provides information about the structure and the stoichiometry of the deposited suboxides. The GIXRD profiles of the TaO_x/HfO_2/TiN stacks, obtained by scanning at an incidence angle of 0.65°, are shown in **Figure 4.1 (a)**. The peak at 42.5° is characteristic of the (200) TiN [156]. The broad TaO_x peaks are shaped by the superposition of multiple contributions; in the figure, I indicated only the crystallographic peaks accounting for more than 30% of the intensity. Their deconvolution, reported in the annex, reveals the coexistence of multiple stoichiometries of TaO_x and Ta metal, indicating that the grown material does not have a unique composition but is rather a complex phase. With increasing sputtering pressures, a clear change in the broad TaO_x peak shape is identified, due to the change in the relative intensity of the contributing peaks: the oxidized components increase at the expenses of the pure Ta peak. These findings indicate that more oxygen is incorporated in the Ta lattice when sputtering at higher pressures, in accordance with previous literature [157].

(a)



(b)

TaO$_x$ properties from XRR analysis

Pressure [µbar]	Thickness crystalline TaO$_x$ [nm]	Density crystalline TaO$_x$ [g/cm^3]	Thickness amorphous TaO$_x$ [nm]	Density amorphous TaO$_x$ [g/cm^3]	Planar TaO$_x$ resistivity [Ω cm]
6	25.2	11.6	2.2	9.3	0.0082
7	26.8	11.0	1.8	9.3	0.0317
8	26.5	10.9	1.5	9.3	0.086
9	26.4	10.7	2.2	9	0.261
10	26.6	10.7	2.3	8.6	0.496
11	4.1	9.9	2.2	8.5	n.a.

Figure 3.1: **(a)** GIXRD profiles of the $TaO_x/HfO_2/TiN$ samples with TaO_x sputtered at increasing chamber pressures from 6 to 11 μbar. The nominal peak positions of the various TaO_x crystalline phases, Ta and TiN, are reported from the reference database [147], with the corresponding space group. **(b)** Table summarizing the TaO_x films thicknesses, densities and resistivities, as determined by XRR and CTLM. **(c)** Measured sheet resistance values plotted as a function of the sputtering chamber pressure. Modified from [94]. These results were obtained within the supervision of Master student Roberto Guido.

Table 4.1 (b) lists the TaO_x films' thicknesses and densities, measured by X-Ray Reflectivity (XRR). The XRR patterns and fits are reported in the annex. From this analysis, we understand that deposition rate is constant and approximately ≈ 0.15 nm/s for all layers, except for the one deposited with a chamber pressure of 11 μbar, where the rate drops to approximately ≈ 0.03 nm/s. Furthermore, the XRR analysis reveals the formation of a ≈ 2 nm thick TaO_x interface layer at the boundary with the HfO_2, with lower density than the bulk TaO_x (see **table** in **4.1 (b)**). With increasing chamber

pressures, the density of hte bulk TaO_x layer gradually decreases, from 11.6 g/cm^3 at 6 μbar to 10.7 g/cm^3 at 10 μbar, and 9.9 g/cm^3 at 11 μbar.
The TaO_x films' resistivities were characterized by Circular Transmission Line Measurements (CTLM), with values already reported in **Figure 2.1 (d)**. The materials' resistive window spans over more than one order of magnitude, from a few k$\Omega/\square$ up to values larger than 100 $k\Omega/\square$.
I converted the sheet resistances into resistivities (ρ) and reported the values in **Table 4.1 (b)**. For the film deposited at 11 μbar it was not possible to quantify the resistivity because the measured sheet resistance exceeded the tool maximum measurable limit of approximately $\sim$ 10 G$\Omega/\square$.

3.2 ReRAM device processing and chemical analysis

The structure of our ReRAM devices is depicted in chapter 2, **Fig. 2.3 (c)** and **Fig. 2.3 (e)**. This device geometry reflects a "*vertical*" design, featuring a single bottom contact, made of n-doped Si, shared among all devices. Each device has isolated top contacts, individual electrical characterization. The Bottom Electrode (BE) consists of a 20 nm thick TiN layer, deposited by plasma-enhanced atomic layer deposition (PE-ALD) at 300°C using a tetrakis-(dimethylamino)titanium (TDMAT) precursor and (N_2, H_2) plasma. Subsequently, a 6 nm thick HfO_2 layer is deposited by PE-ALD at 290°C using a tetrakis-(ethylmethylamino)hafnium (TEMAH) precursor and O_2 plasma. Following this, a thin TaO_x film is deposited by reactive sputtering, as described in Section 3.1. I fabricated multiple chips, depositing TaO_x layers with thicknesses ranging from 20 to 60 nm, to understand their impact on the ReRAM's electrical properties. The TaO_x thin film is covered by a 20 nm thick TiN Top Electrode (TE), deposited by RF sputtering of a TiN target in a mixed (Ar,N_2) plasma. A 50 nm W layer is sputtered on top. The deposition of the W/TiN/TaO_x multilayer stack proceeds without vacuum breaking to avoid uncontrolled oxidation at the reactive interfaces.
Afterwards, a *mesa* device geometry is patterned through inductively coupled plasma (ICP) etching of the W/TiN/TaO_x stack, using a mixed CHF_3 and SF_6 plasma, which stops at the HfO_2 layer. A passivation layer of 100 nm thick SiN_x is grown by plasma-enhanced chemical vapour deposition (PECVD). A via to the device TE is etched by reactive-ion etching (RIE) with a mixed CHF_3 and O_2 plasma. A 100 nm thick W layer is deposited by DC sputtering and then patterned by RIE to define the top electrical contacts of the ReRAM devices. The described process flow is CMOS-compatible and it avoids any lift-off steps, which are not feasible in foundries.

After completing the process flow, we characterized the entire material stack by Transmission Electron Microscopy (TEM). The specific structure that is shown in **Figure 3.2 (a)** is a ReRAM with an approximately $\approx$ 20 nm thick TaO_x layer deposited at 9 μbar. The lamella preparation for the TEM analysis was conducted using the FEI Helios NanoLab 450S Focused Ion Beam (FIB). The TEM analysis was performed with a double spherical aberration corrected JEOL ARM200F operated at 200 kV. Energy-Dispersive-Spectroscopy (EDS) scans were conducted using a liquid-nitrogen-free silicon drift detector.
Figure 3.2 (a) depicts a scanning-TEM (STEM) image in bright field (BF) mode. The contrast is sufficient to identify a thin, dense, amorphous layer at the TaO_x/HfO_2 interface, estimated to be approximately $\approx$ 3 nm thick. Additionally, a thin oxidized TiO_x, approximately $\approx$ 1.5 nm thick, grows at the at the HfO_x/TiN interface. This TiN surface oxidation is attributed to the deposition of HfO_2 in an O_2 plasma atmosphere at 290°C [91].
Figure 3.2 (b) shows the chemical analysis performed after an EDS scan. This analysis confirms that the TaO_x layer grew inhomogeneously: the first $\approx$ 3 nm are amorphous and more oxidized (higher O counts, lower Ta counts), while the remaining $\approx$ 16 nm thick bulk TaO_x layer is crystalline (details not

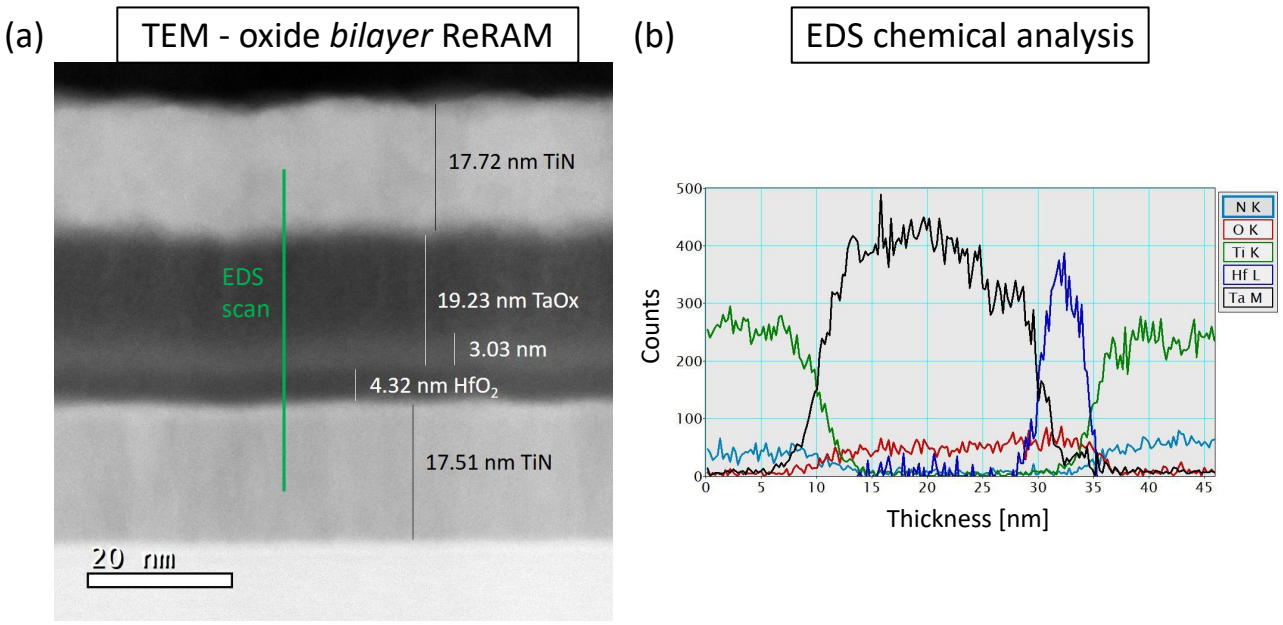


Figure 3.2: **(a)** Bright Field (BF) STEM image of the TiN/TaO_x/HfO_2/TiN ReRAM stack, revealing an amorphous TaO_x phase between the crystalline TaO_x bulk and the HfO_x layer. The green vertical line indicates the region scanned for Electron-Dispersive-Spectroscopy (EDS) chemical characterization, shown in **(b)**. Modified from [94].

shown) and less oxidized (lower O counts, higher Ta counts). The pillar growth of the TiN is believed to be responsible for the projection of TiN roughness to the rest of the stack's morphology, causing an overlap of the Hf and Ta counts at the TaO_x/HfO_2 interface (and, similarly, of the Hf and Ti counts at the HfO_2/TiN interface). Moreover, interdiffusion can not be excluded.

3.3 Comparison between baseline and oxide *bilayer* ReRAM

In this Section, I compare the DC I-V characteristics of a "baseline" ReRAM device, based on the MIM structure (Ti/HfO_2/TiN, with a Ti thickness of $\approx$ 10 nm) to an oxide *bilayer* structure (TiN/TaO_x/HfO_2/TiN, with a TaO_x thickness of $\approx$ 30 nm).
Figure 3.3 (a) and (b) display the DC I-V curves of the filament *forming* process, obtained through the application of a positive voltage sweep. The *bilayer* ReRAM shows lower currents before filament forming and a higher breakdown voltage (V_f) of $\approx$ 5.5 V, compared to $\approx$ 2.5 V of the baseline structure. The lower leakage currents of the *bilayer* structure can be explained by two main reasons. First, the Ti electrode tends to scavenge more O_2 from the HfO_2 layer compared to Ta, due to their different oxygen affinities [118]. Additionally, in the oxide *bilayer*, the top electrode has been already partially oxidized during its deposition due to the oxygen plasma environment.
After the filament is formed, the application of a negative voltage sweep induces a "reset" transition (increase of the resistive state) in the *baseline* device and a "set" transition (decrease of the resistive state) in the bilayer device, as shown in **Figure 3.3 (c) and (d)**. Positive voltages also lead to opposite resistive-switching mechanisms in the two devices, bringing them back to resistive states close

to the post-forming values.
Examinating the shape of these curves reveals that, in the *baseline* device, the conductance changes abruptly as soon as the resistive-switching process initiates, while in the *bilayer* ReRAM this process is gradually sustained by increasing voltages.
We repeated the same I-V measurements five times, and the corresponding curves are displayed in **Figure 3.3 (e) and (f)**. The *baseline* device shows variability in the V_{set}, attributed to the stochastic nature of the *reset* process [93], accompanied by a large variability of the programmed High-Resistive-State (HRS). In contrast, the current curves of the *bilayer* ReRAM almost overlap during cycling, resulting in improved stability of the HRS.

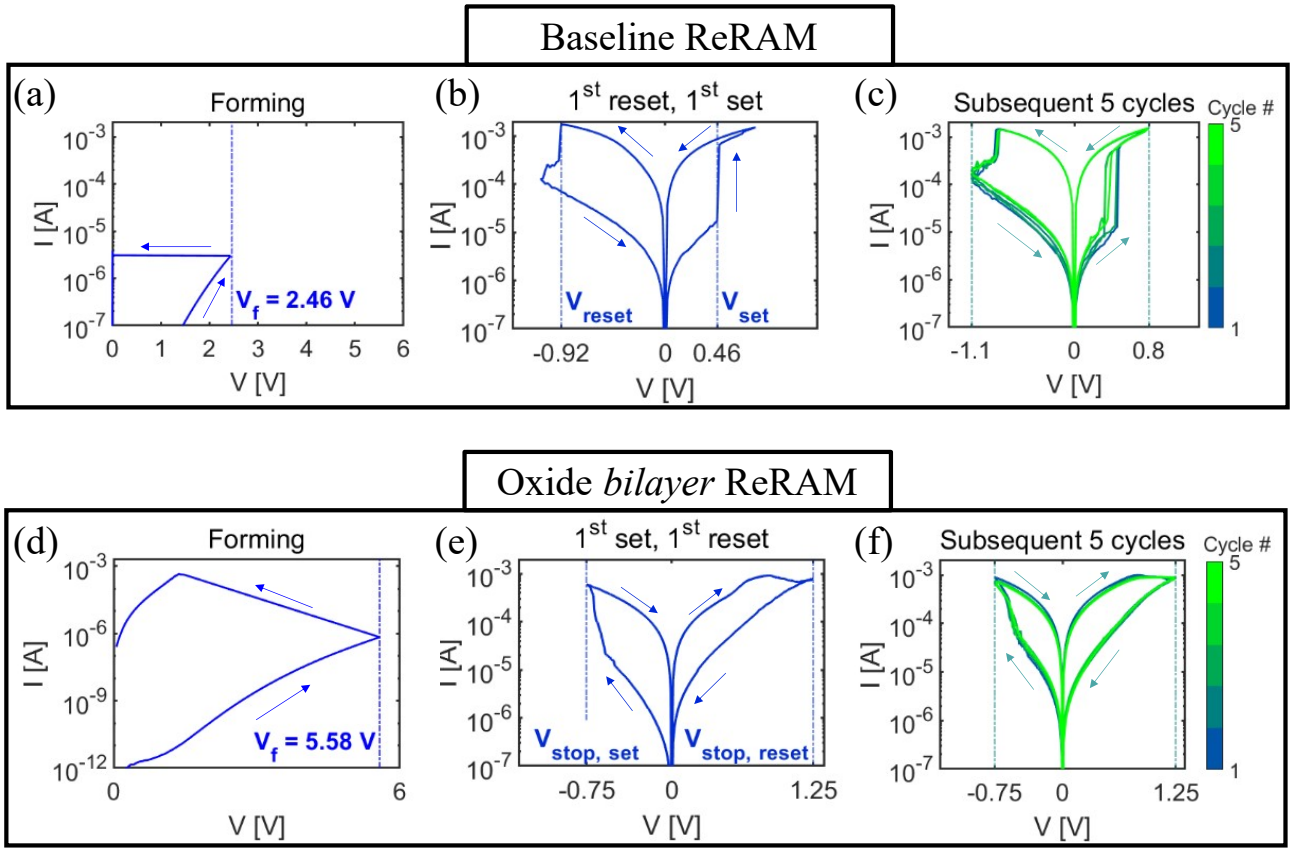


Figure 3.3: Comparison between the DC I-V characteristics of the *baseline* ReRAM (first row), and the oxide *bilayer* ReRAM (second row). **(a)** and **(d)** show the filament *forming* curves. **(b)** and **(e)** display the first *set* and *reset* cycles. **(c)** and **(f)** present the 5 programming cycles following the first *set* and *reset*. Modified from [94].

3.4 Speculative interpretation of the resistive-switching mechanisms

In this Section, we offer a speculative interpretation of the reversed switching polarity between the baseline and the bilayer ReRAM devices. In filamentary ReRAMs, the polarity of the resistive switching mechanisms depends, among other factors, on the WFs of the electrodes [88, 158]. I fabricated an oxide *bilayer* device with a Pt BE, mandatorily forcing a WF asymmetry in the same direction as the one of the *baseline*. However, as displayed in **Figure 3.4**, the switching polarity of the devices with Pt or

TiN BE is the same. Thus, we conclude that the reverse switching polarity between the *bilayer* and the *baseline* ReRAMs cannot be attributed to an inversion of the WF asymmetry.
To explain this property, we propose an interpretation of the nanoionic processes leading to the resistive-switching mechanisms in **Figure 3.4**. For each programming step - filament forming, set and reset - we outline a description of the related microscopic mechanisms.
In **Figures 3.4 (a)** and **Figure 3.4 (b)** we discuss the *forming* process. In our interpretation, the applied voltage generates a conductive filament of oxygen vacancies ($V_O^{\bullet\bullet}$) across the HfO_2 layer, due to oxygen anions (O^{2-}) drifting towards the TaO_x. This process reflects the generally accepted model for filament forming [116]. The abrupt increase of the current upon forming the filament induces a local temperature increase, combined with Coulomb repulsion, causing the O^{2-} to diffuse radially [116]. A more oxidized region ($TaO_{y>x}$) is obtained, dominating the device resistance after *forming*.
Now the device is formed and programmed in a state denoted as "post-forming resistance" (R_{pf}), corresponding to the device's High-Resistive-State (HRS). Next, the device can be *set* to a Low Resistive State (LRS) upon the application of a negative voltage sweep, as shown in **Figure 3.4 (c)**. To prevent excessive currents, I imposed a compliance of 100 μA. Our nanoionic interpretation of the *set* process is depicted in **Figure 3.4 (d)**. Activated by the negative voltage, the O^{2-} formerly diffused in the TaO_y region migrate back to the HfO_2 layer, reducing the volume of the O-rich TaO_y region. In the HfO_2 layer, the $V_O^{\bullet\bullet}$ filament recombine with the incoming O^{2-}. Overall, I measure a decrease in the total device resistance, indicating that, between these two counteracting processes, the reduction of the TaO_y layer dominates over the oxidation of the filament tip. A positive voltage induces *reset* transition,

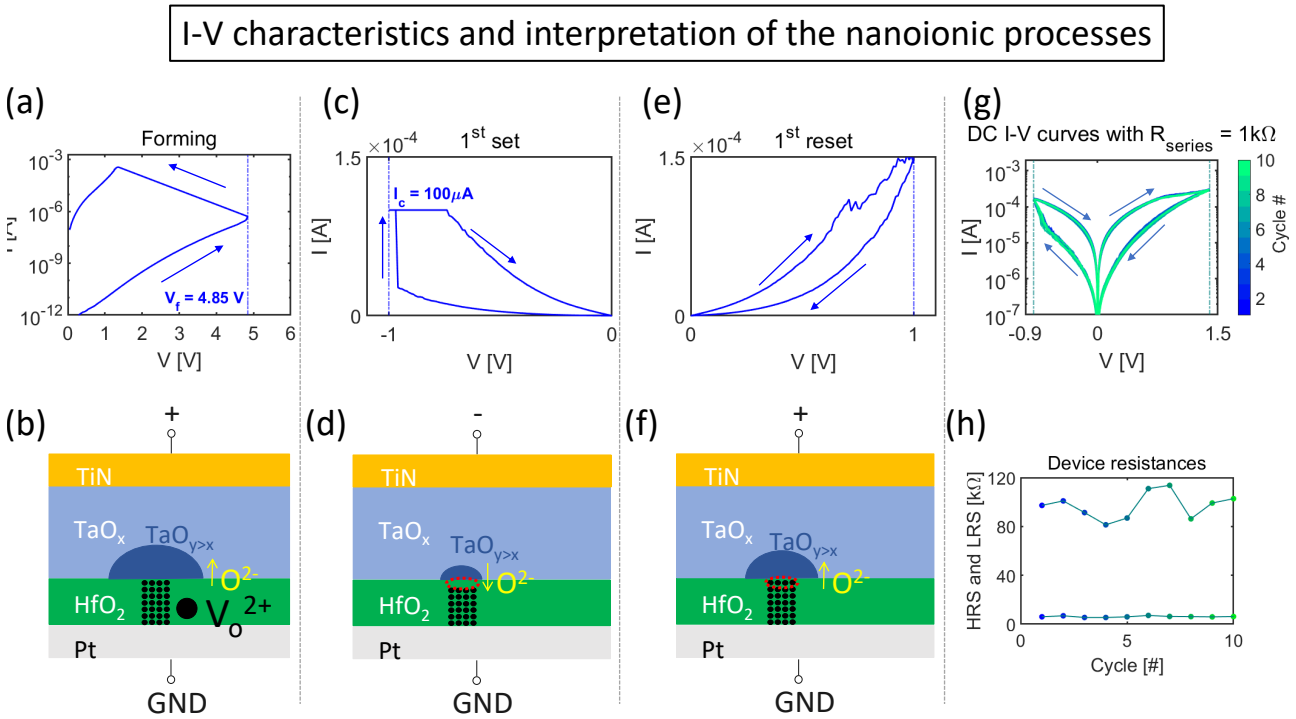


Figure 3.4: DC I-V characteristics of the TiN/TaO_x/HfO_2/Pt *bilayer* ReRAM, including the filament *forming* **(a)**, the first *set* **(c)** and first *reset* **(e)**. Below, in **(b)**, **(d)** and **(f)**, the interpretation of these processes as nanoionic processes occurring at the nanoscale. **g)** The DC bipolar switching of the same *bilayer* stack when a 1 kΩ external series resistor is introduced. **(h)** Resistive states extracted from **(g)**. Modified from [94].

as shown in **Figure 3.4 (e)**. Our interpretation of the *reset* process is depicted in **Figure 3.4 (f)**. O^{2-} anions drift towards the TaO_x layer, due to the applied positive voltage, while the filament gap in the HfO_2 reduces. In this case, since I measure a total increase in the device resistance, the dominant process is the local oxidation of the TaO_x material.
Under these hypotheses, we conclude that the TaO_x layer acts as current bottleneck both in the LRS and in the HRS.

I further explored the DC I-V curves of the Pt-BE device and made a very interesting discovery. This time, I introduced an external series resistor (1 kΩ), using the methods described in chapter 2, Section 2.3. The results are reported in **Figure 3.4 (g)**: you can see a gradual *set* process, sustained by increasing voltages. In **Figure 3.4 (c)**, where no series resistor was included, the *set* is an abrupt process, requiring a current compliance to avoid irreversible breakdown. The only difference in the measurements depicted in **Figure 3.4 (c) and (g)** is the introduction of a 1 kΩ series resistor.
The speculative explanation we have provided is inspired from previous literature. Former research studies on filamentary ReRAM devices showed that the acceleration of the *set* process depends on the previous resistive state [124]. When the device is programmed at very high resistances, the *set* process is accelerated, while it tends to become a more gradual process when the device is prepared at intermediate resistive states [124].
Our Pt-BE oxide *bilayer* is programmed to a relatively high resistance after the filament forming ($R_{HRS,Pt-BE} \approx 305$ kΩ, from **Figure 3.4 (a)**). During the *set* process, illustrated in **Figure 3.4 (d)**, the O-rich TaO_y region shrinks, but maintains the dominant contribution to the device resistance, leading to a local electric field (F) enhancement. The current increase accompanying the *set* process leads to a temperature (T) growth in the same TaO_y region, due to Joule heating. The (F,T) enhancement promotes further O^{2-} migration, ultimately installing a positive (F,T) feedback loop, resulting in an abrupt resistive-switching mechanism [119].
In contrast, as discussed in chapter 1, Section 1.8, a series resistor placed in series dynamically reduces the electric field in the ReRAM device, hindering the installation of a positive (F,T) feedback. This conclusion is sustained by former literature, showing that the integration of series resistors reduces the power dissipated in the device during the *set* process [154].
A 1 kΩ series resistor is sufficient for this purpose, without compromising the resistive switching window, as displayed in **Figure 3.4 (g) and (h)**, where an on/off ratio of approximately 1 order of magnitude is achieved by applying a V_{stop}= +1.4 V.

It may be argued that the gradual switching properties, typical of *interface-type* resistive-switching devices [159], could be attributed to non-filamentary valence-change mechanisms occurring at any of our materials' interfaces. To confirm our hypothesis of filamentary switching, I studied the area dependency of the ReRAM's resistive states, including $(6\ \mu m)^2$, $(12\ \mu m)^2$, $(30\ \mu m)^2$ and $(60\ \mu m)^2$ areas in the statistics. For each of them, I applied 10 bipolar voltage sweeps up to the stop values -1.2 V and +1.6 V. The median LRS and HRS were grouped in boxplot graphs and plotted against the device area, as shown in **Figure 3.5**. We observe that neither the HRS nor the LRS scale with the area. The fluctuations between different cells can be attributed to device-to-device variability and to the relatively small size of the collected statistics.
These results support the hypothesis that the switching mechanisms are filamentary-based, even if they become gradual processes with the introduction of a 1 kΩ series resistor.

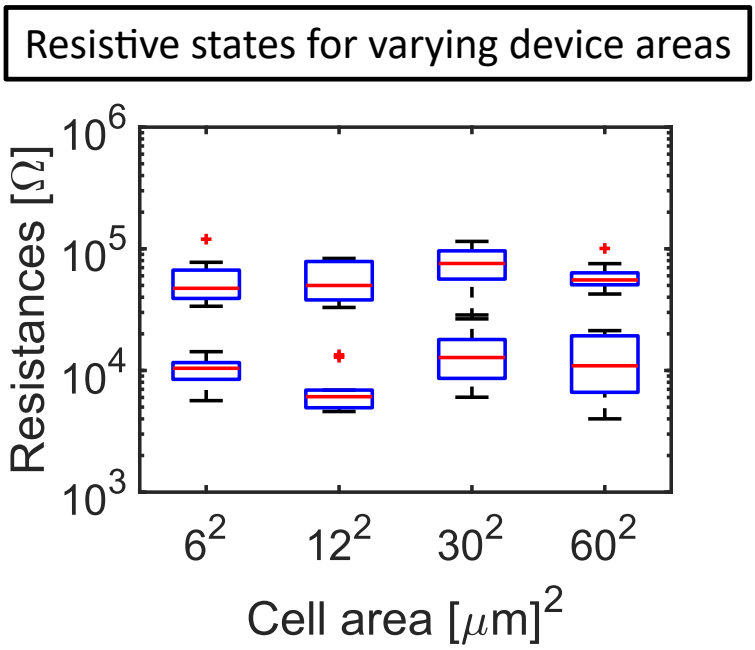


Figure 3.5: Resistive states (LRS and HRS) with varying ReRAM areas. Data was collected from 10 devices for each selected area. Every device was tested under identical electrical conditions ($V_{stop,set}$ = -1.2 V and $V_{stop,reset}$ = +1.6 V). Modified from [94].

3.5 TaO_x resistivity and ReRAM electrical properties

We measured the electrical properties of two ReRAM devices incorporating TaO_x layers with different resistivities. The two resistivities under comparison are: $\rho_{TaO_x} \approx 0.01\ \Omega\ cm$ and $\rho_{TaO_x} \approx 0.25\ \Omega\ cm$. In both cases, the TaO_x film thickness is $\approx$ 30 nm.
In **Figure 3.6 (a) and (b)** we compare the DC I-V characteristics. Both devices show low-stochastic and gradual switching properties, driven by low voltages (<1 V). The main difference between them is that the ReRAM with a more resistive TaO_x material shows higher resistances and a broader ON/OFF ratio.
To validate this result statistically, we measured 10 devices per each TaO_x type, using the same stop voltages. In **Figure 3.6 (c) and (d)** we compare the cumulative distributions of their resistive states. Each device is represented by a pair of lines of the same colour, one for the LRS and the other one for the HRS. The statistics confirms that the more resistive TaO_x improves the HRS values beyond 10 $k\Omega$ and increases the on/off ratio by approximately one order of magnitude. However, increase in resistive states comes at the cost of broader device-to-device variability (more spread-out cumulative distributions), and larger cycling variability (more oblique cumulative distributions).

3.6 TaO_x thickness variations and ReRAM electrical properties

I fabricated three variations of TiN-BE *bilayer* devices with different TaO_x thicknesses ($\approx$ 20, $\approx$ 50 and $\approx$ 60 nm) to understand their impact on the device switching properties. All devices use highly resistive TaO_x material ($\rho_{TaO_x} \approx 0.25\ \Omega\ cm$). **Figure 3.7 (a)** shows that the filament forming voltages remain almost constant with increasing thicknesses (small variations can be attributed to chip-to-chip

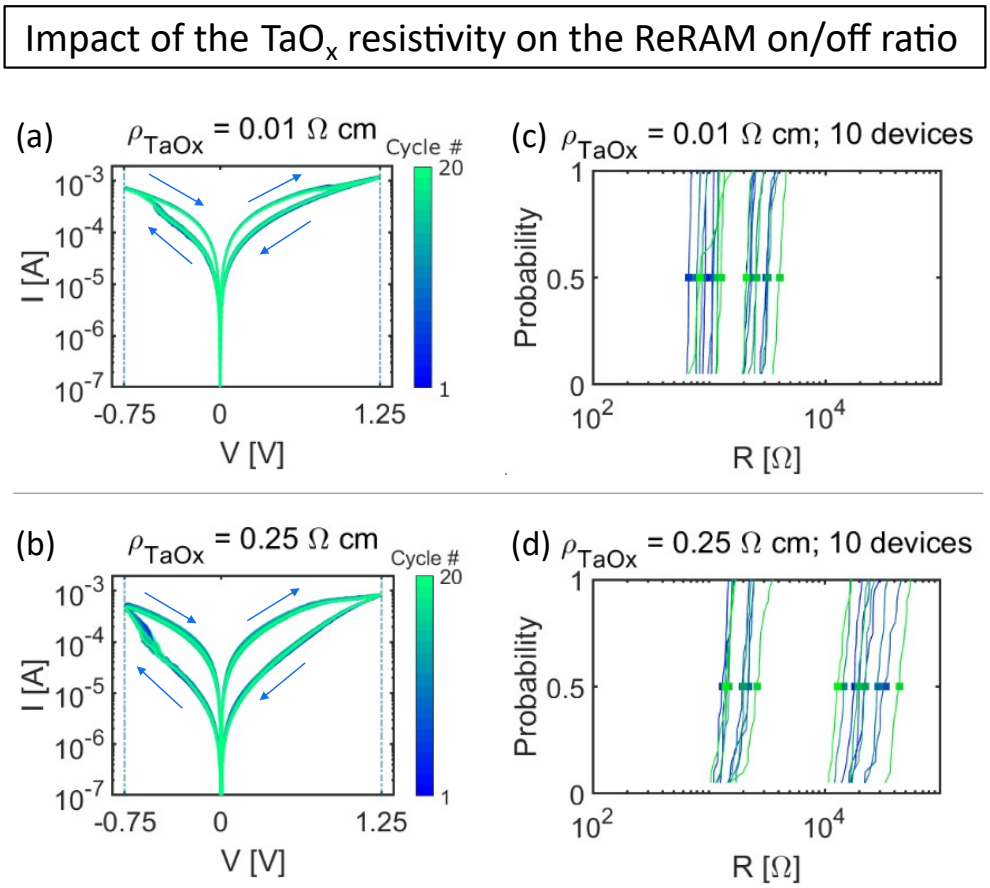


Figure 3.6: ReRAM electrical characteristics for different TaO_x resistivities (first row: $\rho_{TaO_x} \approx 0.01$ Ω cm, second row: $\rho_{TaO_x} \approx 0.25$ Ω cm). **(a)** and **(b)** DC I-V curves. **(c)** and **(d)** Cumulative distributions of the resistive states. Modified from [94].

variability).

An interesting observation is the decrease in the High Resistive States (HRS) for devices with thicker TaO_x layers, as depicted in **Figure 3.7 (b)**. This plot also includes data on the resistive states measured right after filament forming (R_{pf}). Both R_{pf} and HRS values decrease with increasing TaO_x thicknesses, with the R_{pf} values generally higher than the HRS values. This indicates that the non-reversible resistance decrease accompanying the *forming* process is stronger for the ReRAM devices with a thicker TaO_x layer.

Figure 3.8 (a), (b) and (c) compare the DC I-V curves following filament formation. The device with the thinnest TaO_x layer (20 nm) exhibits similar electrical properties to the device with a 30 nm TaO_x layer shown in **Figure 3.6 (b)**). Conversely, the device with a 50 nm TaO_x layer displays an abrupt *set* transition, while the one with the 60 nm thick TaO_x shows multiple abrupt *set* transitions. To understand this trend, I correlated the distributions of the HRS with the R_{pf}. The thicker TaO_x layers result in lower R_{pf} values, due to the TaO_x thermal properties. The thermal coefficient of a conductive TaO_x film depends on its resistivity [160]. Given the resistivity $\rho_{TaO_x} \approx 0.25$ Ω cm, we assume a low thermal coefficient $\lesssim 1$ W/mK [160]. Consequently, the TaO_x acts as a heat confinement layer in our devices [123, 128].

This heat confinement effect, depicted in the nanoscale model in **Figure 3.4 (a)**, becomes more pronounced with thicker TaO_x layers. During filament formation, high temperatures (T) at the

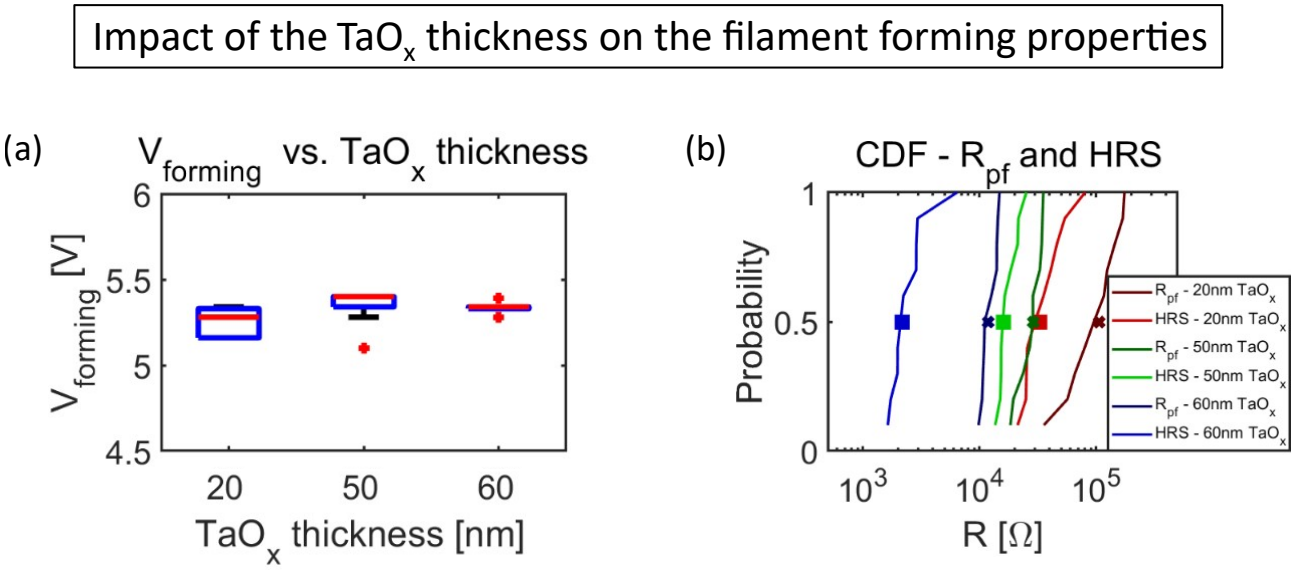


Figure 3.7: **(a)** Filament forming voltages of ReRAM devices with a 20, 50 nm and 60nm thick TaO_x layer. Data shown in boxplots, 10 devices measured per each TaO_x thickness. **(b)** Cumulative-Distribution-Functions (CDF) of the High Resistive States (HRS) and the post-forming resistances (R_{pf}) from ReRAM devices with a 20, 50 nm and 60nm thick TaO_x layer. 10 devices per each TaO_x thickness were measured. Modified from [94].

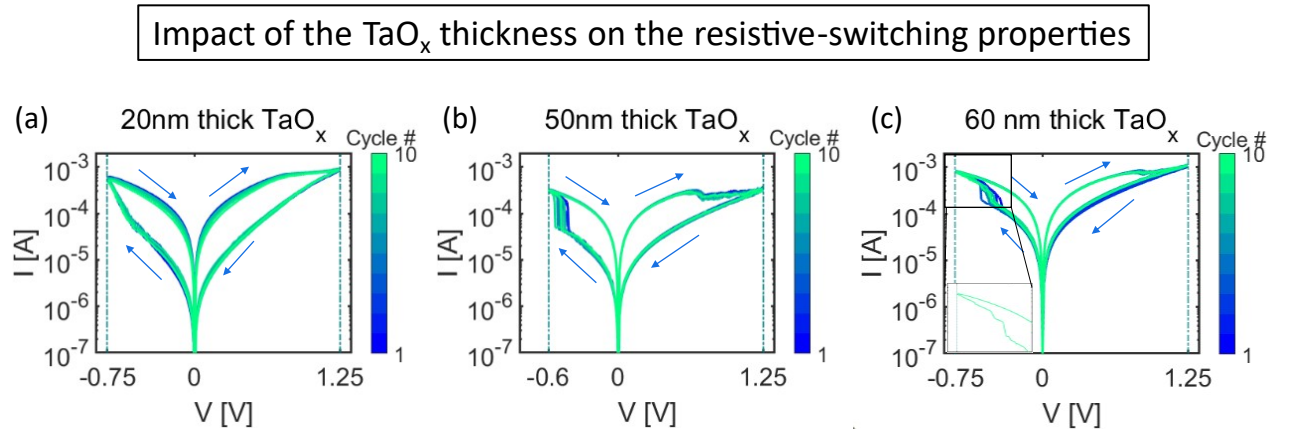


Figure 3.8: DC I-V characteristics from ReRAM devices with **(a)** 20 nm, **(b)** 50 nm, and **(c)** 60nm thick TaO_x layer. Modified from [94].

TaO_x/filament interface promote ion mobility, facilitating the O^{2-} migration from the HfO_2 to the TaO_x, and generating oxygen vacancies ($V_O^{\bullet\bullet}$) in the HfO_2 layer.
With thicker TaO_x layers, the heat confinement increases due to the greater distance from the heat sink (the thermally conductive TiN-TE). Higher local temperatures favor lateral O^{2-} diffusion in the TaO_x layer [161, 162], reducing the measured resistances (the device resistance is measured vertically, see illustration in **Figure 3.4 (b)**).
Enhanced heat confinement also explains the sharper *set* mechanisms observed in **Figure 3.8 (b) and (c)** for devices with thicker TaO_x layers, leading to a rapid temperature surge during the *set* process and enhancing the positive feedback loop between temperature and electric field (T and F) [119, 163]

Thus, the abrupt *set* transitions in devices with 50 and 60 nm thick TaO_x layers are driven by enhanced heat confinement effect, overcoming the gradual switching seen in devices with thinner TaO_x layers. A notable case is the ReRAM device with a 60 nm thick TaO_x, which exhibits multi-step *set* transitions, as shown in the magnified corner of **Figure 3.8 (c)**. This phenomenon could be due to the formation of multiple filaments, as proposed by Wu et al. [128].

3.7 Generalization of the oxide *bilayer* ReRAM design

The oxide *bilayer* concept for the material design of a ReRAM device can be generalized to different materials. For instance, we replaced the conductive-TaO_x layer with a different conductive metal-oxide, denoted as "Metal-Oxide (MO)" in this dissertation.
In **Figure 3.9 (a)**, I compare the forming voltages of the TaO_x- and the MO-ReRAM devices. The latter shows lower forming voltages due to a uniform oxygen reduction across the entire MO layer. In contrast,the TaO_x material grows in two separate phases, with a gradient of oxygen-content along the vertical axis, as detected by TEM and EDS analysis (see **Figure 3.2 (a) and (b)** in Section 3.1. **Figure 3.9 (a)** also shows a logaritmic dependence between the ReRAM area and the filament forming voltage. This trend can be explained by considering the formation of a conductive filament as a stochastic process, involving percolative paths across the metal-oxide stack [164].
Figure 3.9 (b) exhibits 150 DC I-V cycles of an MO-ReRAM device, showing similar resistive-switching properties to the TaO_x-ReRAM devices, such as the gradual change of the resistive state upon increasing voltages and the reduced stochasticity compared to a *baseline* device (see **Figure 3.3 for comparisons**). As illustrated in **Figure 3.9 (c)**, approximately $\sim$ 50 programming cycles are required to "stabilize" the device resistive window.

The analog resistive switching properties are enhanced when the ReRAM devices are programmed by short voltage pulses. **Figure 3.10 (a)** illustrates the programming scheme devised to test the pulsing response. The scheme consists of sending multiple (200 $\times$) voltage pulses of identical polarity and

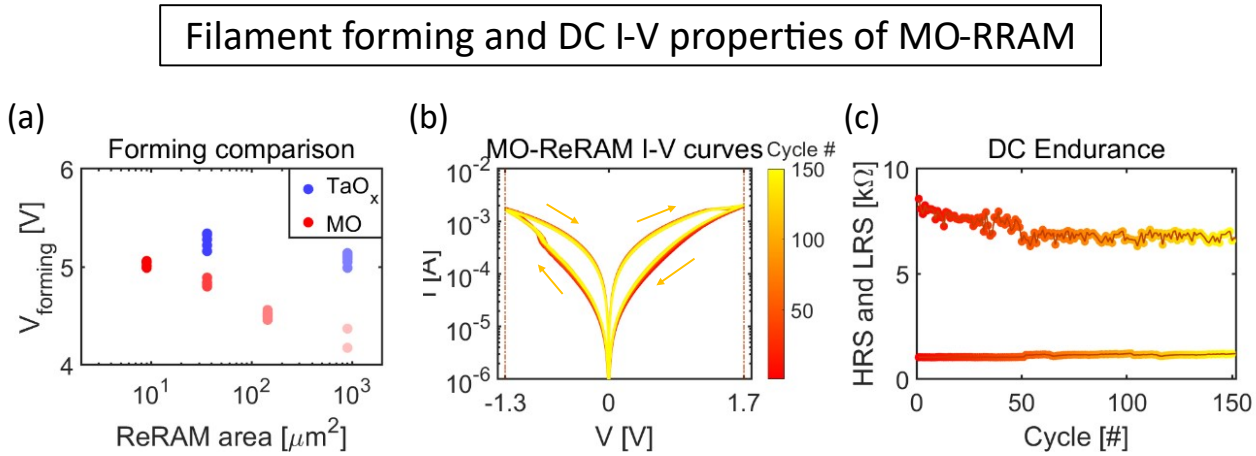


Figure 3.9: **(a)** Comparison between the filament forming voltages of the TaO_x- and the MO-ReRAM devices. **(b)** DC I-V characteristics of the MO-ReRAM devices. **(c)** LRS and HRS values extracted from the DC I-V curves.

amplitude. Afterwards, an identical number of programming pulses is applied in the opposite polarity. This sequence (200 pulses up, 200 pulses down) is iterated numerous times (10 cycles) to check the

stability of the synaptic traces. A pulse duration of 200 ns was selected for both devices, with the following voltage parameters: V_+ = 1.1 V, V_- = -0.95 V, and V_+ = 1.7 V, V_- = -1.4 V for the TaO_x- and the MO-ReRAM, respectively.
Figure 3.10 (b) and **Figure 3.10 (c)** depict the synaptic response of the ReRAM devices. Both structures exhibit analog and stable Long-Term-Potentiation (LTP) and Long-Term-Depression (LTP), driven by identical pulse streams, which is a fundamental requirement for deep learning accelerators based on resistive-switching devices (see Section 1.6). Although the TaO_x-ReRAM is programmed by voltage pulses of lower amplitude, its synaptic trace covers a broader window (approximately a factor 5×), while the MO-ReRAM shows approximately a 3× window. Moreover, the conductive trace of the TaO_x-ReRAM is positioned at lower values, which is beneficial for the energy efficiency of the *forward* and *backward* passes (see Section 1.6).

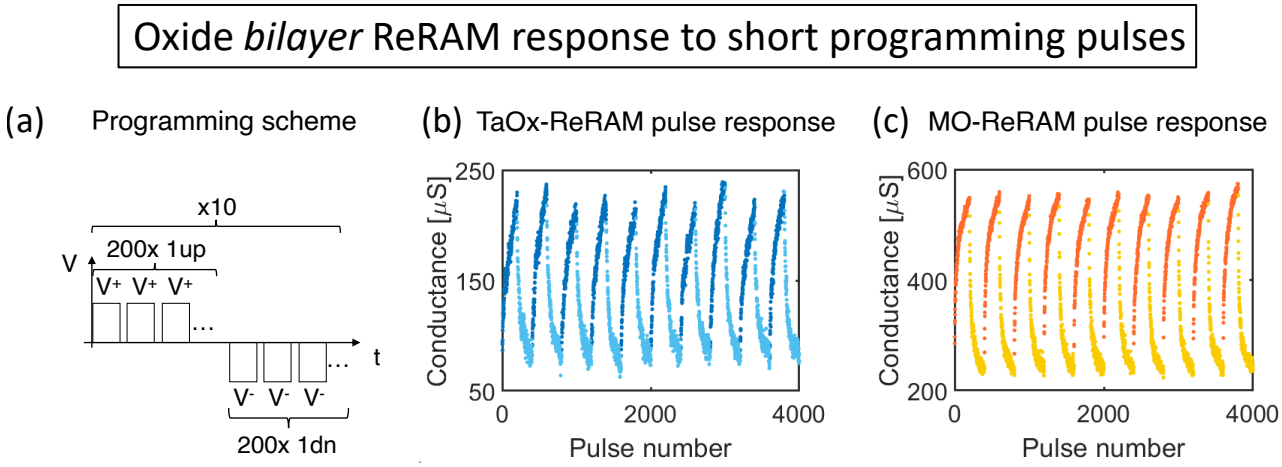


Figure 3.10: **(a)** Programming scheme devised to test the ReRAM response to short programming pulses. **(b)** and **(c)** Analog long-term potentiation and depression cycles for TaO_x- and MO-ReRAM devices. Modified from [94].

3.8 Conclusion

In this chapter, we introduced the material and electrical properties of the ReRAM devices based on a conductive-TaO_x/HfO_2 bilayer stack. We have experimentally demonstrated the impact of the TaO_x thickness and resistivity on the ReRAM electrical properties by fabricating multiple chips with varying TaO_x material properties. We understood that the TaO_x layer acts as a current bottleneck, and more resistive TaO_x material results in higher resistive states in the device. We also learned that a thicker TaO_x layer acts as thermal barrier, accelerating the device resistive-switching mechanisms, which hinders analog programming into intermediate states.
We demonstrated that the oxide *bilayer* design concept can be generalized to other Conductive-Metal-Oxides (CMO), such as our "MO" material. The MO- and the TaO_x- ReRAM devices show very similar DC I-V characteristics and synaptic traces. Compared to *baseline* devices based on the Metal-Insulator-Metal (MIM) structure, the CMO-ReRAM devices exhibits greatly reduced stochasticity and enhanced analog switching properties. Both aspects are critical for the application of deep learning acceleration with analog Resistive-Processing-Units (RPUs) in the BEOL (see Section 1.6).

In the next chapter, we will present further results on ReRAM materials optimization. This time, we

modified the deposition of the HfO_2 layer, which led to increased leakage currents and strongly reduced filament forming voltages. Consequently, the devices could be operated at lower currents, compared to the ones exhibited in this chapter.
I will also present an alternative process flow for the ReRAM devices, enabling precise definition of active areas down to 50 nm $\times$ 50 nm. Finally, I will discuss a more advanced electrical characterization, testing the multistate retention up to 85 °C and the endurance to iterative programming cycles. I extended the electrical pulse characterization measurements, by including a test sequence simulating the *zero-shifting* technique, required as a starting step for the *Tiki-Taka* algorithm (see Section 1.7 for more details).

Chapter 4

ReRAM device scaling and analog properties

In this chapter, I will present further optimized ReRAM materials and processes, enabling the scaling of unit cell geometry and device operational power. I will also evaluate the device compatibility with the novel *Tiki-Taka* algorithm, which requires specific analog resistive-switching properties.
The topics covered in this chapter include:

1. **Geometry scaling**: Modern AI models, such as Convolutional Neural Networks (CNNs) [1, 2] and transformers [3], involve large networks and numerous parameters. These parameters can range from hundreds of millions [65, 66] for CNNs to hudreds of billions for transormer-based models [7, 8].
   The computing workloads associated with AI models can be partially executed through a Resistive-Processing-Unit (RPU), by mapping the model's parameters into the RPU's conductances (see section 1.6). Therefore, millions/billions of resistive-switching devices must be integrated in the Back-End-Of-The-Line of a small chip (with a maximum area of approximately 800 mm^2 [38, 44]). This sets a geometrical constraint on the ReRAM cells' area: in a crossbar array architecture with a 4 F^2 density (where F is the minimum lithographic feature, equal to the device lateral size), an 800 mm^2 chip can host up to 5 billions devices of area 200 nm × 200 nm. In this chapter, we will demonstrate highly accurate ReRAM cell scaling down to 200 nm × 200 nm.

2. **Forming voltage scaling**: Most ReRAM devices, including our oxide *bilayer*, require to limit the current during the filament forming operation, as discussed in Section 3.3. In silicon integrated systems, the smallest established component for independent current compliance is a transistor, such as a Metal-Oxide-Semiconductor-Field-Effect-Transistor (MOSFET). Thus, when discussing technology scalability, we must consider the footprint of a MOSFET device compatible with our ReRAM specifics. The ground rule for MOSFET miniaturization involves reducing both transistor's length and operating voltages linearly [20]. Consequently, recent transistor technologies introduced in the last 20 years operate with a maximum voltage supply approximately $\lesssim$ 1.2 V [165]. However, our ReRAM devices require voltage pulses with an amplitude higher than >5 V to initiate the forming process, which would damage the transistor when applied as V_{DS}.
   Therefore, reducing the ReRAM forming voltage ($V_{forming}$) is a critical challenge for effective technology scalability.

3. **Symmetry point properties**: In the framework of the *Tiki-Taka* algorithm, one RPU array must be initialized through the *zero-shifting* technique (see Section 1.7), driving the ReRAM

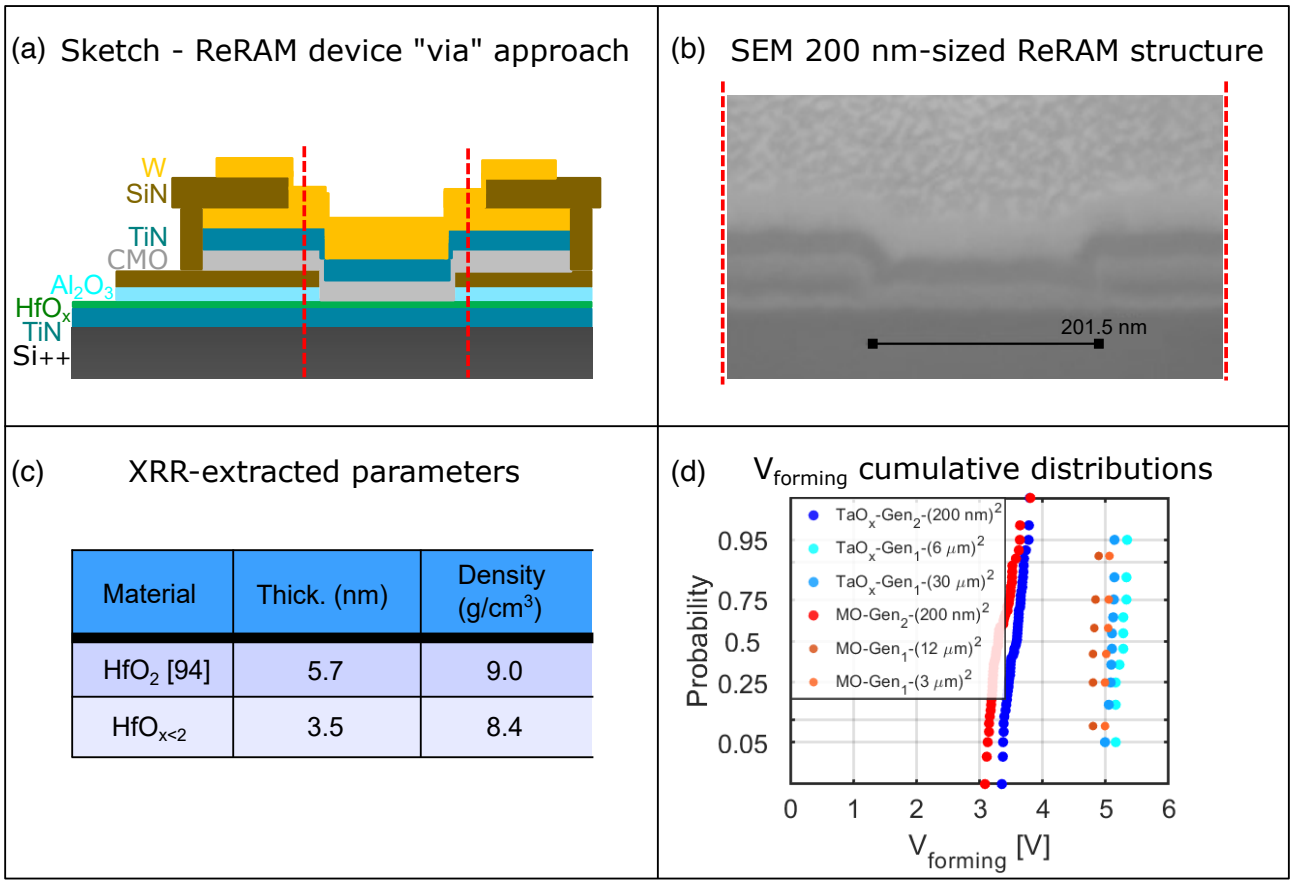
devices to their conductive state known as *symmetry point*. Consequently, the ReRAM devices must be tested under a programming scheme assessing the resistive-switching properties around symmetry point. We will define new Figures of Merit (FoM) to benchmark these properties.

4. **Retention and endurance measurements**: In deep learning, some synaptic connections undergo frequent updates during the training phase, while other parameters may remain idle for long times, even a few days. Therefore, we must check for drift issues in ReRAM devices. To accelerate any ionic diffusion/relaxation process that may alter the conductance states, I tested their retention at 85° C. Additionally, an RPU designed for numerous training tasks should provide sufficient endurance to the programming pulses. I performed endurance measurements to check the stability of the analog response and the symmetry point properties.

5. **Deep learning simulations**: Finally, we modeled the pulse response of our optimized ReRAM devices using a suitable mathematical model accounting for device-to-device variability, improving simulation reliability. A specialized deep learning simulator toolkit uses the device model to compare the training process with an analog RPU system to conventional hardware with digital numeric representation. We performed simulations on complex networks to show real-world applications and benchmarked the novel Tiki-Taka algorithm to highlight the importance of both algorithm and device optimizations.

4.1 Via process and optimized HfO_x layer

In this section, I will present a new batch of ReRAM devices with further optimized materials and processes compared to the structures presented in Section 2.1, **Fig. 2.3 (c) and (d)**. The new devices will be denoted as "*Generation 2*" (Gen_2), while the older ones will be labeled as "*Generation 1*" (Gen_1). A structural representation of the Gen_2-ReRAM devices is depicted in **Figure 4.1 (a)**. The active layers are the Conductive-Metal-Oxide (CMO) (either TaO_x or the "MO"), and the HfO_x, which provide the analog bidirectional switching properties discussed in the previous Chapter 3, Section 3.3. Scanning electron microscopy (SEM) was used to image the cross-section of a $(200\ nm)^2$ RRAM device, as depicted in **Figure 4.1 (b)**, showing accurate control on the unit cell area.
High filament forming voltages (>5 V) were measured in Gen_1-devices [94]. To decrease these voltages, we modified the Plasma-Enhanced Atomic Layer Deposition (PE-ALD) of the HfO_x layer. As discussed in Chapter 2, Section 2.1, the O-plasma duration is an important parameter affecting the material properties of the deposited thin film [146]. Specifically, an insufficient plasma duration may hinder the reaction of the oxygen radicals with the intermediate functional groups, leaving impurities in the deposited layer [146]. This is typically measured by a reduction of the material density, and/or by an increase of the leakage current density [116, 166, 167].
For the fabrication of the Gen_2-ReRAM devices, we reduced the O_2 plasma time from 10 s to 1 s. This led to a reduction of the film density, from 9 to 8.4 g/cm^3, measured by X-Ray Reflectivity (XRR) characterization (see **Figure 4.1 (c)**). Moreover, the number of ALD cycles was reduced from 50 to 34 and the film thickness decreased from 5.7 to 3.5 nm, as confirmed by the XRR measurement (see **Figure 4.1 (c)**). Combining the decrease of the HfO_x thickness and density, we achieve a strong reduction of the $V_{forming}$ values, as depicted in **Figure 4.1 (d)**. In this plot we compare the Cumulative-Distribution-Functions (CDF) of the Gen_1- and Gen_2-ReRAM devices, for varying CMO materials and areas. In the Gen_1-chip, we observed that devices with smaller areas exhibit slightly higher forming voltages [94]. This can be explained by considering the filament forming as a stochastic process, involving percolative paths across the metal-oxide stack [164].
Even if the Gen_2-devices are $\sim$ 1000 times smaller, they show much lower forming voltages (around



Material	Thick. (nm)	Density (g/cm^3)
HfO_2 [94]	5.7	9.0
$HfO_{x<2}$	3.5	8.4

Figure 4.1: **(a)** Sketch of the RRAM cross-section profile. **(b)** SEM cross-section image of a (200 nm)^2 Gen_2-RRAM. **(c)** Comparison between the thickness and density of the HfO_2 and the $HfO_{x<2}$ thin films used in Gen_1- and Gen_2-RRAMs, respectively. **(d)** Cumulative distributions of the forming voltages from Gen_1- and Gen_2-devices, in orange and green, respectively. Various device areas are compared. Modified from [111].

30% less) compared to the Gen_1. We attribute this result to the modified HfO_x properties.
The process flow used for the fabrication of both Gen_1- and Gen_2-devices is CMOS-compatible, and it avoids any lift-off steps, which are not feasible in foundries.

4.2 DC electrical properties of the scaled ReRAM devices

Figure 4.2 (a) and (b) show the DC I-V characteristics of 200 nm × 200 nm TaO_x- and MO-ReRAM devices, respectively. Both structures exhibit low-stochastic and gradual resistive-switching mechanisms, similar to the *Gen_1*-devices presented in Chapter 3. The different switching polarity observed in the MO-ReRAM device is due to a forming voltage pulse with negative polarity (not shown).
Other differences are the resistive window and the operating voltages in the MO-ReRAM devices (compare **Figure 4.2 (b)** and **Figure 3.9 (b)**). The newer devices operate at approximately one order of magnitude lower currents and require around 20% lower voltages to achieve an on/off ratio greater than 10. These current and voltage properties are crucial for deep learning applications, because the power consumed for programming and reading scales linearly with these parameters.

By increasing the $|V_{stop,reset}|$ value up to 1.8 V, we can reset the MO-device to even higher resistances. **Figure 4.2 (c)** shows a reproducible 3-bit programmability in the resistive window (200 kΩ, 5 MΩ). Such high resistances could not be achieved in Gen_1-MO-devices, which exhibited a maximum HRS value of approximately 10 kΩ (see **Figure 3.9 (c)**). Moreover, stressing these devices beyond a $V_{stop,reset}$ of approximately 2 V would result in a permanent short circuit (not shown).
We attribute the higher resistances of Gen_2-devices to reduced filament forming currents. The peak current value during this process can be estimated as: $V_{forming}/R_{pf}$ [151], where R_{pf} is the "post-forming resistance", the resistive state measured after the filament formation is completed. These devices exhibit approximately 30% lower forming voltages (see **Figure 4.1 (d)**) and R_{pf} values in the MΩ range, which are about 1 order of magnitude higher compared to the R_{pf} of Gen_1 devices (see **Figure 3.7 (b)**). Consequently, we expect lower filament forming currents, enabling to operate the device between the resistive states ranging from 10 kΩ to 5 MΩ.

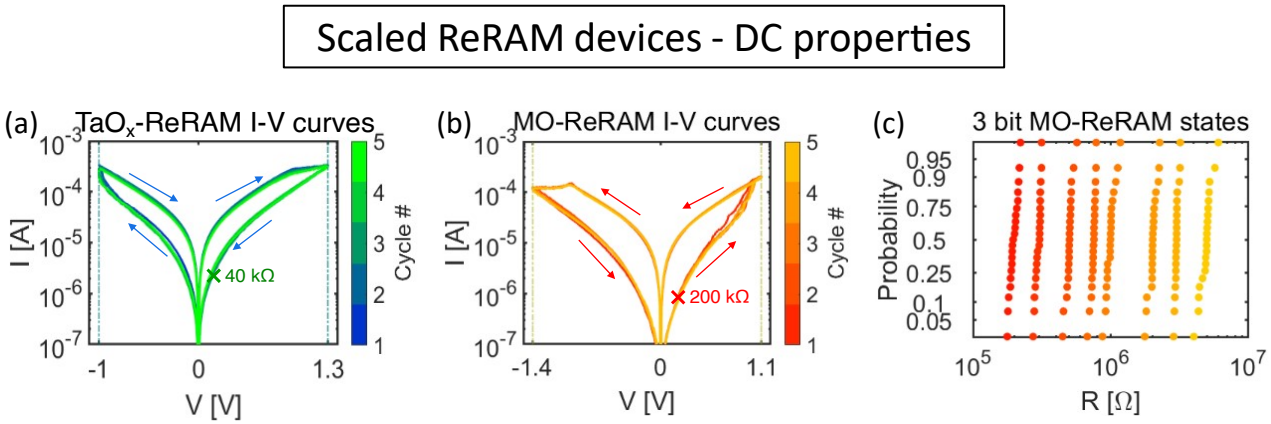


Figure 4.2: **(a)** and **(b)** DC I-V characteristics of the scaled TaO_x- and MO-ReRAM devices, respectively. **(c)** Reproducible 3-bit programmability for an MO-ReRAM device. To achieve these data, we introduced current compliances ranging from 3 μA to 50 μA, improving the programming accuracy.

4.3 Symmetry point properties

A new procedure for testing the analog properties of the optimized ReRAM devices is described in **Figure 4.3 (a)**. This scheme involves concatenating a sequence of 1000 cycles of "1 pulse up - 1 pulse down" to an initial pulse stream of 5 cycles of "200 pulses up - 200 pulses down". The first 2000 pulses measure the entire conductive window and evaluate its stability. The next 2000 pulses implement the *zero-shifting* procedure required by the *'Tiki-Taka'* algorithm (see Chapter 1, Section 1.7). The *zero-shifting* technique forces a resistive-switching device to its *symmetry point*, where updates in both directions are equal [108].
Figure 4.3 (b) and **Figure 4.3 (c)** show the pulse response of TaO_x- and MO-ReRAM devices, respectively, using identical pulsing paramters (V_{set} = 1.5 V, V_{reset} = 1.7 V, t_{pulse} = 1 μs). Both devices exhibit a stable conductive (G) window with many analog intermediate states, followed by the *symmetry point*.

When evaluating a resistive-switching device for the Tiki-Taka algorithm, several key parameters must be considered. To strengthen our benchmark, we collected data from 20 additional devices for both materials.

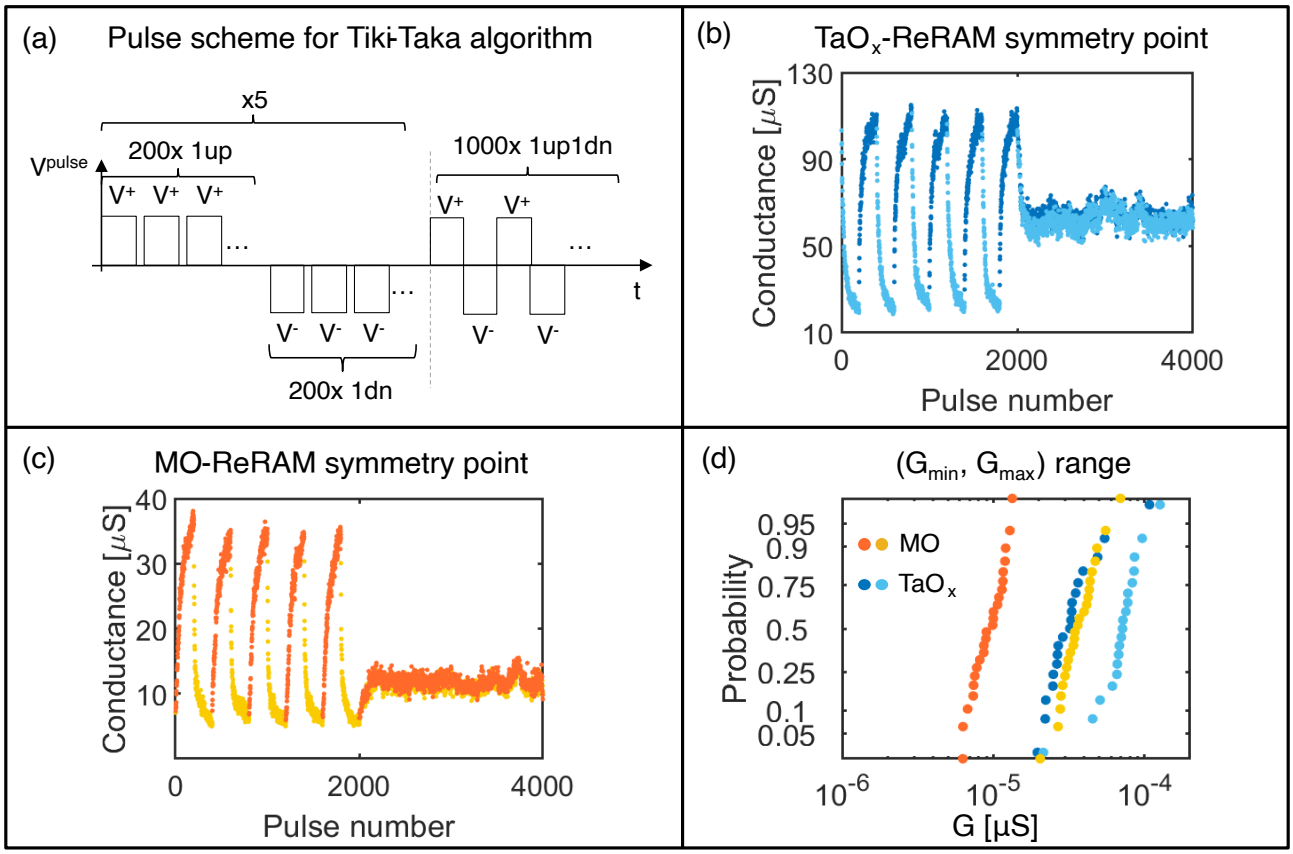


Figure 4.3: **(a)** Pulse scheme used for programming the ReRAM devices through their entire conductive swing and for implementing the zero-shifting procedure. **(b)** Full analog swing and symmetry point of the TaO_x-ReRAM devices. **(c)** Full analog swing and symmetry point of the MO-ReRAM devices. **(d)** Comparison of the conductive states (G_{min}, G_{max}) between the TaO_x- and the MO-ReRAM devices. Modified from [111].

- The first important parameter is the conductive window (G_{min}, G_{max}). Low conductances are required for power-efficient AI model training [108]. **Figure 4.3 (d)** illustrates the (G_{min}, G_{max}) values measured for the two ReRAM materials, plotted as cumulative distribution functions. The MO-ReRAM exhibits lower (G_{min}, G_{max}) values by an approximate factor of 2.5×.

- Another important device parameter is the *number of states*, defined by the equation:

$$N_{states} = \frac{G_{max} - G_{min}}{\overline{\Delta G_{SP}}} \tag{4.1}$$

  Where G_{max} and G_{min} are the maximum and minimum G values extracted from the full G swings, and $\overline{\Delta G_{SP}}$ is the mean value of the conductance updates during the "1 pulse up - 1 pulse down" procedure. N_{states} describes the resolution of gradient accumulation, a critical property for training accuracy. A graphical description of this function is shown in **Figure 3.3 (a)**. To account for cycle-to-cycle variability, we calculate G_{max} and G_{min} values by averaging over 10 full G swings.

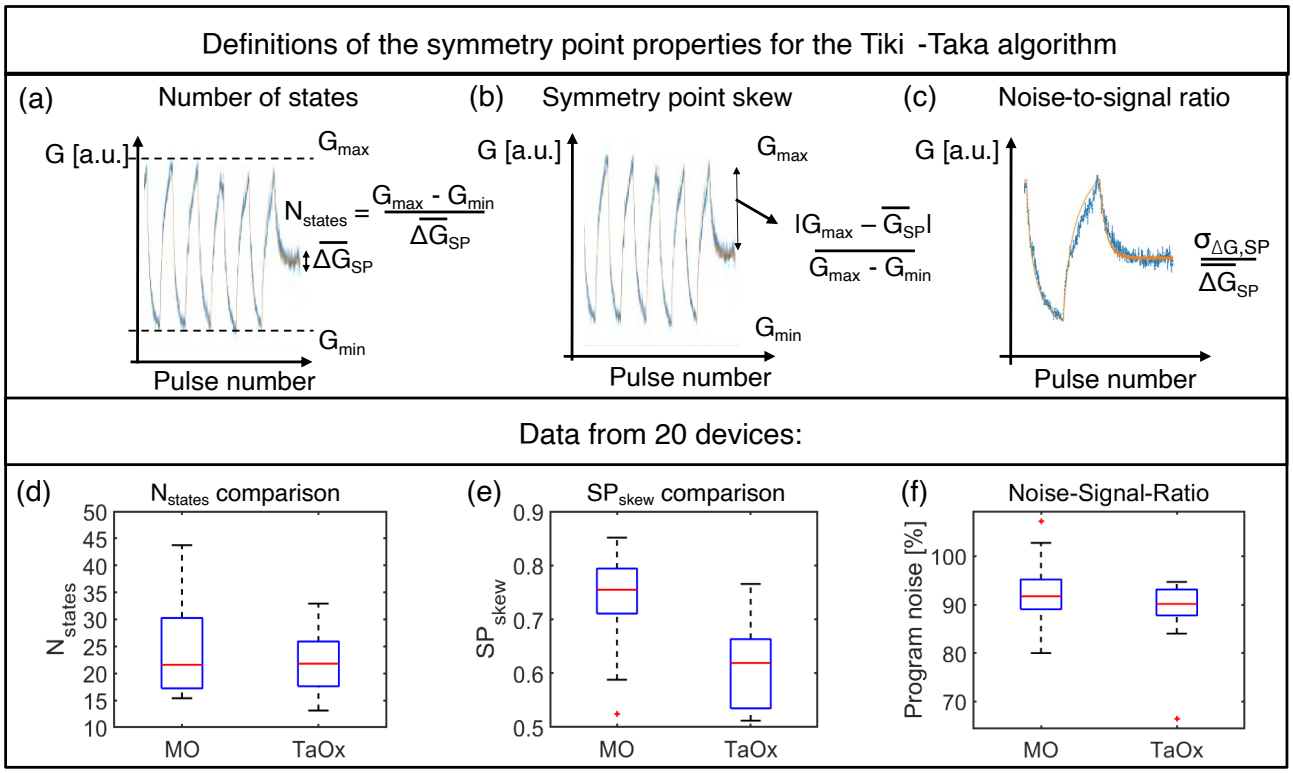


Figure 4.4: **(a)** Definition of the *number of states*, based on the presence of a *symmetry point* within the G swing. **(b)** Definition of the *symmetry point skew*, whose ideal value is 50%, enabling maximum gradient accumulation in both directions. **(c)** Definition of the *noise-to-signal ratio*, indicating the capability to measure minimal changes in the device state. **(d)**, **(e)** and **(f)** Boxplots comparing the *number of states*,*symmetry point skew* and *noise-to-signal ratio*, respectively, between the TaO_x- and MO-ReRAM devices. Modified from [111].

- The *Symmetry Point skew* (SP_{skew}) measures the position of the symmetry point compared to the full G swing. It is defined by the equation:

$$SP_{skew} = \frac{G_{max} - \overline{G_{SP}}}{G_{max} - G_{min}} \tag{4.2}$$

  Where $\overline{G_{SP}}$ is the average G state measured during the "1 pulse up - 1 pulse down" procedure. The optimal *SP skew* is 50%, while higher or lower values indicate a drift towards G_{min} or G_{max}, respectively. An illustration of the SP_{skew} parameter is depicted in **Figure 4.4 (b)**. In extreme cases, an SP_{skew} of 0 or 1 would prevent the neural network from learning because $G_{SP} = G_{max}$ or $G_{SP} = G_{min}$, making it impossible to accumulate positive or negative gradients. For example, the symmetry point of the MO-ReRAM device in **Figure 4.3 (c)** is skewed towards G_{min}, reducing the resolution for negative updates.

- The last important parameter is the *Noise-to-Signal Ratio (NSR)*, which describes the capability to distinguish clear updates from noise. It is defined by the equation:

$$NSR = \frac{\sigma_{\Delta G,SP}}{\overline{\Delta G_{SP}}} \tag{4.3}$$

  Where $\sigma_{G,SP}$ is the standard deviation of the G updates during the "1 pulse up - 1 pulse down" procedure. An *NSR* <1 is required for a successful execution of the *'Tiki-Taka'* algorithm,

indicating the ability to discriminate between positive and negative updates around the *SP*. A graphical explanation of this property is shown in **Figure 4.4 (c)**. The results for these three Figures of Merit (FoM) are reported as boxplots in **Figure 4.4 (d), (e) and (f)** for both the TaO_x- and MO-ReRAM devices.

Overall, we observe similar results for the both materials in terms of N_{states} and *NSR*. The median value for N_{states} is above 20 for both ReRAM types (the minimum threshold for simple deep learning applications is 15 [112]). Regarding the *NSR*, only a few MO-outliers exhibit values beyond 100%, while, generally, our ReRAM devices show clear updates.
The only FoM where the TaO_x devices clearly performs better than the MO-ReRAMs is the SP_{skew}. This can be explained by the more asymmetric (V_{set}, V_{reset}) voltages used for the MO-devices, which enabled to achieve very low G_{min} values. Consequently, we observe an SP_{skew} towards G_{min} in MO-ReRAMs.

4.4 Analog resistive-switching endurance

Device endurance to programming pulses is crucial for real-world applications. As discussed in Chapter 1, Section 1.4, training a neural network can involve millions of updates for single synaptic weights. Therefore, it is essential that the resistive-switching devices maintain stable analog properties over repeated programming operations.
We investigated the endurance to short programming pulses in both the TaO_x- and the MO-ReRAM devices. **Figure 4.5** presents a comprehensive endurance test conducted on an MO-ReRAM device. Initially, the device undergoes the 10×(200 up - 200 down) + 1000× (1 up - 1 down) pulse scheme described in **Figure 4.4 (a)**. The results, shown in **Figure 4.5 (a)**, reveal Long-Term-Potentiation (LTP) and Long-Term-Depression (LTD) curves, defining the device's conductive swing, followed by a centered symmetry point.
Next, we applied alternating (1 up - 1 down) pulses for 10^5 times, followed by repeating the LTP + LTD + Zero Shifting (ZS) programming scheme. The results in **Figure 4.5 (b)** indicate that both the G swing and the symmetry point shifted towards lower conductances, suggesting that more programming cycles after the initial 6000 were needed to stabilize the device's response to a stable conductive swing. The symmetry point remained stable within the first 2000 pulses of **Figure 4.5 (b)**, at values between 10 and 15 μS. We then continued applying (1 up - 1 down) pulses for 10^6 cycles, monitoring the symmetry point position every 10^5 pulses. The results in **Figure 4.5 (c)** show that within the first 10^5 pulses, the symmetry point drifted towards a low conductance state, below the G_{min} of the full G swing. If the $G_{SP} \leq G_{min}$, the resistive-switching device cannot be used for negative gradient accumulation, necessitating the recovery of a more centered symmetry point.
The recovery method consists of applying 10^4 consecutive pulses in the "up" direction only. The result, shown in **Figure 4.5 (d)**, indicates that the device did not react to the first 4000 pulses, and only after this point did Long Term Potentiation (LTP) begin. Subsequently, the conductance trace saturates to a G_{max} value close to the initial $G_{max,fresh}$ (compare with **Figure 4.5 (a)**).

We performed a similar endurance test on a TaO_x-ReRAM device. This time, we did not include the Zero-Shifting (ZS) procedure in the pulsing scheme because we observed instabilities already from the repeated LTP and LTD cycles.
Figure 4.6 (a) depicts the main results from the entire endurance test. We observe a stable G swing between approximately 30 and 200 μS up to more than 10^6 pulses. However, starting from few tens of thousands of pulses, we noticed a collapse of the conductive window, mainly determined by the decrease of the G_{max} value (not shown). This effect is also clearly visible when monitoring the (G_{min},

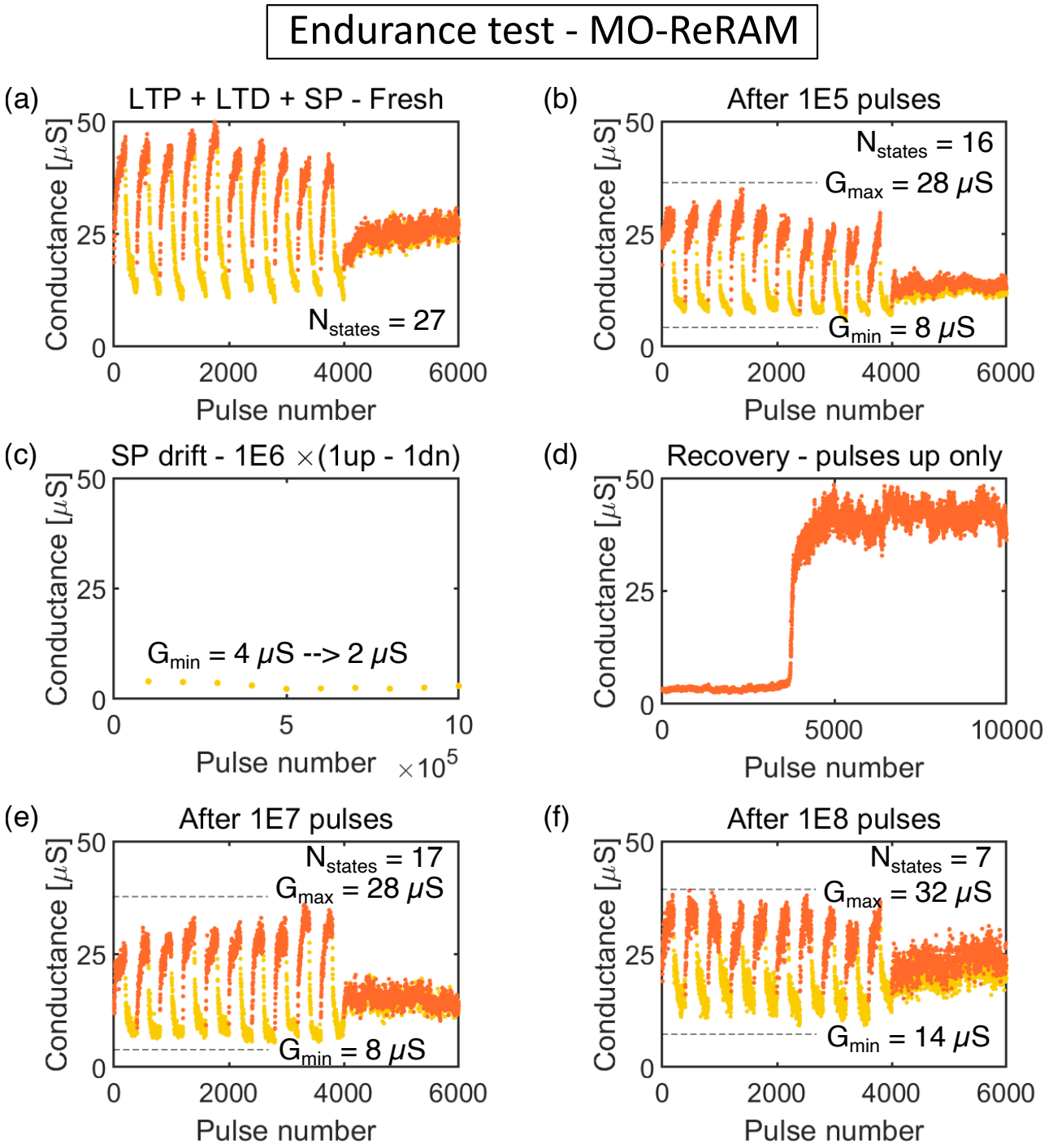


Figure 4.5: **(a)** MO-ReRAM Long-Term-Potentiation (LTP) + Long-Term-Depression (LTD) + Symmetry Point (SP) traces right after the filament forming process (fresh device). **(b)** Same test after having applied $10^5\times$ (1 up - 1 down) pulses. **(c)** Evolution of the symmetry point during $10^6\times$ cycles of (1 up - 1 down) pulses. **(c)** Recovery of the full conductive swing under the application of "only-up" pulses. **(e)** LTP + LTD + ZS after $10^7\times$ cycles of (1 up - 1 down) pulses. **(f)** Same test after $10^8\times$ cycles of (1 up - 1 down) pulses.

G_{max}) values for up to 10^6 programming pulses, as shown in **Figure 4.6 (b)**. The conductive window gradually reduces, leading to a reduction of the N_{states} parameter.
To compensate for this effect, we modified the programming parameters by increasing the applied voltages in both polarities in steps of 100 mV after approximately 10^5 and 10^6 pulses, as indicated in **Figure 4.6 (a)**.

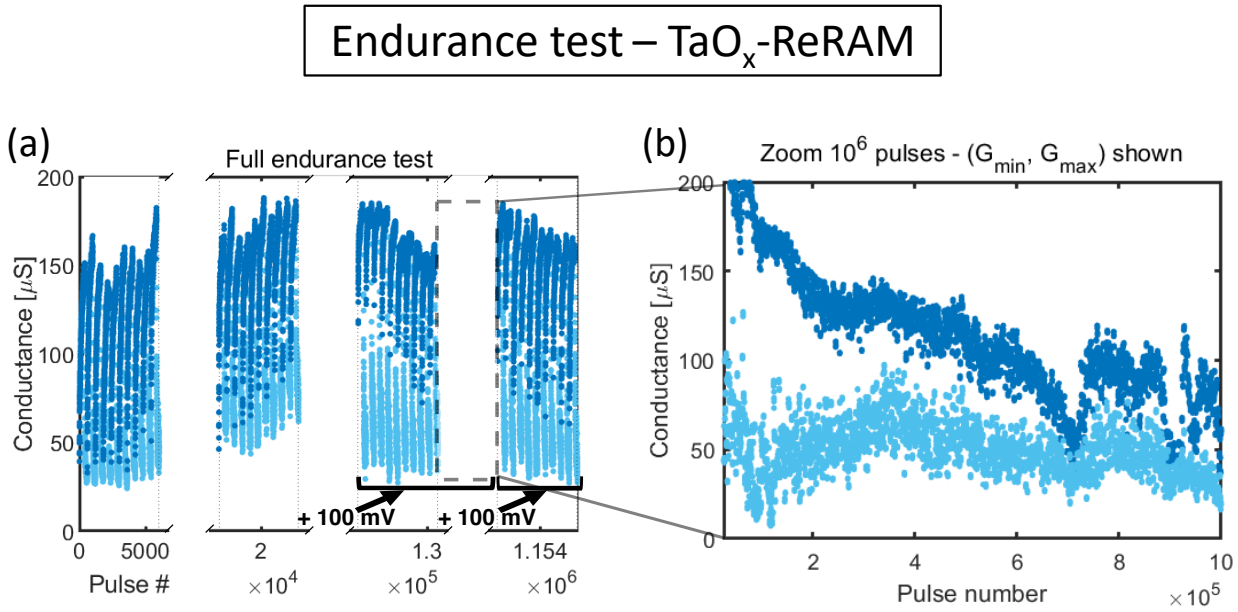


Figure 4.6: **(a)** TaO_x-ReRAM analog response to more than 10^6 programming pulses. Between 10^4 and 10^5 pulses, the conductive window reduces (not shown), requiring a 100 mV increase in voltage amplitude for both polarities to restore a full swing. The same reduction occurs between 10^5 and 10^6 pulses, shown in **(b)**, necessitating a further 100 mV voltage increase. Gradual increases in voltage parameters enable maintaining stable (G_{min}, G_{max}) values and a stable symmetry point up to 10^6 pulses. Modified from [111].

4.5 Retention of the analog conductive states

Retention of the analog conductive states is a crucial requirement for deep learning applications. As discussed in Chapter 1, Section 1.6, training tasks require several iterations involving the processing of large datasets. Consequently, training times can span from a few minutes to several hours [168], depending on factors such as the size of the AI model, the size of the dataset, and the hyperparameters used. In a context where resistive-switching elements map the synaptic weights of a neural network, the temporal stability of the conductive states is essential for enhancing learning accuracy.
We monitored the evolution of the conductive states from TaO_x- and MO-ReRAM device. **Figure 4.7 (a) and (b)** show the results for 32 states (5-bit programmability) at a 1 Hz frequency over a 100-seconds time interval for both device types. Only minor random fluctuations were observed, generally occurring on the timescale of few seconds. These sudden "jumps" can be attributed to Random-Telegraph-Noise, which will be specifically addressed in future work.
Studying the stability of the analog states within the 1 to 100-second interval is insufficient: for deep learning acceleration, the devices must operate at high frequency, ideally in the MHz-GHz regimes, requiring a monitoring on shorter timescales. Such measurements are crucial for optimized ReRAM

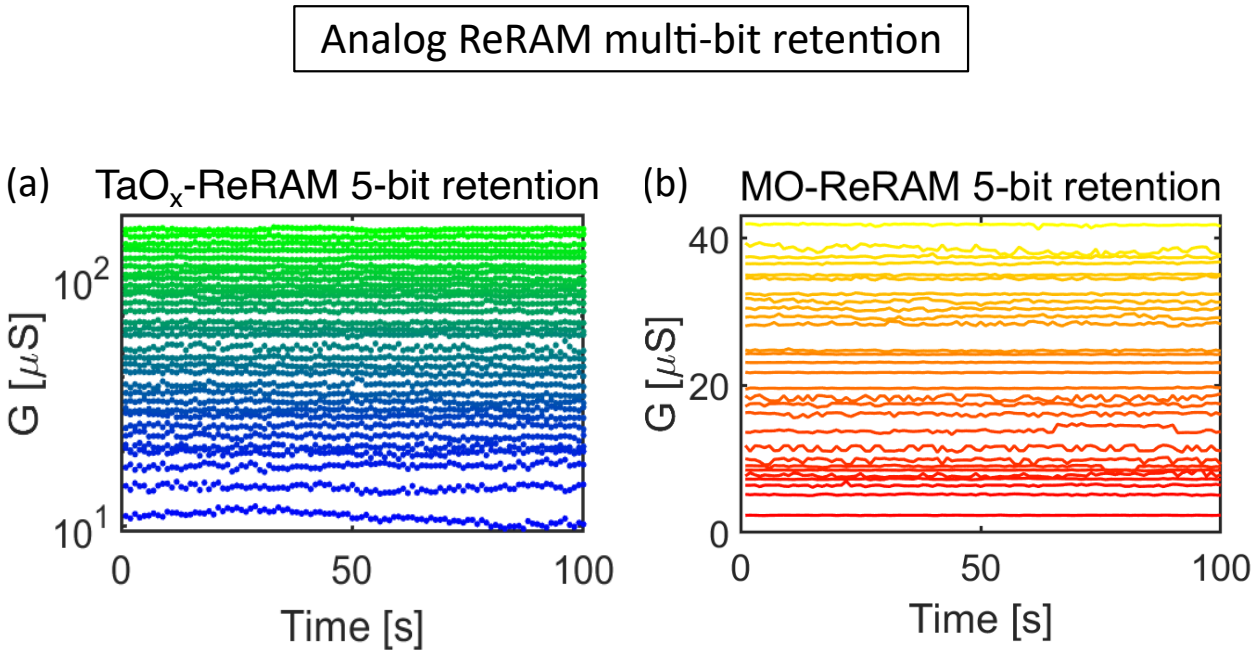


Figure 4.7: **(a)** 100 seconds retention for 5-bit conductive states in TaO_x-ReRAM. **(b)** 100 seconds retention for 5-bit conductive states in MO_x-ReRAM. Modified from [111].

devices based on the oxide *bilayer* structure, which have previously been reported to suffer significant drift issues (up to 50%) within one second after programming [169]. This phenomenon, known as the *relaxation* of the oxygen vacancies, needs to be addressed.
Additionally, synaptic weights may remain idle for longer than 100 seconds in some cases. Therefore, evaluating retention over extended periods is also necessary. Moreover, resistive-switching arrays will be embedded in the BEOL of high-frequency processors, which may experience temperature rises due to high power consumption. Hence, the devices should operate robustly even at high temperatures.
To address these multiple issues, we performed several dedicated measurements. First, we investigated long-time retention for up to 72 hours at a high temperature of 85 ° C. Elevated temperatures generally promote oxygen ion mobility [170], providing a more challenging test for the stability of the programmed states. The results, presented in **Figure 4.8 (a)**, show very stable states under high temperature conditions. Only the low conductive state (G_{min}) exhibited a minor drift ($< 4\%$) towards lower conductances within 72 hours.
Next, we investigated the stability of the programmed states over short timescales. We first applied a sequence of 2000 pulses to program the devices to high conductances (see **Figure 4.8 (b)**), or low conductances (**Figure 4.8 (e)**). Then, we monitored the evolution of the conductance for up to 2 minutes, sampling at high frequency (every 500 μs). **Figure 4.8 (c) and (f)** zoom in on a two-second temporal window, including data from the end of the programming cycle and the beginning of the retention measurement. Negligible *relaxation* was observed for both high and low conductances. To gain a more clear evidence about the absence of relaxation effects, **Figure 4.8 (d) and (g)** illustrate the data sampled in the last 10 ms of the programming cycle and the first 10 ms of the retention measurement. In this short temporal window, the states remained stable after the programming cycle was completed.

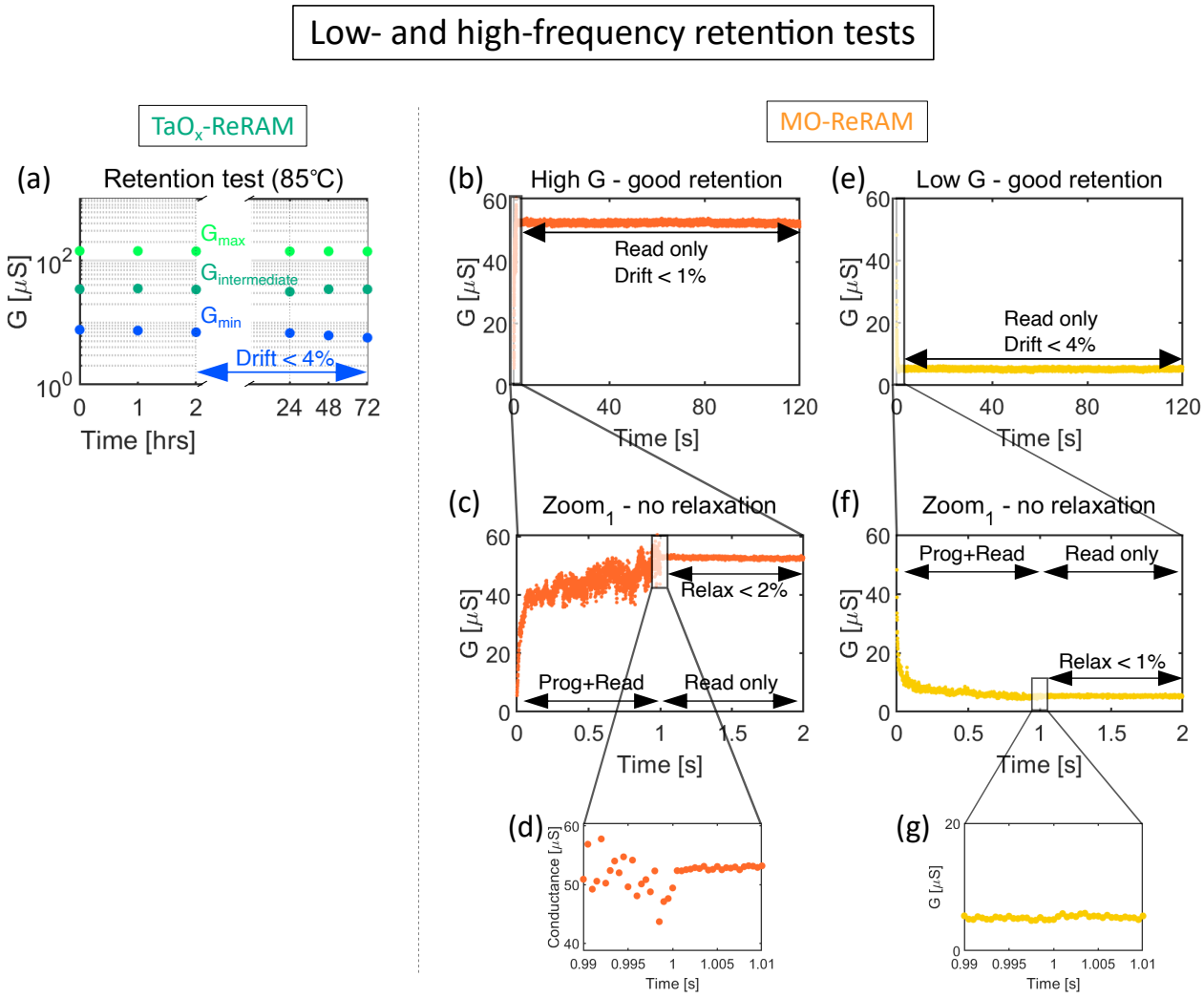


Figure 4.8: **(a)** Retention of the Gmin, G$_{intermediate}$ and G$_{max}$ for 72 hours at 85 ° C. **(b)** Long-Term-Potentiation (LTP) followed by high-frequency (2 kHz) tracking of the programmed high-conductive state. **(c)** Zoom into a two-seconds time interval including the tail of the LTP and the start of the retention measurement to track any potential high-frequency relaxation of the programmed state. **(d)** Further zoom in the 20 ms time interval including the tail of the LTP and the start of the retention measurement. **(e)** Long-Term-Depression (LTD) followed by high-frequency (2 kHz) tracking of the programmed low-conductive state. **(f)** Zoom into a two-seconds time interval including the tail of the LTD and the start of the retention measurement. **(g)** Further zoom into a 20 ms time interval including the tail of the LTD and the start of the retention measurement. Modified from [111].

4.6 Localized delamination of ReRAM capping layers corroborating filamentary hypothesis

In Chapter 3, Section 3.4, we discussed our interpretation of the analog resistive-switching mechanisms in our analog ReRAM devices. Area-dependent measurements (see **Figure 3.5**) confirmed the filamentary nature of these resistive-switching properties. However, a direct observation of the filament size/area would provide stronger evidence to support our model.

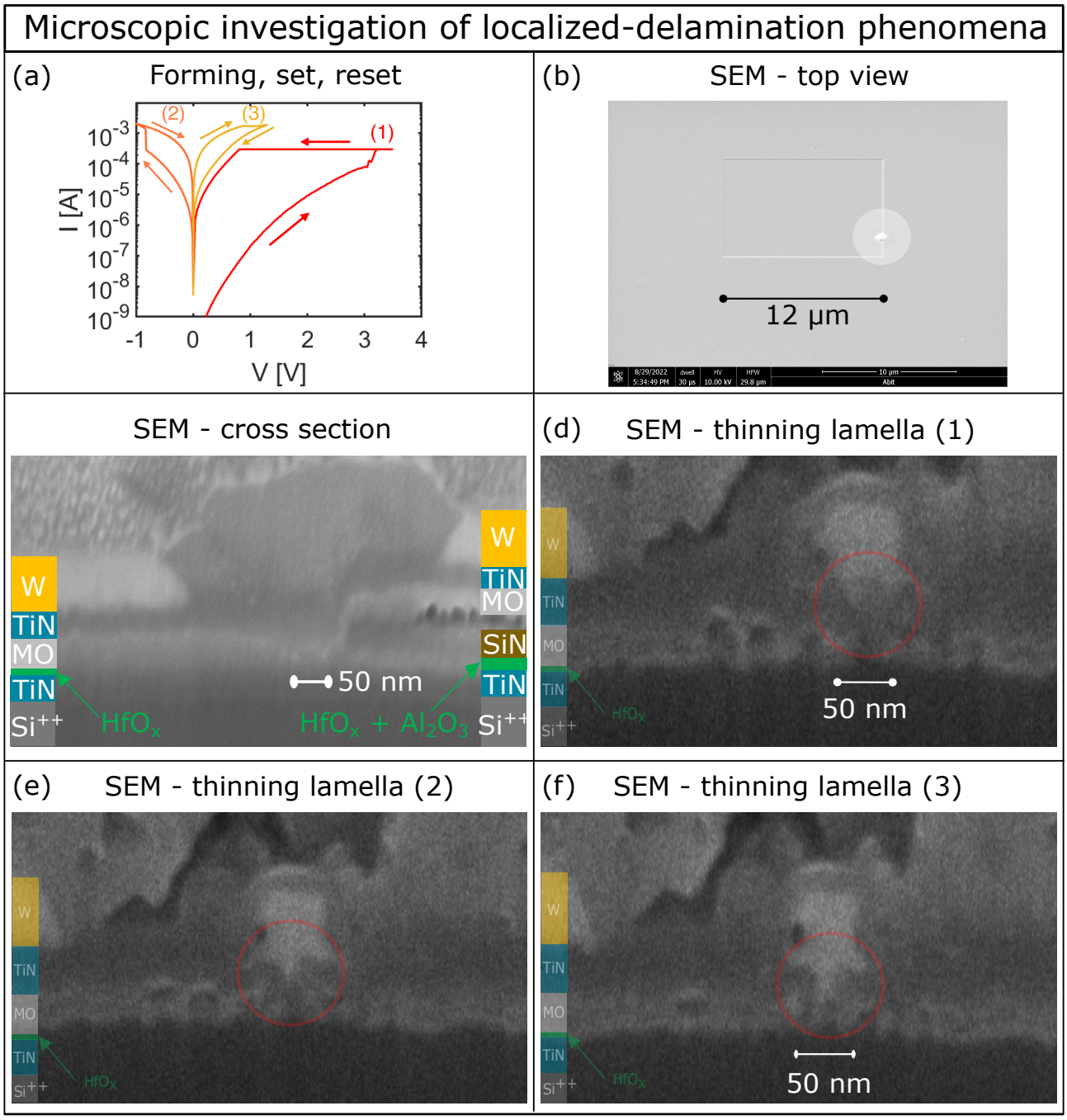


Figure 4.9: **(a)** Positive forming, set and reset voltage sweeps for an MO-ReRAM device of active area of $(12\ \mu m)^2$. Following these electrical measurements, the top surface of the device was investigated using Scanning-Electron Microscopy (SEM). **(b)** SEM top view of the device, showing a "bump" located at the *via* edge. **(c)** Cross-section of the ReRAM device imaged by SEM after Focused-Ion-Beam (FIB) cutting at the "bump" location. Structural modifications in the MO-material composition are highlighted. **(d)**, **(e)**, **(f)** SEM images of another ReRAM device during the lamella-thinning procedure. The images were sequentially taken while approaching the center of the "bump" area, highlighting a high-contrast "filament plug" that appears to connect the capping W layer to the MO material.

By applying positive filament-forming voltages to the MO-ReRAM devices, we observe reversible resistive-switching properties (set + reset transitions), as depicted in **Figure 4.9 (a)**. Compared to the case of negative forming polarity, presented in **Figure 4.2 (b)**), higher currents were measured for similar stop voltages, approximately 10× higher. This significantly impacted the structural properties of our ReRAM devices. Examination of the top surface of the device's active area revealed the formation of a localized "bump" at its edge, as illustrated in **Figure 4.9 (b)**.
We used Focused-Ion-Beam (FIB) and Scanning-Electron Microscopy (SEM) to cut and image a cross-section of the ReRAM device at the location of the "bump". **Figure 4.9 (c)** shows that the W layer capping the ReRAM layer stack (TiN + MO + HfO_x + TiN) delaminated. We attribute this effect to a combination of mechanical stress from DC sputtering and thermal stress induced by $\sim$ mA currents confined in $\sim$ nm-sized filaments.
Given the limited resolution of this investigation technique, significant quantitative analysis is not possible. However, examining the *via* edge, there is a region (approximately 50 nm in lateral size) in the MO layer where the material appears structurally modified, as indicated by a different color contrast compared to the rest. It is plausible that the filament sform at the edge of the *via* structure, a region of greater electric field enhancement.
To further explore this phenomenon, we prepared another ReRAM device from the same chip in a similar state by applying currents in the $\sim$ mA regime (not shown). We cut a lamella for Transmission-Electron-Microscopy (TEM) analysis. The results from the TEM analysis will come in a future work. During the lamella thinning process, we approached the center of the "bump", where we expected a filament to be localized. **Figure 4.9 (d), (e) and (f)** depict SEM images obtained while approaching the center, where a small "filament plug" (approximately 20 nm in lateral size) between the MO layer and the TiN top electrode appears. However, due to the limited contrast of the SEM technique at this magnification point, we cannot make additional statements about the nature of such "filament". Further high-resolution microscopy, such as TEM, and chemical analysis, such as Electron-Dispersive Spectroscopy (EDS) are required to better classify the material properties in this complex scenario.
Future work will also evaluate the reproducibility and reliability of the *via*-approach processing, where we observed such delamination issues for high ReRAM currents, approximately in the $\sim$ mA range.

4.7 Device response model and deep learning simulations

To estimate the performance of deep learning accelerators based on our ReRAM devices, we created mathematical models of their analog synaptic traces and conducted neural network training simulations. We utilized open-source models of resistive-switching devices available on the *'aihwkit'* [171] platform developed by IBM. Specifically, we employed the *'SoftBounds'* model to fit the analog device updates. The fitting provess involves minimizing the average deviation of a parametric function from the experimental data traces, using Powell's conjugate direction method [172] (lmfit toolbox [173]). An example of such a fit is displayed in **Figure 4.10 (a)**. We used analog traces from 20 devices per ReRAM type (TaO_x- and MO-) to account for device-to-device variability. Some examples are shown in **Figure 4.10 (b)**. To further enlarge our dataset and enhance robustness to statistical variations, we transitioned from the *SoftBounds* model, characterized by numerous variables and parameters (such as ΔG_{mean}, G_{min} and G_{max}), to a new device model defined by only two variables: *number of states* and *asymmetry*. For clarity, it is important to note that the variable *number of states* is different from the parameter N_{states} described in 4.3. The *number of states* is derived from the analytical *'SoftBounds'* model (we take the measured device trace, fit it using the *'SoftBounds'* model, and then use the model to derive the variable *number of states*), whereas N_{states} was directly derived from the measured G_{min}, G_{max}, ΔG_{SP} parameters (no model is involved in the calculation of N_{states}).

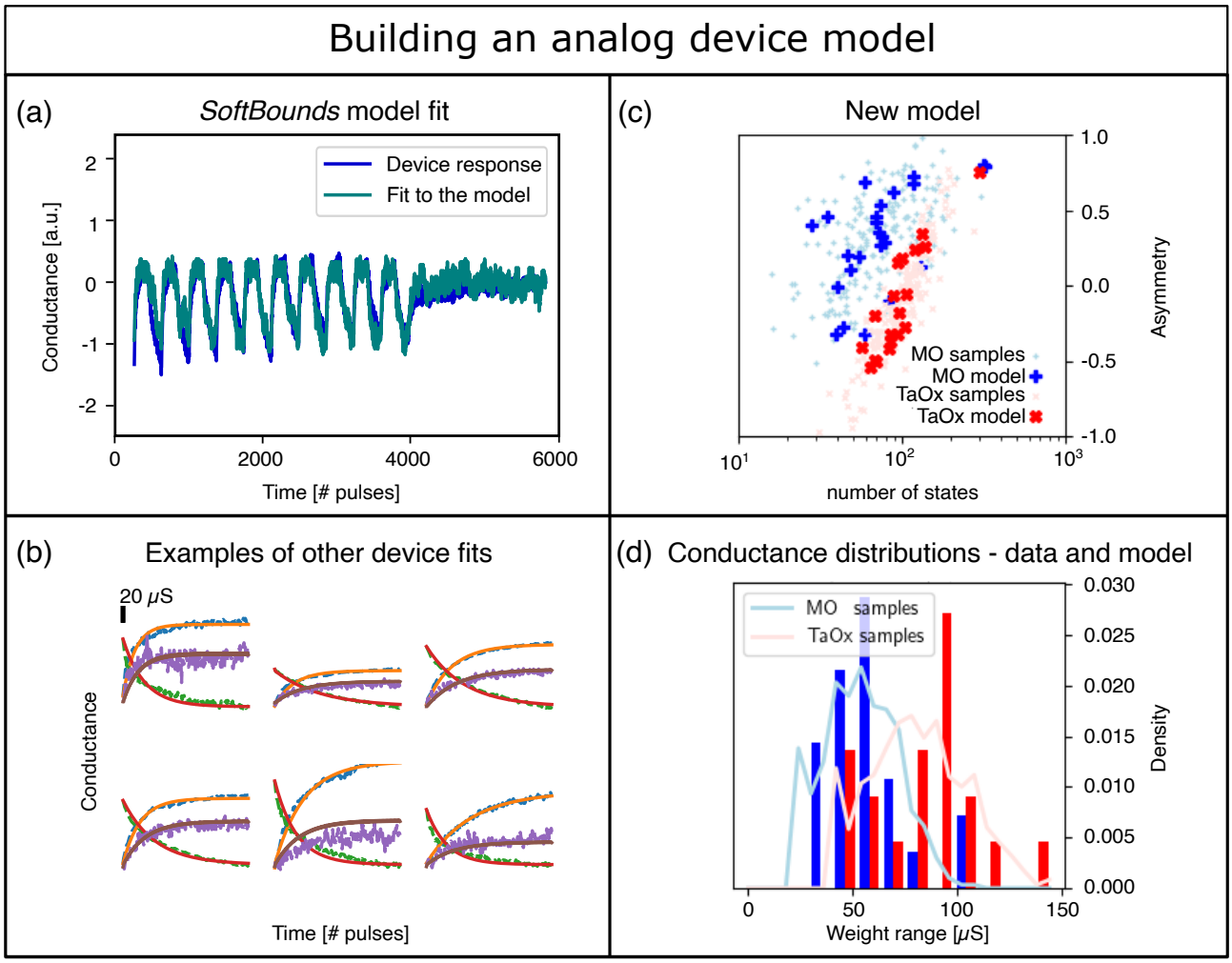


Figure 4.10: **(a)** *'SoftBounds'* model fitting the synaptic trace of an analog ReRAM device. **(b)** Collection of other device fits used to include variability in the device model. **(c)** 2D plot displaying the variables: *asymmetry* and *number of states*, derived from the *'SoftBounds'* models of the TaO_x- and MO-ReRAM devices, in red and blue, respectively. From these models, we created a multivariate Gaussian function, from which we sampled additional data for the deep learning simulations. **(d)** Distribution of the conductive states calculated from the *SoftBounds* models of the TaO_x- and MO-ReRAM devices, in red and blue, respectively. The conductive distributions obtained from sampling the multivariate Gaussian function for: *asymmetry* and for: *number of states* are also shown. Modified from [111].

We created a 2D Gaussian distribution based on the mean value and the covariance of the variables (*number of states*, asymmetry), from which we sampled additional data points to expand our dataset. The results are displayed in **Figure 4.10 (c)**.
We can summarize the advantages introduced by the new device model depicted as follows:

- **Realistic simulations:** Device-to-device variability was accounted for when generating the 2D Gaussian model, resulting in more realistic simulations compared to previous studies.
- **Reliable parameters:** Since the measured conductive window is obtained from a limited number of programming pulses per direction (200), it might be insufficient for the saturation of the

curves. The model, however, is based on the slope of the curves and thus returns more reliable parameters than the measurements.

Using this model, we simulated neural network training and compared a digital processor with 32-bit floating point (FP) numerical resolution to a resistive processor based on our analog devices. In the first simulation, we selected a simple network, a three-layer Fully-Connected Network (3-FCN), trained on the "Modified National Institute of Standards and Technology" (MNIST) dataset [174]. **Figure 4.11 (a)** compares the training errors between the digital and the analog processors. It is important to specify that the digital processor was trained using a conventional algorithm, Stochastic Gradient Descent (SGD), while for the analog processor, we used a novel version of the Tiki-Taka algorithm (*TTv2*, see Chapter 1, Section 1.7). The best training accuracy was achieved by our TaO_x-ReRAM devices (97.4%), approximately 1% better than the MO-ReRAM devices. After 100 training epochs, only a 0.7% difference in accuracy was observed between the digital and analog processors, demonstrating that analog signal processing also delivers accurate training of AI models. We extended the simulations

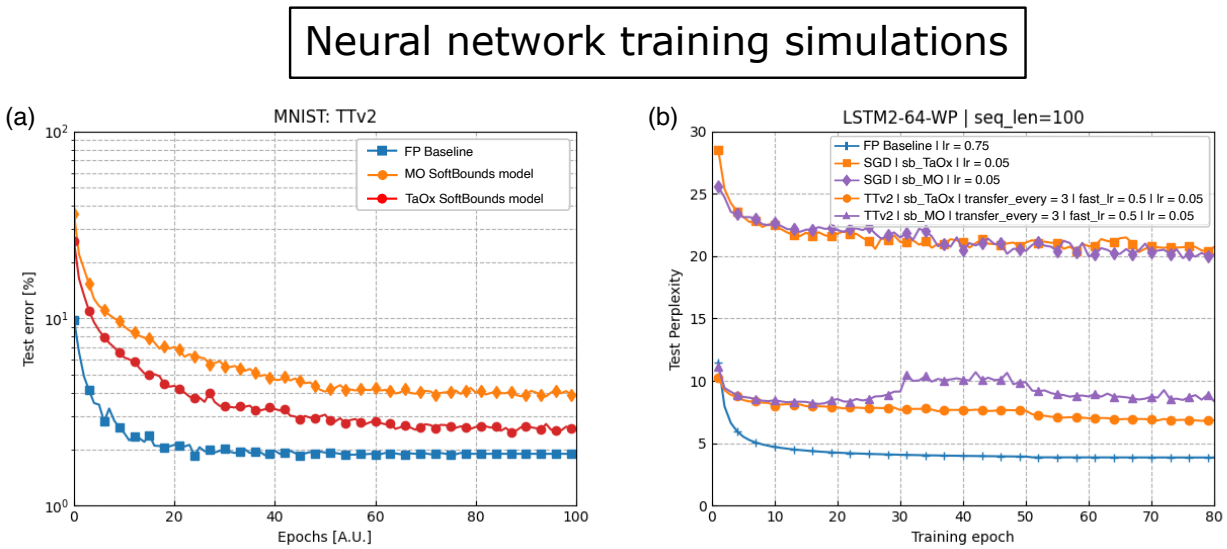


Figure 4.11: **(a)** Simple neural network training simulation. Network: 3-layer Fully-Connected-Network (FCN). Application: image classification using the MNIST dataset. Simulated processors: digital CPU using 32-bits Floating Point (FP) resolution (in blue) and a cascade of three Resistive-Processing-Units (RPUs) of analog ReRAM devices (in orange). The digital processor was trained with the SGD algorithm, while the analog processor used *TTv2*. **(b)** Training simulation for a larger network. Network type: 2-layer Long-Short-Term-Memory (LSTM) network with a hidden vector size of 64. Application: character prediction using text from *'War and Peace'* (*LSTM2-64-WP*). Input sequence length: 100 characters (*'seq_len'* = 100). Key parameters: for the SGD algorithm, the learning rate (*'lr'*) is equal to 0.75 for the digital processor (blue lines), and 0.05 for the analog processor (orange squares for TaO_x- and violet rhombuses for MO-ReRAMs). For *TTv2*: learning rate for matrix A (*'fast_lr'*) is 0.5, and learning rate for hidden matrix H (*'lr'*) is 0.05. Gradient accumulation undergoes 3 iterations before updating matrix H $transfer_{every} = 3$. More details on the *TTv2* algorithm can be found in [109], or in Chapter 1, Section 1.7. Modified from [111].

to a more complex network composed of two-stacked Long-Short-Term-Memory blocks with a hidden vector size of 64, followed by a fully-connected layer. This network was used for character prediction in sequences of words made of 100 characters, extracted from the *'War and Peace'* novel. The results are

shown in **Figure 4.11 (b)**, where the exponential of the cross-entropy loss, the test perplexity [175], is plotted over 80 training epochs. In this simulation, we included an additional comparison, benchmarking the *TTv2* algorithm against SGD for the analog devices as well. We observed that *TTv2* largely outperformed the SGD algorithm by more than 30%. Moreover, the TaO_x-based analog processor once again approached the performance of the digital reference. This result demonstrates that *TTv2* training with optimized ReRAM devices generalizes well to more complex AI tasks than just learning to classify the MNIST dataset.

4.8 Analog ReRAM benchmark

In this chapter, we have discussed the electrical properties of the scaled ReRAM devices, focusing on analog-switching aspects such as symmetry point, programming endurance, and retention of the analog states. Now, we will combine the collected measurements and make a benchmark for the targeted deep learning application.
Table 4.12 below compares the main properties of our ReRAM to literature. The results are depicted using a colormap with shades of green and red. Dark green indicates comparable or even better results than recent, State-Of-The-Art (SOTA) literature [92, 112, 130, 133], while red refers to results that are still far from the best technologies by more than 1 order of magnitude (when applicable). Lighter

	Gen_1-TaO_x-ReRAM	**Gen_1-MO-ReRAM**	**Gen_2-TaO_x-ReRAM**	**Gen_2-MO-ReRAM**
Area [µm²]	36	36	0.04	0.04
$V_{forming}$ [V]	5.6	5.1	3.6	3.3
(G_{min}, G_{max}) [µS]	(50, 250)	(200, 600)	(30, 70)	(10, 40)
(V_{set}, V_{reset}) [V]	(1, 1.1)	(1.4, 1.7)	(1.5, 1.7)	(1.5, 1.7)
I_{max} (= I_{reset}) [µA]	600	2000	300	100
($t_{pulse,prog}$) [µs]	0.2	0.2	1	1
N_{states} (min, max)	n.a.	(30, 40)	(17, 30)	(18, 26)
SP_{skew} (min, max) [%]	n.a.	(67, 73)	(53, 66)	(71, 79)
NSR (min, max) [%]	n.a.	(98, 128)	(88,93)	(89, 95)
Analog endurance (number of programming pulses)	n.a.	n.a.	$> 10^6$** $> 10^7$***	$> 10^7$* $> 10^8$***
Analog states retention	n.a.	n.a.	32 states, 100 s @ RT < 5% drift, 72 h @ 85°C	32 states, 100 s @ RT < 2 % relaxation after programming

Figure 4.12: Benchmark table for the analog ReRAM devices based on the oxide *bilayer* materials.

shades indicate intermediate results. We can summarize the comparison into three categories:

- **General device properties:** In the upper part of the table (from the spec: "Area" to: "$t_{pulse,prog}$") we listed the geometrical and electrical properties of broad interest, which are critical for many applications including storage, deep learning inference, and training. These properties were compared to the results published in *Nature* by Yao et al. [133] and by Wan et al. [130], since these works represent the SOTA technologies in the ReRAM field.
  We observe that our Gen_2-ReRAM devices achieve similar and even better properties in all fields,

except for the $t_{pulse,prog}$. However, in our experiments, the programming time was limited by the electrical setup, not the switching speed of the devices (see Chapter 2, Section 2.3 for more details). Using newer equipment we applied 50 ns long programming pulses (the same duration used in [130, 133]), leading to very similar electrical properties. These results will be presented at the next Device Research Conference (DRC) 2024.

- **Analog properties for deep learning acceleration with the Tiki-Taka algorithm:** In the central part of the table (from row label: "N_{states}" to row label: "NSR"), I summarized the analog properties of specific interest in the field of deep learning acceleration with the Tiki-Taka algorithm. We compared our work to the results published in the *International Electron Devices Meeting* by Gong et al. [112] and in *Scientific Reports* by Abedin et al. [92], which are the latest studies about ReRAM characterization for deep learning acceleration. Again, we observe that our Gen_2-ReRAM devices achieve similar/or better properties in all the fields under comparison.

- **Analog switching endurance and retention of the analog states:** In the lower part of the table, I reported the reliability aspects for the analog programming, which are critical for many applications including storage, deep learning inference and training.

  For the endurance of the analog programming with identical voltage pulses we could not find similar works to compare our devices; however, we know from simulations what are the targets for a reliable deep learning acceleration. Ideally, *TTv2* requires the devices to maintain a stable conductive window and a constant symmetry point (requirement labeled with one asterisk *) for **more than** 10^7 pulses, which is the limit demonstrated by our MO-devices 4.5. Thus, we could not label this cell in green. However, recent algorithm improvements relaxed the device specs compared to *TTv2* [107]. For this reason, and considering that the MO-ReRAM maintains analog properties under identical pulsing conditions for more than 10^8 updates (result denoted with three asterisk ***), we colored the cell in orange. Differently, the TaO_x-ReRAM required incremental voltages (a result that we indicate with two asterisks **) to mainatin stable analog properties up to just 10^6 pulses. Thus, we colored in red the TaO_x-cell: incremental voltage amplitudes depending on the fatigue of the array elements is incompatible with the whole RPU concept for deep learning acceleration (see Chapter 1, Section 1.6 for more details).

  The retention of the analog states, instead, is excellent in both ReRAM types. Compared to the oxide *bilayer* devices presented in Xu et al. at IEDM 2020 [169], showing >50% relaxation effects within 1 second after programming, our devices show a negligible drift of the programmed states (< 2%). The long-term retention is also excellent (< 4% drift at 85° C).

4.9 Conclusion

In conclusion, we demonstrated that our oxide *bilayer* ReRAM devices are scalable in terms of both geometry and electrical parameters, such as $V_{forming}$, G states, I_{write}.
We defined and discussed the analog properties of interest for the Tiki-Taka algorithm. We also characterized the reliability of the analog properties, including endurance measurements and retention measurements at different timescales.
Compared to other technologies discussed in recent high-impact literature, our devices show comparable/better general properties. This validation encouraged us to advance our research work by integrating the scaled ReRAM devices in the Back-End-Of-Line (BEOL) of a foundry-processed chip, comprising Complementary-Metal-Oxide-Semiconductor (CMOS) circuits. Together with our partner

collaborator, professor B. Barranco-Linares from the “Instituto de Microelectronica de Sevilla” (IMSE), we realized a flexible chip design enabling to advance our research on device development, as well as demonstrating real-world applications with crossbar array architectures.
The next chapter is devoted to discussing the CMOS integration and the results from the crossbar array programming.

Chapter 5

Back-End-Of-Line (BEOL) integration and crossbar array measurements

In this chapter, I will present the BEOL integration of our optimized ReRAM devices into a silicon chip comprising Complementary-Metal-Oxide-Semiconductor (CMOS) circuits. The chip technology is 350 μm and the design includes stand-alone 1-Transistor-1-Resistor (1T1R) structures and crossbar arrays. We optimized the electrical parameters for the measurement of 1T1R devices, leading to high-yield and high-performing resistive-switching properties. These optimized parameters enabled us to demonstrate analog properties and a uniform conductive window within a crossbar array of size 5×4.
We also demonstrated parallel crossbar updates through coinciding pulses applied at the array terminals, as required for the acceleration of deep learning tasks with a resistive processing unit.

5.1 BEOL integration of ReRAM devices

The integration of our ReRAM technology into the BEOL of a silicon chip comprising CMOS circuits was a complex project that required diverse design competences and advanced fabrication capabilities. This shared research effort was led by myself and my colleague Dr. A. La Porta from IBM, with support from prof. B. Barranco-Linares' team at the "Instituto de Microelectronica de Sevilla" (IMSE) and the foundry "*X-FAB*".
In the following section, I will outline the motivations behind this effort, provide a general overview of the chip design properties and the ReRAM integration process, and detail my specific contributions to such project.

5.1.1 Motivation

The primary motivation for integrating ReRAM devices onto CMOS circuits is to address the practical requirements for real-world applications. Our resistive-switching devices necessitate an initial filament forming operation, which can lead to an irreversible breakdown unless a current compliance is imposed by an additional device (see Chapter 1, Section 1.8 for details).
In the earlier chapters of this dissertation, specifically in Chapter 3 and Chapter 4, we utilized an external series resistor as a fast-responding current compliance component (see Chapter 2, Section 2.3 for details about such method). This approach, however, is impractical when scaling from device development to large crossbar arrays. Due to the stochastic nature of the filament forming voltage, the current flow through the ReRAM device and the series resistor, dictated by Ohm's law, varies. This variability can result in inconsistent conductive windows and programming parameters across the

devices in the array.
To mitigate these sources of variability, it is essential to integrate a voltage-controlled, active component that can perform the current compliance role: a transistor.

While it could be argued that previous literature includes examples of passive crossbar arrays (= arrays of 1R devices only) used for processing AI workloads, such as neural network weight updates [176], our research demands reduced device variability and improved scalability. This is particularly important for demonstrating parallel analog updates. Achieving this goal requires the fabrication of active crossbar arrays, which embed a selector with each resistive-switching element, thereby creating arrays of 1T1R devices.

5.1.2 General overview of the chip design properties and ReRAM integration

The CMOS transistor technogy is realized in the Front-End-Of-Line (FEOL) of a silicon substrate. **Figure 5.1 (a)** displays a picture of the 200 mm wafer received from the foundry "*X-FAB*", which performed the FEOL processing of our chips and the BEOL fabrication up to three metal levels. The wafer consists of multiple reticles, each approximately 1 cm × 1 cm, repeated across the entire wafer area. Every reticle incorporates the chip design depicted in **Figure 5.1 (b)**.
This chip design consists of a combined ReRAM and CMOS layout. My colleague A. La Porta and I developed the ReRAM layout, whose details will be outlined in the next section. The circuit design was created by H. Erfanijazi and prof. Barranco-Linares from the IMSE, following our specifications about the ReRAM device requirements. We asked for:

- **Transistors capable of handling operational voltages up to 5 V and delivering currents up to 10 mA**, necessary for resetting highly-conductive ReRAM devices such as our Gen_1-MO-ReRAM (see Chapter 3, Section 3.7. These transistors are n-type Field-Effect-Transistors (n-FETs), characterized by a width (W_{big}) of 10 μm, a channel length (L_{big}) of 0.6 μm, and 5× gate fingers.

- **Smaller transistors providing lower variability in the low current range: (100 nA, 10 μA).** To force a compliance current in the range (100 nA, 10 μA), our big transistors require a gate voltage below 500 mV. For such voltage values, the transistor operatres in the sub-threshold regime, characterized by an exponential dependency between gate-source voltage and drain-source current. This could lead to large device-to-device variability in the forming current.
  For this reason, we asked the CMOS designers to also include in the design smaller transistors operating above threshold in the (100 nA, 10 μA) current range. These smaller transistors are n-FETs characterized by a width (W_{small}) of 4.5 μm, a channel length (L_{small}) of 1.2 μm, and 1× gate finger.

- **Each transistor is paired with an adjacent transistor of identical sizing, serving as a current mirror** for more accurate control of the ReRAM current during device operations.
  However, the results presented in this dissertation never involved the utilization of the current mirror transistor. For this reason, even if our devices are designed as a 2T1R structure, we will refer to 1T1R devices and measurements only.

Figure 5.1 (c) depicts an example of a small, current-mirror transistor layout (in the right) and a ReRAM *mesa* (in the left) from the combined layout. Since these components are very distant in the horizontal plane (> 20 μm), we combined two separate images, explaining the dotted separation between them.

In **Figure 5.1 (c)** we drew a horizontal line connecting the ReRAM "mesa" to the n-FET components. Along this line we cut a finished chip (after integrating the ReRAM device in the BEOL) with a Focused-Ion-Beam (FIB), and imaged the cross-section of the whole BEOL with Scanning-Electron-Microscopy (SEM). In the left of **Figure 5.1 (d)** we see the ReRAM device fabricated on top of the last metal level (M3), and on the right, we see the metal plugs for vertical connection across the three metal layers (M1, M2, M3).
Figure 5.1 (e) displays the process flow designed by A. La Porta and me for integrating the ReRAM devices on the wafers received from *X-FAB*. The wafers were delivered with an M3 coating above a foundry dielectric, as shown in **Figure 5.1 (e)**. Our process flow starts from this M3 level and consists of 11 lithographic steps, including both e-beam and optical lithography methods, such as mask fabrication and exposure.

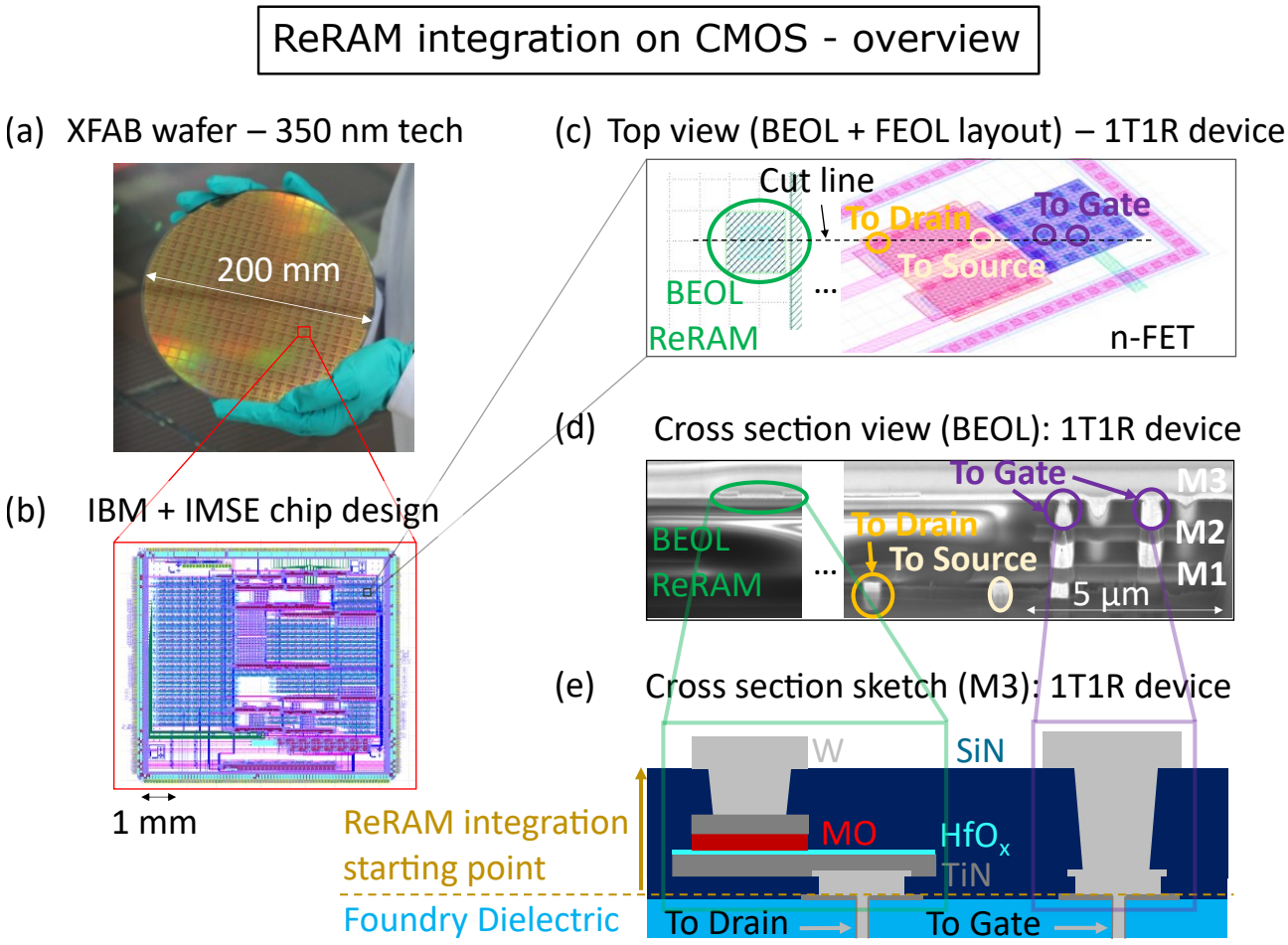


Figure 5.1: **(a)** Picture of the 200 mm wafer received from $X-FAB$ including complete processing of the FEOL and the BEOL up to the metal level $M3$. **(b)** Layout of the chip designed by IBM (ReRAM integration) and IMSE (CMOS design), which is replicated across multiple reticles in the 200 mm wafer. **(c)** Zoom in a region of the layout where a 1T1R device is drawn. On the left, the ReRAM device *mesa*, defining its active area. On the right: gate, source and drain regions of two n-FETs of matched sizes. One of them acts as a voltage-controlled transistor, the other one is a current mirror. **(d)** Cross sectional view of the BEOL. The metal levels M1, M2 and M3 are processed by $X-FAB$, while we integrated the ReRAM on top of them. **(e)** Cross sectional illustration of the ReRAM device structure processed above M3.

5.1.3 Layout for ReRAM integration on CMOS

Figure 5.2 displays the entire layout designed for the integration of the ReRAM devices on CMOS. I

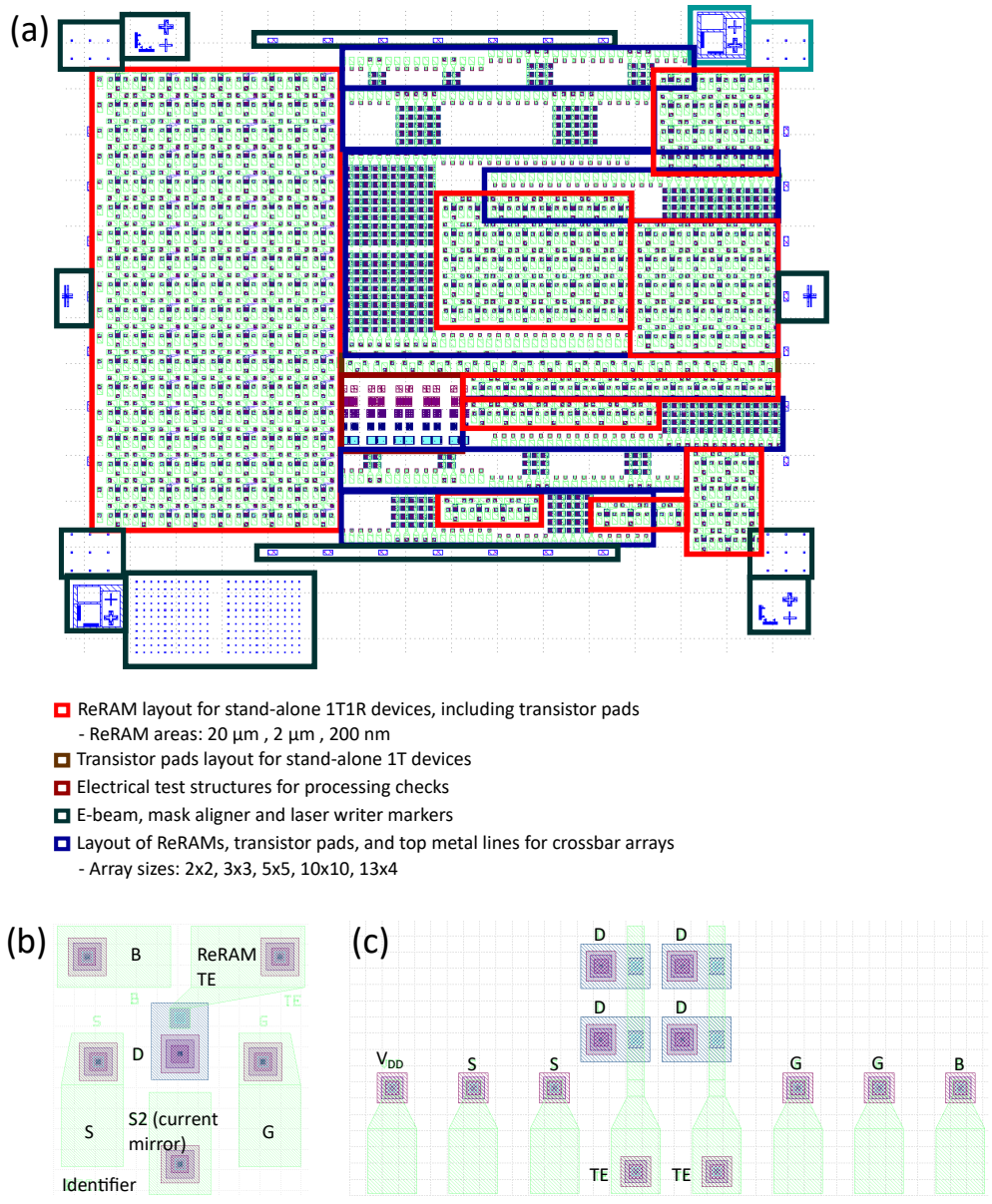


Figure 5.2: **(a)** Layout designed for the integration of the ReRAM devices on CMOS. The layout contains 1T1R stand-alone devices (highlighted in red), 1T stand-alone devices (highlighted in brown), electrical test structures for validating the processing steps (in red berry), lithography markers (in black) and crossbar arrays of 1T1R devices (in blue). **(b)** Zoom-in on a 1T1R structure, indicating the pads to the transistor Body (B), Gate (G), Source (S) terminals, and the current mirror Source (S2). The transistor Drain (D) is connected to the ReRAM bottom electrode through the definition of a rectangular island colored in blue. The access to the ReRAM Top Electrode (TE) is given by the pad labeled as TE. **(c)** Zoom-in on a 2×2 crossbar array. When activated, the V_{DD} pad disconnects the internal logic circuits to enable external control of the array through external biasing of the array pads.

performed this work under the supervision of my colleague A. La Porta. Among the most important features present in the layout, we highlighted:

- **Stand-alone 1T1R devices:** Around 30% of the chip area is covered by 1T1R devices, highlighted in red in **Figure 5.2 (a)**. We included different ReRAM active areas of (20 μm)^2, (2 μm)^2, and (200 nm)^2 to study the scalability of the 1T1R devices. Half of the devices have a large transistor, and the other half a small transistor to enable either high reset currents or precise low-forming-currents. Approximately 10% of the 1T1R devices include a 1kΩ series resistor, fabricated as an integrated polysilicon layer, and another 10% have a 10 kΩ series resistor, to study their impact on the resistive-switching properties.
  The different ReRAM areas, transistor sizes, and series resistor values enable exploring the impact of key design aspects on the 1T1R electrical properties. The purpose is to optimize the device response through the optimization of such parameters for a second tape-out including larger crossbar arrays.
  More details about the 1T1R devices are displayed in **Figure 5.2 (b)**, showing a zoomed picture of the 1T1R layout. The pads for probing the transistor (Body (B), Source (S), Gate (G), and Source 2 (S2), which activates the current mirror functionality) are visible. The Drain node (D) is short-circuited by design to the ReRAM bottom electrode. To access the 1T1R device, the Top Electrode (TE) of the ReRAM devices must also be biased through a dedicated pad.
  Inside each pad, concentric squares and a set of green small dots at their center indicate where the CMOS designers should locate the transistor terminals, as the green dots define metal pillars connecting the FEOL to the BEOL through the metal layers M1 and M2.

- **Stand-alone 1T devices:** Single transistor structures, highlighted in brown. These will be electrically characterized to compare the current-voltage characteristics of the 1T devices with the 1T1R structures, enabling the decoupling of the properties between the transistor and the ReRAM.

- **Electrical test structures:** Highlighted in red berry color. Our process-flow is fully CMOS- and foundry-compatible, avoiding noble metals and lift-off processes. Therefore, the design includes test structures to electrically measure whether every etching step has been successful.

- **Lithography markers:** The design includes electron-beam, laser writer and mask aligner markers, enabling both optical and e-beam lithography, highlighted in black. Only the processing steps where sub-μm feature are patterned, such as the *'mesa'* definition, will be performed by e-beam lithography, which involves longer exposure times and critical resist development steps. All the larger features will be instead defined by optical lithography.

- **Layout of crossbar arrays:** The design includes crossbar arrays of 1T1R devices of sizes: 2×2, 3×3, 5×5, 10×10, 13×4, all highlighted in blue. The motivation is to gradually increase the measurement complexity, eventually targeting small-scale, real-world applications with the largest arrays (10×10 and 13×4).
  Figure 5.2 (c) displays the zoom inside a 2×2 array. The V_{DD} pad is critical to define the mode of operation of the array. When floating, the array is controlled by internal circuitry such as a pulse-pattern generator. When biased at 5 V, draining a current of approximately 10 mA, the logic circuits are disabled, and the array can be controlled by an external multichannel waveform

generator. All results presented in this dissertation were obtained by controlling the crossbar arrays under the latter operating mode.

5.2 DC characteristics of the single 1T structures

Before discussing the electrical characteristics of the 1T1R devices, it is useful to present the current-

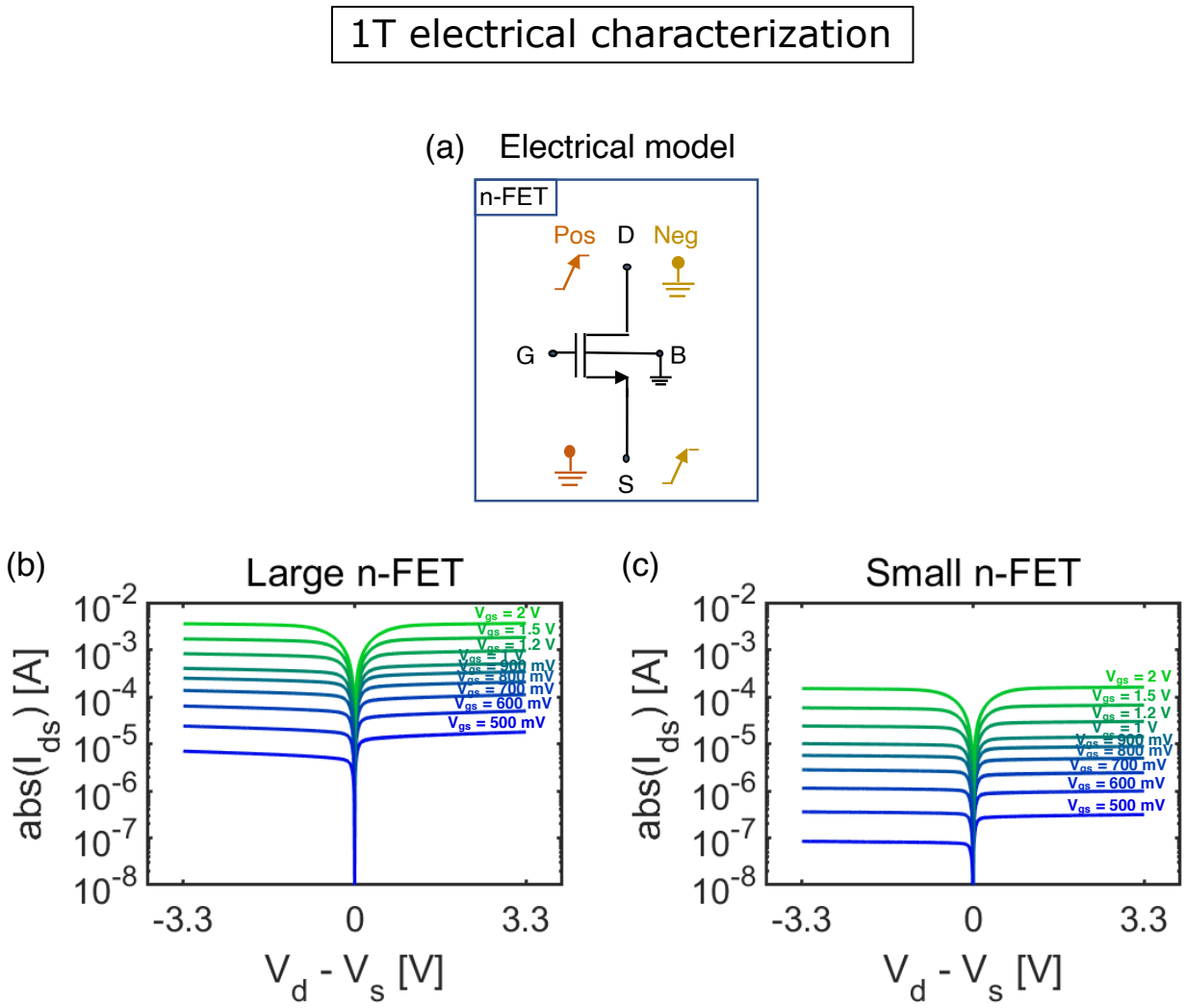


Figure 5.3: **(a)** Electrical equivalent scheme of the DC measurements performed on the 1T device. We apply a positive voltage sweep at the the drain (source) side, while keeping the source (drain) at ground. **(b)** (I_{ds} vs. V_{ds}) characteristics of big transistor (geometrical details in main text). **(c)** (I_{ds} vs. V_{ds}) characteristics of small transistor (geometrical details in main text).

voltage properties of the single 1T structures (n-FET). Knowing the transistor's (I_{ds} vs. V_{ds}) characteristics is essentail for analyzing the transistor's operating point in the 1T1R measurements.
Figure 5.3 (a) depicts the electrical equivalent scheme for the measurements performed on a 1T device. The bipolar current-voltage characteristics are obtained by performing a positive voltage sweep at the drain (source) side, while keeping the source (drain) at ground. This method avoids negative

voltage biases for both the drain and the source terminal, thus preventing the forward biasing the body-source *pn* junction.
The (I_{ds} vs. V_{ds}) characteristics are displayed in **Figure 5.3 (b) and (c)**, for big and small transistors, respectively. We applied increasing gate voltages (V_g), from 500 mV to 2 V. As expected from simulations, we observed a difference in the current values between the two transistor sizes by approximately a factor 20×. The asymmetry in the current curves is attributed to asymmetric technology properties between source and drain terminal of the n-FET.

5.3 Tuning the operating regime of 1T1R devices by adjusting the filament forming current

5.3.1 High forming current - DC characteristics

Figure 5.4 (a) depicts the top view of a 1T1R device. The layout includes large pads, each larger than 60 μm × 100 μm, facilitating easy electrical probing with measurement tips. One pad connects the ReRAM Top Electrode (TE), while the other pads connect the Body (B), Source (S), and Gate (G) terminals of the transistor.
Figure 5.4 (b) illustrates the cross section of a ReRAM device (2 μm × 2 μm) integrated on CMOS. The device's active area is defined by a *mesa* structure, similarly to the Gen_1 devices, to prevent the W delamination observed in Gen_2- devices (when biased with currents in the mA range, see Chapter 4, Section 4.6). The ReRAM stack consists of TiN/HfO_x/MO/TiN/W, with the HfO_x layer optimized as described in Chapter 4, Section 4.1. The bottom electrode is a 20 nm-thick TiN layer, selectively removed by Reactive-Ion-Etching (RIE) to electrically isolate the ReRAM device, except where it connects to the transistor's drain terminal (rectangle colored in green, labeled "*D*" at the center of **Figure 5.4 (a)**).
Figure 5.4 (c) describes the method applied for the electrical characterization of the 1T1R device. Only positive or zero voltage biases are applied to the 1T1R terminals to prevent forward biasing of the *pn* junction between the body and the source/drain regions. During a reset voltage sweep, the body-source ($|V_{bs}|$) voltage value increases, causing a *body effect*. To minimize this issue, the transistor is fully activated by applying a gate voltage equal to 5 V during the reset operation.
Figure 5.4 (d) shows 100 subsequent DC bipolar I-V curves obtained using a gate voltage (V_g) equal to 1 V during the forming/set operations, resulting in a compliance current of around 100 μA. The electrical characteristics are similar to those of the optimized 1R devices discussed in Chapter 4, Section 4.2, indicating that the CMOS integration did not affect the ReRAM properties. To increase the on/off ratio, we can either increase the V_g during the set transition or the stop amplitude of the reset voltage sweep, or both. **Figure 5.4 (e)** shows the device on/off ratio over 100 DC write/erase cycles. There is only a minor reduction in the on/off ratio over cycling, indicating the device robustness to severe electrical stress such as DC programming.

5.3.2 Low forming current - synaptic characteristics

When the gate voltage is reduced from 1 V to 0.7 V, the compliance current of the filament forming process drops to approximately 2 μA. The reduction of the ReRAM current leads to a milder resistive-switching process, as demonstrated in **Figure 5.5 (a)**. Following this step, we conducted Long-Term-Potentiation (LTP) and Long-Term-Depression (LTD) measurements by applying a train of 128× programming pulses for potentiation and 128× pulses for depression. **Figure 5.5 (b)** displays the synaptic characteristic from these measurements. Key observations from the results include:

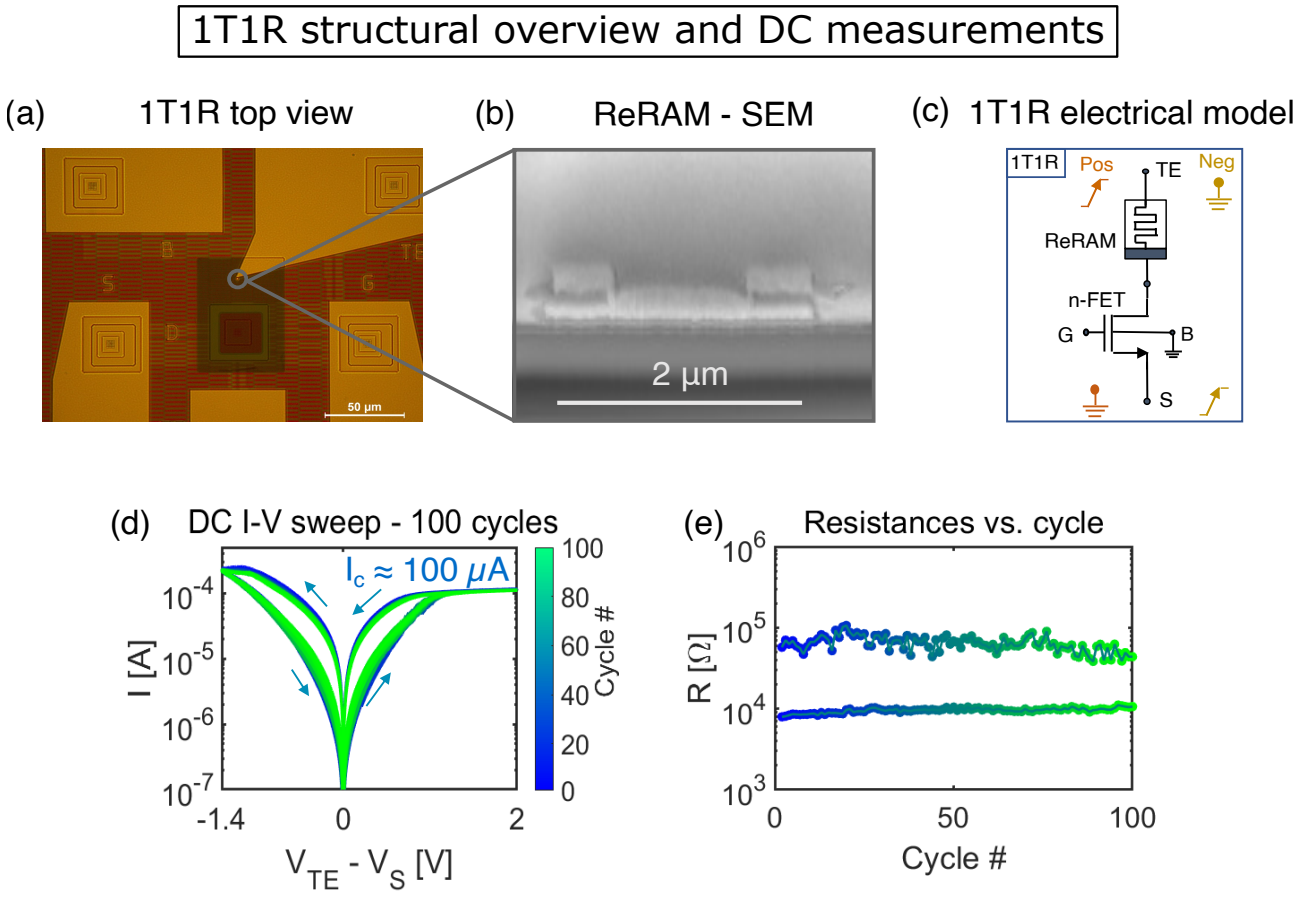


Figure 5.4: **(a)** Top view of a 1T1R device. The ReRAM top electrode is connected to a pad labeled as "*TE*". The pads connected to the transistor terminals are labeled as "*B*", "*S*", "*G*". An additional pad connects to a second transistor that acts as a current mirror. **(b)** Cross sectional view of a 2 μm $\times$ 2 μm ReRAM device integrated on CMOS. **(c)** Electrical equivalent scheme for the DC measurements performed on the 1T1R device. This scheme outlines the method for obtaining bipolar I-V characteristics by applying voltage sweeps to the "TE" and "S" pads while keeping the other pad grounded. **(d)** Bipolar DC I-V characteristics of the 1T1R device. The gate bias is set to 1 V during the set transition to achieve a compliance current of approximately 110 μA. The gate bias is raised to 5 V for the reset transition to ensure sufficient reset current. **(e)** Low-Resistive-State (LRS) and High-Resistive-State (HRS) measured during 100 DC program/erase cycles, showing only a minor reduction of the on/off ratio.

- **Analog switching behavior at very low conductances:** The device exhibits nearly analog switching behavior, characterized by very low conductances - approximately three orders of magnitude lower compared to the best 1R devices discussed in Chapter 4, Section 4.3).

- **Higher stop voltages:** The measurements were performed with slightly higher stop voltages ($V_{stop,max}$ = 2.4 V) compared to the maximum $V_{stop,max}$ = 1.8 V used in the 1R measurements. The same pulse duration was applied in both cases. The higher resistive-switching voltages observed in low-conductive ReRAM devices can be attributed to reduced Joule heating, hindering the drift of oxygen anions.

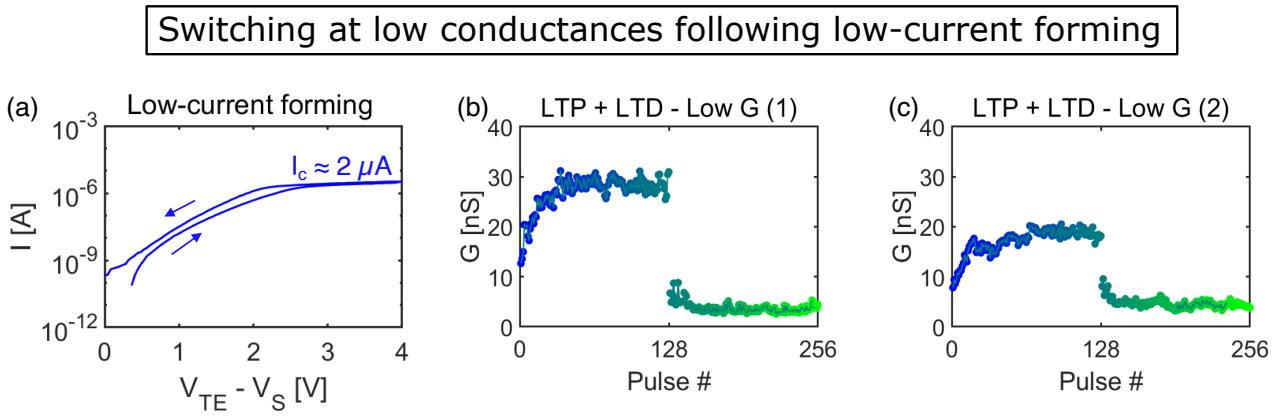


Figure 5.5: **(a)** Filament forming voltage sweep with a gate voltage of 0.7 V, resulting in a compliance current of approximately 2 μA. **(b)** Analog Long-Term-Potentiation (LTP) and Long-Term-Depression (LTD) characteristics following pulse trains with an amplitude of V_{set} = 2.4 V and V_{reset} = 2.3 V, respectively. The pulse duration is t_{pulse} = 1 μs. **(c)** Enhanced analog LTP and LTD characteristics achieved by reducing both the set and reset voltages by 100 mV.

- **Reduced on/off ratio:** To improve the analog device response, from **Figure 5.5 (b)** to **Figure 5.5 (c)** we reduced the set and reset stop voltages by 100 mV. Consequently, the on/off ratio decreased, but the resulting synaptic curves showed improved analog behavior in both directions.

Although requiring higher voltages and exhibiting a reduced on/off ration, the ReRAM devices formed with a low compliance current exhibited fast and analog resistive-switching properties in a conductive window ranging from few nS, to few tens of nS. This result marks a fundamental advancement for energy-efficient analog processing in deep learning applications.
However, it is crucial to address a significant drawback: the low-current filament forming method is accompanied by substantial device-to-device variability in the conductive window, ranging from (1 nS to 1 μS) (not shown). This variability presents a challenge for reliable performance in array-level experiments and suggests that further development is needed before this method can be effectively applied to larger arrays.

5.4 1T1R programming methods towards array measurements

5.4.1 Half-bias programming scheme

The conventional algorithm implementing backpropagation through parallel crossbar array updates is described in Chapter 1, Section 1.6 (see **Figure 1.7**). The proposed implementation assumes a crossbar array with single 1R elements subjected to bipolar voltage pulses. In a 1T1R configuration, where the 1T component is an n-FET transistor, we should avoid negative voltage pulses to prevent forward biasing the bulk-source *pn* junction. Therefore, our programming scheme must include only positive voltage values.
Figure 5.6 (a) shows the electrical scheme for performing Long-Term-Potentiation (LTP) and Long-Term-Depression (LTD) on a 1T1R device. For LTP, we decrease the voltage at the Source Line (SL) terminal from a DC value, denoted as V_{HB}, down to 0 V (where HB stands for Half-Bias). Simultaneously, we apply a pulse at the BL terminal, increasing the voltage from V_{HB} to 2×V_{HB}. For LTD, we pulse the BL to decrease its voltage from V_{HB} to ground, while increasing the SL voltage from V_{HB} to 2×V_{HB}.

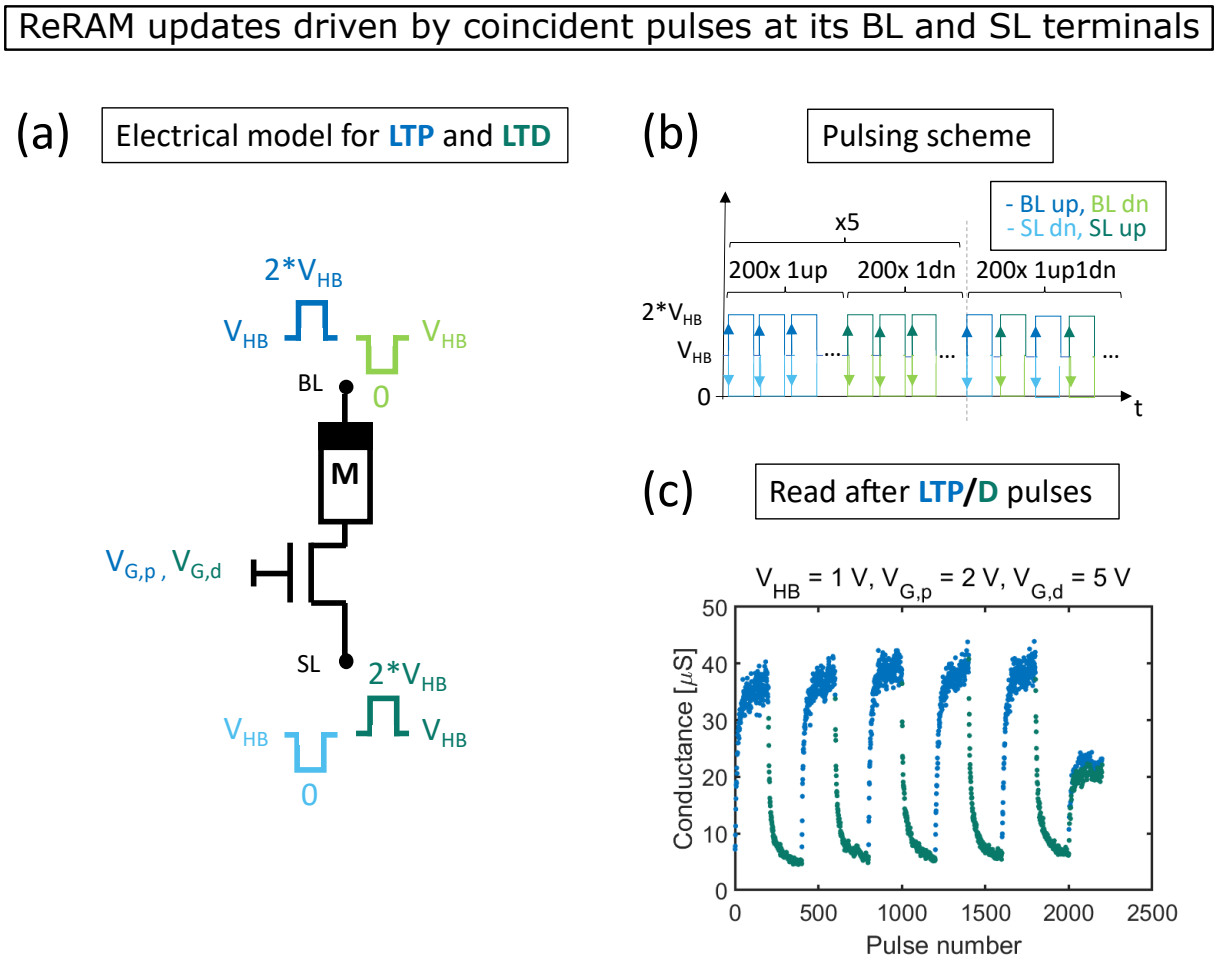


Figure 5.6: **(a)** Illustration of the *Half-Bias* programming scheme. Parameters related to Long-Term-Potentiation (LTP) are shown in blue tones, while green is used for Long-Term-Depression (LTD). **(b)** Pulsing scheme devised to test LTP, LTD and Zero-Shifting (ZS) following the *Half-Bias* pulsing method. **(c)** 1T1R conductive states measured after each programming pulse.

The LTP and LTD cycles occur in four different phases (see Chapter 1, Section 1.6 for details). Consequently, V_{HB} can differ between the LTP and LTD processes, and different gate voltages can also be applied.

We tested the 1T1R device response to the *Half-Bias* pulsing conditions, as illustrated in **Figure 5.6 (b)**. We alternated between 200× LTP pulses and 200× LTD pulses, repeated five times. This sequence was followed by 1000× alternating 1-up and 1-down pulses. The complete pulsing scheme aims to assess the analog properties of the 1T1R device and test the *zero-shifting* method required by the Tiki-Taka algorithm.

The results, shown in **Figure 5.6 (c)**, demonstrate analog properties similar to the ones exhibited by the optimized 1R devices (compare **Figure 5.6 (c)** to **Figure 4.3 (c)** from Chapter 4). In this experiment, we used triangular-shaped voltage pulses with rising and falling times of 1 μs (see Methods Chapter 2, Figure 2.8 for more details) due to the limited bandwidth of our measurement setup.

In conclusion, the *Half-Bias* programming scheme, which involves two coincident pulses with voltages always greater than 0 V, yields analog properties comparable to those achieved with single positive or

negative pulses. This validation is crucial for ensuring the effectiveness of 1T1R devices in the targeted array operations.

5.4.2 Stochastic conductance updates with the half-bias scheme

We extended the *Half-Bias* programming scheme to include tests of stochastic updates, which are

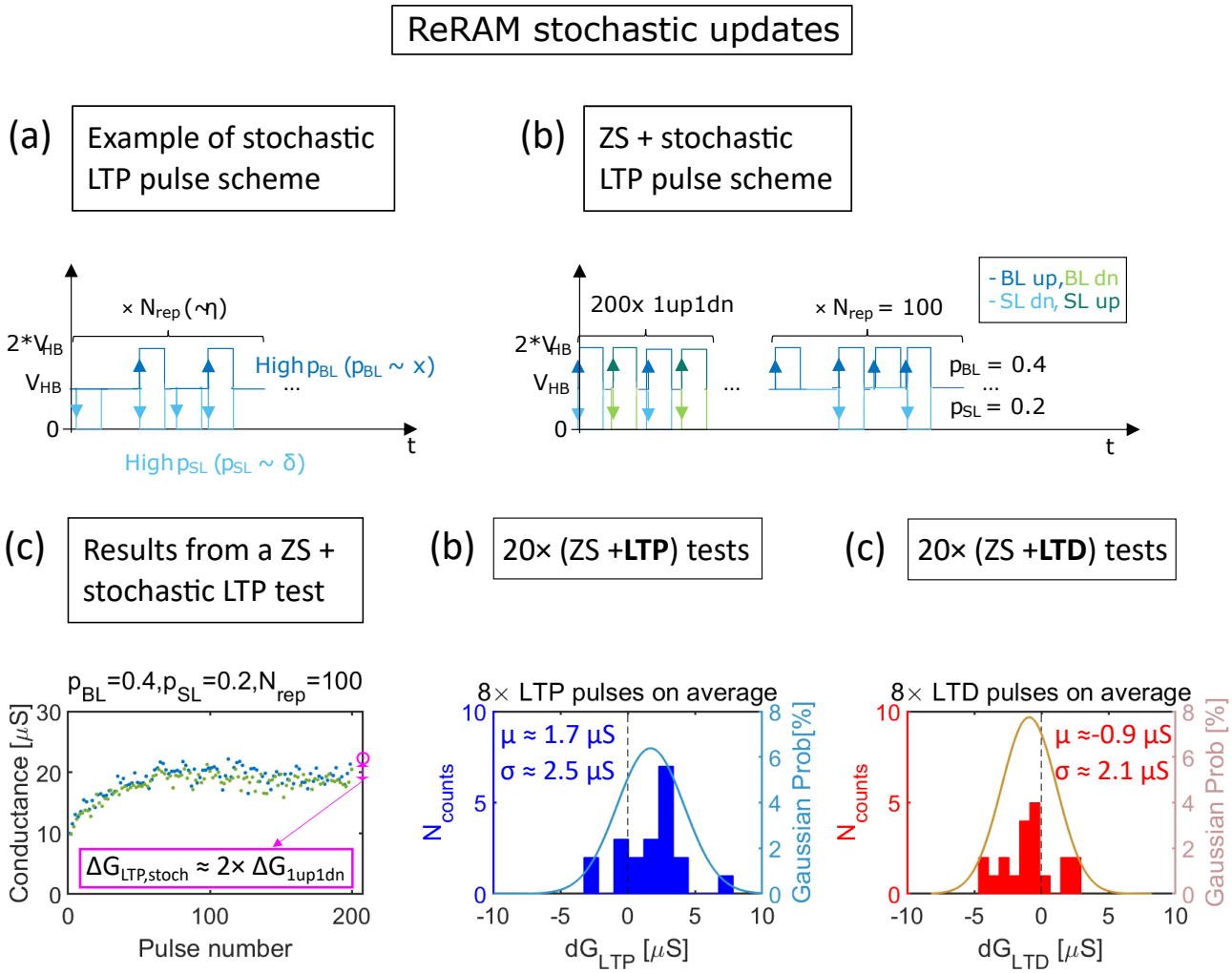


Figure 5.7: **(a)** Example of a stochastic LTP process performed using the *Half-Bias* programming scheme. The probabilities p_{BL} and p_{SL} are determined by the neural network functions x and δ, while the pulse stream length N_{rep} depends on the learning rate of the backpropagation algorithm. **(b)** Full implementation of a stochastic LTP test, which includes the zero-shifting (ZS) initialization. **(c)** Results from a (ZS + stochastic LTP) test with $p_{BL} = 0.4$, $p_{SL} = 0.2$, $N_{rep} = 100$. The conductance circled in purple is measured after the LTP process is over. **(d)** Distribution of 20× stochastic LTP processes performed with identical parameters ($p_{BL} = 0.4$, $p_{SL} = 0.2$, $N_{rep} = 100$, averaging 8 programming pulses). Results are reported as hystograms (in blue), fitted by a Gaussian probability density function (light blue). Mean value and standard deviation of the conductance updates are labeled as μ and σ. **(e)** Results from 20× stochastic LTD processes performed with identical parameters ($p_{BL} = 0.4$, $p_{SL} = 0.2$, $N_{rep} = 100$, averaging 8 programming pulses). Results are reported as hystograms (in red) fitted by a Gaussian probability density function (light red).

essential for realistic neural network training algorithms. **Figure 5.7 (a)** illustrates an example of a pulse scheme for implementing a cycle of stochastic Long-Term-Potentiation (LTP) updates. The pulsing probability at the BL (p_{BL}) is proportional to the input activation function x, while the pulsing frequency at the SL (p_{SL}) depends on the backpropagated error δ. The length of the pulse stream (N_{rep}) is proportional to the learning rate of the Stochastic Gradient Descent (SGD) algorithm. More details on this concept can be found in Chapter 1, Section 1.6.
To meet the requirements of the Tiki-Taka algorithm, we prepared the devices at their symmetry point before performing the stochastic half-bias updates. Therefore, we devised a complete electrical pulse scheme, including 200× programming pulses with alternating polarities to drive the device state to a symmetry point, followed by stochastic LTP (LTD) pulses, as illustrated in **Figure 5.7 (b)**.
Figure 5.7 (c) presents the results from a stochastic LTP test, using a potentiating pulse stream with probabilities: $p_{BL} = 0.4$ and $p_{SL} = 0.2$, and length: $N_{rep} = 100$. With these parameters we expected an average of 8 pulse coincidences, and, as desired, we measured a positive conductance update. The amount of the stochastic LTP update was approximately twice the average ΔG measured during the zero-shifting procedure.
We repeated the same (ZS + stochastic LTP) test 20 times. **Figure 5.7 (d)** shows the distribution of the ΔG values measured after each stochastic potentiation process. In most cases, we observed a positive conductance update, as expected, with a mean value of approximately 1.7 μS. However, there was a large spread in the distribution ($\sigma > \mu$), with some negative updates also occurring. We attribute this large spread to the variability in the stochastic pulsing probability and the inherent programming noise of the ReRAM device.
Figure 5.7 (e) shows the results from a similar stochastic LTD test, again performing the ZS initialization first. In most cases we achieved the desired negative conductance changes, but the conductance updates were widely spread (μ = -0.9 μS, σ = 2.1 μS), with few updates in the wrong direction.
The measured programming noise and the undesired update directions may represent an important obstacle to the acceleration of deep learning tasks with resistive processors based on our ReRAM technology. Future work will focus on mitigating this issue by further optimizing the ReRAM material properties and/or by increasing the algorithm robustness to the programming noise.

5.4.3 Resistance update suppression of non-targeted RRAM cells

When programming a resistive-switching device embedded in a crossbar array architecture using the pulse coincidence method, disturbance pulses affect all elements sharing the same row and column. In the *Half-Bias* method, the amplitude of a disturbance pulse equals V_{HB}. **Figure 5.8 (a)** describes the electrical testing conditions used to evaluate the impact of half-bias pulses on non-targeted cells. **Figure 5.8 (b)** illustrates a complete pulsing scheme designed to detect any potential cumulative effect from subsequent disturbance pulses. We apply 5× consecutive half-bias pulses at the BL terminal, alternated by 5× half-bias pulses at the SL terminal. This pattern is repeated for 200 times.
We first target a ReRAM device programmed to a low conductive state of approximately 10 μS. **Figure 5.8 (c)** depicts the results of this experiment: in this case, the disturbance pulses have no effect on the ReRAM state in this condition.
In **Figure 5.8 (d)**, we repeat the same test for a ReRAM device programmed to a high conductive state of approximately 40 μS. This time, the device conductance is slightly modified by the disturbance pulses. These pulsing activity changes the device state in the expected direction: a half-bias pulse applied at the BL terminal increases the ReRAM conductance, while pulses applied to the SL terminal decrease the device conductance. Zooming-in on **Figure 5.8 (d)** (inset figure in the left), we clearly see the effect of the disturbance pulses. However, the maximum change of the device conductace is

very little (below 4% compared to its initial state).

In conclusion, since our ReRAM devices exhibit strong non-linearity between the programming

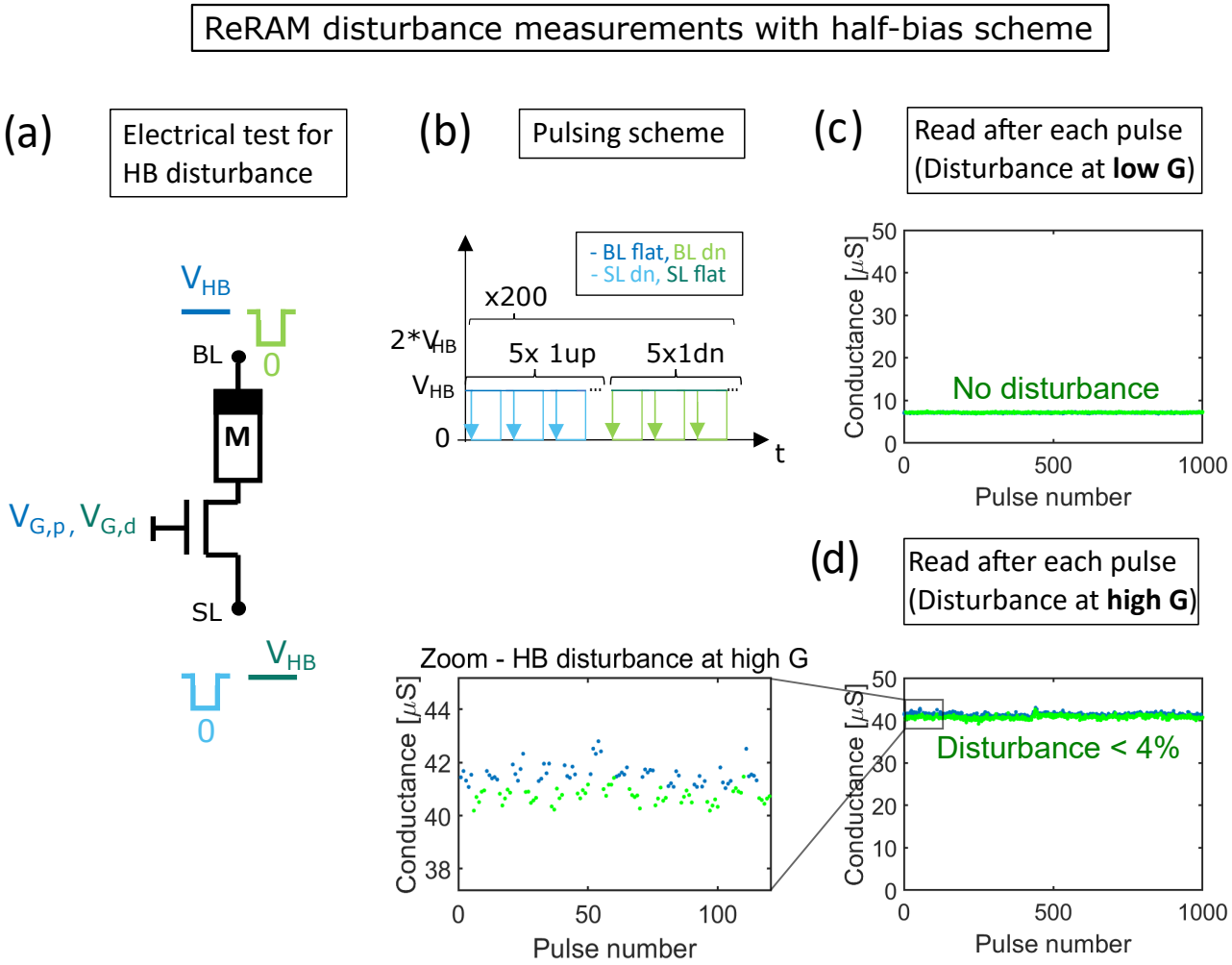


Figure 5.8: **(a)** Disturbance test description: non-coincident half-bias pulses are applied at the 1T1R terminals. Pulses at the SL are shown in blue, while BL pulses are colored in green. **(b)** Pulsing scheme designed to test the cumulative impact of multiple disturbance pulses affecting the same device. **(c)** Disturbance measurements from a ReRAM device programmed at a low conductance (approximately 10 μS). No change in conductance is observed following the disturbance pulses. **(d)** Disturbance measurements from a ReRAM device programmed at a high conductance (approximately 50 μS). The disturbance pulses slightly affect the device conductance, as highlighted in the zoomed-in figure on the left.

voltage amplitude and the change in conductance, the disturbance caused by half-bias pulses is almost negligible compared to the full G swing.

Nonetheless, we have addressed this minor issue by proposing an alternative programming scheme. The new scheme is based on the coincidence of pulses at the Word Line (WL) and SL, theoretically preventing from any impact of the disturbance pulses. The idea is depicted in **Figure 5.9**. Each quadrant of the figure represents a different phase of the parallel update scheme (++, +-, -+, −) (please see Chapter 1, Section 1.6 for more details regarding the conventional implementation of parallel crossbar array updates). **Figure 5.9 (a)** describes the case $x > 0$, $\delta > 0$ (++ phase). The targeted

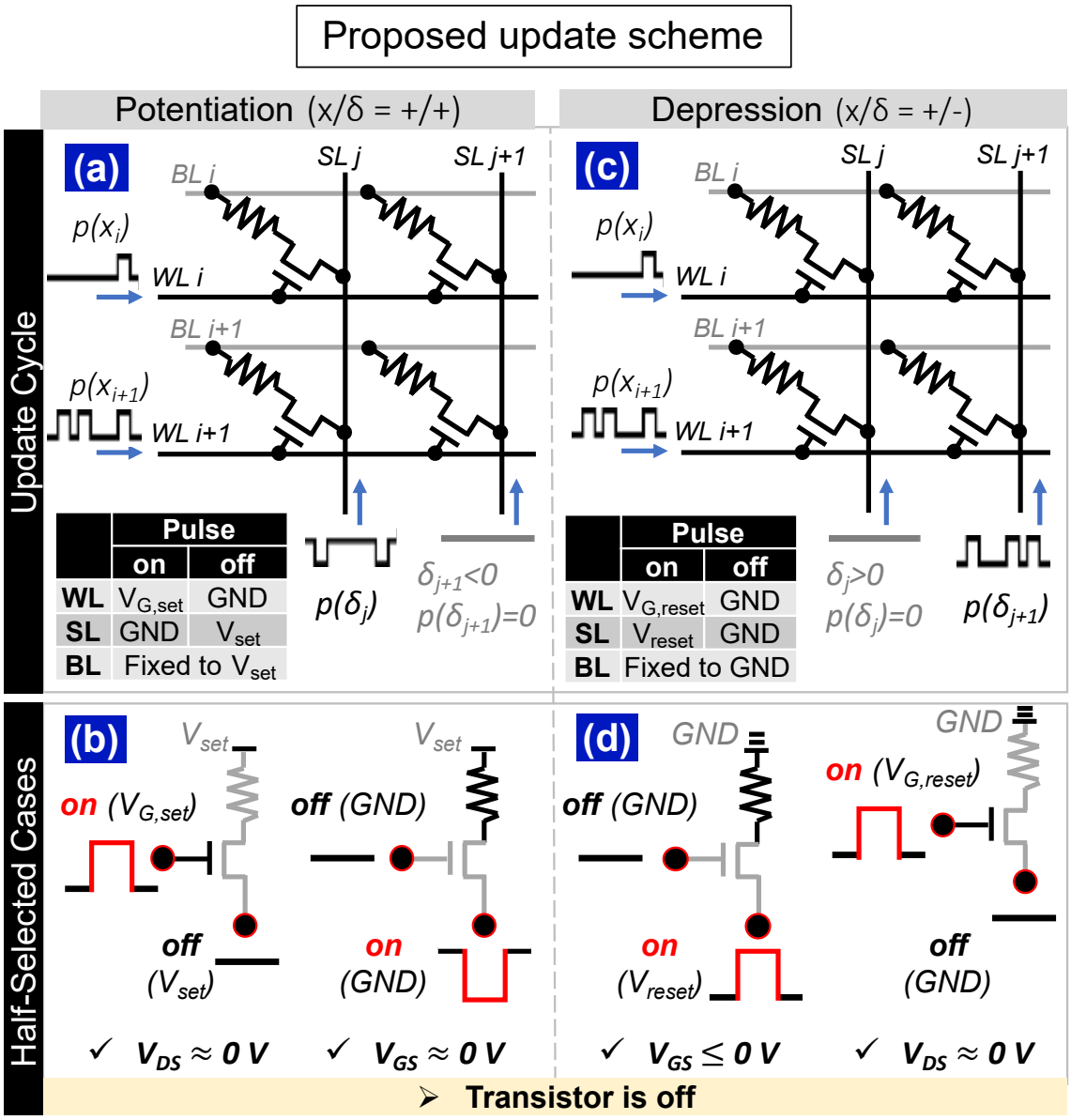


Figure 5.9: **(a)** Proposed potentiation cycle with coincident pulses occurring at the Word Lines (WL) and Source Lines (SL). The table summarizes the voltage drop scenarios at the WL, SL and Bit Line (BL) when the pulsing activity is on or off (first and second columns, respectively). **(b)** Voltage drop scenarios for non-targeted cells during a potentiation cycle. **(c)** Implementation of a depression cycle. **(d)** Voltage drop scenarios for non-targeted cells during a depression cycle.

cells are updated by coincident pulses of opposite polarity occurring at the WL and the SL. The pulses applied at the WL increase the WL voltage from ground (GND) to $V_{G,set}$. The pulses applied at the SL decrease the SL voltage from V_{set} to GND. Under these conditions, the non-targeted cell either undergoes a 0 V drop between SL and BL (case $V_{DS} = 0$ V from **Figure 5.9 (b)**), or there is a 0 V drop between gate and source of the transistor (case $V_{GS} = 0$ V **Figure 5.9 (b)**).

Figure 5.9 (c) describes a depression phase, such as: x >0 and δ ((+-) phase). The targeted cells are updated by coincident pulses of opposite polarity occurring at the WL and the SL. The pulses

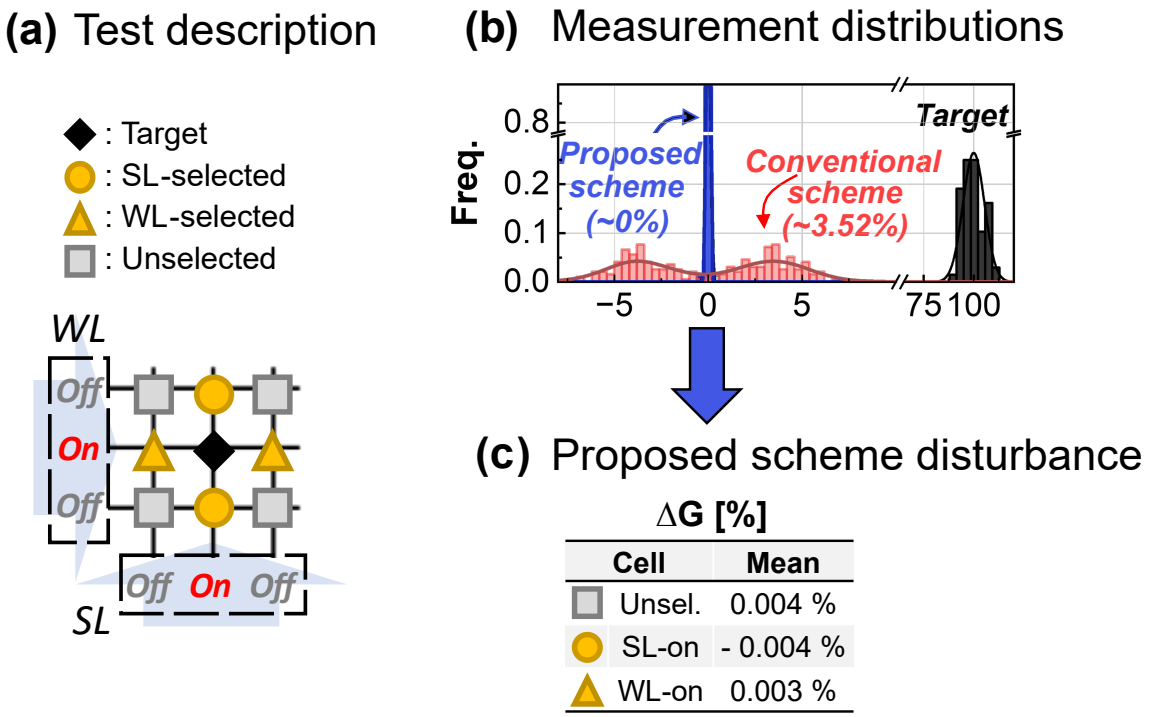


Figure 5.10: **(a)** Testing the new parallel update scheme compared to the conventional Half-Bias scheme. The target cell is surrounded by neighboring devices potentially affected by disturbance pulses. **(b)** Comparison of the relative change in conductance between the target device (in black) and neighboring cells. Results are shown as histograms, with the conventional scheme (in red) showing a 3.5% average change, while the proposed scheme (in blue) shows almost zero disturbance. **(c)** Precise values of disturbance registered in neighboring cells under the new proposed scheme, indicating infinitesimal changes.

applied at the WL increase the WL voltage from ground (GND) to $V_{G,reset}$. The pulses applied at the SL increase the SL voltage from GND to V_{reset}. Every BL is biased at V = GND. Under these conditions, the non-targeted cell undergoes either a negative voltage drop between SL and BL (case $V_{GS} < 0$ V from **Figure 5.9 (d)**), or there is a 0 V drop between SL and BL (case $V_{DS} = 0$ V from **Figure 5.9 (d)**).

In conclusion, the proposed scheme should result in zero disturbance for non-targeted cells. When a disturbance pulse occurs at any of their terminals, it is either a pulse at the gate terminal with V_{DS} at zero, or a pulse at the SL terminal with the transistor off. The phases (-+) and (–) are similar to the cases already outlined.

We tested the new parallel update scheme and we compared it to the conventional *Half-Bias* scheme. **Figure 5.10 (a)** describes the test conducted to compare the two schemes. The target cell is surrounded by neighboring devices potentially affected by disturbance pulses. The ones positioned in the same row share the same WL, hence will have the transistors on. The cells positioned in the same column share the same SL, which is subjected to programming pulses in both schemes.

Figure 5.10 (b) compares the relative change of conductance between the target device and the

neighboring cells. The results are shown as histograms, defining a probability density function similar to a Gaussian distribution. The conventional scheme determines a relative conductance change of 3.5% on average, while the proposed scheme generates almost zero disturbance with a tight distribution. **Figure 5.10 (c)** lists the precise values of the disturbance registered in all neighboring cells under the new proposed scheme. All of them are infinitesimal changes, indicating that none of the neighboring cells is affected by any disturbance issue.
This result is very significant towards the main research goal of parallel array updates.

5.5 Array measurements - parallel crossbar updates

Figure 5.11 illustrates the test designed to demonstrate parallel and analog crossbar updates. We

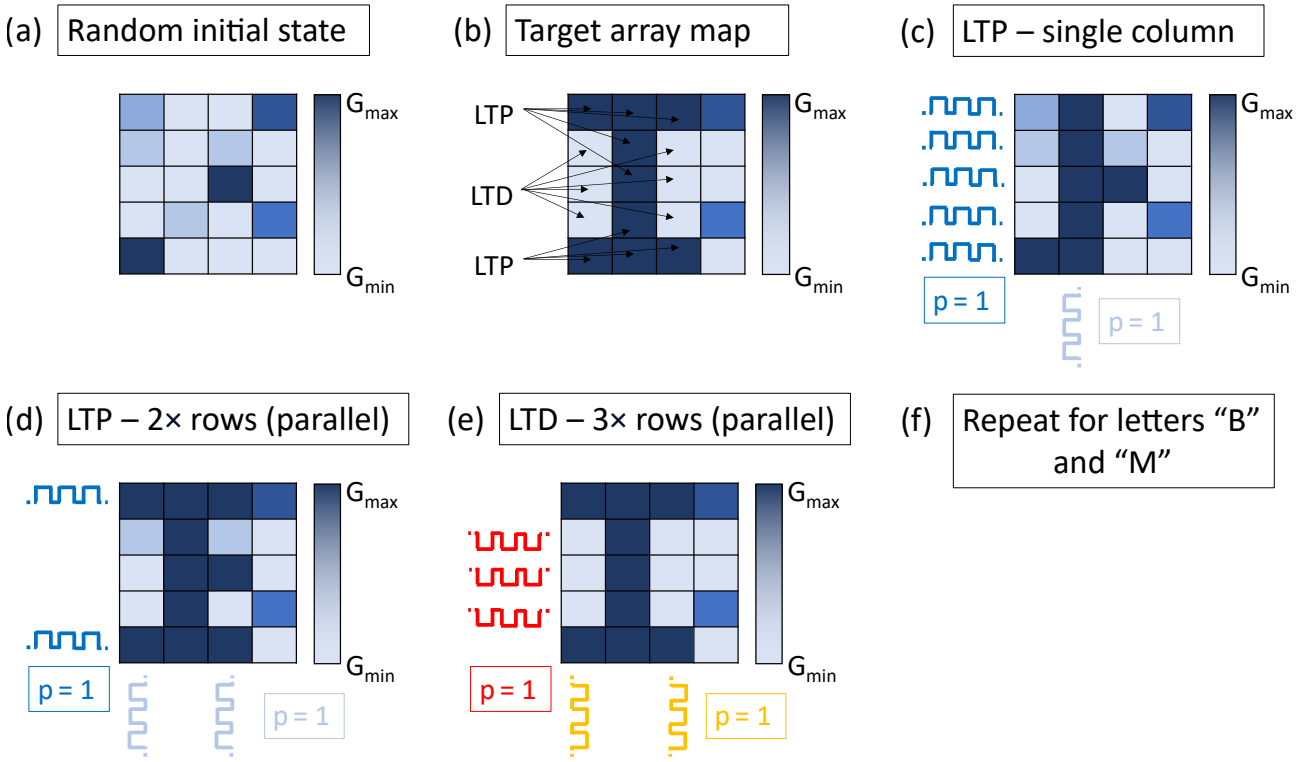


Figure 5.11: Illustration of the experiment designed to test parallel and analog crossbar updates. **(a)** Array elements initialized in random conductive states. **(b)** Target conductance map for the letter *"I"* and type of update (LTP or LTD) required for each array element. **(c)** Implementation of the LTP pulse stream for the central column of the array. **(d)** Implementation of the LTP pulse stream for the first and last column of the array. **(e)** Implementation of the LTD pulse stream for the second, third and fourth column of the array, finalizing the mapping of the letter *"I"* onto the array conductances. **(f)** Repetition of the same experiment for the letters *"B"* and *"M"*.

wire bonded a 5×5 crossbar array, and selected a sub-group of 5×4 cells for programming tests. As depicted in **Figure 5.11 (a)**, the 5×4 matrix is initially set to random conductive states. Our goal is to sequentially map the letters *"I"*, *"B"* and *"M"* onto the array based on the conductance values, by potentiating and/or depressing the required cells. **Figure 5.11 (b)** depicts the target array map for the letter *"I"*. Certain cells need positive updates (LTP), others require negative updates (LTD), while the last column remains unchanged. To achieve this, we apply a pulse stream with deterministic

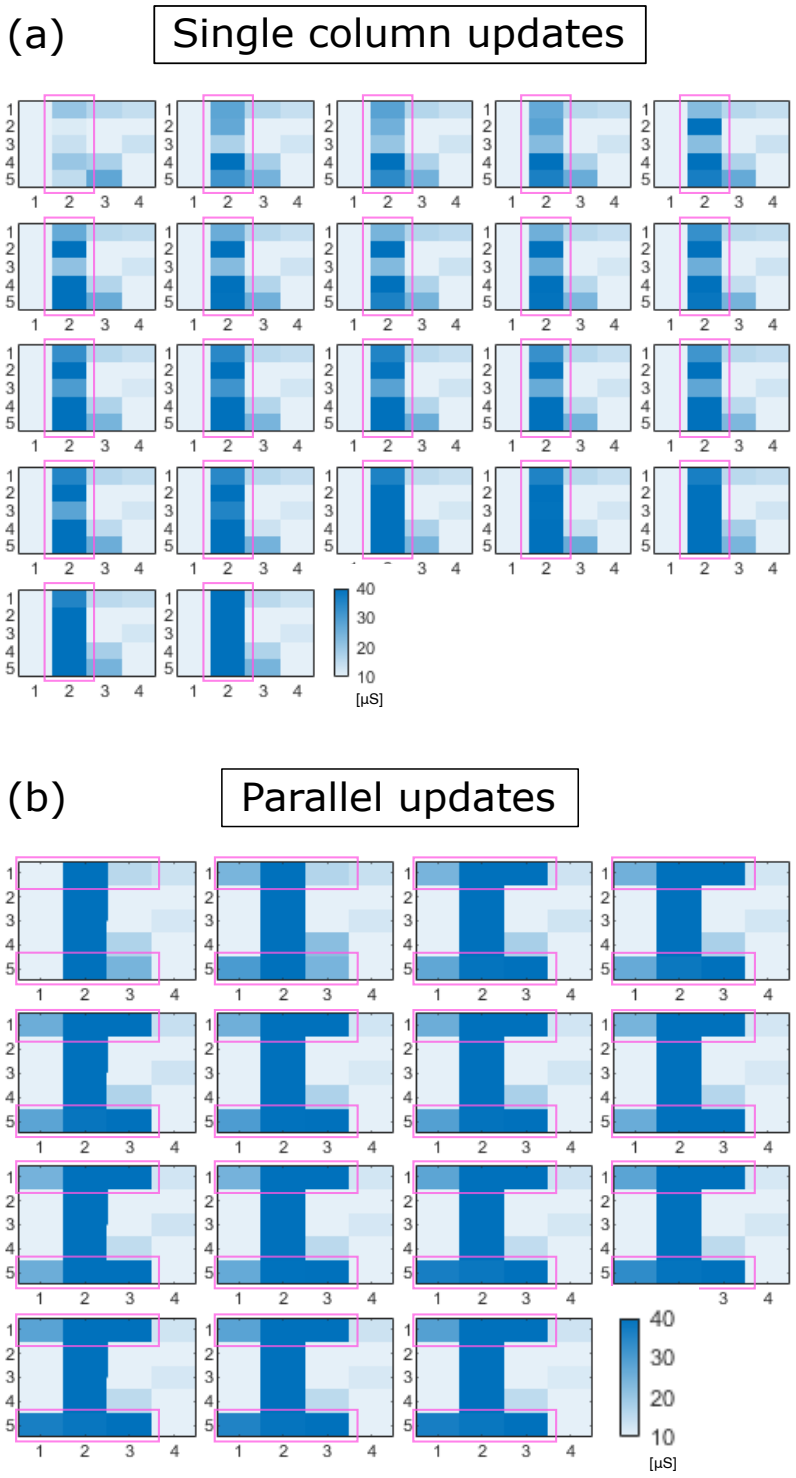


Figure 5.12: Results from the experiment for the letter *I*. **(a)** Analog LTP of the first targeted line of the array (second column from left). **(b)** Analog LTP of the first and fifth row (from top) of the array (outlined in purple). This is a demonstration of parallel and analog LTP performed by coincident programming pulses.

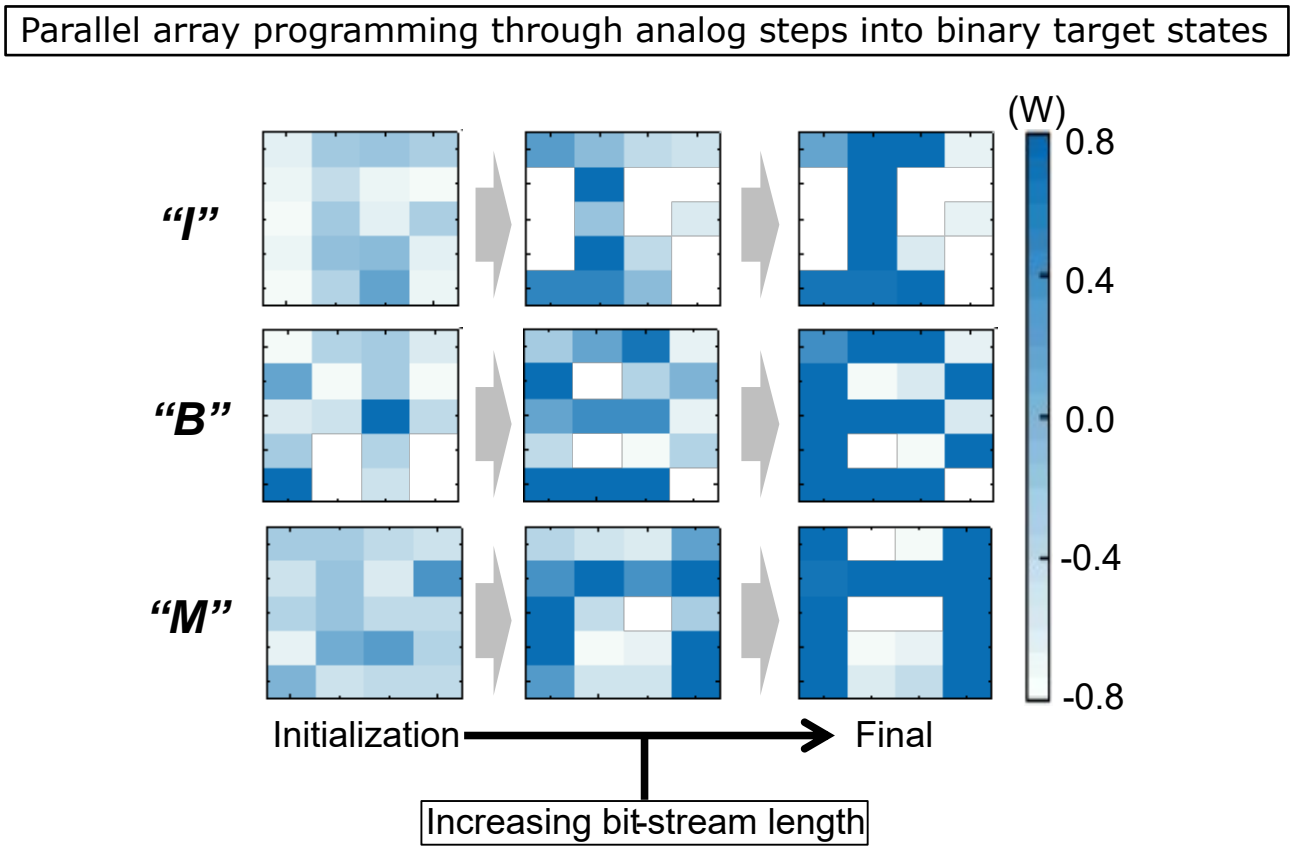


Figure 5.13: Summary of the results from the experiments performed for the three letters (I, B and M). We demonstrated parallel and analog array updates enabling to map the IBM letters into the array conductances.

probability ($p_{SL} = p_{BL} = 1$) until the targeted cells reach the desired conductance. **Figure 5.11 (c)** outlines the pulse scheme for the first LTP cycle, potentiating the second column of the array with coincident programming pulses occurring at every row and at the second column. This constitutes a single-line update.
The LTP cycle is completed by potentiating the first and fifth rows, as depicted in **Figure 5.11 (d)**. This step constitutes a parallel crossbar update, where multiple lines and columns are simultaneously addressed.
Next, we perform LTD on the elements from the second, third and fourth column requiring to be updated in the negative direction, as shown in **Figure 5.11 (e)**. This also constitutes a parallel update of the crossbar.
A similar procedure is repeated for the letters *"B"* nad *"M"*.
The results from the LTP cycles performed to map the letter *"I"* are displayed in **Figure 5.12**. Each quadrant shows the heat map of the array conductances measured after every pulse stream. Initially, the array elements are set to random states (top-left quadrant). The second column (circled in purple) is gradually potentiated in analog steps (**Figure 5.12 (a)**) and then we shift the target to rows number 1 and 5 (**Figure 5.12 (b)**). The latter is an example of parallel crossbar updates, with simultaneous programming pulses applied to rows 1 and 5, and columns 1 and 3. Finally, we perform an LTD cycle on rows number 2, 3 and 4 (not shown). The final heatmap is displayed in the top-right quadrant of **Figure 5.13**, with the color map rescaled to non-dimensional units called "Weights" (W).

Figure 5.13 shows the temporal evolution of the experiments involving all the three *IBM* letters. In every experiment, we demonstrated parallel and analog updates in both potentiation and depression directions. Most targeted cells are updated with just a few ($< 5\times$) pulses; though achieving high contrast between highly-potentiated and strongly-depressed cells in the whole array required more programming cycles ($> 20\times$). This discrepancy can be attributed to device-to-device variability and will be addressed in future works.
To summarize, this experiment demonstrated the parallel and analog crossbar arrays towards binary conductance states, representing the letters *I*, *B* and *M*. The binary states were achieved through deterministic pulse streams, until a high contrast (G_{max} - G_{min}) was achived. This process significantly differs from backpropagation, which requires performing the outer product $\boldsymbol{x} \otimes \delta^{(i)}$ in $4\times$ stochastic programming cycles of fixed length, delivering accurate array updates (see Chapter 1, Section 1.6).
Nevertheless, our experiment successfully demonstrated parallel and analog crossbar array updates, which is a fundamental prerequisite for the acceleration of deep learning workloads with resistive-arrays.

5.6 Conclusion

In this Chapter, we described the integration of our optimized ReRAM devices on a 350 nm CMOS technology. The 1T1R devices were carefully characterized, exploring different operational regimes. We found that our ReRAM technology exhibits fast resistive-switching properties (<1 μs), even when operating at very low conductances (10, 50 nS), while also preserving analog properties. Further research is required to demonstrate the reproducibility of such operations.
We extended our measurements, from stand-alone 1T1R devices, to the crossbar arrays. First, we studied the conventional pulsing scheme for implementing the backpropagation algorithm (*Half-Bias* scheme). We proposed a new method for reducing the disturbance effect caused by single pulses occurring at the non-targeted cells. The new scheme reduces the disturbance to zero.
Finally, we demonstrated that our technology enables parallel and analog crossbar array updates. Further tests are required to evaluate the achievable resolution (in number of bits). Next, we will test the crossbar arrays for small-scale, real-world applications, involving the backpropagation algorithm.

Chapter 6

Conclusions and outlook

This chapter summarizes the main scientific achievements presented in this dissertation and provides an outlook for future research advancements.

6.1 Conclusions

In this dissertation, we address three significant challenges in the field of deep learning acceleration with resistive-switching devices. The first involves material and device optimization to establish a ReRAM technology with excellent analog resistive-switching properties. The second challenge is scaling down the device geometry and operational parameters, such as forming voltage and read/program currents, to enable integration into advanced electronic platforms and demonstrate high efficiency in real-world applications. Finally, we worked towards realizing small-scale prototypes of resistive-switching accelerators to demonstrate the potential of our optimized technology at the system level. We achieved parallel and analog crossbar updates, fulfilling the requirements for training neural networks with resistive processors.
The results presented in this dissertation are structured into three Chapters:

- **Materials optimization for analog ReRAM devices:** ReRAM devices designed with a baseline structure of metal/insulator/metal exhibit stochastic and abrupt resistive-switching properties. Reduced stochasticity and analog switching properties can be achieved by replacing the conventional design with a metal/insulating-metal-oxide/conductive-metal-oxide/metal stack [128].
  The physical explanation for the upgraded properties is still debated, leading to an unclear path on how to further optimize the materials and the devices to achieve even greater properties.
  To address this issue, we focused on the conductive-metal-oxide layer, which imparts the analog properties to ReRAM devices, and we created multiple device variations with different material properties.
  Firstly, we developed a DC sputtering recipe for fabricating conductive-TaO_x thin films. We determined the processing parameters to gradually tune the resistivity of conductive-TaO_x thin films from approximately 0.01 to 0.5 Ω·cm. These TaO_x layers were integrated into device structures comprising a $TiN/HfO_2/TaO_x/TiN$ material stack, creating a metal/insulating-metal oxide/conductive-metal-oxide/metal ReRAM device.
  We tested several variations, adjusting the TaO_x thickness and resistivity. The best performance, characterized by low conductances, large on/off ratios, and analog updates, was observed in the devices with a thin TaO_x layer (20 nm thick) and a higher resistivity (0.25 Ω·cm).
  We continued our device optimization work by addressing the challenge of scaling down the

ReRAM device geometry and operational currents/voltages, critical for compatibility with advanced CMOS nodes.
We developed a process flow enabling high precision patterning of sub-μm devices, demonstrating reproducible results down to geometries as small as 200 nm $\times$ 200 nm.
We modified our ALD deposition process for the HfO_x layer, targeting an increase in defect density and a reduction in thickness. This led to a significant decrease in the filament forming voltage for the $(200\ \text{nm})^2$ devices, from approximately 5 V, to about 3.3 V. Moreover, the lower forming voltages led to a softer filament forming process, enabling the device to operate at higher resistances up to few MΩ. High resistances are functional for reducing power consumption in both inference and training workloads.

- **Device qualification for training neural networks with the novel Tiki-Taka algorithm:** We designed and conducted a set of electrical measurements to test device operations and reliability for real-world applications. The Tiki-Taka algorithm mitigates the linearity and symmetry requirements of resistive-switching devices used for training applications, but it requires them to operate around the *symmetry point*, an equilibrium state reached when alternating programming pulses of opposite polarity.
  We conducted statistical studies of the analog properties exhibited around the symmetry point, which enabled us to create a mathematical model of the device response accounting for device-to-device variability. This model enabled realistic simulations of training workloads. The simulation results indicate that our optimized ReRAM devices can process neural network training workloads for image classification and character prediction tasks, yielding minimal accuracy drop compared to baseline digital processors.
  Regarding reliability aspects, we conducted retention measurements at high temperature (85° C), which resulted in negligible relaxation and drift processes, and endurance measurements of the analog properties (up to 10^8 programming pulses).

- **Back-End-Of-Line (BEOL) integration and crossbar array measurements:** The integration of the ReRAM devices onto CMOS integrated circuits, creating 1T1R structures and crossbar array architectures, is required to demonstrate functional crossbar arrays at large scale, as needed for real-world applications. We led a joint project involving multiple partner institutions, realizing a chip layout (without the CMOS components) and fabricating the ReRAM devices above the integrated-circuit substrates provided by X-FAB.
  In the 1T1R devices, the transistor accurately controls the ReRAM current, enabling the formation of conductive filaments with currents as low as 1 μA. This soft forming process enabled us to demonstrate analog synaptic characteristics in the ultra-low conductive range of (5, 40) nS.
  The 1T1R devices were tested under the conventional *Half-Bias (HB)* scheme for parallel array updates, exhibiting excellent analog properties. To suppress the disturbance effect of HB pulses on non-selected devices, we proposed an alternative parallel updates method. This new method eliminates the disturbance while preserving the analog switching properties. We tested the parallel and analog update functionality in a 5×4 crossbar array using a customized pulsing scheme. This test qualitatively showed that our technology enables the desired functionality. A quantitative assessment of the updates accuracy requires testing the crossbar under some sort of task where an output error is defined.

6.2 Outlook

The methods and the results described in this dissertation represent a foundational step in the research on resistive processors for deep learning acceleration. To further advance our resistive technology and demonstrate a processor capable of handling deep learning workloads, several key aspects still need to be addressed:

1. **Improve the physical understanding of device properties:** We performed electrical measurements and structural characterizations on devices based on different material stack to provide a speculative interpretation of the resistive-switching mechanisms. To enhance our understanding, we need to complement the experiments on materials and devices with physical modeling, additional microscopy and spectroscopy measurements, and noise analysis:

   - Finite-element-method and analytical device models: These models will provide a quantitative description of the analog resistive-switching mechanisms and offer explanations for failure mechanisms such as limited programming endurance. Additionally, modeling work will guide further optimizations of the ReRAM stack, enhancing device properties.
     Significant contributions by Falcone et al. [120, 127] have modeled the spreading resistance role of the conductive-Metal-Oxide material, which is critical for analog properties. Additionally, an analytical description for both electronic and ionic current mechanisms was provided, delivering a deeper understanding of static and dynamic processes in the devices. Future work must complement the physical model with the dynamical response to programming pulses across different timescales. This will deliver a more complete description of the resistive-switching properties, leading to the first compact model of analog ReRAMs to be used for integrated circuit simulations.

   - Local electronic microscopy and spectroscopy characterization: This is a promising research path for improving the undestanding on the physical device properties. For instance, performing TEM and EDS experiments on the ReRAM devices exhibiting local delamination issues could reveal valuable information on the structural and chemical properties of the conductive filament.
   - Noise measurements and analysis: In Chapter 4 we presented relaxation and retention measurements, monitoring the evolution of the conductive states at different timescales, with sampling intervals of 500 μs and one second. Conducting a frequency analysis of the conductive states' fluctuations will provide a noise spectral profile, which can offer valuable insights into the filament structure and the type of current transport occurring within the device [177].

2. **Conduct parallel crossbar updates implementing the backpropagation algorithm:** Our ReRAM technology enabled parallel crossbar updates through analog steps, demonstrated by applying deterministic programming pulses (probability = 1) coinciding at two rows and two columns of a crossbar array.
   To show a real-world application based on the parallel updates concept, we need to extend these measurements to deep learning workloads, where the probability of the pulsing activity is dictated by the backpropagation algorithm.
   In such an experiment, we will analyze the resolution of the weight updates and the processing efficiency. By comparing these performances to a digital accelerator such as a GPU, we will demonstrate the impact of our achievements in ReRAM technology development.

List of publications

1. Choi, W.*, **Stecconi, T.***, ... & Offrein, B.
   Write-Disturb-Tolerant, Fully Parallel Stochastic Update of 1T-1ReRAM (1T1R) Crossbar Array for Analog Deep Learning Accelerators. Manuscript in preparation. * = shared first author

2. **Stecconi, T.**, Bragaglia, V., Rasch, M. J., Carta, F., Horst, F., Falcone, D. F., ... & Offrein, B. (2024). Analog Resistive Switching Devices for Training Deep Neural Networks with the Novel Tiki-Taka Algorithm. *Nano Letters, 24(3), 866-872.*

3. Falcone, D. F., Menzel, S., **Stecconi, T.**, Galetta, M., La Porta, A., Offrein, B. J., Bragaglia, V. (2024). Analytical modelling of the transport in analog filamentary conductive-metal-oxide/HfO x ReRAM devices. *Nanoscale Horizons, 9(5), 775-784.*

4. **Stecconi, T.**, Guido, R., Berchialla, L., La Porta, A., Weiss, J., Popoff, Y., ... & Bragaglia, V. (2022). Filamentary TaOx/HfO2 ReRAM Devices for Neural Networks Training with Analog In-Memory Computing. *Advanced electronic materials, 8(10), 2200448.*

Conference talks

1. Falcone, D. F., Menzel, S., **Stecconi, T.**, La Porta, A., Carraria-Martinotti, L., Offrein, B. J., & Bragaglia, V. (2023, May). Physical modeling and design rules of analog Conductive Metal Oxide-HfO 2 ReRAM. In 2023 *IEEE International Memory Workshop (IMW) (pp. 1-4). IEEE.*

2. **Stecconi, T.**, Falcone, D. F., Popoff, Y., Horst, F., Offrein, B. J., & Bragaglia, V. (2023, April). Conductive-TaOx/HfOx Based RRAM Devices with Enhanced Properties for Training Deep Neural Networks. In: *Materials Research Society (MRS) Spring Meeting.*

3. Falcone, D. F., Popoff, Y., **Stecconi, T.**, Bragaglia, V., Sousa, M., Horst, F., ... & Offrein, B. J. (2023, April). Optimization of materials and interfaces for low power analog Conductive Metal Oxide/HfO2 ReRAM artificial synapses. *Materials Research Society (MRS) Spring Meeting.*

4. Bragaglia, V., Syed, G. S., Jonnalagadda, V. P., Kersting, B., Luchtenveld, J., **Stecconi, T.**, ... & Offrein, B. J. (2023, April). Mitigating Analog Memory Non-Idealities for In-Memory Inference and Learning of DNN. In *Materials Research Society (MRS) Spring Meeting.*

5. Galetta, M., Falcone, D. F., Menzel, S., La Porta, A., **Stecconi, T.**, ... & Bragaglia, V. (2024, September). Compact Model of Conductive-Metal-Oxide/HfOx Analog Filamentary ReRAM Devices. In 2024 *European Solid-State Electronics Research Conference (ESSDERC).*

6. Hohenheim, J., **Stecconi, T.**, Liu, Z. D., & Palopoli, P. (2022, November). A Trainable Sequence Learner that Learns and Recognizes Two-Input Sequence Patterns. In: *TENCON 2022-2022 IEEE Region 10 Conference (TENCON) (pp. 1-5). IEEE.*

7. **Stecconi, T.**, Popoff, Y., Guido, R., Falcone, D., Halter, M., Sousa, M., ... & Bragaglia, V. (2022, September). Role of Conductive-Metal-Oxide to HfO x, Interfacial Layer on the Switching Properties of Bilayer TaO x/HfO x ReRAM. In: *ESSDERC 2022-IEEE 52nd European Solid-State Device Research Conference (ESSDERC) (pp. 297-300). IEEE.*

8. **Stecconi, T.**, Popoff, Y., Guido, R., Halter, M., Falcone, D., La Porta, A., ... & Bragaglia, V. (2022, June). Equivalent electrical circuit modelling of a TaO x/HfO x based RRAM with optimized resistance window and multilevel states. In: *2022 Device Research Conference (DRC) (pp. 1-2). IEEE.*

9. **Stecconi, T.**, Popoff, Y., Falcone, D. F., Guido, R., La Porta, A., Horst, F., ... & Bragaglia, V. (2022, May). Filamentary-based TaOx/HfO2 memristive synapses. In: *European Materials Research Society (e-MRS) Spring Meeting.*

10. **Stecconi, T.**, Popoff, Y., Falcone, D. F., Guido, R., La Porta, A., Horst, F., ... & Bragaglia, V. (2022, May). Scalable Conductive Metal-Oxide/HfO2-Based Bilayer ReRAMs for Analog In-Memory Computing. In: *Materials Research Society (MRS) Spring Meeting.*

11. Covi, E., Lin, Y. H., Wang, W., **Stecconi, T.**, Milo, V., Bricalli, A., ... & Ielmini, D. (2019, November). A volatile RRAM synapse for neuromorphic computing. In: *2019 26th IEEE International Conference on Electronics, Circuits and Systems (ICECS) (pp. 903-906). IEEE.*

List of abbreviations

- Artificial Intelligence (AI)
- Deep Neural Network (DNN)
- Convolutional Neural Networks (CNN)
- Back-End-Of-the-Line (BEOL)
- Front-End-Of-the-Line (FEOL)
- Resistive Processing Units (RPUs)
- Neural Networks (NN)
- Conductive-Metal-Oxide (CMO)
- Resistive-switching Random Access Memory (ReRAM), (RRAM)
- Complementary-Metal-Oxide-Semiconductor (CMOS)
- 1-Transistor-1-ReRAM (1T1R)
- Vector-Matrix Multiplications (VMM)
- Gate-All-Around Field-Effect-Transistor (GAAFET)
- International Roadmap for Devices and Systems (IRDS)
- International Electron Devices Meeting (IEDM)
- In-Memory Computing (IMC)
- Internet of Things (IoT)
- Fully-Connected Network (FCN)
- Multiply-Accumulate (MAC)
- Graphics Processing Unit (GPU)
- Central Processing Unit (CPU)
- Streaming Multiprocessors (SM)
- Dynamic Random-Access Memory (DRAM)

- High Bandwidth Memory (HBM)
- Extreme Ultra-Violet (EUV)
- Static Random-Access Memory (SRAM)
- Figures of Merit (FoM)
- Tera-OPerationS per second [TOPS]
- Tera-OPerationS per second per Watt [TOPS/W]
- FinField-Effect-Transistor (FinFET)
- 1-Transistor-1-Capacitor (1T1C)
- Bit-Cost-Scalable (BiCS)
- Non-Volatile Memory (NVM)
- Ferroelectric Random-Access Memory (FeRAM)
- Field-Effect-Transistors (FeFET)
- Ferroelectric Tunnel Junction (FTJ)
- Phase Change Memory (PCM)
- Magnetic Tunnel Junction (MTJ)
- Storage Class Memory (SCM)
- Pulse-Width-Modulation (PWM)
- current-voltage (I-V)
- Conductance (G)
- Metal-Insulator-Metal (MIM)
- Electro-Chemical Random-Access-Memory (ECRAM)
- Symmetry Point (SP)
- Tiki-Taka (TT)
- Tiki-Taka version 2 (TTv2)
- Transition Metal Oxide (TMO)
- Oxygen vacancy ($V_O^{\bullet\bullet}$
- work functions (WF)
- Finite Element Method (FEM)
- State-Of-The-Art (SOTA)

- Direct Current (DC)
- Radio-Frequency (RF)
- Atomic-Layer-Deposition (ALD)
- Reactive-Ion-Etching (RIE)
- Inductive-Coupled-Plasma (ICP)
- Plasma-Enhanced ALD (PEALD)
- tetra(ethylmethylamido)hafnium (TEMAH)
- ethylmethylamido (NEtMe)
- hydroxyl (OH)
- hafnia (HfO_2)
- Scanning-Electron-Microscopy (SEM)
- Atomic-Force-Microscopy (AFM)
- Grazing-Incidence X-Ray Diffraction (GIXRD)
- X-Ray Reflectivity (XRR)
- Circular-Transmission-Line-Measurements (C-TLM)
- Graphical User Interface (GUI)
- Device Under Test (DUT)
- Surface Mount Resistor (SMD)
- Arbitrary Waveform Generator (AWG)
- Pulse Measure Unit (PMU)
- Source Measure Unit (SMU)
- Plastic Leaded Chip Carriers (PLCC)
- tetrakis-(dimethylamino)titanium (TDMAT)
- Top Electrode (TE)
- Bottom Electrode (BE)
- Electron-Dispersive-Spectroscopy (EDS)
- Scanning-TEM (STEM)
- High-Resistive-State (HRS)
- Low Resistive State (LRS)

- electric field (F)
- temperature (T)
- Post-forming resistance (R_{pf})
- Forming voltage (V_f), ($V_{forming}$)
- Oxygen anion (O^{2-})
- Transmission-Electron-Microscopy (TEM)
- Long-Term-Potentiation (LTP)
- Long-Term-Depression (LTD)
- Metal-Oxide (MO)
- Symmetry Point (SP)
- Zero-Shifting (ZS)
- Metal-Oxide-Semiconductor-Field-Effect-Transistor (MOSFET)
- Generation 1 (Gen_1)
- Generation 2 (Gen_2)
- Cumulative-Distribution-Functions (CDF)
- Noise-to-Signal Ratio (NSR)
- Focused-Ion-Beam (FIB)
- Modified National Institute of Standards and Technology (MNIST)
- Stochastic Gradient Descent (SGD)
- Floating Point (FP)
- Long-Short-Term-Memory (LSTM)
- Instituto de Microelectronica de Sevilla (IMSE)
- International Business Machines (IBM)
- n-type Field-Effect-Transistors (n-FETs)
- Body (B)
- Source (S)
- Gate (G)
- Drain (D)
- Metal 1 (M1)

- Metal 2 (M2)
- Metal 3 (M3)
- Half-Bias (HB)
- Ground (GND)

List of Figures

1.1 IRDS strategy on the future of semiconductor technologies for AI 2
1.2 Derivation of the deep learning formulas 3
1.3 GPU layout 4
1.4 Overview of the established silicon memory technologies that have been applied to IMC 6
1.5 Overview of the main emerging NVM devices for deep learning applications 8
1.6 Ilustration of the concept about utilizing crossbar arrays of resistive-switching devices as VMM and inference accelerators 10
1.7 Implementation example of the outer product operation through stochastic pulse streams applied at the RPU terminals 12
1.8 Implementation example of the Tiki-Taka algorithm 14
1.9 Electrodynamics concepts about the resistive-switching mechanisms in filamentary ReRAM devices, comparing a MIM structure to the M/CMO/TMO/M 17
1.10 Research status comparison between inference and training accelerators based on resistive-switching devices 20

2.1 DC sputtering setup, process description. Control in the TaO_x thin film stoichiometry through optimizing the reactive sputtering power. 22
2.2 HfO_2 PEALD cycle breakdown 23
2.3 Dry etching tools comparison: RIE vs ICP 24
2.4 From *mesa* to *via* approach: SEM images and AFM measurements 25
2.5 XRD and XRR setups, measurement and data analysis examples 27
2.6 CTLM on TaO_x thin films 28
2.7 Setup pulsed measurements 30
2.8 Array measurement setup 32

3.1 TaO_x GIXRD patterns and table of XRR-measured material properties 34
3.2 TEM image and Electron-Dispersive-Spectroscopy (EDS) analysis of ReRAM stack . . 36
3.3 Baseline vs. oxide *bilayer* ReRAM: DC I-V characteristics comparison 37
3.4 Interpretation of the resistive-switching measurements as nanoionic processes 38
3.5 Resistive states dependence on the device area 40
3.6 Impact of the TaO_x resistivity on the ReRAM electrical properties 41
3.7 Impact of the TaO_x thickness on the filament forming and resistive states of the ReRAM devices 42
3.8 Impact of the TaO_x thickness on the resistive-switching properties of the ReRAM devices 42
3.9 Comparison between two oxide *bilayer* ReRAM devices: filament forming and DC I-V characteristics 43
3.10 Comparison between two oxide *bilayer* ReRAM devices: pulsed electrical characteristics 44

4.1 SEM of (200 nm)^2-ReRAM device, optimized HfO_x material properties, and forming voltage comparison with previous ReRAM generation . . . 49
4.2 DC I-V characteristics and multistate programming of the scaled ReRAM devices . . . 50
4.3 Zero-shifting and symmetry point of the optimized ReRAM devices . . . 51
4.4 Device Figures of Merit (FoM) for the Tiki-Taka algorithm . . . 52
4.5 MO-ReRAM endurance to analog programming . . . 54
4.6 TaO_x-ReRAM endurance to analog programming . . . 55
4.7 Multibit programmability and retention . . . 56
4.8 Analog states long term retention at high temperature and short term relaxation . . . 57
4.9 Scanning-Electron-Microscopy (SEM) investigation of filamentary structures in the ReRAM devices . . . 58
4.10 Analog ReRAM model . . . 60
4.11 Neural network training simulations . . . 61
4.12 Benchmark table . . . 62

5.1 Overview of the ReRAM integration in the BEOL of a CMOS chip: wafer, chip design, layout, process integration check and description . . . 67
5.2 Layout for ReRAM integration on CMOS . . . 68
5.3 n-FET current-voltage characteristics (I_{ds} vs. V_{ds}) . . . 70
5.4 1T1R structural overview and DC characteristics . . . 72
5.5 Highly-resistive, analog device response to short programming pulses following low-current forming process . . . 73
5.6 1T1R updates driven by coincident pulses occurring at the TE and S terminals . . . 74
5.7 Measurements of stochastic ReRAM updates . . . 75
5.8 Measurement of the disturbance induced by half-bias programming pulses . . . 77
5.9 Proposed update scheme based on the wordline-sourceline pulse coincidence . . . 78
5.10 Comparison of disturbances between conventional and proposed update schemes . . . 79
5.11 Parallel crossbar updates - experiment description . . . 80
5.12 Examples of parallel crossbar array updates - letter "I" . . . 81
5.13 Multiple tests for the parallel crossbar array updates . . . 82

Bibliography

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks." [Online]. Available: http://code.google.com/p/cuda-convnet/

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 770–778, dec 2015. [Online]. Available: https://arxiv.org/abs/1512.03385v1

[3] A. Vaswani, G. Brain, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All you Need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[4] E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Modern Deep Learning Research," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 09, pp. 13 693–13 696, apr 2020. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/7123

[5] T. B. B. Danny Hernandez, "Measuring the Algorithmic Efficiency of Neural Networks," 2020. [Online]. Available: https://openai.com/research/ai-and-efficiency

[6] D. Amodei and D. Hernandez, "AI and compute." [Online]. Available: https://openai.com/blog/ai-and-compute/,2018

[7] L. Floridi and M. Chiriatti, "GPT-3: Its Nature, Scope, Limits, and Consequences," *Minds and Machines*, vol. 30, no. 4, pp. 681–694, dec 2020. [Online]. Available: https://link.springer.com/article/10.1007/s11023-020-09548-1

[8] S. Samsi, D. Zhao, J. McDonald, B. Li, A. Michaleas, M. Jones, W. Bergeron, J. Kepner, D. Tiwari, and V. Gadepally, "From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference," *2023 IEEE High Performance Extreme Computing Conference, HPEC 2023*, 2023.

[9] A. C. Elster, T. A. Haugdahl, V. Kindratenko, and A. Elster, "Nvidia Hopper GPU and Grace CPU Highlights," *Computing in Science and Engineering*, vol. 24, no. 2, pp. 95–100, 2022.

[10] W. Luo, R. Fan, Z. Li, D. Du, Q. Wang, and X. Chu, "Benchmarking and Dissecting the Nvidia Hopper GPU Architecture," feb 2024. [Online]. Available: https://arxiv.org/abs/2402.13499v1

[11] NVIDIA Corporation, "GP100 Pascal [Whitepaper]," *Whitepaper*, p. 45, 2017. [Online]. Available: https://www.nvidia.com/object/pascal-architecture-whitepaper.html{%}0Ahttps://www.nvidia.com/object/pascal-architecture-whitepaper.html

[12] Nvidia, "NVIDIA H100 Tensor Core GPU Architecture [White paper]," *Whitepaper*, 2023. [Online]. Available: https://resources.nvidia.com/en-us-tensor-core/gtc22-whitepaper-hopper

[13] H. Takato, K. Sunouchi, N. Okabe, A. Nitayama, K. Hieda, and F. Masuoka, "Impact of Surrounding Gate Transistor (SGT) for Ultra-High-Density LSI's," *IEEE Transactions on Electron Devices*, vol. 38, no. 3, pp. 573–578, 1991.

[14] S. V. Kalinin, D. A. Poteryaev, D. I. Ostertak, and M. A. Kuznetsov, "A New Revolution in Logic Silicon IC Technology: GAA FETs are Replacing FinFETs," *Proceedings of the 2023 IEEE 16th International Scientific and Technical Conference Actual Problems of Electronic Instrument Engineering, APEIE 2023*, pp. 160–165, 2023.

[15] K. J. Kuhn, A. Murthy, R. Kotlyar, and M. Kuhn, "(Invited) Past, Present and Future: SiGe and CMOS Transistor Scaling," *ECS Transactions*, vol. 33, no. 6, p. 3, oct 2010. [Online]. Available: https://iopscience.iop.org/article/10.1149/1.3487530https://iopscience.iop.org/article/10.1149/1.3487530/meta

[16] Q. Zhao, Z. Xie, Y. P. Peng, K. Wang, H. Wang, X. Li, H. Wang, J. Chen, H. Zhang, and X. Yan, "Current status and prospects of memristors based on novel 2D materials," *Materials Horizons*, vol. 7, no. 6, pp. 1495–1518, jun 2020. [Online]. Available: https://pubs.rsc.org/en/content/articlehtml/2020/mh/c9mh02033khttps://pubs.rsc.org/en/content/articlelanding/2020/mh/c9mh02033k

[17] N. Fu, Y. Liu, X. Ma, and Z. Chen, "EUV Lithography: State-of-the-Art Review," *Journal of Microelectronic Manufacturing*, vol. 2, no. 2, pp. 1–6, 2019. [Online]. Available: https://www.researchgate.net/publication/334136595{_}EUV{_}Lithography{_}State-of-the-Art{_}Review

[18] G. E. Moore, "Cramming More Components onto Integrated Circuits."

[19] I. R. Committee and Others, "The International Roadmap for Devices and Systems," 2022. [Online]. Available: https://irds.ieee.org/images/files/pdf/2022/2022IRDS{_}MM.pdf

[20] R. H. Dennard, F. H. Gaensslen, H. N. Yu, V. L. Rideout, E. Bassous, and A. R. Leblanc, "Design of Ion-Implanted MOSFET's With Very Small Physical Dimensions," *IEEE Journal of Solid-State Circuits*, vol. 9, no. 5, pp. 256–268, 1974.

[21] M. M. Shulaker, T. F. Wu, M. M. Sabry, H. Wei, H. S. Wong, and S. Mitra, "Monolithic 3D integration: A path from concept to reality," *Proceedings -Design, Automation and Test in Europe, DATE*, vol. 2015-April, pp. 1197–1202, 2015.

[22] S. Datta, S. Dutta, B. Grisafe, J. Smith, S. Srinivasa, and H. Ye, "Back-End-of-Line Compatible Transistors for Monolithic 3-D Integration," *IEEE Micro*, vol. 39, no. 6, pp. 8–15, 2019.

[23] M. Koyanagi, "Heterogeneous 3D integration - Technology enabler toward future super-chip," *Technical Digest - International Electron Devices Meeting, IEDM*, 2013.

[24] G. Van Der Plas and E. Beyne, "Design and Technology Solutions for 3D Integrated High Performance Systems," *IEEE Symposium on VLSI Circuits, Digest of Technical Papers*, vol. 2021-June, jun 2021.

[25] S. Das, A. Chen, and M. Marinella, "Beyond cmos," in *2021 IEEE International Roadmap for Devices and Systems Outbriefs*. IEEE, 2021, pp. 1–129.

[26] M. Marinella, S. Agarwal, A. Chen, and S. Das, "Beyond CMOS and Emerging Research Materials IRDS Panel IEDM 2020." 2020.

[27] A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," *IEEE Access*, vol. 7, pp. 53 040–53 065, 2019.

[28] C. A. Goodfellow lan, Bengio Yoshua, "Deep Learning - Ian Goodfellow, Yoshua Bengio, Aaron Courville - Google Books," 2016.

[29] Z. Lin, R. Memisevic, and K. Konda, "How far can we go without convolution: Improving fully-connected networks," nov 2015. [Online]. Available: https://arxiv.org/abs/1511.02580v1

[30] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2323, 1998.

[31] Q. Liu and Y. Wu, "Supervised Learning," *Encyclopedia of the Sciences of Learning*, pp. 3243–3245, 2012. [Online]. Available: https://link.springer.com/referenceworkentry/10.1007/978-1-4419-1428-6{_}451

[32] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature 1986 323:6088*, vol. 323, no. 6088, pp. 533–536, 1986. [Online]. Available: https://www.nature.com/articles/323533a0

[33] G. Batra, Z. Jacobson, S. Madhav, A. Queirolo, and N. Santhanam, "Artificial-intelligence hardware: New opportunities for semiconductor companies," *McKinsey and Company, January*, vol. 2, 2019.

[34] W. Jeon, G. Ko, J. Lee, H. Lee, D. Ha, and W. W. Ro, "Deep learning with GPUs," *Advances in Computers*, vol. 122, pp. 167–215, jan 2021.

[35] NVIDIA, "NVIDIA A100 tensor core GPU architecture," *Volume 1.0: Whitepaper*, vol. 1, no. 2020, p. 82, 2020.

[36] V. Mehta and M. Milakov, "Optimizing CUDA Applications for the Volta/Turing Architecture," in *GPU Techonology Conference Israel (GTC Israel)*, 2018.

[37] K. C. Chun, Y. K. Kim, Y. Ryu, J. Park, C. S. Oh, Y. Y. Byun, S. Y. Kim, D. H. Shin, J. G. Lee, B. K. Ho, M. S. Park, S. J. Cho, S. Woo, B. M. Moon, B. Kil, S. Ahn, J. H. Lee, S. Y. Kim, S. K. Choi, J. S. Jeong, S. G. Ahn, J. Kim, J. J. Kong, K. Sohn, N. S. Kim, and J. B. Lee, "A 16-GB 640-GB/s HBM2E DRAM with a Data-Bus Window Extension Technique and a Synergetic On-Die ECC Scheme," *IEEE Journal of Solid-State Circuits*, vol. 56, no. 1, pp. 199–211, jan 2021.

[38] A. Arunkumar, E. Bolotin, B. Cho, U. Milic, E. Ebrahimi, O. Villa, A. Jaleel, C. J. Wu, and D. Nellans, "MCM-GPU: Multi-chip-module GPUs for continued performance scalability," *Proceedings - International Symposium on Computer Architecture*, vol. Part F1286, pp. 320–332, jun 2017.

[39] T. Groves, B. Brock, Y. Chen, K. Z. Ibrahim, L. Oliker, N. J. Wright, S. Williams, and K. Yelick, "Performance Trade-offs in GPU Communication: A Study of Host and Device-initiated Approaches," *Proceedings of PMBS 2020: Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems, Held in conjunction with SC 2020: The*

International Conference for High Performance Computing, Networking, Storage and Analysis, pp. 126–137, nov 2020.

[40] M. Horowitz, "1.1 Computing's energy problem (and what we can do about it)," in *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, vol. 57, 2014.

[41] M. A. Zidan, J. P. Strachan, and W. D. Lu, "The future of electronics based on memristive systems," *Nature Electronics 2017 1:1*, vol. 1, no. 1, pp. 22–29, jan 2018. [Online]. Available: https://www.nature.com/articles/s41928-017-0006-8

[42] W. A. Wulf and S. A. McKee, "Hitting the Memory Wall: Implications of the Obvious," *Computer Architecture News*, vol. 23, no. 1, 1995.

[43] M. Ishida, "APX (Advanced Package X)-Advanced organic technology for 2.5 D interposer," in *2014 CPMT Seminar, Latest Advances in Organic Interposers*, 2014, pp. 27–30.

[44] "IRDS 2022: Lithography - IEEE IRDS." [Online]. Available: https://irds.ieee.org/editions/2022/irds-2022-lithography

[45] "NVIDIA Blackwell Architecture Technical Overview." [Online]. Available: https://resources.nvidia.com/en-us-blackwell-architecture

[46] "AMD INSTINCT MI300X ACCELERATOR Leading-Edge, industry-standard accelerator module for generative AI, training, and high-performance computing Leading-Edge Discrete GPU for AI and HPC," 2023. [Online]. Available: https://www.top500.org/lists/top500/2023/11/

[47] S. Naffziger, N. Beck, T. Burd, K. Lepak, G. H. Loh, M. Subramony, and S. White, "Pioneering chiplet technology and design for the AMD EPYC and Ryzen processor families: Industrial product," *Proceedings - International Symposium on Computer Architecture*, vol. 2021-June, pp. 57–70, jun 2021.

[48] A. I. Vlasov, I. V. Gudoshnikov, V. P. Zhalnin, A. T. Kadyr, and V. A. Shakhnov, "Market for memristors and data mining memory structures for promising smart systems," *Entrepreneurship and Sustainability Issues*, vol. 8, no. 2, p. 98, 2020.

[49] S. Salahuddin, K. Ni, and S. Datta, "The era of hyper-scaling in electronics," *Nature Electronics 2018 1:8*, vol. 1, no. 8, pp. 442–450, aug 2018. [Online]. Available: https://www.nature.com/articles/s41928-018-0117-x

[50] S. Li, D. Niu, K. T. Malladi, H. Zheng, B. Brennan, and Y. Xie, "DRISA: A DRAM-based reconfigurable in-situ accelerator," *Proceedings of the Annual International Symposium on Microarchitecture, MICRO*, vol. Part F1312, pp. 288–301, oct 2017. [Online]. Available: https://dl.acm.org/doi/10.1145/3123939.3123977

[51] M. Technology and I. Jcooke, "NAND Flash 101: An Introduction to NAND Flash and How to Design it into Your Next Product." [Online]. Available: www.forward-insights.com/

[52] S. Yu, H. Jiang, S. Huang, X. Peng, and A. Lu, "Compute-in-Memory Chips for Deep Learning: Recent Trends and Prospects," *IEEE Circuits and Systems Magazine*, vol. 21, no. 3, pp. 31–56, jul 2021.

[53] T. Yamauchi, Y. Yamaguchi, T. Kono, and H. Hidaka, "Embedded flash technology for automotive applications," *Technical Digest - International Electron Devices Meeting, IEDM*, pp. 28.6.1–28.6.4, jan 2017.

[54] Q. Dong, M. E. Sinangil, B. Erbagci, D. Sun, W. S. Khwa, H. J. Liao, Y. Wang, and J. Chang, "A 351TOPS/W and 372.4GOPS Compute-in-Memory SRAM Macro in 7nm FinFET CMOS for Machine-Learning Applications," *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, vol. 2020-Febru, pp. 242–244, feb 2020.

[55] S. Xie, C. Ni, A. Sayal, P. Jain, F. Hamzaoglu, and J. P. Kulkarni, "16.2 eDRAM-CIM: Compute-In-Memory Design with Reconfigurable Embedded-Dynamic-Memory Array Realizing Adaptive Data Converters and Charge-Domain Computing," *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, vol. 64, pp. 248–250, feb 2021.

[56] G. Fredeman, D. W. Plass, A. Mathews, J. Viraraghavan, K. Reyer, T. J. Knips, T. Miller, E. L. Gerhard, D. Kannambadi, C. Paone, D. Lee, D. J. Rainey, M. Sperling, M. Whalen, S. Burns, R. R. Tummuru, H. Ho, A. Cestero, N. Arnold, B. A. Khan, T. Kirihata, and S. S. Iyer, "A 14 nm 1.1 Mb embedded DRAM macro with 1 ns access," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 1, pp. 230–239, jan 2016.

[57] J. Zhang, Z. Wang, and N. Verma, "A machine-learning classifier implemented in a standard 6T SRAM array," *IEEE Symposium on VLSI Circuits, Digest of Technical Papers*, vol. 2016-Septe, sep 2016.

[58] W. S. Khwa, J. J. Chen, J. F. Li, X. Si, E. Y. Yang, X. Sun, R. Liu, P. Y. Chen, Q. Li, S. Yu, and M. F. Chang, "A 65nm 4Kb algorithm-dependent computing-in-memory SRAM unit-macro with 2.3ns and 55.8TOPS/W fully parallel product-sum operation for binary DNN edge processors," *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, vol. 61, pp. 496–498, mar 2018.

[59] H. Valavi, P. J. Ramadge, E. Nestler, and N. Verma, "A Mixed-Signal Binarized Convolutional-Neural-Network Accelerator Integrating Dense Weight Storage and Multiplication for Reduced Data Movement," *IEEE Symposium on VLSI Circuits, Digest of Technical Papers*, vol. 2018-June, pp. 141–142, oct 2018.

[60] X. Si, J. J. Chen, Y. N. Tu, W. H. Huang, J. H. Wang, Y. C. Chiu, W. C. Wei, S. Y. Wu, X. Sun, R. Liu, S. Yu, R. S. Liu, C. C. Hsieh, K. T. Tang, Q. Li, and M. F. Chang, "24.5 A Twin-8T SRAM Computation-In-Memory Macro for Multiple-Bit CNN-Based Machine Learning," *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, vol. 2019-Febru, pp. 396–398, mar 2019.

[61] H. Mori, W. C. Zhao, C. E. Lee, C. F. Lee, Y. H. Hsu, C. K. Chuang, T. Hashizume, H. C. Tung, Y. Y. Liu, S. R. Wu, K. Akarvardar, T. L. Chou, H. Fujiwara, Y. Wang, Y. D. Chih, Y. H. Chen, H. J. Liao, and T. Y. J. Chang, "A 4nm 6163-TOPS/W/b 4790-TOPS/mm2/b SRAM Based Digital-Computing-in-Memory Macro Supporting Bit-Width Flexibility and Simultaneous MAC and Weight Update," *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, vol. 2023-Febru, pp. 132–134, 2023.

[62] Nvidia, "NVIDIA TURING GPU ARCHITECTURE Graphics Reinvented NVIDIA Turing GPU Architecture."

[63] T. Song, H. Kim, W. Rim, Y. Kim, S. Park, C. Park, M. Hong, G. Yang, J. Do, J. Lim, S. Lee, I. Kim, S. Baek, J. Jung, D. Ha, H. Jang, T. Lee, C. H. Park, B. Kwon, H. Jung, S. Cho, Y. Choo, and J. S. Choi, "A 7nm FinFET SRAM macro using EUV lithography for peripheral repair analysis," *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, vol. 60, pp. 208–209, mar 2017.

[64] G. Yeap, X. Chen, B. R. Yang, C. P. Lin, F. C. Yang, Y. K. Leung, D. W. Lin, C. P. Chen, K. F. Yu, D. H. Chen, C. Y. Chang, S. S. Lin, H. K. Chen, P. Hung, C. S. Hou, Y. K. Cheng, J. Chang, L. Yuan, C. K. Lin, C. C. Chen, Y. C. Yeo, M. H. Tsai, Y. M. Chen, H. T. Lin, C. O. Chui, K. B. Huang, W. Chang, H. J. Lin, K. W. Chen, R. Chen, S. H. Sun, Q. Fu, H. T. Yang, H. L. Shang, H. T. Chiang, C. C. Yeh, T. L. Lee, C. H. Wang, S. L. Shue, C. W. Wu, R. Lu, W. R. Lin, J. Wu, F. Lai, P. W. Wang, Y. H. Wu, B. Z. Tien, Y. C. Huang, L. C. Lu, J. He, Y. Ku, J. Lin, M. Cao, T. S. Chang, S. M. Jang, H. C. Lin, Y. C. Peng, J. Y. Sheu, and M. Wang, "5nm CMOS Production Technology Platform featuring full-fledged EUV, and High Mobility Channel FinFETs with densest 0.021μm2 SRAM cells for Mobile SoC and High Performance Computing Applications," *Technical Digest - International Electron Devices Meeting, IEDM*, vol. 2019-Decem, dec 2019.

[65] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," pp. 7464–7475, jul 2022. [Online]. Available: https://arxiv.org/abs/2207.02696v1

[66] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, sep 2014. [Online]. Available: https://arxiv.org/abs/1409.1556v6

[67] I. Bhati, M. T. Chang, Z. Chishti, S. L. Lu, and B. Jacob, "DRAM Refresh Mechanisms, Penalties, and Trade-Offs," *IEEE Transactions on Computers*, vol. 65, no. 1, pp. 108–121, jan 2016.

[68] A. Teman, P. Meinerzhagen, A. Burg, and A. Fish, "Review and classification of gain cell eDRAM implementations," *2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel, IEEEI 2012*, 2012.

[69] S. Kim, Z. Li, S. Um, W. Jo, S. Ha, J. Lee, S. Kim, D. Han, and H. J. Yoo, "16.5 DynaPlasia: An eDRAM In-Memory-Computing-Based Reconfigurable Spatial Accelerator with Triple-Mode Cell for Dynamic Resource Switching," *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, vol. 2023-Febru, pp. 256–258, 2023.

[70] S. Kim and H. J. Yoo, "An Overview of Computing-in-Memory Circuits With DRAM and NVM," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 71, no. 3, pp. 1626–1631, mar 2024.

[71] Z. Chen, X. Chen, and J. Gu, "15.3 A 65nm 3T Dynamic Analog RAM-Based Computing-in-Memory Macro and CNN Accelerator with Retention Enhancement, Adaptive Analog Sparsity and 44TOPS/W System Energy Efficiency," *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, vol. 64, pp. 240–242, feb 2021.

[72] S. Ha, S. Kim, D. Han, S. Um, and H. J. Yoo, "A 36.2 dB High SNR and PVT/Leakage-Robust eDRAM Computing-In-Memory Macro With Segmented BL and Reference Cell Array," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 69, no. 5, pp. 2433–2437, may 2022.

[73] S. Kim, S. Um, W. Jo, J. Lee, S. Ha, Z. Li, and H. J. Yoo, "Scaling-CIM: An eDRAM-based In-Memory-Computing Accelerator with Dynamic-Scaling ADC for SQNR-Boosting and Layer-wise Adaptive Bit-Truncation," *Digest of Technical Papers - Symposium on VLSI Technology*, vol. 2023-June, 2023.

[74] S. Yu, "Semiconductor Memory Devices and Circuits," *Semiconductor Memory Devices and Circuits*, apr 2022. [Online]. Available: https://www.taylorfrancis.com/books/mono/10.1201/9781003138747/semiconductor-memory-devices-circuits-shimeng-yu

[75] H. Tanaka, M. Kido, K. Yahashi, M. Oomura, R. Katsumata, M. Kito, Y. Fukuzumi, M. Sato, Y. Nagata, Y. Matsuoka, Y. Iwata, H. Aochi, and A. Nitayama, "Bit Cost Scalable technology with and plug process for ultra high density flash memory," *Digest of Technical Papers - Symposium on VLSI Technology*, pp. 14–15, 2007.

[76] W. Shim, H. Jiang, X. Peng, and S. Yu, "Architectural Design of 3D NAND Flash based Compute-in-Memory for Inference Engine," *ACM International Conference Proceeding Series*, pp. 77–85, sep 2020. [Online]. Available: https://dl.acm.org/doi/10.1145/3422575.3422779

[77] W. Shim and S. Yu, "Technological Design of 3D NAND-Based Compute-in-Memory Architecture for GB-Scale Deep Neural Network," *IEEE Electron Device Letters*, vol. 42, no. 2, pp. 160–163, feb 2021.

[78] H. T. Lue, K. C. Wang, and C. Y. Lu, "Introduction of 3D AND-type flash memory and it's applications to computing-in-memory (CIM)," *VLSI-TSA 2021 - 2021 International Symposium on VLSI Technology, Systems and Applications, Proceedings*, apr 2021.

[79] H. T. Lue, G. R. Lee, T. H. Yeh, T. H. Hsu, C. R. Lo, C. L. Sung, W. C. Chen, C. T. Huang, K. Y. Shen, M. Y. Wu, P. Tseng, M. F. Hung, C. J. Chiu, K. Y. Hsieh, K. C. Wang, and C. Y. Lu, "3D AND: A 3D stackable flash memory architecture to realize high-density and fast-read 3D NOR flash and storage-class memory," *Technical Digest - International Electron Devices Meeting, IEDM*, vol. 2020-Decem, pp. 6.4.1–6.4.4, dec 2020.

[80] M. L. Wei, H. T. Lue, S. Y. Ho, Y. P. Lin, T. H. Hsu, C. C. Hsieh, Y. C. Li, T. H. Yeh, S. H. Chen, Y. H. Jhu, H. P. Li, H. W. Hu, C. H. Hung, K. C. Wang, and C. Y. Lu, "Analog Computing in Memory (CIM) Technique for General Matrix Multiplication (GEMM) to Support Deep Neural Network (DNN) and Cosine Similarity Search Computing using 3D AND-type NOR Flash Devices," *Technical Digest - International Electron Devices Meeting, IEDM*, vol. 2022-Decem, pp. 3331–3334, 2022.

[81] A. Khakifirooz, E. Anaya, S. Balasubrahmanyam, G. Bennett, D. Castro, J. Egler, K. Fan, R. Ferdous, K. Ganapathi, O. Guzman, C. W. Ha, R. Haque, V. Harish, M. Jalalifar, O. W. Jungroth, S. T. Kang, G. Karbasian, J. Y. Kim, S. Li, A. S. Madraswala, S. Maddukuri, A. Mohammed, S. Mookiah, S. Nagabhushan, B. Ngo, D. Patel, S. K. Poosarla, N. V. Prabhu, C. Quiroga, S. Rajwade, A. Rahman, J. Shah, R. S. Shenoy, E. T. Menson, A. Tankasala, S. K. Thirumala, S. Upadhyay, K. Upadhyayula, A. Velasco, N. K. B. Vemula, B. Venkataramaiah, J. Zhou, B. M. Pathak, and P. Kalavade, "A 1.67Tb, 5b/Cell Flash Memory Fabricated in 192-Layer Floating Gate 3D-nand Technology and Featuring a 23.3 Gb/mm2 Bit Density," *IEEE Solid-State Circuits Letters*, vol. 6, pp. 161–164, 2023.

[82] W. Jung, H. Kim, D.-B. Kim, T.-H. Kim, N. Lee, D. Shin, M. Kim, Y. Rho, H.-J. Lee, Y. Hyun, J. Park, T. Kim, H. Kim, G. Lee, J. Lee, J. Jang, J. Park, S. Kim, S. C. Jeon, S. Kim, J.-H.

Song, M.-S. Kim, T. Lee, B.-K. Chun, T. Kim, Y. G. Lee, H. Lee, S. Lee, H. Lee, D. Cho, S.-W. Nam, Y. Kim, K. Yoon, Y. Lee, S. Kim, J. Hwang, R. Song, H. Jang, J. Son, H. Jeon, M. Lee, M. Lee, K. Kim, E. Lee, M. Lee, S. Jo, C. H. Kim, J. C. Park, K. Yun, S. Seol, J.-H. Cho, S. Lee, J.-Y. Lee, and S.-H. Hur, "13.3 A 280-Layer 1Tb 4b/cell 3D-NAND Flash Memory with a 28.5Gb/mm2 Areal Density and a 3.2GB/s High-Speed IO Rate," pp. 236–237, mar 2024.

[83] J. F. Scott, C. A. Araujo, and L. D. McMillan, "Ferroelectric memory applications," *Ultrasonics Symposium Proceedings*, vol. 1, pp. 299–308, 1989.

[84] H. Ishiwara, "Ferroelectric random access memories," *Journal of Nanoscience and Nanotechnology*, vol. 12, no. 10, pp. 7619–7627, oct 2012.

[85] E. Y. Tsymbal and H. Kohlstedt, "Tunneling across a ferroelectric," *Science*, vol. 313, no. 5784, pp. 181–183, jul 2006. [Online]. Available: https://www.science.org/doi/10.1126/science.1126230

[86] M. Wuttig and N. Yamada, "Phase-change materials for rewriteable data storage," *Nature Materials 2007 6:11*, vol. 6, no. 11, pp. 824–832, 2007. [Online]. Available: https://www.nature.com/articles/nmat2009

[87] S. Parkin, X. Jiang, C. Kaiser, A. Panchula, K. Roche, and M. Samant, "Magnetically engineered spintronic sensors and memory," *Proceedings of the IEEE*, vol. 91, no. 5, pp. 661–679, 2003.

[88] R. Waser, R. Dittmann, G. Staikov, and K. Szot, "Redox-Based Resistive Switching Memories - Nanoionic Mechanisms, Prospects, and Challenges," *Advanced Materials*, vol. 21, no. 25-26, pp. 2632–2663, jul 2009. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/adma.200900375https://onlinelibrary.wiley.com/doi/abs/10.1002/adma.200900375https://onlinelibrary.wiley.com/doi/10.1002/adma.200900375

[89] A. Sebastian, M. L. Gallo, G. W. Burr, S. Kim, M. BrightSky, and E. Eleftheriou, "Tutorial: Brain-inspired computing using phase-change memory devices," *Journal of Applied Physics*, vol. 124, no. 11, p. 111101, sep 2018. [Online]. Available: https://aip.scitation.org/doi/abs/10.1063/1.5042413

[90] L. Begon-Lours, M. Halter, D. D. Pineda, Y. Popoff, V. Bragaglia, A. L. Porta, D. Jubin, J. Fompeyrine, and B. J. Offrein, "A BEOL Compatible, 2-Terminals, Ferroelectric Analog Non-Volatile Memory," *2021 5th IEEE Electron Devices Technology and Manufacturing Conference, EDTM 2021*, apr 2021.

[91] M. Halter, L. Bégon-Lours, V. Bragaglia, M. Sousa, B. J. Offrein, S. Abel, M. Luisier, and J. Fompeyrine, "Back-End, CMOS-Compatible Ferroelectric Field-Effect Transistor for Synaptic Weights," *ACS Applied Materials & Interfaces*, vol. 12, no. 15, pp. 17 725–17 732, apr 2020. [Online]. Available: https://pubs.acs.org/doi/abs/10.1021/acsami.0c00877

[92] M. Abedin, N. Gong, K. Beckmann, M. Liehr, I. Saraf, O. Van der Straten, T. Ando, and N. Cady, "Material to system-level benchmarking of CMOS-integrated RRAM with ultra-fast switching for low power on-chip learning," *Scientific Reports 2023 13:1*, vol. 13, no. 1, pp. 1–10, sep 2023. [Online]. Available: https://www.nature.com/articles/s41598-023-42214-x

[93] D. Ielmini, "Resistive switching memories based on metal oxides: mechanisms, reliability and scaling," *Semiconductor Science and Technology*, vol. 31, no. 6, p. 063002, may 2016. [Online]. Available: https://iopscience.iop.org/article/10.1088/0268-1242/31/6/063002https://iopscience.iop.org/article/10.1088/0268-1242/31/6/063002/meta

[94] T. Stecconi, R. Guido, L. Berchialla, A. L. Porta, J. Weiss, Y. Popoff, M. Halter, M. Sousa, F. Horst, D. Dávila, U. Drechsler, R. Dittmann, B. J. Offrein, V. Bragaglia, T. Stecconi, R. Guido, L. Berchialla, A. L. Porta, J. Weiss, Y. Popoff, M. Halter, M. Sousa, F. Horst, D. Dávila, U. Drechsler, B. J. Offrein, V. Bragaglia, and R. Dittmann, "Filamentary TaOx/HfO2 ReRAM Devices for Neural Networks Training with Analog In-Memory Computing," *Advanced Electronic Materials*, vol. 8, no. 10, p. 2200448, oct 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/aelm.202200448https://onlinelibrary.wiley.com/doi/abs/10.1002/aelm.202200448https://onlinelibrary.wiley.com/doi/10.1002/aelm.202200448

[95] P. Yao, H. Wu, B. Gao, S. B. Eryilmaz, X. Huang, W. Zhang, Q. Zhang, N. Deng, L. Shi, H.-S. P. Wong, and H. Qian, "Face classification using electronic synapses," *Nature Communications 2017 8:1*, vol. 8, no. 1, pp. 1–8, may 2017. [Online]. Available: https://www.nature.com/articles/ncomms15199

[96] S. Kim, J. A. Ott, T. Ando, H. Miyazoe, V. Narayanan, J. Rozen, T. Todorov, M. Onen, T. Gokmen, D. Bishop, P. Solomon, K. T. Lee, M. Copel, and D. B. Farmer, "Metal-oxide based, CMOS-compatible ECRAM for Deep Learning Accelerator," *Technical Digest - International Electron Devices Meeting, IEDM*, vol. 2019-Decem, dec 2019.

[97] G. W. Burr, B. N. Kurdi, J. C. Scott, C. H. Lam, K. Gopalakrishnan, and R. S. Shenoy, "Overview of candidate device technologies for storage-class memory," *IBM Journal of Research and Development*, vol. 52, no. 4-5, pp. 449–464, 2008.

[98] B. Govoreanu, G. S. Kar, Y. Y. Chen, V. Paraschiv, S. Kubicek, A. Fantini, I. P. Radu, L. Goux, S. Clima, R. Degraeve, N. Jossart, O. Richard, T. Vandeweyer, K. Seo, P. Hendrickx, G. Pourtois, H. Bender, L. Altimime, D. J. Wouters, J. A. Kittl, and M. Jurczak, "10x10nm 2 Hf/HfO x crossbar resistive RAM with excellent performance, reliability and low-energy operation," *Technical Digest - International Electron Devices Meeting, IEDM*, 2011.

[99] S. Raoux, G. W. Burr, M. J. Breitwisch, C. T. Rettner, Y. C. Chen, R. M. Shelby, M. Salinga, D. Krebs, S. H. Chen, H. L. Lung, and C. H. Lam, "Phase-change random access memory: A scalable technology," *IBM Journal of Research and Development*, vol. 52, no. 4-5, pp. 465–479, 2008.

[100] D. Ha and H. S. Kim, "Prospective Innovation of DRAM, Flash, and Logic Technologies for Digital Transformation (DX) Era," *Digest of Technical Papers - Symposium on VLSI Technology*, vol. 2022-June, pp. 417–418, 2022.

[101] N. Chandrasekaran, N. Ramaswamy, and C. Mouli, "Memory technology: Innovations needed for continued technology scaling and enabling advanced computing systems," *Technical Digest - International Electron Devices Meeting, IEDM*, vol. 2020-Decem, pp. 10.1.1–10.1.8, dec 2020.

[102] A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, and E. Eleftheriou, "Memory devices and applications for in-memory computing," *Nature Nanotechnology 2020 15:7*, vol. 15, no. 7, pp. 529–544, mar 2020. [Online]. Available: https://www.nature.com/articles/s41565-020-0655-z

[103] R. Khaddam-Aljameh, M. Martemucci, B. Kersting, M. Le Gallo, R. L. Bruce, M. Brightsky, and A. Sebastian, "A Multi-Memristive Unit-Cell Array with Diagonal Interconnects for In-Memory Computing," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 68, no. 12, pp. 3522–3526, dec 2021.

[104] F. Aguirre, A. Sebastian, M. Le Gallo, W. Song, T. Wang, J. J. Yang, W. Lu, M. F. Chang, D. Ielmini, Y. Yang, A. Mehonic, A. Kenyon, M. A. Villena, J. B. Roldán, Y. Wu, H. H. Hsu, N. Raghavan, J. Suñé, E. Miranda, A. Eltawil, G. Setti, K. Smagulova, K. N. Salama, O. Krestinskaya, X. Yan, K. W. Ang, S. Jain, S. Li, O. Alharbi, S. Pazos, and M. Lanza, "Hardware implementation of memristor-based artificial neural networks," *Nature Communications 2024 15:1*, vol. 15, no. 1, pp. 1–40, mar 2024. [Online]. Available: https://www.nature.com/articles/s41467-024-45670-9

[105] S. Jain, H. Tsai, C. T. Chen, R. Muralidhar, I. Boybat, M. M. Frank, S. Wozniak, M. Stanisavljevic, P. Adusumilli, P. Narayanan, K. Hosokawa, M. Ishii, A. Kumar, V. Narayanan, and G. W. Burr, "A Heterogeneous and Programmable Compute-In-Memory Accelerator Architecture for Analog-AI Using Dense 2-D Mesh," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 31, no. 1, pp. 114–127, 2023.

[106] T. Gokmen and Y. Vlasov, "Acceleration of Deep Neural Network Training with Resistive Cross-Point Devices: Design Considerations," *Frontiers in Neuroscience*, vol. 0, no. JUL, p. 333, 2016.

[107] M. J. Rasch, F. Carta, O. Fagbohungbe, and T. Gokmen, "Fast offset corrected in-memory training," mar 2023. [Online]. Available: https://arxiv.org/abs/2303.04721v1

[108] T. Gokmen and W. Haensch, "Algorithm for Training Neural Networks on Resistive Device Arrays," *Frontiers in Neuroscience*, vol. 14, p. 103, feb 2020. [Online]. Available: www.frontiersin.org

[109] T. Gokmen, "Enabling Training of Neural Networks on Noisy Hardware," *Frontiers in Artificial Intelligence*, vol. 4, p. 699148, sep 2021.

[110] W. Wu, H. Wu, B. Gao, N. Deng, S. Yu, and H. Qian, "Improving Analog Switching in HfOx-Based Resistive Memory with a Thermal Enhanced Layer," *IEEE Electron Device Letters*, vol. 38, no. 8, pp. 1019–1022, aug 2017.

[111] T. Stecconi, V. Bragaglia, M. J. Rasch, F. Carta, F. Horst, D. F. Falcone, S. C. ten Kate, N. Gong, T. Ando, A. Olziersky, and B. Offrein, "Analog Resistive Switching Devices for Training Deep Neural Networks with the Novel Tiki-Taka Algorithm," *Nano Letters*, vol. 24, no. 3, pp. 866–872, jan 2024. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/acs.nanolett.3c03697

[112] N. Gong, M. J. Rasch, S. C. Seo, A. Gasasira, P. Solomon, V. Bragaglia, S. Consiglio, H. Higuchi, C. Park, K. Brew, P. Jamison, C. Catano, I. Saraf, F. F. Athena, C. Silvestre, X. Liu, B. Khan, N. Jain, S. McDermott, R. Johnson, I. Estrada-Raygoza, J. Li, T. Gokmen, N. Li, R. Pujari, F. Carta, H. Miyazoe, M. M. Frank, D. Koty, Q. Yang, R. Clark, K. Tapily, C. Wajda, A. Mosden, J. Shearer, A. Metz, S. Teehan, N. Saulnier, B. J. Offrein, T. Tsunomura, G. Leusink, V. Narayanan, and T. Ando, "Deep learning acceleration in 14nm CMOS compatible ReRAM array: device, material and algorithm co-optimization," *Technical Digest - International Electron Devices Meeting, IEDM*, vol. 2022-Decem, pp. 3371–3374, 2022.

[113] D. Ielmini and H.-S. P. Wong, "In-memory computing with resistive switching devices," *Nature Electronics 2018 1:6*, vol. 1, no. 6, pp. 333–343, jun 2018. [Online]. Available: https://www.nature.com/articles/s41928-018-0092-2

[114] M. Lanza, H.-S. P. Wong, E. Pop, D. Ielmini, D. Strukov, B. C. Regan, L. Larcher, M. A. Villena, J. J. Yang, L. Goux, A. Belmonte, Y. Yang, F. M. Puglisi, J. Kang, B. Magyari-Köpe, E. Yalon, A. Kenyon, M. Buckwell, A. Mehonic, A. Shluger, H. Li, T.-H. Hou, B. Hudec, D. Akinwande, R. Ge, S. Ambrogio, J. B. Roldan, E. Miranda, J. Suñe, K. L. Pey, X. Wu, N. Raghavan, E. Wu, W. D. Lu, G. Navarro, W. Zhang, H. Wu, R. Li, A. Holleitner, U. Wurstbauer, M. C. Lemme, M. Liu, S. Long, Q. Liu, H. Lv, A. Padovani, P. Pavan, I. Valov, X. Jing, T. Han, K. Zhu, S. Chen, F. Hui, and Y. Shi, "Recommended Methods to Study Resistive Switching Devices," *Advanced Electronic Materials*, vol. 5, no. 1, p. 1800143, jan 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/aelm.201800143https://onlinelibrary.wiley.com/doi/abs/10.1002/aelm.201800143https://onlinelibrary.wiley.com/doi/10.1002/aelm.201800143

[115] F. A. Kröger and H. J. Vink, "Relations between the concentrations of imperfections in solids," *Journal of Physics and Chemistry of Solids*, vol. 5, no. 3, pp. 208–223, may 1958.

[116] A. Padovani, L. Larcher, O. Pirrotta, L. Vandelli, and G. Bersuker, "Microscopic modeling of HfOx RRAM operations: From forming to switching," *IEEE Transactions on Electron Devices*, vol. 62, no. 6, pp. 1998–2006, jun 2015.

[117] Y. Guo and J. Robertson, "Materials selection for oxide-based resistive random access memories," *Applied Physics Letters*, vol. 105, no. 22, p. 223516, dec 2014. [Online]. Available: https://aip.scitation.org/doi/abs/10.1063/1.4903470

[118] WonjooÂ Kim, StephanÂ Menzel, D. J.Â Wouters, YuzhengÂ Guo, JohnÂ Robertson, BerndÂ Roesgen, RainerÂ Waser, and VikasÂ Rana, "Impact of oxygen exchange reaction at the ohmic interface in Ta 2 O 5 -based ReRAM devices," *Nanoscale*, vol. 8, no. 41, pp. 17 774–17 781, oct 2016. [Online]. Available: https://pubs.rsc.org/en/content/articlehtml/2016/nr/c6nr03810ghttps://pubs.rsc.org/en/content/articlelanding/2016/nr/c6nr03810g

[119] S. Larentis, F. Nardi, S. Balatti, D. C. Gilmer, and D. Ielmini, "Resistive switching by voltage-driven ion migration in bipolar RRAMPart II: Modeling," *IEEE Transactions on Electron Devices*, vol. 59, no. 9, pp. 2468–2475, 2012.

[120] D. F. Falcone, S. Menzel, T. Stecconi, A. L. Porta, L. Carraria-Martinotti, B. J. Offrein, and V. Bragaglia, "Physical modeling and design rules of analog Conductive Metal Oxide-HfO2ReRAM," *2023 IEEE International Memory Workshop, IMW 2023 - Proceedings*, 2023.

[121] A. Hardtdegen, C. La Torre, F. Cuppers, S. Menzel, R. Waser, and S. Hoffmann-Eifert, "Improved switching stability and the effect of an internal series resistor in HfO2/TiOx Bilayer ReRAM Cells," *IEEE Transactions on Electron Devices*, vol. 65, no. 8, pp. 3229–3236, aug 2018.

[122] S. Menzel, M. Waters, A. Marchewka, U. Böttger, R. Dittmann, and R. Waser, "Origin of the Ultra-nonlinear Switching Kinetics in Oxide-Based Resistive Switches," *Advanced Functional Materials*, vol. 21, no. 23, pp. 4487–4492, dec 2011. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/adfm.201101117https://onlinelibrary.wiley.com/doi/abs/10.1002/adfm.201101117https://onlinelibrary.wiley.com/doi/10.1002/adfm.201101117

[123] D. C. Sekar, B. Bateman, U. Raghuram, S. Bowyer, Y. Bai, M. Calarrudo, P. Swab, J. Wu, S. Nguyen, N. Mishra, R. Meyer, M. Kellam, B. Haukness, C. Chevallier, H. Wu, H. Qian, F. Kreupl, and G. Bronner, "Technology and circuit optimization of resistive RAM for low-power, reproducible operation," *Technical Digest - International Electron Devices Meeting, IEDM*, vol. 2015-Febru, no. February, pp. 28.3.1–28.3.4, feb 2015.

[124] F. Cüppers, S. Menzel, C. Bengel, A. Hardtdegen, M. von Witzleben, U. Böttger, R. Waser, and S. Hoffmann-Eifert, "Exploiting the switching dynamics of HfO2-based ReRAM devices for reliable analog memristive behavior," *APL Materials*, vol. 7, no. 9, p. 091105, sep 2019. [Online]. Available: https://aip.scitation.org/doi/abs/10.1063/1.5108654

[125] J. Woo, K. Moon, J. Song, S. Lee, M. Kwak, J. Park, and H. Hwang, "Improved synaptic behavior under identical pulses using AlOx/HfO2 bilayer RRAM array for neuromorphic systems," *IEEE Electron Device Letters*, vol. 37, no. 8, pp. 994–997, aug 2016.

[126] Y. Zhao, P. Huang, Z. Chen, C. Liu, H. Li, B. Chen, W. Ma, F. Zhang, B. Gao, X. Liu, and J. Kang, "Modeling and optimization of bilayered TaOitalicx RRAM based on defect evolution and phase transition effects," *IEEE Transactions on Electron Devices*, vol. 63, no. 4, pp. 1524–1532, apr 2016.

[127] D. F. Falcone, S. Menzel, T. Stecconi, M. Galetta, A. L. Porta, B. J. Offrein, and V. Bragaglia, "Analytical modelling of the transport in analog filamentary conductive-metal-oxide/HfOx ReRAM devices," *Nanoscale Horizons*, vol. 9, no. 5, pp. 775–784, apr 2024. [Online]. Available: https://pubs.rsc.org/en/content/articlehtml/2024/nh/d4nh00072bhttps://pubs.rsc.org/en/content/articlelanding/2024/nh/d4nh00072b

[128] W. Wu, H. Wu, B. Gao, N. Deng, S. Yu, and H. Qian, "Improving Analog Switching in HfOx-Based Resistive Memory with a Thermal Enhanced Layer," *IEEE Electron Device Letters*, vol. 38, no. 8, pp. 1019–1022, aug 2017.

[129] M. Le Gallo, R. Khaddam-Aljameh, M. Stanisavljevic, A. Vasilopoulos, B. Kersting, M. Dazzi, G. Karunaratne, M. Brändli, A. Singh, S. M. Müller, J. Büchel, X. Timoneda, V. Joshi, M. J. Rasch, U. Egger, A. Garofalo, A. Petropoulos, T. Antonakopoulos, K. Brew, S. Choi, I. Ok, T. Philip, V. Chan, C. Silvestre, I. Ahsan, N. Saulnier, V. Narayanan, P. A. Francese, E. Eleftheriou, and A. Sebastian, "A 64-core mixed-signal in-memory compute chip based on phase-change memory for deep neural network inference," *Nature Electronics 2023 6:9*, vol. 6, no. 9, pp. 680–693, aug 2023. [Online]. Available: https://www.nature.com/articles/s41928-023-01010-1

[130] W. Wan, R. Kubendran, C. Schaefer, S. B. Eryilmaz, W. Zhang, D. Wu, S. Deiss, P. Raina, H. Qian, B. Gao, S. Joshi, H. Wu, H. S. Wong, and G. Cauwenberghs, "A compute-in-memory chip based on resistive random-access memory," *Nature 2022 608:7923*, vol. 608, no. 7923, pp. 504–512, aug 2022. [Online]. Available: https://www.nature.com/articles/s41586-022-04992-8

[131] S. Ambrogio, P. Narayanan, A. Okazaki, A. Fasoli, C. Mackin, K. Hosokawa, A. Nomura, T. Yasuda, A. Chen, A. Friz, M. Ishii, J. Luquin, Y. Kohda, N. Saulnier, K. Brew, S. Choi, I. Ok, T. Philip, V. Chan, C. Silvestre, I. Ahsan, V. Narayanan, H. Tsai, and G. W. Burr, "An analog-AI chip for energy-efficient speech recognition and transcription," *Nature 2023 620:7975*, vol. 620, no. 7975, pp. 768–775, aug 2023. [Online]. Available: https://www.nature.com/articles/s41586-023-06337-5

[132] W. Song, M. Rao, Y. Li, C. Li, Y. Zhuo, F. Cai, M. Wu, W. Yin, Z. Li, Q. Wei, S. Lee, H. Zhu, L. Gong, M. Barnell, Q. Wu, P. A. Beerel, M. S. W. Chen, N. Ge, M. Hu, Q. Xia, and J. J. Yang, "Programming memristor arrays with arbitrarily high precision for analog computing," *Science*, vol. 383, no. 6685, pp. 903–910, feb 2024. [Online]. Available: https://www.science.org/doi/10.1126/science.adi9405

[133] P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang, W. Zhang, J. J. Yang, and H. Qian, "Fully hardware-implemented memristor convolutional neural network," *Nature 2020 577:7792*, vol. 577, no. 7792, pp. 641–646, jan 2020. [Online]. Available: https://www.nature.com/articles/s41586-020-1942-4

[134] Y. Kim, T. Gokmen, H. Miyazoe, P. Solomon, S. Kim, A. Ray, J. Doevenspeck, R. S. Khan, V. Narayanan, and T. Ando, "Neural network learning using non-ideal resistive memory devices," *Frontiers in Nanotechnology*, vol. 4, p. 1008266, oct 2022.

[135] Y. Li, T. P. Xiao, C. H. Bennett, E. Isele, A. Melianas, H. Tao, M. J. Marinella, A. Salleo, E. J. Fuller, and A. A. Talin, "In situ Parallel Training of Analog Neural Network Using Electrochemical Random-Access Memory," *Frontiers in Neuroscience*, vol. 15, p. 636127, apr 2021. [Online]. Available: www.frontiersin.org

[136] E. J. Fuller, S. T. Keene, A. Melianas, Z. Wang, S. Agarwal, Y. Li, Y. Tuchman, C. D. James, M. J. Marinella, J. J. Yang, A. Salleo, and A. A. Talin, "Parallel programming of an ionic floating-gate memory array for scalable neuromorphic computing," *Science*, vol. 364, no. 6440, pp. 570–574, may 2019. [Online]. Available: https://www.science.org/doi/10.1126/science.aaw5581

[137] Y. Li, S. Kim, X. Sun, P. Solomon, T. Gokmen, H. Tsai, S. Koswatta, Z. Ren, R. Mo, C. C. Yeh, W. Haensch, and E. Leobandung, "Capacitor-based cross-point array for analog neural network with record symmetry and linearity," *Digest of Technical Papers - Symposium on VLSI Technology*, vol. 2018-June, pp. 25–26, oct 2018.

[138] L. Gao, I. T. Wang, P. Y. Chen, S. Vrudhula, J. S. Seo, Y. Cao, T. H. Hou, and S. Yu, "Fully parallel write/read in resistive synaptic array for accelerating on-chip learning," *Nanotechnology*, vol. 26, no. 45, p. 455204, oct 2015. [Online]. Available: https://iopscience.iop.org/article/10.1088/0957-4484/26/45/455204https://iopscience.iop.org/article/10.1088/0957-4484/26/45/455204/meta

[139] J. D. Plummer, *Silicon VLSI technology: fundamentals, practice and modeling*. Pearson Education India, 2009.

[140] T. Heisig, R. Dittmann, and U. Simon, "Spectromicroscopic investigation of local redox processes in resistive switching transition metal oxides," Fachgruppe Chemie, Tech. Rep., 2022.

[141] J. D. Plummer and P. B. Griffin, "Integrated Circuit Fabrication: Science and Technology," *Integrated Circuit Fabrication*, nov 2023.

[142] C. S. Hwang, *Atomic layer deposition for semiconductors*. Springer, 2014.

[143] R. G. Gordon, "ALD precursors and reaction mechanisms," *Atomic Layer Deposition for Semiconductors*, pp. 15–46, 2013.

[144] X. Liu, S. Ramanathan, A. Longdergan, A. Srivastava, E. Lee, T. E. Seidel, J. T. Barton, D. Pang, and R. G. Gordon, "ALD of Hafnium Oxide Thin Films from Tetrakis(ethylmethylamino)hafnium and Ozone," *Journal of The Electrochemical Society*, vol. 152, no. 3, p. G213, jan 2005. [Online]. Available: https://iopscience.iop.org/article/10.1149/1.1859631https://iopscience.iop.org/article/10.1149/1.1859631/meta

[145] T. Faraz, H. C. Knoops, M. A. Verheijen, C. A. Van Helvoirt, S. Karwal, A. Sharma, V. Beladiya, A. Szeghalmi, D. M. Hausmann, J. Henri, M. Creatore, and W. M. Kessels, "Tuning Material Properties of Oxides and Nitrides by Substrate Biasing during

Plasma-Enhanced Atomic Layer Deposition on Planar and 3D Substrate Topographies," *ACS Applied Materials and Interfaces*, vol. 10, no. 15, pp. 13 158–13 180, apr 2018. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/acsami.8b00183

[146] M. Lapteva, M. Lapteva, M. Lapteva, V. Beladiya, V. Beladiya, V. Beladiya, S. Riese, P. Hanke, F. Otto, T. Fritz, P. Schmitt, P. Schmitt, O. Stenzel, A. Tünnermann, A. Tünnermann, A. Szeghalmi, and A. Szeghalmi, "Influence of temperature and plasma parameters on the properties of PEALD HfO₂," *Optical Materials Express, Vol. 11, Issue 7, pp. 1918-1942*, vol. 11, no. 7, pp. 1918–1942, jul 2021. [Online]. Available: https://opg.optica.org/viewmedia.cfm?uri=ome-11-7-1918{&}seq=0{&}html=truehttps://opg.optica.org/abstract.cfm?uri=ome-11-7-1918https://opg.optica.org/ome/abstract.cfm?uri=ome-11-7-1918

[147] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, "Commentary: The Materials Project: A materials genome approach to accelerating materials innovation," *APL Materials*, vol. 1, no. 1, p. 011002, jul 2013. [Online]. Available: https://aip.scitation.org/doi/abs/10.1063/1.4812323

[148] M. Yasaka and Others, "X-ray thin-film measurement techniques," *The Rigaku Journal*, vol. 26, no. 2, pp. 1–9, 2010.

[149] J. H. Klootwijk and C. E. Timmering, "Merits and limitations of circular TLM structures for contact resistance determination for novel III-V HBTs," *IEEE International Conference on Microelectronic Test Structures*, pp. 247–252, 2004.

[150] G. K. Reeves and H. B. Harrison, "Obtaining the Specific Contact Resistance from Transmission Line Model Measurements," *IEEE Electron Device Letters*, vol. 3, no. 5, pp. 111–113, 1982.

[151] P. R. Shrestha, D. M. Nminibapie, J. P. Campbell, J. T. Ryan, D. Veksler, H. Baumgart, and K. P. Cheung, "Analysis and control of RRAM overshoot current," *IEEE Transactions on Electron Devices*, vol. 65, no. 1, pp. 108–114, jan 2018.

[152] S. Menzel, U. Böttger, M. Wimmer, and M. Salinga, "Physics of the Switching Kinetics in Resistive Memories," *Advanced Functional Materials*, vol. 25, no. 40, pp. 6306–6325, oct 2015. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/adfm.201500825https://onlinelibrary.wiley.com/doi/abs/10.1002/adfm.201500825https://onlinelibrary.wiley.com/doi/10.1002/adfm.201500825

[153] S. Ambrogio, V. Milo, Z. Q. Wang, S. Balatti, and D. Ielmini, "Analytical Modeling of Current Overshoot in Oxide-Based Resistive Switching Memory (RRAM)," *IEEE Electron Device Letters*, vol. 37, no. 10, pp. 1268–1271, oct 2016.

[154] S. Tirano, L. Perniola, J. Buckley, J. Cluzel, V. Jousseaume, C. Muller, D. Deleruyelle, B. De Salvo, and G. Reimbold, "Accurate analysis of parasitic current overshoot during forming operation in RRAMs," *Microelectronic Engineering*, vol. 88, no. 7, pp. 1129–1132, jul 2011.

[155] National Instruments, "DAQ Analog Output Series NI 6711..." no. December, 2007.

[156] R. Sun, K. Makise, W. Qiu, H. Terai, and Z. Wang, "Fabrication of (200)-oriented TiN films on Si (100) substrates by DC magnetron sputtering," *IEEE Transactions on Applied Superconductivity*, vol. 25, no. 3, jun 2015.

[157] C. M. M. Rosário, B. Thöner, A. Schönhals, S. Menzel, A. Meledin, N. P. Barradas, E. Alves, J. Mayer, M. Wuttig, R. Waser, N. A. Sobolev, and D. J. Wouters, "Metallic filamentary conduction in valence change-based resistive switching devices: the case of TaOx thin film with x ˜ 1," *Nanoscale*, vol. 11, no. 36, pp. 16 978–16 990, sep 2019. [Online]. Available: https://pubs.rsc.org/en/content/articlehtml/2019/nr/c9nr05285bhttps://pubs.rsc.org/en/content/articlelanding/2019/nr/c9nr05285b

[158] K.-J. Lee, L.-W. Wang, T.-K. Chiang, and Y.-H. Wang, "Effects of Electrodes on the Switching Behavior of Strontium Titanate Nickelate Resistive Random Access Memory," *Materials 2015, Vol. 8, Pages 7191-7198*, vol. 8, no. 10, pp. 7191–7198, oct 2015. [Online]. Available: https://www.mdpi.com/1996-1944/8/10/5374/htmhttps://www.mdpi.com/1996-1944/8/10/5374

[159] A. Sawa and R. Meyer, "Interface-Type Switching," *Resistive Switching*, pp. 457–482, jan 2016. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/9783527680870.ch16https://onlinelibrary.wiley.com/doi/abs/10.1002/9783527680870.ch16https://onlinelibrary.wiley.com/doi/10.1002/9783527680870.ch16

[160] C. D. Landon, R. H. Wilke, M. T. Brumbach, G. L. Brennecka, M. Blea-Kirby, J. F. Ihlefeld, M. J. Marinella, and T. E. Beechem, "Thermal transport in tantalum oxide films for memristive applications," *Applied Physics Letters*, vol. 107, no. 02, jul 2015.

[161] D. B. Strukov, F. Alibart, and R. Stanley Williams, "Thermophoresis/diffusion as a plausible mechanism for unipolar resistive switching in metal-oxide-metal memristors," *Applied Physics A: Materials Science and Processing*, vol. 107, no. 3, pp. 509–518, jun 2012. [Online]. Available: https://link.springer.com/article/10.1007/s00339-012-6902-x

[162] T. Heisig, K. Lange, A. Gutsche, K. T. Goß, S. Hambsch, A. Locatelli, T. O. Mentes, F. Genuzio, S. Menzel, and R. Dittmann, "Chemical Structure of Conductive Filaments in Tantalum Oxide Memristive Devices and Its Implications for the Formation Mechanism," *Advanced Electronic Materials*, vol. 2100936, p. 2100936, 2022.

[163] R. Islam, S. Qin, S. Deshmukh, Z. Yu, C. KöroÇ§lu, A. I. Khan, K. Schauble, K. C. Saraswat, E. Pop, and H. S. Wong, "Improved gradual resistive switching range and 1000 x on/off ratio in HfOxRRAM achieved with a Ge2Sb2Te5thermal barrier," *Applied Physics Letters*, vol. 121, no. 8, p. 82103, aug 2022. [Online]. Available: /aip/apl/article/121/8/082103/2834070/Improved-gradual-resistive-switching-range-and

[164] C. Pennetta, Z. Gingl, L. B. Kiss, and L. Reggiani, "Biased percolation and electrical breakdown," *Semiconductor Science and Technology*, vol. 12, no. 9, p. 1057, sep 1997. [Online]. Available: https://iopscience.iop.org/article/10.1088/0268-1242/12/9/001https://iopscience.iop.org/article/10.1088/0268-1242/12/9/001/meta

[165] B. Yu, L. Chang, S. Ahmed, H. Wang, S. Bell, C. Y. Yang, C. Tabery, C. Ho, Q. Xiang, T. J. King, J. Bokor, C. Hu, M. R. Lin, and D. Kyser, "FinFET scaling to 10 nm gate length," *Technical Digest - International Electron Devices Meeting*, pp. 251–254, 2002.

[166] Z. Fang, X. P. Wang, J. Sohn, B. B. Weng, Z. P. Zhang, Z. X. Chen, Y. Z. Tang, G. Q. Lo, J. Provine, S. S. Wong, H. S. Wong, and D. L. Kwong, "The role of ti capping layer in HfOx-Based RRAM Devices," *IEEE Electron Device Letters*, vol. 35, no. 9, pp. 912–914, 2014.

[167] Y. Bum Yoo, J. Ho Park, K. Ho Lee, H. Woo Lee, K. Moon Song, S. Jong Lee, and H. Koo Baik, "Solution-processed high-k HfO 2 gate dielectric processed under softening temperature of polymer substrates." [Online]. Available: www.rsc.org/MaterialsC

[168] J. Rasley Microsoft Samyam Rajbhandari Microsoft Olatunji Ruwase Microsoft Yuxiong He Microsoft, "DeepSpeed: System Optimizations Enable Training Deep Learning Models with Over 100 Billion Parameters," *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, vol. 2020-Novem, pp. 3505–3506, nov 2020. [Online]. Available: https://doi.org/10.1145/3394486.3406703

[169] F. Xu, B. Gao, Y. Xi, J. Tang, H. Wu, and H. Qian, "Atomic-device hybrid modeling of relaxation effect in analog RRAM for neuromorphic computing," *Technical Digest - International Electron Devices Meeting, IEDM*, vol. 2020-Decem, pp. 13.2.1–13.2.4, dec 2020.

[170] M. Halter, L. Bégon-Lours, M. Sousa, Y. Popoff, U. Drechsler, V. Bragaglia, and B. J. Offrein, "A multi-timescale synaptic weight based on ferroelectric hafnium zirconium oxide," *Communications Materials 2023 4:1*, vol. 4, no. 1, pp. 1–8, feb 2023. [Online]. Available: https://www.nature.com/articles/s43246-023-00342-x

[171] M. J. Rasch, D. Moreda, T. Gokmen, M. Le Gallo, F. Carta, C. Goldberg, K. El Maghraoui, A. Sebastian, and V. Narayanan, "A Flexible and Fast PyTorch Toolkit for Simulating Training and Inference on Analog Crossbar Arrays," *2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems, AICAS 2021*, jun 2021.

[172] M. J. D. Powell, "An efficient method for finding the minimum of a function of several variables without calculating derivatives," *The Computer Journal*, vol. 7, no. 2, pp. 155–162, jan 1964. [Online]. Available: https://dx.doi.org/10.1093/comjnl/7.2.155

[173] M. Newville, T. Stensitzki, D. B. Allen, and A. Ingargiola, "LMFIT: Non-Linear Least-Square Minimization and Curve-Fitting for Python," sep 2014. [Online]. Available: https://zenodo.org/record/11813

[174] L. Deng, "The MNIST database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[175] Y. Bengio, R. Ducharme, and P. Vincent, "A Neural Probabilistic Language Model," in *Advances in Neural Information Processing Systems*, T. Leen, T. Dietterich, and V. Tresp, Eds., vol. 13. MIT Press, 2000. [Online]. Available: https://proceedings.neurips.cc/paper{_}files/paper/2000/file/728f206c2a01bf572b5940d7d9a8fa4c-Paper.pdf

[176] M. Prezioso, F. Merrikh-Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev, and D. B. Strukov, "Training and operation of an integrated neuromorphic network based on metal-oxide memristors," *Nature 2015 521:7550*, vol. 521, no. 7550, pp. 61–64, may 2015. [Online]. Available: https://www.nature.com/articles/nature14441

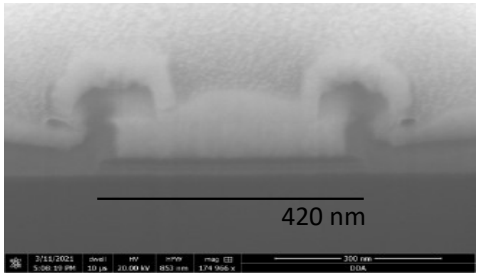
[177] B. Sánta, Z. Balogh, L. Pósa, D. Krisztián, T. N. Török, D. Molnár, C. Sinkó, R. Hauert, M. Csontos, and A. Halbritter, "Noise Tailoring in Memristive Filaments," *ACS Applied Materials and Interfaces*, vol. 13, no. 6, pp. 7453–7460, feb 2021. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/acsami.0c21156

Annexe - ReRAM device development overview

1st generation ReRAM devices:

- Min. area: (500 nm)^2
- Conductance range: (50, 500) µS
- Forming voltage: > 5 V

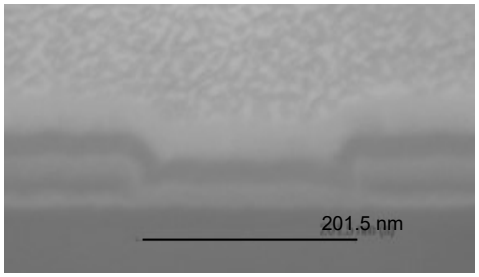
SEM – Gen_1 ReRAM



2nd generation ReRAM devices:

- Area: down to (200 nm)^2
- Conductance range: (5, 50) µS
- Forming voltage: < 4 V

SEM – Gen_2 ReRAM



3rd generation
1-Transistor-1-ReRAM devices:

- BEOL integration on CMOS circuits
- Crossbar arrays of 1T1R devices
- Electrically similar to Gen^2

SEM – Gen_3 ReRAM

