UC Santa Barbara
UC Santa Barbara Electronic Theses and Dissertations

Title

3D Integrated Memristive Devices for Memory and Computing

Permalink

https://escholarship.org/uc/item/15j4w2ct

Author

Adam, Gina Cristina

Publication Date

2015

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Santa Barbara

3D Integrated Memristive Devices for Memory and Computing

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Electrical and Computer Engineering

by

Gina Cristina Adam

Committee in charge:

Professor Dmitri Strukov, Chair

Professor Bob York

Professor Jon Schuller

Professor Sumita Pennathur

Professor Wei Wu (University of Southern California)

December 2015

The dissertation of Gina Cristina Adam is approved.

_____

Bob York

_____

Jon Schuller

_____

Sumita Pennathur

_____

Wei Wu

_____

Dmitri Strukov, Committee Chair

November 2015

3D Integrated Memristive Devices for Memory and Computing

DEDICATION

To the two women who opened my eyes to science and engineering

– my chemistry teacher Cecilia Vasile and my undergraduate mentor, Prof. Mihaela Albu –

with the promise to continue to spread their enthusiasm for science to future generations.

ACKNOWLEDGEMENTS

I want to thank many people for supporting me while pursuing my PhD and writing this dissertation. None of this work would have been possible without my advisor, Prof. Dmitri Strukov. He provided challenging projects and ideas that trained me to be persistent and rigorous in my research pursuits. He offered me guidance and feedback when needed, while also giving me freedom to pursue my own ideas and interests. It has been an honor working with Prof. Strukov and his group.

Next, I would like to thank my dissertation committee members, Prof. Bob York, Prof. John Schuller, Prof. Sumita Pennathur, and Prof. Wei Wu, for taking time to be at my qualifying and defense exams and for providing challenging questions and useful feedback.

I would also like to express my gratitude to the staff of the Nanotech cleanroom facility at UCSB for all of the technical support that they have given me, including, Brian Thibeault, Bill Mitchel, Adam Abrahamsen, Brian Lingg, Mike Silva, Luis Zuzunga and Tom Reynolds.

Special thanks to the department graduate student advisor, Val de Veyra, and to all the staff at the Office of International Student and Scholars for making my accommodation to the U.S. academic environment very smooth.

Next, I would like to thank all of my current and past colleagues in the Strukov group for making this journey a pleasant experience, including Dr. Mirko Prezioso, Dr. Bhaswar Chrakrabarti, Dr. Fabien Alibart, Dr. Ligang Gao, Dr. Farnood Merrikh-Bayat, Elham Zamanidoost and Xinjie Guo. Our discussions, lunches in the lab and sailing trips for the past five years are all very special to me. Special thanks to Advait Madhavan for his constant friendship in times of joy and in times of need. Last but not least, a big thank you to Brian Hoskins for always supporting me in lab and in life.

Finally, I would like to thank my parents who supported me to pursue my dreams far away from home. My achievements are possible because of their sacrifices.

VITA OF GINA CRISTINA ADAM

December 2015

EDUCATION

Bachelor of Science in Electronics Engineering, Universitatea Politehnica Bucuresti, Romania, June 2010.

Master of Science in Electrical and Computer Engineering, University of California, Santa Barbara, March 2012.

Master of Arts in Teaching and Learning, University of California, Santa Barbara, December 2015 (expected).

Doctor of Philosophy in Electrical and Computer Engineering, University of California, Santa Barbara, December 2015 (expected)

PROFESSIONAL EMPLOYMENT

Fall 2015: Teaching Assistant, Department of Electrical and Computer Engineering, University of California, Santa Barbara.

Summer 2012: Mirzayan Fellow, National Academy of Engineering.

2010-2015: Research Assistant, Department of Electrical and Computer Engineering, University of California, Santa Barbara.

PUBLICATIONS

Adam, G. C., Hoskins, B. D., & Strukov, D. B. Material Implication Logic Constraints for Adder Implementation in Memristor Crossbar. *in preparation.*

Chakrabarti B., Lastras-Montano M., **Adam G. C.**, Hoskins B.D., Shkabko A., Cheng K., Strukov D., High-precision tunable memristive devices with low current operation vertically integrated on 0.5 μm CMOS technology. *in preparation.*

Adam, G. C., Hoskins, B. D., Prezioso, M., & Strukov, D. B. (2015). Three-Dimensional Stateful Material Implication Logic. *arXiv preprint arXiv:1509.02986.*

Prezioso M., Kataeva I., Merrikh-Bayat F., Hoskins B., **Adam G.C.**, Sota T., Likharev K., & Strukov D.B. "Modeling and Implementation of Firing-Rate Neuromorphic-Network Classifiers with Bilayer Pt/Al$_2$O$_3$/ TiO$_{2-x}$ /Pt Memristors", *IEDM 2015*.

Prezioso, M., Merrikh-Bayat, F., Hoskins, B. D., **Adam, G. C.**, Likharev, K. K., & Strukov, D. B. (2015). Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature*, *521*(7550), 61-64.

AWARDS

2015 - UCSB ECE Departmental Summer Dissertation Fellowship 2015

2013-14 - P.E.O. International Peace Scholarship

2013 - Lockheed Martin Scholarship (Society for Women Engineers)

2012 - Mirzayan Science and Technology Policy Fellowship, U.S. National Academies

2012 - UCSB NNIN Research Experience for Teachers - Fellowship for Mentors

2010 - International Fulbright Science and Technology Award

ABSTRACT

3D Integrated Memristive Devices for Memory and Computing

by

Gina Cristina Adam

Traditionally, increased speed and lower power consumption in modern electronics has been achieved through aggressive transistor scaling and more elegant architectural designs. Another pathway that has been explored extensively in the recent years is the use of new computational devices. Memristors ("memory resistors") are novel two terminal electronic devices based on resistive switching, a physical phenomena where a dielectric rapidly changes its resistance under strong applied voltage due to coupled ionic and electronic transport. These two terminal devices are highly scalable - recent work demonstrating sub-10nm structures - and have demonstrated high endurance and low power consumption.

Memristors have shown potential for breakthrough applications because of their intrinsic capability for both nonvolatile memory storing and material implication-based logic. Memristors are now actively investigated for non-volatile memory applications and energy-efficient hardware implementations of artificial neural networks. Another potential application is logic-in-memory that can provide a potential new way forward for opening the von Neumann bottleneck. The von Neumann bottleneck has been growing narrower over the years, as CPU speed and memory have been increasing much faster than the bandwidth

between them can accommodate. One promising approach to circumvent this problem is logic-in-memory computing, where computation is performed in the memory itself, significantly reducing traffic between the CPU and the memory subsystem. The most practical implementation of logic-in-memory utilizes electronic devices that can perform both storage and logic while being monolithically integrated into existing CMOS technology and memristors are a prominent candidate for this technology. In 2009, a logic-in-memory approach implementing material implication logic with memristors was proposed by Hewlett Packard. Three dimensional monolithically-stacked memristor layers with inter-layer material implication logic capability would provide increased density and throughput.

The objective of this dissertation is to advance of the state of the art for material implication logic through three research goals. Our first goal was to develop a fabrication pathway for monolithical vertical integration of memristors in order to implement 3D memories. This allowed us to experimentally test logic-in-memory systems. Our second goal was to determine memristor device and circuit constraints for implementing material implication logic and explore circuit and device level solutions to increase robustness of operation. Our final goal was to combine these two efforts together and demonstrate reliable material implication logic in vertically stacked memristors. To this end, we fabricated and successfully tested monolithically stacked memristive structures implemented with TiO_2-based memristors. We also developed an optimized circuit configuration able to perform material implication with maximum tolerance to device variations. This allowed us to demonstrate, for the first time, hundreds of successful three-dimensional data manipulation cycles using material implication. An inter-layer NAND gate with the inputs and output in different device layers was implemented with 94% yield. This high yield demonstrates the potential for using the inter-layer stateful logic gates in larger circuits for in-memory logic.

This implementation also opens the way through aggressive scaling to achieve one of the Feynman Grand Challenges - the construction of a functional nano-scale 8-bit adder in 50x50x50nm for which a circuit implementation is proposed.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter I

Introduction

A. Computing with Memory

A.1. The von Neumann bottleneck

In 1945, John von Neumann [1] was suggesting the concept of stored-program computer, a revolutionary idea that offered unparallel flexibility over the hard-wired computers of the 1940s and allowed the computing industry to evolve to what it is today. In the von Neumann architecture, the memory unit stores both the program instructions and the data in the random-access memory (RAM). As shown in Figure 1.1, a central bus connects the central processing unit (CPU) and the RAM and is used for both instruction fetch and data operations.

Due to this shared bus, the program memory and data memory cannot be accessed at the same time which limits the effective processing speed. This problem was described in 1977 by John Backus in his ACM Turing Award lecture [2]:

Surely there must be a less primitive way of making big changes in the store than by pushing vast numbers of words back and forth through the von Neumann bottleneck. Not only is this tube a literal bottleneck for the data traffic of a problem, but, more importantly, it is an intellectual bottleneck that has kept us tied to word-at-a-time thinking instead of encouraging us to think in terms of the larger conceptual units of the task at hand. Thus programming is basically planning and detailing the enormous traffic of words through the von Neumann bottleneck, and much of that traffic concerns not significant data itself, but where to find it."

Figure 1.1. The Von Neumann bottleneck due to the shared bus used for program and data fetching from memory.
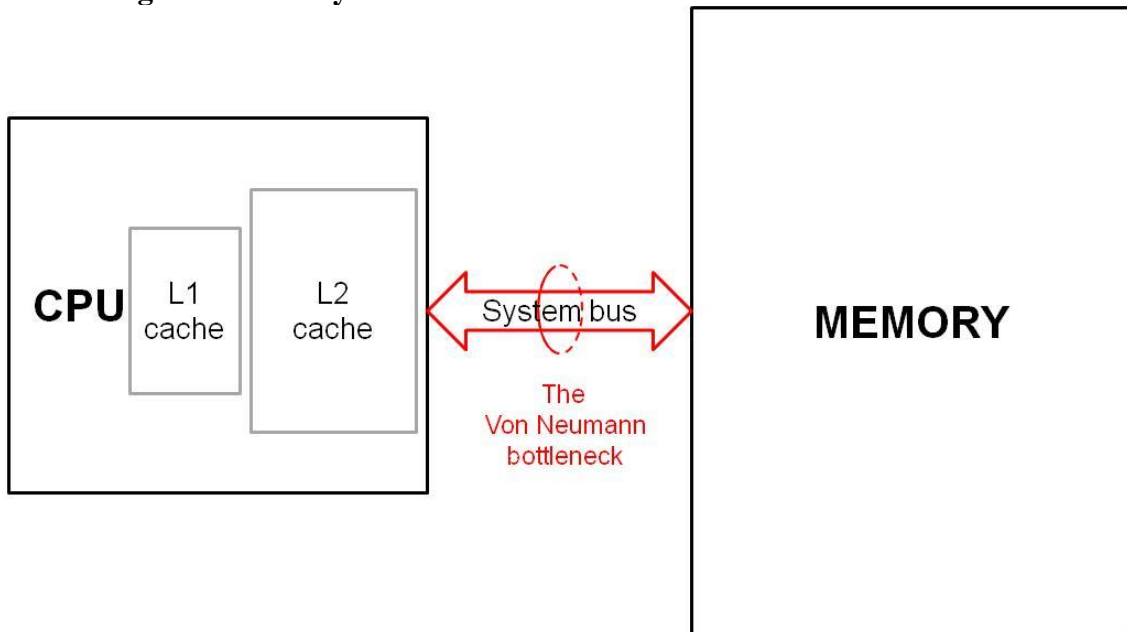
The von Neumann bottleneck problem has become more acute over the years, since the CPU speed and memory size have increased much faster than the throughput between them. Figure 1.2 shows how the evolution of the performance gap between the CPU and the DRAM grows by 50% every year[3]. The CPU has doubled its speed every 1.5 years following Moore's law, leaving behind the DRAM which doubled its speed only every 10

2

years. A 1996 report [4] estimated that the CPU had to wait on average three out of four cycles for the data to be transferred to or from memory. With the multi-threading and increase in performance gap, this average has gotten worse.

Figure 1.2. The performance gap caused by the difference in speed between CPU and DRAM. (adapted from [3], page 73)
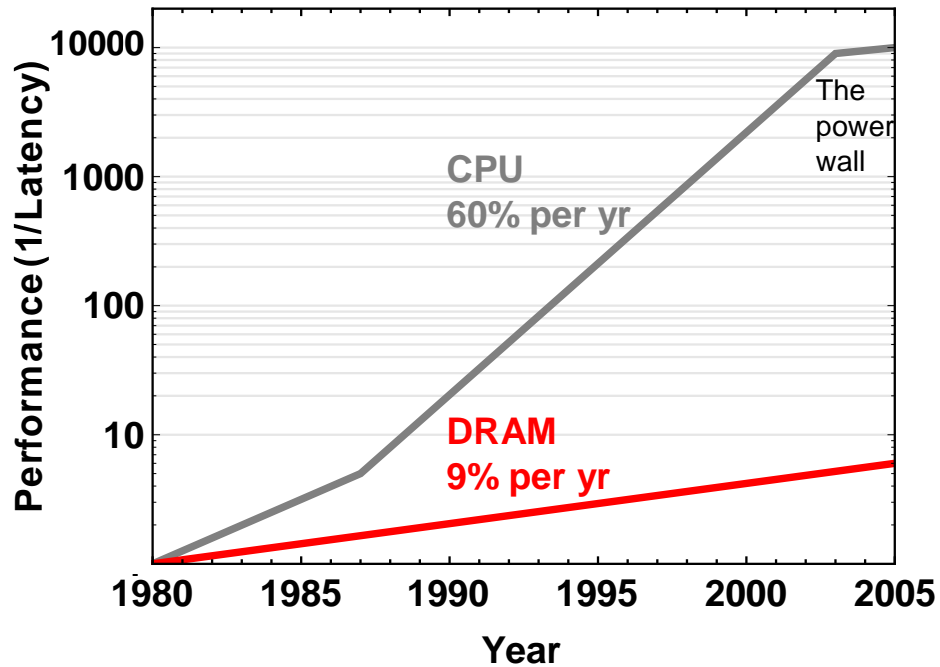
Several mechanisms have been used to alleviate the von Neumann bottleneck. The modified Harvard architecture provides separate caches for data and instructions, which significantly increases the effective processing speed but only if the needed information is already in the cache. Branch predictor algorithms are useful but only in the case when the prediction made is correct. Parallel computing using multiple processing cores has been used as another mechanism to partially overcome this problem, but it only works if the software sequence is highly parallelizable according to Amdahl's law [5]. Implementation of a high number of parallel CPU cores allowed for a bigger datasets and more complex problems to be tackled, further increasing the aggregate peak bandwidth and deepening the von Neumann

3

bottleneck. According to [3], the Intel Core i7 with four cores and 3.2 GHz clock rate can generate a total peak bandwidth of 409.6 GB/sec, in comparison with the DRAM main memory that has a peak bandwidth of only 25 Gb/sec.

A.2. Non-von Neumann architectures

The von Neumann bottleneck is an inherent problem of the von Neumann architecture and temporary solutions as the one presented above can provide only a small gain. Non-von Neumann architectures have been proposed that do not suffer from this bottleneck.

The idea of content addressable memory (CAM), also known as associative memory, stems from the "word recognition unit" proposed by Dudley Allen Buck in 1955 [6]. By comparison with the RAM where the user provides a memory address and the RAM returns the data word stored at that address, the CAM uses the data word provided by the user to search all its memory. If the word is found, one or more storage addresses are returned. The CAM searches for its entire memory in one operation and therefore is much faster than RAM in all search applications. However, the CAM adoption is limited by the prohibitive production costs and power consumption, because each bit needs additional comparison circuitry. CAM has been so far used in specialized applications such as in network routing, cache controllers, database engines and data compression. A particularly exciting application is the artificial neural network that can provide recursive capabilities (Hopfield Network) and shows unmatched performance in pattern classification and recognition [7, 8].

Another idea proposed has been the reconfigurable system, capable of being physically reconfigured on the fly as needed by the data operations or transfer. These characteristics allow for larger flexibility and insure fault tolerance, since bad sectors can be easily reconfigured to be avoided. The reconfigurable systems are naturally able to perform more

functions and are more robust than their von Neumann counterparts. The reconfigurable systems are based on blocks or cells that can process and store data independently. The connection between the blocks is not predetermined and can be configured as needed, which can be provide a significant speed advantage for complex problems that require a high degree of flexibility and parallelism such as signal processing, speech recognition, cryptography, computer hardware emulation, etc.. The most prominent of such implementation is the field programmable gate array (FPGA). However, historically large overhead real-estate required for configuration of FPGAs increased their energy consumption and area footprint, making them less desirable for general applications [9].

Another approach has been logic-in-memory. One implementation is based on small logic and memory cells with memory devices and logic elements distributed in close vicinity to each other. In order to take full advantage of this implementation, the memory devices should have very short access time (<10ns), very high endurance and small dimensions compared to the existing CMOS. An example is the full adder-circuit [10] implemented with 34 transistors and 4 MTJs with offers significant performance with ~20% less power consumption and area usage since the MTJs are stacked on the transistor layer.

Another possibility is using electronic devices that can provide both storage and logic capabilities. One example is a quantum cellular automata [11, 12] that uses five quantum dots occupied by two electrons. It computes the information based on the coulomb repulsion between the electrons that create a strongly polarized ground state with a highly non-linear response and bistable staturation to electrostatic stimuli. It also stores the state automatically, allowing for power efficient computation. Another possibility is to store the memory state in a different variable than charge. The spin has been explored as a state variable in devices for

5

processing and storing information, with the all-spin logic device with buit-in memory fabricated by Behin-Aein et. al as a notable example [13].

The resistance value as a state variable is another option that has been investigated. A example of devices that naturally store state as resistance value are the memristors. These two-terminal devices show excellent scalability and thanks to their memory properties, are actively investigated for non-volatile memory applications and power efficient hardware implementation of artificial neural networks. Moreover, they have been shown to be able to perform a special type of logic, called material implication [14]. The next section will describe in more detail what memristors and material implication are and the advantages, pitfalls and potential applications for memristor-based material implication.

B. Stateful Logic with Memristors

B.1. Memristors and state of the art

Memristive devices are switches with a variable resistance (Figure 1.3a). Chua [15] introduced in 1971 in circuit theory the ideal memristor as the 4th passive electronic component, with its resistance directly dependent on the flux of charge passing through the device. Memristive devices [16] are a broader class of devices where the resistance can be dependent of a set of state variables. The three characteristics defining a memristive device [17] are: 1) pinched hysteresis loop (Figure 1.3b); 2) hysteresis area decrease with frequency; 3) stable non-hysteretic characteristic at infinite frequency.

Figure 1.3. Memristor. (a) Symbol; (b) Hysteresis curve and (c) Two-terminal design
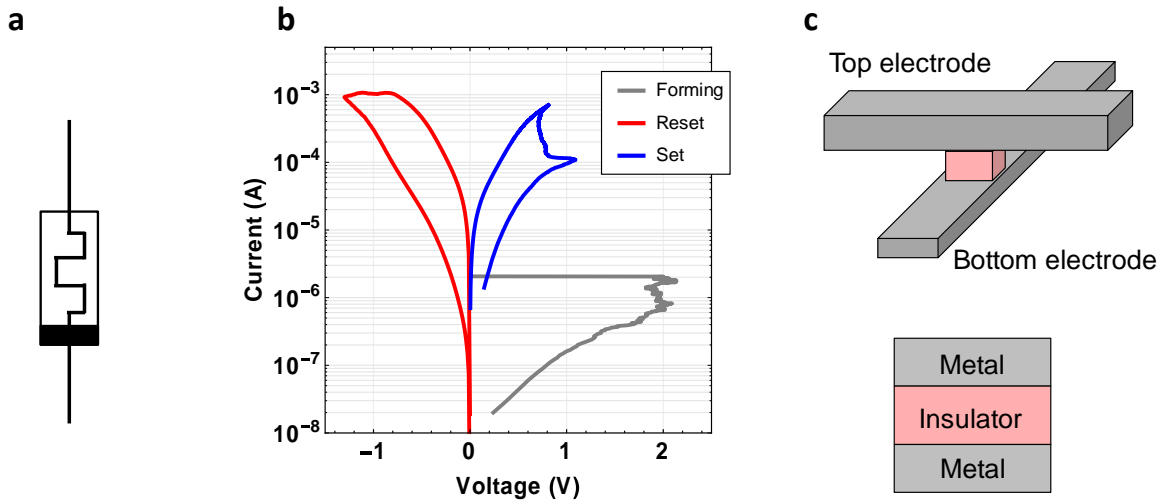
The memristive behavior has been observed in a variety of device geometries, based two terminal or three terminal devices, with a vertical or a planar configuration, based on thin films or nanotubes. The geometry most useful for its scaling potential is the two-terminal on [Figure 1.3c]. No matter the geometry, the memristor has at its core a material capable of exhibiting resistive switching. The resistive switching behavior with the characteristic pinched hysteresis was observed since 1960s particularly in thin films of transition metal oxides. Sandia Corp. reported such behavior in anodized aluminum-based films [18]. By 1970, Sliva's review [19] summarizes results related to resistive switching in a variety of materials such as metal oxides (PbO, CuO, TiO_2, Fe_2O_3, V_2O_5, HgO, Al_2O_3 and Ta_2O_3), organics (saran wrap, phthalocyanines and polystyrene) and other inorganic materials. Forty years later, the review from Yang [20] lists a far more comprehensive list of materials that exhibit this behavior. It is unclear what material is the best for what applications.

Two types of switching has been observed: 1) bipolar, which requires opposite voltage polarities for switching ON-set or OFF-reset and 2) unipolar, where the device can be switched between ON and OFF with the same voltage polarity [21]. The mechanisms behind

each type of switching are not certainly confirmed, but the studies so far seem to suggest that the electric field controls the bipolar switching while the unipolar switching is based on Joule heating. Sometimes both switching modes can be observed in the same device. It is desired to engineer the device structure and select the device materials in order to have a predominant switching mode. The devices with an active layer of transition metal oxide (such as TiO_2, Ta_2O_5, HfO_2, etc) are believed to have as mobile species oxygen vacancies and typically have bipolar switching. Devices such as electrochemical metallization memory, conductive bridging RAM or atomic switches have as mobile species a metallic cation and typically have unipolar switching.

The device structure and material selection are important to eliminate electroforming. The electroforming step is one-time application of high voltage in order to partially break down the active material. It is desired to eliminate electroforming due to circuit constraints. Due to high electric fields and Joule heating in the electroforming step, the active material can locally change its phase which accounts for switching. It has proved hard to identify the actual switching material, but attempts have been made to identify what mobile species (electrons or ions) and the exact location of the switching.

Memristors have shown potential for breakthrough applications because of their intrinsic capability for both nonvolatile memory storing and material implication-based logic. These two terminal devices are highly scalable, recent work [22] demonstrating sub-10nm structures, and 3D monolithically-stacked memristor layers promise to provide increased density. Plenty of applications have been suggested for the memristor devices, from non-volatile memories that can compete with the flash technology in terms of speed and energy consumption while offering larger storage space to artificial neural networks capable of

ultra-fast pattern recognition and image processing. An exciting application is logic-in-memory based on stateful computing using material implication and memristors.

B.2. Material implication summary

Whitehead and Russell [23] presented in 1910 four fundamental logic operations: AND, OR, NOT and material implication (p IMP q or "p→q"). p implies q means that if p is true, then q is true. The implication is true in all cases, except when p is true and q is false. Its truth table and Venn diagram is presented in Figure 1.4. The material implication and the false operation constitute c, so any Boolean function can be implemented using sequences of these two operations.

Figure 1.4. Implication logic (a) Truth table; (b) Venn diagram.
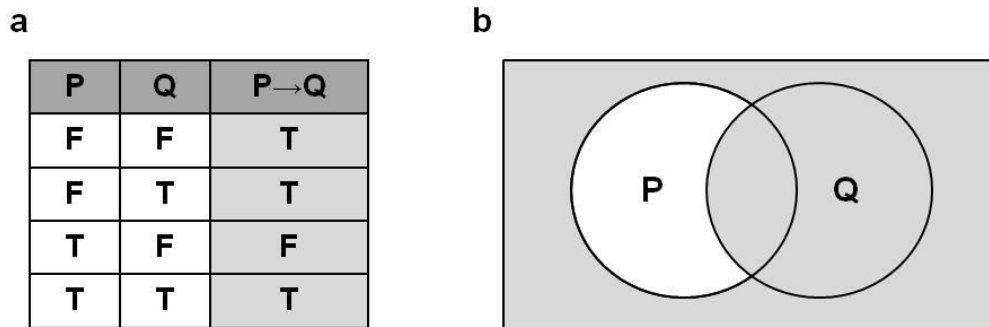
In 1936, Shannon [24] invented digital electronics by showing that the first three logic operations describe by Whitehead and Russell – AND, OR and NOT- can be implemented in a simple fashion using a few number of electronic switches. These three logic operations form a universal Boolean set as well. With the first fabrication of a solid-state transistor in 1947 at Bell Labs and the first solid-state integrated circuit in 1959, the era

of large-scale digital electronics became possible. The AND, OR and NOT gates are still at the core of the digital circuits in use today.

On the other hand, little effort has been spent investigating material implication-based logic. In 2010, Borghetti et al [25] from Hewlett Packard Research Labs has shown that a system of two memristors and a resistor can naturally implement a material implication gate. More details are presented in the following sub-section.

B.3. Stateful logic with memristors

In Borghetti's work, implication operation Q←P IMP Q was implemented in one layer, e.g., part of the circuitry consisting of load resistor of conductance G_L, and memristors P and Q (Figure 1. 5a), by applying simultaneously specific voltage pulses ("clock") to memristors Q and P. When the clock signal is applied, the resistive states of P and Q dictate the voltage on the common electrode and as a result the bias across memristor Q. In the case when memristors P and Q are in the OFF state, the bias drop on Q is larger than the device switching ON threshold so that the device Q turns on. However, if originally memristor P is in the ON state and Q is in the OFF state, the bias across memristor Q is too small for setting so the device maintains its state. The resulting logic operation is described by the truth table in Figure 1.5b and equivalent to Q ← (NOT P) OR Q. Since the material implication and the false operation are a universal Boolean set, any Boolean function can be computed using a small number of memristors by performing sequences of material implication and RESET operations.

Figure 1.5. Implication logic with memristors (a) Circuit implementation (see Chapter 4.A. for detailed measurement setup); (b) Truth table.

a

b

$$q' \leftarrow p \text{ IMP } q$$

p	q	q'
0	0	1
0	1	1
1	0	0
1	1	1

on ≡ "1"
off ≡ "0"

P	Q	Q'
off	off	on
off	on	on
on	off	off
on	on	on
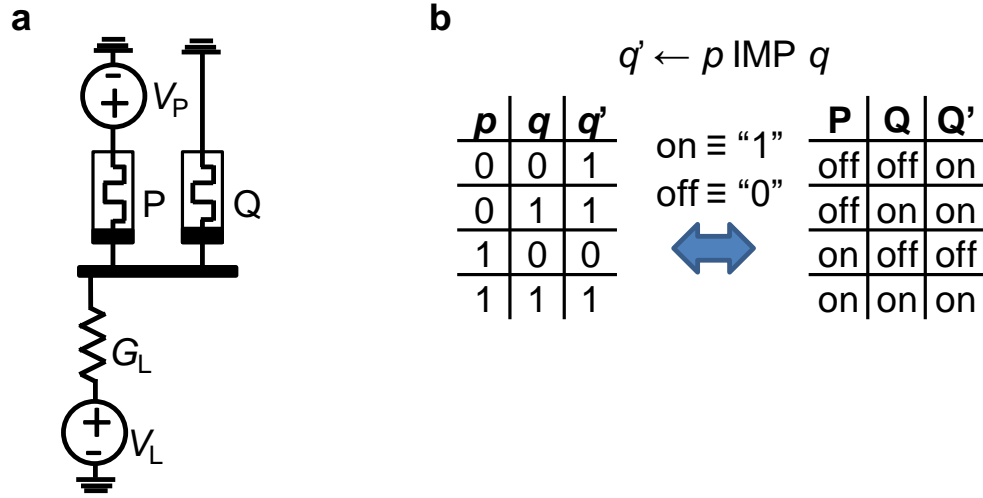
Borghetti and his team used for their experiment transition metal oxide memristors showing bipolar switching. Since then, similar results has been proved in a variety of other systems such as unipolar memristors [26], complementary resistive switches [27], magnetically enhanced memristors [28], biological systems [29-30], CMOS technology [31] and magnetic tunnel junctions [32].

Other memristor-based logic approaches have been suggested, such as memristor ratioed logic [33] and memristor staking logic [34]. These logic approaches are not stateful, meaning the result of the computation is not stored in resistance value of the memristor, but rather represented as a voltage signal, which means that they are not suitable for logic-in-memory applications.

B.4. Advantages and disadvantages of stateful logic with memristors

The stateful logic implemented with memristors has several potential advantages that makes it attractive for novel computing applications. The most important advantage is the statefulness offered by the immediate latching of the computation result as a value stored in

a memristor's resistance. This allows for non-volatile computing, an attractive feature in the context of energy scavenging applications that need to work with an intermittent power supply. Memristive material implication is good for digital logic, because it is insensitive to small or even significant variations in the resistivity states of the memristors, as long as the rations of resistances stay large. Moreover, memristors are two terminal devices with potential for extreme scaling and high density integration on CMOS chips. Hybrid memristor/CMOS integration promises to reduce the physical distance between memory and main processing unit, increase speed and storage. By implementing material implication capabilities in the memristor layer, it is possible to have extremely high bandwidth computing. Memory-related computation can be performed in memory itself without having to spend significant time accessing the CMOS below. Several theoretical studies have predicted significantly higher performance and energy-efficiency for memristor-based IMP logic circuits and very similar concepts over conventional approaches for high-throughput computing applications [35-36].

As any technology, the memristor-based material implication also suffers from pitfalls. Because the state is stored as a resistance, additional circuitry is required to convert it to voltage as needed by the CMOS circuitry [33]. Since the voltage on the common electrode varies with the state of the memristors, an extra CMOS keeper circuit might also be needed to ensure a constant voltage over the switching device Q during the switching [37].

Each IMPLY operation implemented with memristors requires two steps in the original implementation: 1) reset output; 2) material implication between input and output. For a NAND gate implemented with implication logic, one reset and two material implication steps are needed. Therefore the more complex the Boolean functions are, the lengthier the computation sequence becomes. As solutions to alleviate this problem, Lehtonen and

Poikonen proposed parallelism [37] and multi-input implication together with complementary representation of variables [38]. Material implication logic requires external CMOS circuitry to perform in correct order the necessary sequence of implication and reset steps. In order to keep the external circuitry to a minimum, it is best to use it to drive multiple computations in parallel. Pipelining [39] is another way to reduce the apparent computational time, by increasing the throughput of data. This strategy can be combined with the multi-input operation, by applying the same conditional voltage on multiple input memristors at the same time. The complementary representation of variables can simplify the Boolean sequence with the trade-off of space.

Because the state variable is stored as resistance instead of voltage, memristor-based material implication is intrinsically a single-output operation. When the duplicates of a state are needed, a fairly complex copy operation is required. Each copy operation takes two reset and two material implication steps and one auxiliary memristor apart from the input and output memristors. The first implication step between input and auxiliary, sets the auxiliary to the negation of input, while the second implication step between auxiliary and output, sets the output to input value. If the complementary representation of variables is used, the copy operation can be performed in only one implication step. More elegant solutions were proposed. Kim, Shin and Kang [39] suggested a modified AND operation that allows the simultaneous execution of multiple operations with the possibility of duplicating the output state in only two steps.

Since the memristors are passive elements, signal degradation is another problem in memristor-based material implication. The switching to an incomplete ON state limits the length of the maximum achievable Boolean sequence. Memristors with low resistance ratio between ON and OFF degrade the signal the most, so improvements in the memristor

13

technology can alleviate this problem. Levy et al [40] estimates that a ratio of 10^4 enable systems with 10^6 cells with an output degradation of only 10%. Refreshing circuitry can be used to restore the state to a true ON.

The half-select problem is inherent to the crossbar architecture. Sneak currents arise when the memristors exhibit fairly linear characteristics and no selector is used. For memristor-based material implication, the problem might be even more acute since the common node has a variable voltage as needed for the conditional switching. The stray currents that arise can cause computational errors. By using highly non-linear memristor devices or memristors with incorporated selectors, the sneak currents can be kept to a minimum. Lehtonen and Poikonen [37] proposed an alternative operation based on similar principles, called converse non-implication that can only be implemented using rectifying memristors.

B.5. Applications for memristor-based stateful logic

The investigation of logic performed using memristor devices is recent and it is yet unclear what type of architecture would benefit the most from the memory and logic capabilities of these devices.

The CMOL architecture is based on a hybrid CMOS/memristor system. Add-on layers of memristor crossbars are fabricated on top of CMOS circuitry. The CMOL is ideal for reconfigurable computing. The classic FPGA architecture can be easily adapted to the CMOL, with the advantage that the reconfiguration bits can be reprogrammed using memristor-based material implication. The Field Programmable Nanowire Interconnect (FPNI) is such an example suggested by Kim and Shin [39]. Lehtonen [37] proposed a

cellular type of network called Cellular Neural/Nanoscale Networks [CNN], also with reconfigurable features and more extensive use of memristor-based material implication. The memristor crossbar layers would perform material implication in fully parallel fashion, with a dedicated memristor crossbar layer to store the control signals corresponding the each computational sequence.

When the CMOL hybrid architecture is used for non-volatile memory storage, memristor-based material implication can be used for logic-in-memory computation required in the memristor layers (look-up tables, adders, error correcting operations, etc) [40-42]. The mapping of which devices perform logic and which store data can be changed dynamically as needed during the memory operation [32]. Levy et al. [40] proposes a highly dense Akers array architecture that supports memristor-based logic-in-memory computation. Each Akers cell has two anti-serial memristors (CRS cell) and four transistors. Each cell can store one bit in the pair of memristors and can perform a primitive Boolean function. An array of such cells can realize any Boolean function and naturally performs efficient bit sorting. It offers fully parallel functionally therefore it can be used reduce the computational load of the CPU and alleviate the von Neumann bottleneck.

Another exciting application is the use of memristor-based implication logic in content addressable memories (CAM). Zheng and Shing [43] proposed such an application called mTCAM using the flexible ternary CAM architecture. Each mTCAM cell has 2 memristors and 5 transistors. Memristors are programmed individually to insure high impedance between the search lines. The proposed mTCAM offers much higher storage capabilities and non-volatility with similar latency and energy consumption.

More research is required to understand the capabilities of memristor-based implication logic and its potential applications. Improvements in device design and fabrication, circuit design and architectures are needed to make it a viable technological option.

C. Dissertation scope

This dissertation is devoted to the development of stacked memristive devices for material implication logic. The motivation is to determine the circuit and device constrains and to experimentally show reliable and variation-intolerant multi-cycle operation - a requirement for technological adoption. We developed an optimized circuit configuration able to perform material implication with maximum tolerance to variations. We fabricated different monolithically stacked arrays and crossbar structures using TiO_2-based memristors that operated free of thermal crosstalk. This allowed us to demonstrate for the first time hundreds of successful three-dimensional data manipulation cycles using material implication and explore other experimental challenges that may affect future technologies based on this type of logic.

Chapter II is dedicated to circuit optimization. The first part of the chapter presents an improved circuit design. The basic circuit suggested for memristor-based material implication is based on two-memristor and one resistor (Figure 1.5a above). A central component in this design is the load resistance which modulates the common node's potential and enables the conditional switching. Selecting the best possible value of load resistance is desirable, since it allows for operation with the highest distinguishability of states after switching. Using analytical and numerical methods, it is shown that the optimized circuit for memristor-based material implication has a current source as load. This

circuit design also has the advantage of only two external variables (I_L and V_P) in comparison with the original circuit that had three (V_{set}, V_{cond} and R_L). In the second part of this chapter, it is investigated the minimum device requirements and optimal circuit parameters required for implementation of material implication in a memristor crossbar.

Chapter III describes the two fabrication pathways developed in order to achieve stacked memristor-based structures. A stacked array of 2x2 TiO_2-based memristors was fabricated using lift-off techniques. CMP polishing was used to insure a smooth surface for the top layer of memristors. Another more industrial-relevant pathway was developed using ion milling techniques for improved manufacturability. A 10x10 crossbar was fabricated for demonstration purposes. The ion-milling based fabrication pathway is highly flexible and adaptable for further integration on CMOS chips.

Chapter IV describes the first experimental demonstration of three-dimensional data manipulation using material implication. The stacked structures from Chapter III were used to demonstrate an inter-layer NAND gate with the inputs and output in different device layers. This NAND gate showed 94% yield which proves the potential for using the inter-layer stateful logic gates in larger circuits, such as digital memories and hybrid programmable logic. Chapter IV also describes the first experimental demonstration of a half adder circuit implemented using memristor-based material implication entirely in a monolithically integrated memristor structure.

Chapter V describes how memristor-based material implication can opens the way to achieve one of the Feynman Grand Challenges - the construction of a functional nano-scale 8-bit adder in 50x50x50nm. Two possible alternatives are describes. The first part is dedicated to a sequential full adder that requires only 6 memristors. The inputs and outputs

are written and read sequentially for each bit. The second part describes an 8-bit adder that stores all the inputs and outputs. The read is performed only once, at the end of the 8th bit.

A summary of the main conclusions of this dissertation and future work are presented in Chapter VI.

References Chapter 1:

1. Neumann, V. J., 1945. *First Draft of a Report on the EDVAC. University of Pennsylvania, Moore School of Electrical Engineering*.

2. Backus, J. (1978). Can programming be liberated from the von Neumann style?: a functional style and its algebra of programs. *Communications of the ACM*, *21*(8), 613-641.

3. Hennessy, J. L., & Patterson, D. A. (2011). *Computer architecture: a quantitative approach*. Elsevier.

4. Bell, G., Sites, R., Dally, W., Ditzel, D., & Patt, Y. (1996). Architects Look to Processors of Future. *Microprocessor Report, Microdesign Resources*, *10*(10)

5. Amdahl, G. M. (1967). Validity of the single processor approach to achieving large scale computing capabilities. *Proceedings of the 1967 Spring Joint Computer Conference* (pp. 483-485). ACM.

6. TRW Computer Division, 1963, p. 17.

7. Karayiannis, N., & Venetsanopoulos, A. N. (2013). *Artificial neural networks: learning algorithms, performance evaluation, and applications* (Vol. 209). Springer Science & Business Media.

8. Prezioso, M., Merrikh-Bayat, F., Hoskins, B. D., Adam, G. C., Likharev, K. K., & Strukov, D. B. (2015). Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature*, *521*(7550), 61-64.

9. Hauck, S., & DeHon, A. (2010). *Reconfigurable computing: the theory and practice of FPGA-based computation*. Morgan Kaufmann.

10. Matsunaga, S., Hayakawa, J., Ikeda, S., Miura, K., Hasegawa, H., Endoh, T., ... & Hanyu, T. (2008). Fabrication of a nonvolatile full adder based on logic-in-memory architecture using magnetic tunnel junctions. *Applied Physics Express*, *1*(9), 091301.

11. Lent, C. S., Tougaw, P. D., Porod, W., & Bernstein, G. H. (1993). Quantum cellular automata. *Nanotechnology*, *4*(1), 49

12. Cowburn, R. P., & Welland, M. E. (2000). Room temperature magnetic quantum cellular automata. *Science*, *287*(5457), 1466.

13. Behin-Aein, B., Datta, D., Salahuddin, S., & Datta, S. (2010). Proposal for an all-spin logic device with built-in memory. *Nature nanotechnology*, *5*(4), 266-270.

14. Borghetti, J., Snider, G. S., Kuekes, P. J., Yang, J. J., Stewart, D. R., & Williams, R. S. (2010). 'Memristive'switches enable 'stateful'logic operations via material implication. *Nature*, *464*(7290), 873-876.

15. Chua, L. O. (1971). Memristor-the missing circuit element. *Circuit Theory, IEEE Transactions on*, *18*(5), 507-519.

16. Chua, L. O., & Kang, S. M. (1976). Memristive devices and systems.*Proceedings of the IEEE*, *64*(2), 209-223.

17. Adhikari, S. P., Sah, M. P., Kim, H., & Chua, L. O. (2013). Three fingerprints of memristor. *Circuits and Systems I: Regular Papers, IEEE Transactions on*,*60*(11), 3008-3021.

18. Gibbons, J. F., & Beadle, W. E. (1964). Switching properties of thin NiO films.*Solid-State Electronics*, *7*(11), 785-790.

19. Sliva, P. O., Dir, G., & Griffiths, C. (1970). Bistable switching and memory devices. *Journal of Non-Crystalline Solids*, *2*, 316-333.

20. Yang, J. J., Strukov, D. B., & Stewart, D. R. (2013). Memristive devices for computing. *Nature nanotechnology*, *8*(1), 13-24.

21. Sawa, A. (2008). Resistive switching in transition metal oxides. *Materials today*, *11*(6), 28-36.

22. Govoreanu, B., Kar, G. S., Chen, Y. Y., Paraschiv, V., Kubicek, S., Fantini, A., ... & Jurczak, M. (2011, December). 10× 10nm 2 Hf/HfO x crossbar resistive RAM with excellent performance, reliability and low-energy operation. In *Electron Devices Meeting (IEDM), 2011 IEEE International* (pp. 31-6). IEEE.

23. Whitehead, A. N., & Russell, B. (1910). *Principia mathematica*.

24. Shannon, C. (1937), "A Symbolic Analysis of Relay and Switching Circuits," unpublished MS Thesis, Massachusetts Institute of Technology.

25. Borghetti, J., Snider, G. S., Kuekes, P. J., Yang, J. J., Stewart, D. R., & Williams, R. S. (2010). 'Memristive'switches enable 'stateful'logic operations via material implication. *Nature*, *464*(7290), 873-876.

26. Sun, Xianwen, et al. "Unipolar memristors enable "stateful" logic operations via material implication." Applied Physics Letters 99.7 (2011): 072101.

27. Rosezin, R., et al. "Crossbar logic using bipolar and complementary resistive switches." Electron Device Letters, IEEE 32.6 (2011): 710-712.

28. Prezioso, M., et al. "A Single-Device Universal Logic Gate Based on a Magnetically Enhanced Memristor." Advanced Materials 25.4 (2013): 534-538.

29. Elstner, Martin, et al. "Molecular logic with a saccharide probe on the few-molecules level." Journal of the American Chemical Society 134.19 (2012): 8098-8100.

30. Jiang, Yanan, et al. "Highly-Efficient gating of solid-state nanochannels by DNA supersandwich structure containing ATP aptamers: A nanofluidic IMPLICATION logic device." Journal of the American Chemical Society 134.37 (2012): 15395-15401.

31. Roa, Elkim, Wu-Hsin Chen, and Byunghoo Jung. "Material implication in CMOS: A new kind of logic." Design Automation Conference (DAC), 2012 49th ACM/EDAC/IEEE. IEEE, 2012.

32. Mahmoudi, Hiwa, et al. "Implication logic gates using spin-transfer-torque-operated magnetic tunnel junctions for intrinsic logic-in-memory." Solid-State Electronics 84 (2013): 191-197.

33. Kvatinsky, S., Wald, N., Satat, G., Kolodny, A., Weiser, U. C., & Friedman, E. G. (2012, August). MRL—memristor ratioed logic. In *Cellular Nanoscale Networks and Their Applications (CNNA), 2012 13th International Workshop on* (pp. 1-6). IEEE.

34. Shen, W. C., Tseng, Y. H., Chih, Y. D., & Lin, C. J. (2011). Memristor logic operation gate with share contact RRAM cell. *Electron Device Letters, IEEE*, *32*(12), 1650-1652.

35. Hamdioui, S., Xie, L., Nguyen, H. A. D., Taouil, M., Bertels, K., Corporaal, H., ... & van Lunteren, J. (2015, March). Memristor based computation-in-memory architecture for data-intensive applications. In Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition (pp. 1718-1725). EDA Consortium.

36. Prezioso, M., Merrikh-Bayat, F., Hoskins, B. D., Adam, G. C., Likharev, K. K., & Strukov, D. B. (2015). Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature*, *521*(7550), 61-64.

37. Lehtonen, E., Poikonen, J. H., & Laiho, M. (2012, August). Applications and limitations of memristive implication logic. In *Cellular Nanoscale Networks and Their Applications (CNNA), 2012 13th International Workshop on* (pp. 1-6). IEEE.

38. Lehtonen, E., Poikonen, J., & Laiho, M. (2012, May). Implication logic synthesis methods for memristors. In *Circuits and Systems (ISCAS), 2012 IEEE International Symposium on* (pp. 2441-2444). IEEE.

39. Kim, K., Shin, S., & Kang, S. M. (2011, May). Stateful logic pipeline architecture. In *Circuits and Systems (ISCAS), 2011 IEEE International Symposium on* (pp. 2497-2500). IEEE.

40. Levy, Y., Bruck, J., Cassuto, Y., Friedman, E. G., Kolodny, A., Yaakobi, E., & Kvatinsky, S. (2014). Logic operations in memory using a memristive Akers array. *Microelectronics Journal*, *45*(11), 1429-1437.

41. Paul, S., & Bhunia, S. (2012). A scalable memory-based reconfigurable computing framework for nanoscale crossbar. *Nanotechnology, IEEE Transactions on*, *11*(3), 451-462.

42. Linn, E., Rosezin, R., Tappertzhofen, S., Böttger, U., & Waser, R. (2012). Beyond von Neumann—logic operations in passive crossbar arrays alongside memory operations. *Nanotechnology*, *23*(30), 305205.

43. Zheng, L., Shin, S., & Kang, S. M. S. (2014). Memristor-based ternary content addressable memory (mTCAM) for data-intensive computing. *Semiconductor Science and Technology*, *29*(10), 104010.

Chapter II

Circuit optimization for memristor stateful logic

This chapter describes an optimized circuit configuration that can perfom memristor implication logic with maximum margin to device variations. The motivation for this work was that so far prohibitively large device variability in the most prospective memristor-based circuits has limited experimental demonstrations to simple gates and circuits and just a few cycles of operations. Determining the circuit configuration most tolerant to variations is of utmost importance for reliable multi-cycle multi-gate operation.

The margins of operation were first investigated for two isolated devices, the conditional and the switching devices. The second part of the chapter is dedicated to exploring the margins of operation and most advantageous biasing scheme for a crossbar with N x N devices.

A. Motivation

Implication operation Q←P IMP Q implemented with memristors is based on conditional switching of one device, which is dependent on having specific voltage drops across the devices involved for correct functionality. Therefore significant set threshold voltage variations (Fig. 2.1) is a major challenge for implementing IMP logic.

Figure 2.1. Memristor device variation (a) A sketch of simplified (linear) I-V switching curve for a memristor. The thick (thin) solid lines show schematically an I-V curve with average (maximum and minimum) set and reset thresholds. The inset shows experimental setup. (b) Switching I-Vs showing 100 cycles of operation for a characteristic device and the corresponding cycle-to-cycle set switching voltage statistics.
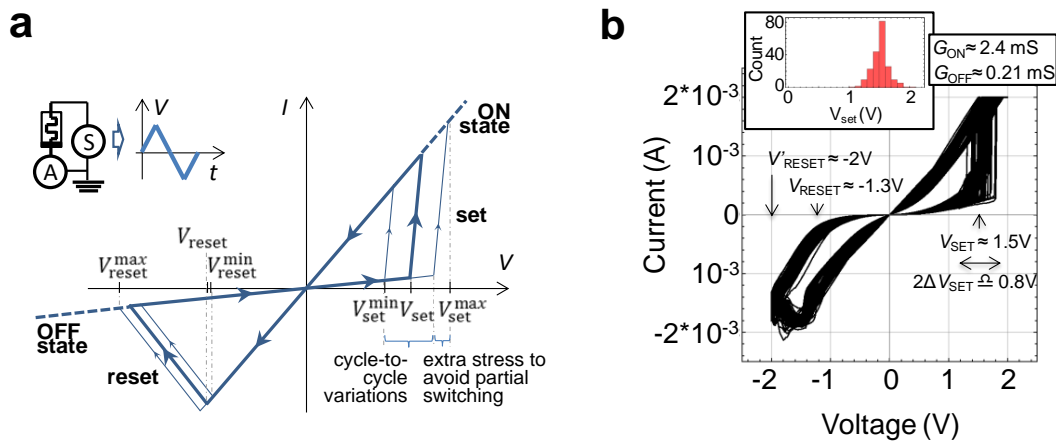
Therefore, it is natural to choose circuit parameters (i.e. G_L, V_L, V_P) that maximize the range of variations, also referred as margins, which can be tolerated without comprising the correctness of logic operation. G_L plays a crucial role in the current functionality of the system, by modulating the potential at the node and the current flow. The best possible value of G_L satisfies with the largest safety margin the voltage constrains mentioned above. This is of particular importance in systems where devices have high behavioral variation, which is the case currently for the memristor technology.

Some earlier works suggested choosing G_L between the ON and OFF conductance values [1,2] of the memristors performing the stateful logic, with a value $\sqrt{G_{ON}G_{OFF}}$ being the most cited [3,4]. However, our simple analysis of IMP logic operation showed in the next section proves that set margins monotonically increase as the load conductance decreases which leads to a modified circuit diagram.

B. Two device case

1. Analytical investigation

The optimal circuit parameters V_P, V_L and G_L, which result in the largest set margins could be derived analytically for the memristors with linear *I-V* (Fig.2.1a). Let us first consider an IMP circuit with specific "parallel" configuration of memristors (Fig. 2.2a).

Figure 2.2. (a) Parallel and (b) anti-parallel polarity configuration for memristor-based IMP logic (see Chapter 4.A. for detailed measurement setup).
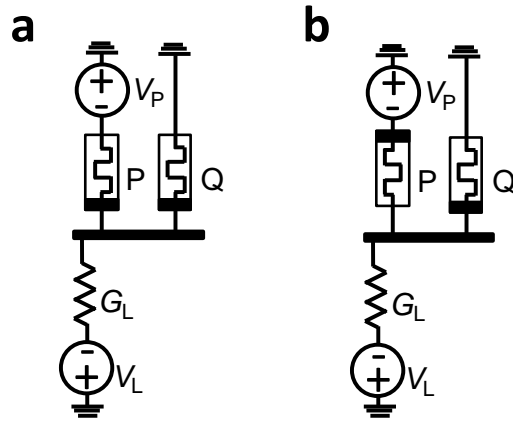
Assuming for convenience that $V_Q = 0$, the proper operation of the material implication logic circuit shown on Figs. 1a, c require that device Q is set only when both P and Q are in the OFF state, i.e.

$$-V_C|_{\mathrm{P=OFF,Q=OFF}} \geq V_{\mathrm{set}}^{\max} \qquad (1)$$

$$-V_C|_{\mathrm{OTHERS}} < V_{\mathrm{set}}^{\min} \qquad (2)$$

where

$$V_C = \frac{-G_L V_L - G_P V_P}{G_L + G_P + G_Q} \qquad (3)$$

is a voltage on the common electrode. Device P should not be disturbed during the IMP operation, i.e.

$$(V_P - V_C)|_{\mathrm{ANY}} < V_{\mathrm{set}}^{\min} \qquad (4)$$

$$(V_P - V_C)|_{\mathrm{ANY}} > V_{\mathrm{reset}}^{\min} \qquad (5)$$

Equations 1, 2, 4, and 5 define 12 inequalities in total. To eliminate redundant inequalities, let us first note that $V_L \geq 0$ does not have valid solutions, while $V_P \geq 0$ always results in sub-optimal margins. Assuming $V_P < 0$ and $V_L < 0$ and that memristors P and Q are characterized by the same parameters $V_{\mathrm{set}}^{\min}$, $V_{\mathrm{set}}^{\max}$, $V_{\mathrm{reset}}^{\min}$, $V_{\mathrm{reset}}^{\max}$, G_{ON}, G_{OFF} (a more general case is discussed later) only three conditions must be considered, namely:

- voltage drop on device Q, when Q and P are in the OFF states, is larger than $V_{\mathrm{set}}^{\max}$,
- voltage drop on device Q, when Q and P are in the ON and OFF states, respectively, is smaller than $V_{\mathrm{set}}^{\min}$, and
- voltage drop on device P, when Q and P are in the OFF states, is smaller than $V_{\mathrm{set}}^{\min}$.

Therefore, the largest set margins and the corresponding optimal parameters can be found by solving the following equations:

26

$$\frac{-V_{\text{P}}G_{\text{OFF}} - V_{\text{L}}G_{\text{L}}}{2G_{\text{OFF}} + G_{\text{L}}} = V_{\text{set}}^* + \Delta_{\text{ideal}} \tag{6}$$

$$\frac{-V_{\text{P}}G_{\text{ON}} - V_{\text{L}}G_{\text{L}}}{G_{\text{OFF}} + G_{\text{ON}} + G_{\text{L}}} = V_{\text{set}}^* - \Delta_{\text{ideal}} \tag{7}$$

$$\frac{V_{\text{P}}(G_{\text{ON}} + G_{\text{L}}) - V_{\text{L}}G_{\text{L}}}{2G_{\text{OFF}} + G_{\text{L}}} = V_{\text{set}}^* - \Delta_{\text{ideal}} \tag{8}$$

where

$$V_{\text{set}}^* = (V_{\text{set}}^{\max} + V_{\text{set}}^{\min})/2 \tag{9}$$

Here, Δ_{ideal} is a set margin for the binary zero-variations (i.e. ideal for the considered application) memristors for which $V_{\text{set}}^* = V_{\text{set}}^{\max} = V_{\text{set}}^{\min}$ (Fig. 2.3a). Accounting for variations in set switching threshold and analog switching, a more relevant for our case margin is

$$\Delta = \Delta_{\text{ideal}} - (V_{\text{set}}^{\max} - V_{\text{set}}^{\min})/2 \tag{10}$$

From Eqs. (7-9) V_{P}, V_{L} and Δ_{ideal} are

$$\Delta_{\text{ideal}} = V_{\text{set}}^* \frac{G_{\text{ON}} - G_{\text{OFF}}}{2G_{\text{L}} + 3G_{\text{ON}} + G_{\text{OFF}}} \tag{11}$$

$$V_{\text{P}} = -2\Delta_{\text{ideal}} \tag{12}$$

$$V_{\text{L}} = -2V_{\text{set}}^* \frac{G_{\text{L}}^2 + 2G_{\text{L}}(G_{\text{ON}} + G_{\text{OFF}}) + G_{\text{OFF}}(3G_{\text{ON}} + G_{\text{OFF}})}{G_{\text{L}}(2G_{\text{L}} + 3G_{\text{ON}} + G_{\text{OFF}})} \tag{13}$$

According to Eq. 10 Δ_{ideal} is monotonically decreasing with G_{L} (Fig. 2.3b) and the maximum margins are achieved for $G_{\text{L}} = 0$.

$$\Delta_{\text{ideal}} = V_{\text{set}}^* \frac{G_{\text{ON}}/G_{\text{OFF}} - 1}{3G_{\text{ON}}/G_{\text{OFF}} + 1} \tag{14}$$

$$I_L \equiv V_L G_L = -2V_{\text{set}}^* G_{\text{OFF}} \tag{15}$$

Figure 2.3. (a) A diagram showing definition of margins in the context of set transition (b) The set margins as a function of load conductance for several representative ON-to-OFF conductance ratios. For convenience, margins and load conductances are normalized with respect to mid-range set voltages V_{set}^* and G_{ON}, respectively. Solid dots show margins for previously proposed optimal load conductance G_L', while solid triangles are margins which were obtained with numerical simulations using experimental device characteristics. The solid and dashed horizontal lines denote the maximum and the actual set margins, respectively, when taking into account experimental data.
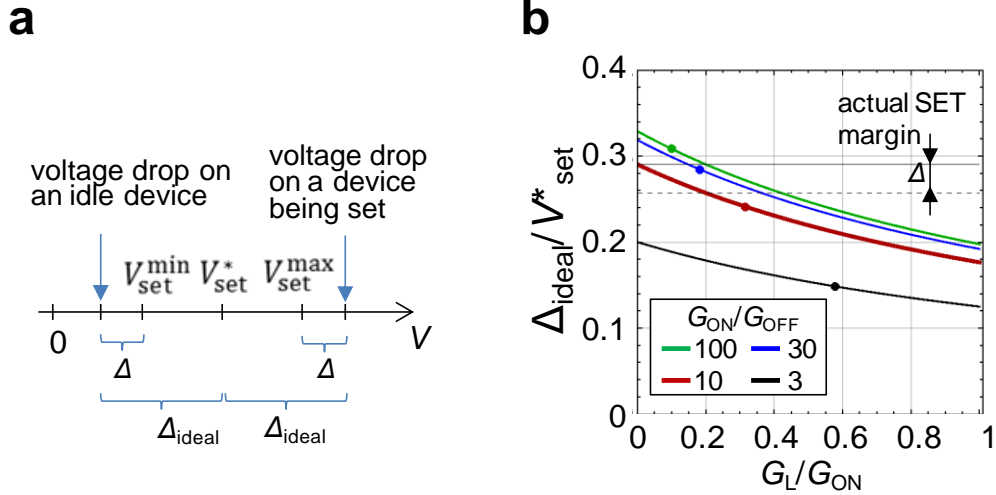
The largest margins are for GL = 0, which cannot be implemented with the original circuit, though can be easily realized by replacing the load resistance and voltage source with a current source (Fig. 2.4). The transition from the original circuit with earlier suggested G_L'

to the modified one with an optimized current source I_L increased set margins by more than 20% (Fig. 2.3b). Such a boost in variation tolerance was critical for our experimental setup by allowing it to cope with virtually all experimentally observed variations (Chapter 3, Fig. 3.11).

Figure 2.4. Modified IMP logic circuit with memristors in parallel configuration. A current source replaces the load resistor in the original configuration (see Chapter 4.A. for detailed measurement setup).
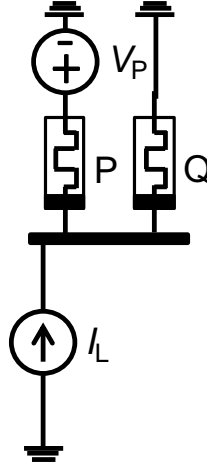
For devices with large ON-to-OFF conductance ratio, Eq. 13 can be approximated with very simple formula

$$\Delta_{\text{ideal}} \approx V_{\text{set}}^*/3. \tag{16}$$

It is instructive to compare IMP logic margins with those of passive crossbar memories. For example, let us consider the most optimal $V/3$-baising scheme,[1] and assume that voltages V and 0 are applied on the lines leading to the selected device, and $V/3$, and $2V/3$ on the corresponding lines leading to the remaining devices. Assuming that voltage across the selected device is $V = V_{\text{set}}^* + \Delta_{\text{memory}}$, while it is $V/3 = V_{\text{set}}^* - \Delta_{\text{memory}}$ across all other devices, it is straightforward to show that the margins for crossbar memory are

$$\Delta_{\text{memory}} = V_{\text{set}}^*/2. \tag{17}$$

Thus voltage margins for memory circuits are more relaxed as compared to those of IMP logic. In principle, a somewhat larger IMP logic set margins can be obtained by not enforcing full switching, e.g. by defining V_{set}^{max} as the largest set threshold voltage due to cycle-to-cycle variations. However, in this case, the ON-to-OFF ratio will get reduced with every IMP logic operation, which is not desirable.

The analysis above is for a specific IMP logic based on memristors with identical linear static *I-V* characteristics. It is straightforward to extend it to a more general case by using specific to memristors Q and P parameters in Eqs. (S6-S8), such as different set and reset threshold voltages for the top and bottom devices, which is the case relevant to the implemented circuit. For example, a more general set of equations for parallel configuration shown on Fig. S6a, which is more convenient to solve for Δ directly, is

$$\frac{-V_P G_{OFF}-V_L G_L}{2G_{OFF}+G_L} = V_{Q\,set}^{max} + \Delta, \quad \frac{-V_P G_{ON}-V_L G_L}{G_{OFF}+G_{ON}+G_L} = V_{Q\,set}^{min} - \Delta, \quad \frac{V_P(G_{ON}+G_L)-V_L G_L}{2G_{OFF}+G_L} = V_{P\,set}^{min} - \Delta \quad (18)$$

from which the actual margin for $G_L = 0$ is

$$\Delta = \frac{(G_{ON}+G_{OFF})\left(V_{Q\,set}^{min}-V_{Q\,set}^{max}\right) + (G_{ON}-G_{OFF})V_{P\,set}^{min}}{3G_{ON}+G_{OFF}} \quad (19)$$

For anti-parallel configuration shown on Fig. 2.2b, the set of equation is

$$\frac{-V_P G_{OFF}-V_L G_L}{2G_{OFF}+G_L} = V_{Q\,set}^{max} + \Delta, \quad \frac{-V_P G_{ON}-V_L G_L}{G_{OFF}+G_{ON}+G_L} = V_{Q\,set}^{min} - \Delta, \quad \frac{V_P(G_{ON}+G_L)-V_L G_L}{2G_{OFF}+G_L} = -(V_{P\,set}^{min} - \Delta) \quad (20)$$

and the actual margin for $G_L = 0$ is

$$\Delta^{anti} = \frac{(G_{ON}+G_{OFF})(V_{Q\,set}^{min}-V_{Q\,set}^{max}) - (G_{ON}-G_{OFF})V_{P\,reset}^{min}}{3G_{ON}+G_{OFF}} \quad (21)$$

It should be noted that, in principle, IMP logic can also be implemented using a memristor's reset transition, i.e. assuming that logic states "0" and "1" are represented by the ON and OFF states instead. However, this would not be helpful in our case, because the gradual reset transition presents even larger problem. Because $-V_{reset}^{min} > V_{set}^{min}$ typically holds for the considered devices (see Chapter 3 for more details), from Eqs. 19 and 21 margins for parallel case are smaller, which is why this case is considered more in detail. Margins and optimal parameters for the remaining parallel and anti-parallel configurations that were experimentally demonstrated in Chapter 3, are similar to those described above with the only difference is that the signs for V_P and I_L are negative.

2. Numerical simulations

Analytical approach can be also utilized for IMP logic based on the memristors with more realistic nonlinear static I-V by using G_{ON} and G_{OFF} measured at large (close to switching threshold) voltages. A more accurate approach, however, is to solve inequalities Eqs. (S1-S5) numerically.

A damped Newton-Raphson-based solver, implemented in Mathematica 10 using the FindRoot function, was used to solve for the currents in the system with two memristor and a RLoad. The solver utilized a fitting of non-linear I-V characteristics from real memristor devices. The fitting was done on log-log data using a polynomial function of 7th degree. The fitting function shows a good fit with R2 > 0.999 and is forced to pass through zero, since the current should be zero if the applied voltage is zero. The solver has 99.97% convergence for 22,000 generated points. The 6 points that did not converge in 100 iterations were discarded.

Table 2.1 describes all the 16 constraints that have to be imposed on the voltage drops of devices P and Q in order to obtain the desired state for each specific case. Device Q is assumed to be the device switching and retaining the result of the implication logic operation. Device P serves as an enabling device allowing for the voltage drop on Q to be modulated according to its state and therefore, facilitating the conditional switching of P. The device Q has to have a high enough voltage drop in case 1 (case where both devices are OFF) in order to set to ON state. Hence the voltage drop on Q should be higher than $V_{SET} + \Delta$ of the device, but lower than a protection voltage called $V_{SET\ max}$ above which the device might get damaged. While device Q is switching, the device P should not be perturbed since the device Q switching is dependent on the memristance value of P. The voltage drop on P should be bounded between the $V_{SET} - \Delta$ and $V_{RESET} - \Delta$, in order to avoid the device to turn ON or, respectively, to turn more OFF. In all the other cases 2-4, both devices P and Q should be under non perturbing conditions with voltage drops between $V_{SET} - \Delta$ and $V_{RESET} - \Delta$, in order to not damage the logic values stored in these devices.

Table 2.1. Required voltage constraints for device P and device Q for each of the four logic cases in the material implication operation. Shaded with gray are the two critical cases.

		P	Q	P*	Q*	Constraints	
1	OFF-OFF	0	0	0	1	P not perturbed Q sets	$V_{reset} - \Delta < V_{drop\ P} < V_{set} - \Delta$ $V_{set} + \Delta < V_{drop\ Q} < V_{set\ max}$
2	OFF-ON	0	1	0	1	P not perturbed Q not perturbed	$V_{reset} - \Delta < V_{drop\ P} < V_{set} - \Delta$ $V_{reset} - \Delta < V_{drop\ Q} < V_{set} - \Delta$
3	ON-OFF	1	0	1	0	P not perturbed Q not perturbed	$V_{reset} - \Delta < V_{drop\ P} < V_{set} - \Delta$ $V_{reset} - \Delta < V_{drop\ Q} < V_{set} - \Delta$
4	ON-ON	1	1	1	1	P not perturbed Q not perturbed	$V_{reset} - \Delta < V_{drop\ P} < V_{set} - \Delta$ $V_{reset} - \Delta < V_{drop\ Q} < V_{set} - \Delta$

Figure 2.5. Fitting to experimental data used for numerical simulations.
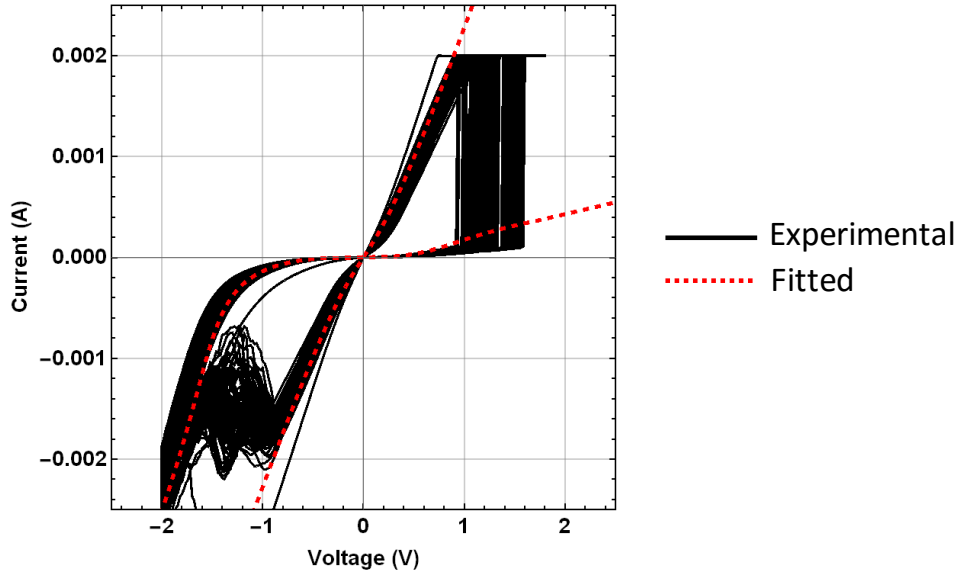
Figure 2.6. The area of acceptable voltages increases with decreasing G_L.
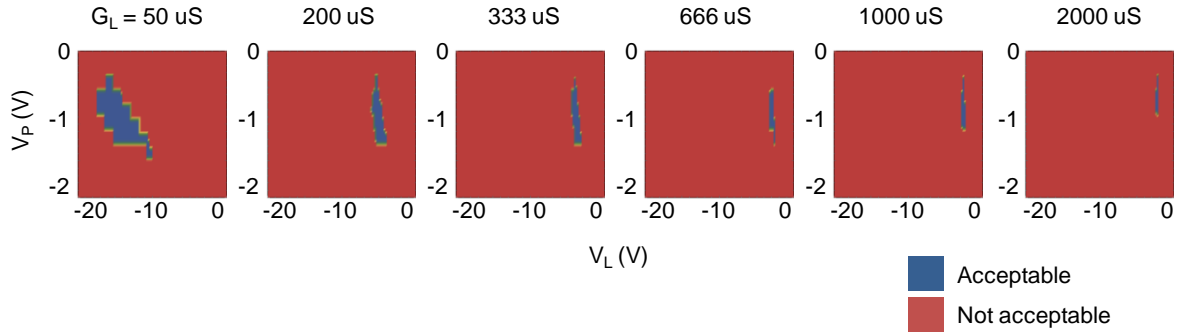
By fitting experimental *I-V* curves (Fig. 2.5) and using Mathematica's Newton-Raphson-based solver, graphical plots were derived showing acceptable ranges of Vp and V_L for various G_Ls in the case of ideal devices requiring zero conditional switching margin to variations. From Fig. 2.6, the area of acceptable voltages increases as the G_L decreases confirming the analytical results.

By introducing a non-zero switching margin term in the constrains, the area of the acceptable region decreases. The highest value of margin for a particular G_L is considered the value at which the acceptable region vanishes in the graph (Fig. 2.7). This last acceptable point provides the optimal values for V_P and V_L.

Figure 2.7. The area of acceptable voltages decreases with increasing margin required.
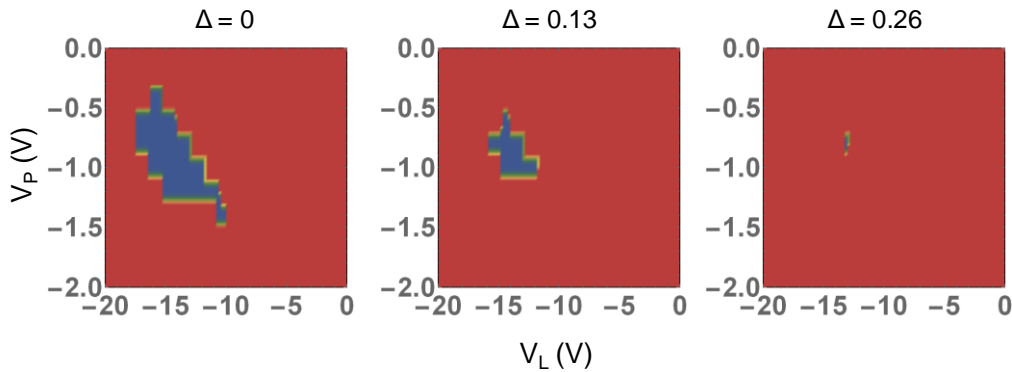
The margins calculated from a numerical simulations for a specific IMP logic are shown on Fig. 2.8 and are in fairly good agreement with simple analytical model for a system with an ON-to-OFF conductance ratio of ~10. A step of 0.01V was used which limits the accuracy of the graphical method.

Figure 2.8. Analytical linear case results vs. numerical non-linear case results.
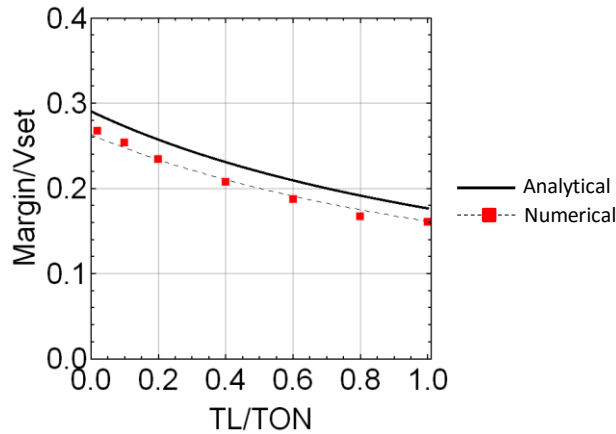
C. N x N crossbar case

A crossbar of n x n devices with integrated selectors is analyzed. In order to perform an analytical determination, the integrated memristor/selector devices in the crossbar can be assumed to have a piecewise linear characteristics (Fig 2.9a).The crossbar is lumped into distinct components as described in Fig. 2.9b.

$$I_{ON}[V] = \begin{cases} G_{ON} * (V + V_{th}) - G_{SEL} * V_{th}, \text{for } V < -V_{th} \\ G_{SEL} * V, \text{for } V \in [-V_{th}, V_{th}] \\ G_{ON} * (V - V_{th}) + G_{SEL} * V_{th}, \text{for } V > V_{th} \end{cases} \tag{1}$$

$$I_{OFF}[V] = \begin{cases} G_{OFF} * (V + V_{th}) - G_{SEL} * V_{th}, \text{for } V < -V_{th} \\ G_{SEL} * V, \text{for } V \in [-V_{th}, V_{th}] \\ G_{OFF} * (V - V_{th}) + G_{SEL} * V_{th}, \text{for } V > V_{th} \end{cases} \tag{2}$$

Figure 2.9. Analytical linear case for an NxN memristor array. (a) Assumed linear model with selector before V_{th}. (b) Lumped model for the NxN memristor array. (c) Constraints imposed on the switching device, idle device and the unused devices.
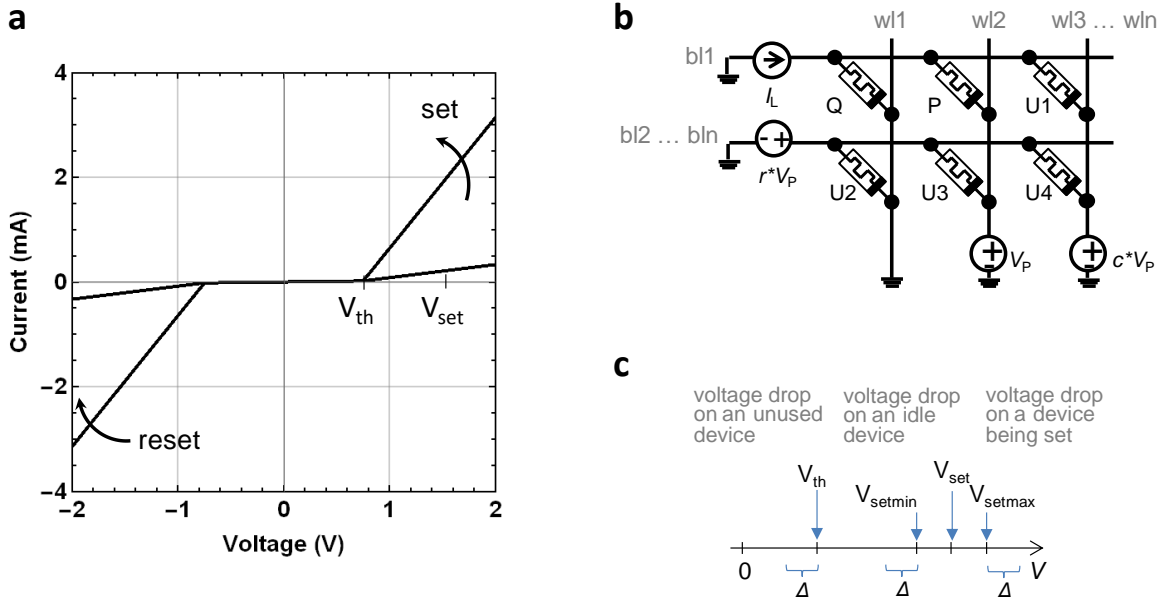
Through the selection of parameters c and r, the voltage drops on the unused devices U1 to U4 are between $-V_{th}$ and V_{th}, therefore the states of these devices is always masked by

35

the OFF state of the selector (Fig. 2.9c). These strict conditions are necessary to insure that the power consumption stays low, as it will be shown later in Fig. 2.11b.

Using Kirchoff's current law on bl1, the following equation is valid:

$$I_L = I_Q + I_P + (n - 1) * I_{U1} \tag{3}$$

The voltage drops on all devices by respect to V_Q are:

$$V_P = V_Q - V_{cond} \tag{4}$$

$$V_{U1} = V_Q - c * V_{cond} \tag{5}$$

$$V_{U2} = r * V_{cond} \tag{6}$$

$$V_{U3} = (r - 1) * V_{cond} \tag{7}$$

$$V_{U4} = (r - c) * V_{cond} \tag{8}$$

Similar to the two device case, only two out of four cases are important: when both Q and P are OFF and when Q is OFF and P is ON. The constraints for the remaining cases are automatically satisfied. When the P is ON, it is assumed that the voltage drop on P is high enough to be above V_{th}. By substituting eq. 4 and 5 into eq. 3 and solving for voltage drop on Q, the following results are obtained:

36

$$V_Q\big|_{P=OFF,Q=OFF} = \frac{I_L + G_{OFF}(1+c\,(n-2))V_{cond}}{n * G_{OFF}} \qquad (9)$$

$$V_Q\big|_{P=ON,Q=OFF} = \frac{I_L + G_{OFF}(c\,(n-2)V_{cond}-V_{th})+G_{ON}(V_{cond}+V_{th})}{G_{ON}+(n-1)*G_{OFF}} \qquad (10)$$

The margin and the external parameters I_L, V_{cond}, c and r are determined from the following system of constrains using the results from eq. 9 and 10 and eq. 4-6:

$$V_Q\big|_{P=OFF,Q=OFF} == V^*_{set} + \Delta \qquad (11)$$

$$V_P\big|_{P=OFF,Q=OFF} == V^*_{set} - \Delta \qquad (12)$$

$$V_Q\big|_{P=ON,Q=OFF} == V^*_{set} - \Delta \qquad (13)$$

$$V_{U1}\big|_{P=OFF,Q=OFF} == V_{th} - \Delta \qquad (14)$$

$$V_{U2}\big|_{P=OFF,Q=OFF} == V_{th} - \Delta \qquad (15)$$

The determined margin is not dependent on the selector OFF state value:

$$\Delta = \frac{(\frac{G_{ON}}{G_{OFF}}-1)(V^*_{set}-V_{th})}{\frac{3G_{ON}}{G_{OFF}}+2n-3} \qquad (16)$$

The necessary external parameters are determined using the equations below. The results from this generalized analysis can be used by using $V_{th}=0$, $n=2$ and $G_{SEL}=G_{OFF}$ to determine the two device case calculated in our previous work:

$$I_L = \frac{V_{th}\left[6\frac{G_{SEL}}{G_{OFF}}\frac{G_{ON}}{G_{OFF}}+2\frac{G_{SEL}}{G_{OFF}}(2n-3)+2\frac{G_{ON}}{G_{OFF}}(2n-7)+n^2-6n-7\right]-V_{set}^*\left[\frac{G_{ON}}{G_{OFF}}(n-8)-5n+8\right]}{\frac{3G_{ON}}{G_{OFF}}+2n-3}G_{OFF}^2 \quad (17)$$

$$V_{cond} = \frac{2\left(\frac{G_{ON}}{G_{OFF}}-1\right)(V_{set}^*-V_{th})}{\frac{3G_{ON}}{G_{OFF}}+2n-3} \quad (18)$$

$$c = \frac{5\left(\frac{G_{ON}}{G_{OFF}}-1\right)+2n}{2\left(\frac{G_{ON}}{G_{OFF}}-1\right)} \quad (19)$$

$$r = \frac{\left(\frac{G_{ON}}{G_{OFF}}+1\right)(V_{set}^*-4V_{th})+2nV_{th}}{2\left(\frac{G_{ON}}{G_{OFF}}-1\right)(V_{set}^*-V_{th})} \quad (20)$$

The voltage drops on devices U4 has to be between -V_{th} and V_{th} which imposes a constraint on V_{th}. If this constraint is satisfied, the voltage drop on devices U3 is automatically satisfied to be between -V_{th} and V_{th}.

$$V_{U4}|_{P=OFF,Q=OFF} == -V_{th} + \Delta \quad (21)$$

$$V_{th}^{min} = \frac{7\frac{G_{ON}}{G_{OFF}}+2n-7}{13\frac{G_{ON}}{G_{OFF}}+6n-13} \quad (22)$$

with $V_{th}^{min} \sim 0.538\, V_{set}$ for large ON/OFF ratios.

38

Figure 2.10. Margin for the case for an NxN memristor array. (a) Normalized margin as a function of V_{th}/V_{set}. Operational margin decreases with increasing V_{th}/V_{set} so for optimum performance is achieved as $V_{th} = V_{thmin}$. It was assumed that $G_{ON} = 1/400$, $G_{ON}/G_{OFF} = 10$ and $G_{OFF}/G_{sel} = 10$. (b) Margin as a function of G_{ON}/G_{OFF}. High ON/OFF ratio > 1000 is crucial in implementing a large scale system. It was assumed that $G_{ON} = 1/400$, $G_{OFF}/G_{sel} = 10$ and $V_{th}/V_{set} = 0.55$.
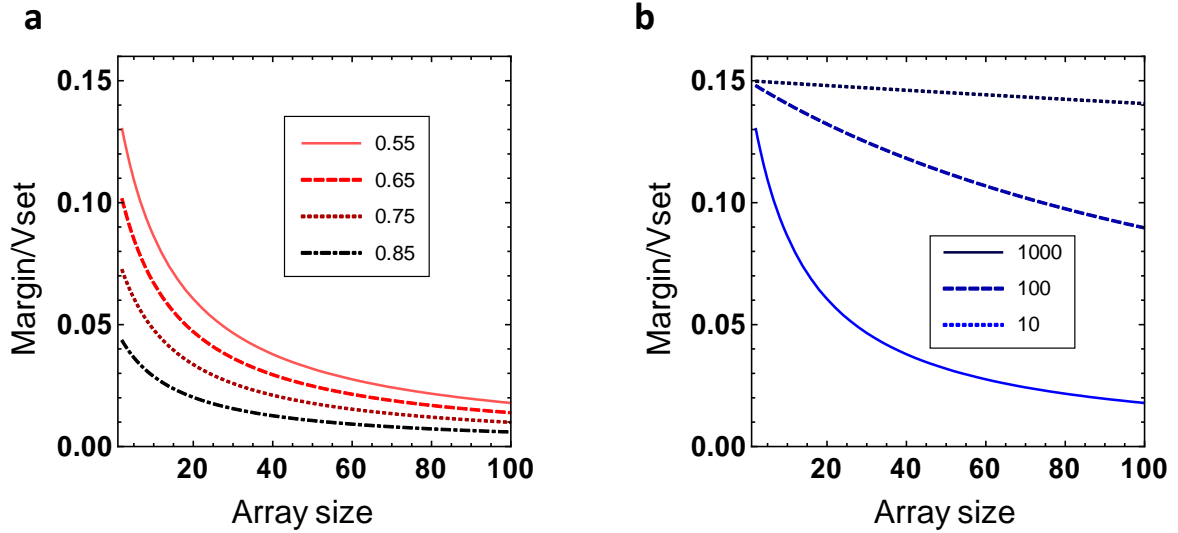
Figure 2.11. Current and power consumption. (a) Normalized current consumption as a function of different conditions. By imposing the constraint that the voltage drop on the unused devices $<V_{th}$, the current consumption is ~10% lower. Lower selector conductances make a difference in the current consumption at low array sizes. (b) Normalized power consumption as a function of different selector conductances G_{OFF}/G_{SEL}. Higher selector conductances keep the power consumption low at high array sizes. It was assumed that $G_{ON} = 1/400$, $G_{ON}/G_{OFF} = 10$ and $V_{th}/V_{set} = 0.55$.

39
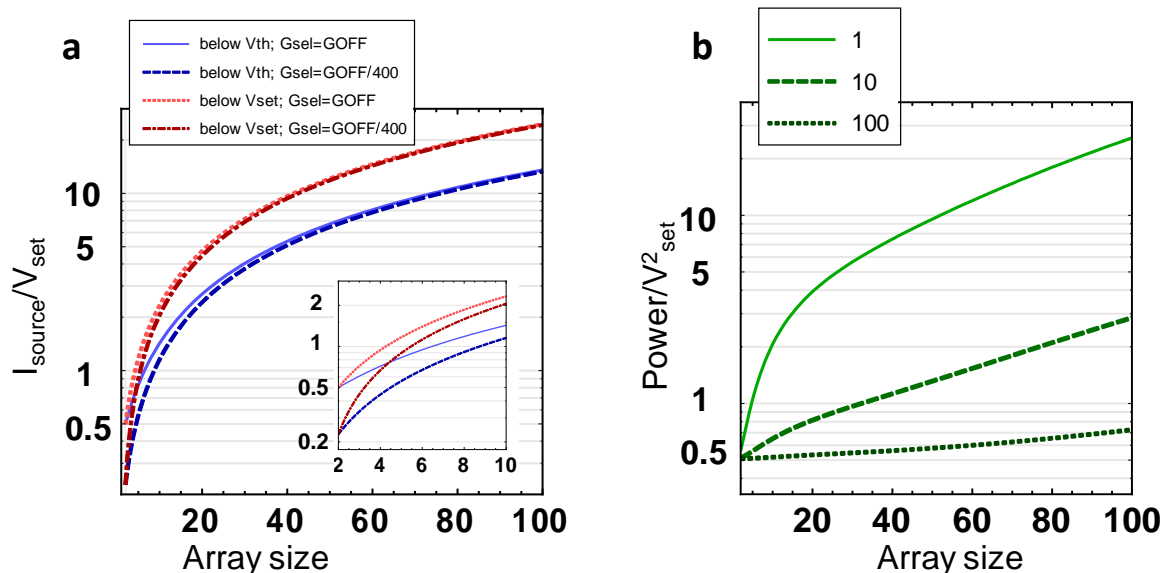
Figure 2.10a shows that a V_{th} close to $V_{th\ min}$ (~0.55 V_{set}) is needed to operate at maximum margin. At lower V_{th}, the system will function incorrectly since some unused devices will switch. For higher V_{th}, the margin will get smaller and the tolerance to device variation will disappear. In order to maintain a decent operational margin, a high ON/OFF ratio > 1000 is needed (Fig 2.10b). For an ON/OFF ratio of 10, which is the typically the case for current memristors, the margin decreases to below .05 V_{set} at an array size of 20. The non-linearity between the OFF curve of the device and the OFF curve of the selector plays no role in the margin, but drastically influences the power consumption (Fig. 2.11).

These theoretical results are needed to inform how to experimentally apply the novel current-based circuit framework for implication logic performed in a crossbar.

References for Chapter 2

1. Borghetti, J., Snider, G. S., Kuekes, P. J., Yang, J. J., Stewart, D. R., & Williams, R. S. (2010). 'Memristive'switches enable 'stateful' logic operations via material implication. *Nature*, *464*(7290), 873-876.

2. Kvatinsky, S., Kolodny, A., Weiser, U. C., & Friedman, E. G. (2011, October). Memristor-based IMPLY logic design procedure. In *Computer Design (ICCD), 2011 IEEE 29th International Conference on* (pp. 142-147). IEEE.

3. Lehtonen, E., Poikonen, J. H., & Laiho, M. *Memristor Networks.* (Adamatzky, A. & Chua, L. ed.) 603 (Springer International Publishing, Switzerland, 2014).

4. Kvatinsky,S., Satat,G., Wald,N., Friedman,E.G., Kolodny A.,Weiser,U.C. (2014). Memristor-based material implication (IMPLY) logic: Design principles and methodologies. *IEEE Transactions on VLSI Systems, 22*(10), 2054-2066.

Chapter III

Monolithically stacked memristor fabrication

3D stacked circuits allow for a much higher density as compared to the planar case. CMOS circuitry integrated with memristor multi-layers providing in-memory computing capabilities can offer a viable solution to the von Newmann bottleneck.

Memristors were fabricated in the bottom layer, then another layer of memristors were monolithically integrated directly above, with stacked pair of devices sharing a common middle electrode (Figure 3.1a). Successful stacked devices were fabricated using lift-off based techniques. In order to improved manufacturability, a different fabrication based on metal ion milling was developed and shows excellent potential for the multi-layer stacking of large memristor crossbar arrays.

The stacked structures were further used in Chapter 4 to show reliable multi-cycle 3D implication and implement implication-based gates and circuits.

41

Figure 3.1. Schematics of stacked memristor structures (a) Stacked memristor arrays were fabricated using lift-off techniques and (b) Stacked memristor crossbars require improved manufacturability achievable through an ion milling fabrication flow. Red denotes bottom devices, while with blue are the top ones. The middle electrode is shared between the bottom and top devices in a pair.
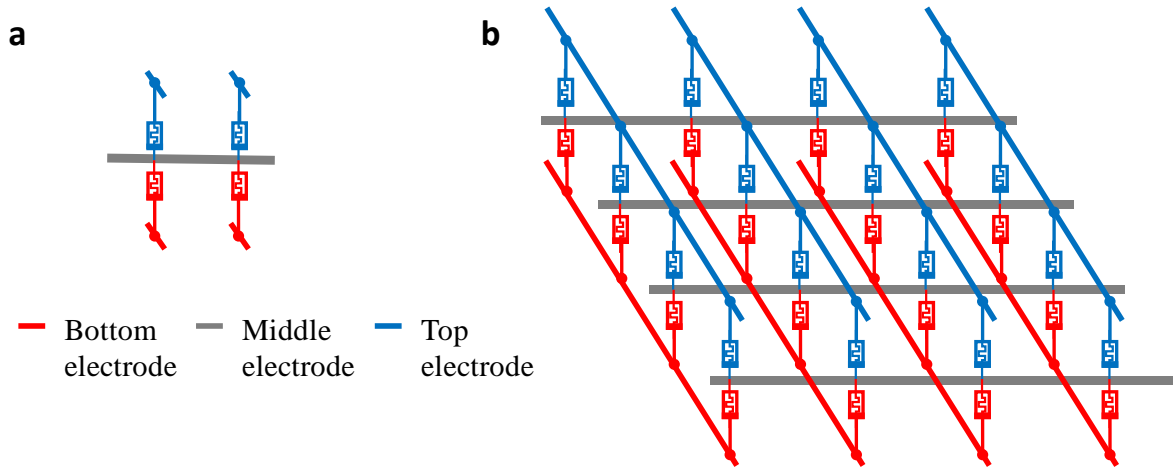
A. *Lift-off based fabrication*

1. Desired structure

This section presents the stacked memristor arrays fabricated using lift-off techniques. The desired device structure for the stacked memristors is presented in Figure 3.2. The major steps involved in fabrication are: patterning of Ta/Pt bottom electrode by e-beam evaporation and lift-off; patterning of bottom Al_2O_3/TiO_{2-x} device and Ti/Pt middle electrode by reactive sputtering and lift-off; planarization by chemical mechanical polishing and etch-back of plasma-deposited sacrificial silicon oxide; and, patterning of top Al_2O_3/TiO_{2-x} device and Ti/Pt top electrode by reactive sputtering and lift-off. All the steps will be presented in the next sections in detail. The next two sections focus on the most challenging steps: the selection of the deposition method for the TiO_{2-x} switching layer and the planarization of the bottom device to reduce step height for top device.

Figure 3.2. The Al_2O_3/TiO_{2-x} memristor circuit: fabrication details. A cartoon of device's cross-section showing the material layers and their corresponding thicknesses.
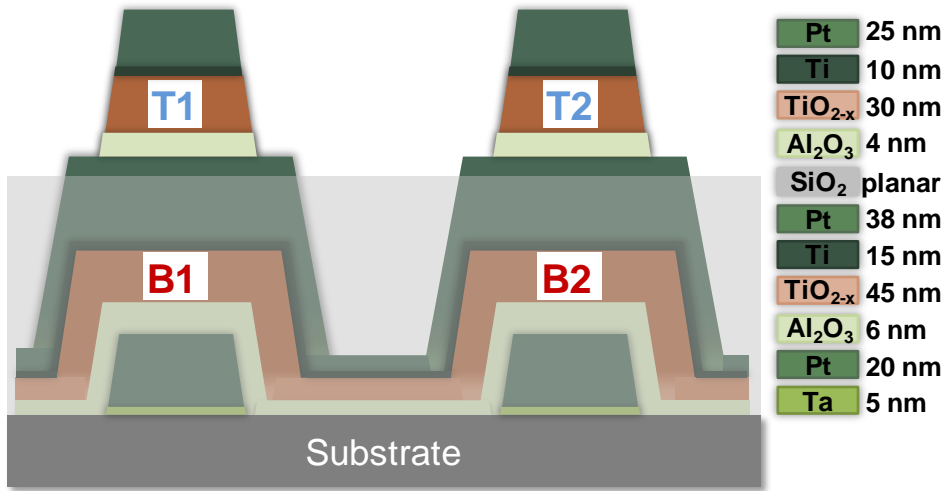
2. Choice of switching layer

Two deposition methods were attempted for the TiO_2 switching layer. Firstly, stacked memristors were fabricated using ALD-grown TiO_2 from TTIP precursor at 200ºC using H_2O. The switching layers were deposited in blanket and electrodes were evaporated ex-situ. Crystallite growth was observed in the TiO_2 films, a behavior also observed by Reiners et al [1]. The presence of these crystallites increased the chance of shorts and reduced the reproducibility of stacked memristors.

A closer look at Figure 3.3 shows that the crystallites are denser on the bottom electrode than on the middle electrode. There are several stacks of materials of interest across the entire structure. In the final fabricated structure, the bottom electrode in the region outside the device is covered with two layers of TiO_2. The middle electrode outside the device is patterned and deposited on one layer of TiO_2 and has the second layer of TiO_2 on top of it. The empty space between patterned features is covered with two TiO_2 layers.

Figure 3.3. Vertically stacked memristors based on ALD-grown TiO₂. (a) AFM image showing the heavy presence of crystallite growth over the entire surface of the device and adjacent areas. (b) Cross-section showing an average height of the crystallites of ~20-30nm and 200nm bunny-ear formations around the top electrodes.
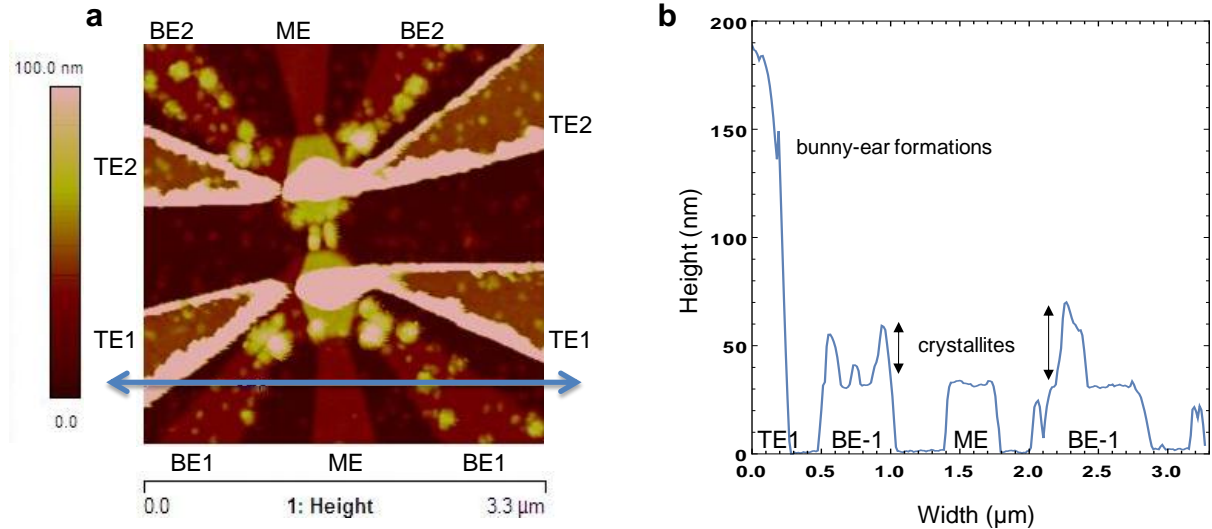
An investigation was carried to understand the substrate influence on the crystallite growth. The substrates of interest were thermal SiO_2 vs. Ti/Pt (5/25nm evaporated on Si/SiO_2 substrate using e-beam evaporator). One batch used pristine substrates with no exposure to photoresist (PR) and another batch had substrates first coated with photoresist SPR-995-0.9, then immediately cleaned with 1165 and 10min of descum by active oxygen dry etching at 350ºC.

Figure 3.4 (a) and (b) shows that the one layer films grown on Ti/Pt have enhanced crystallite density in comparison with the films grown on SiO_2 substrate. The number of crystallites increases when the second layer of TiO_2 is deposited (c and d). The worst crystallite growth is on the double TiO_2 layer contaminated with PR. Further cleaning of the PR-contaminated samples in active oxygen plasma at 350ºC had no effect.

Figure 3.4. Crystallite formation of ALD-grown TiO$_2$ on different substrates. Tests have been performed on both pristine surfaces and surfaces cleaned after PR coating (a) One layer of 30nm of ALD-TiO$_2$ on thermal SiO$_2$ shows almost no crystallites for both types of surfaces; (b) One layer of 30nm ALD-TiO$_2$ on Pt shows few crystallites for both types of surfaces; (c) Two layers of 30nm each of ALD-TiO$_2$ on thermal SiO$_2$ show few crystallites for the pristine surface and high number of crystallites for the PR contaminated surface. (d) Two layers of 30nm each of ALD-TiO$_2$ on Pt show few large crystallites for the pristine surface and high number for the PR contaminated surface.
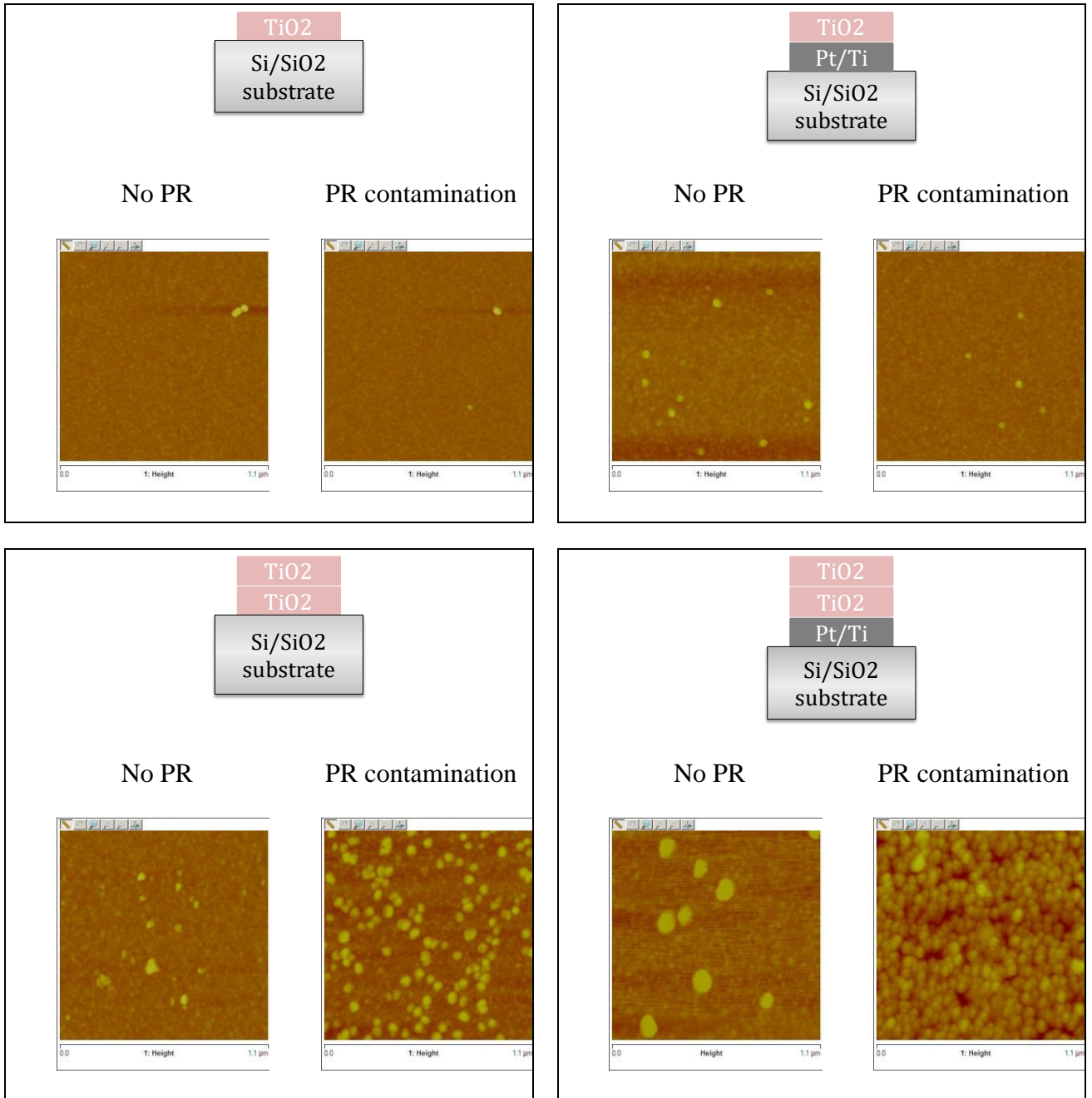
Reiners et al [1] presented that the thickness and the temperature of growth influence the crystallite growth. They have found that these crystallites are made of crystalline TiO_2 in the brookite and rutile form. Reiners suggested that the growth of crystallites happens because of an accumulation of hygroxilic groups at the nucleation spots on the surface.

In order to try to avoid the crystallite growth in the stacked memristive devices, two options were identified. A first path would have been to engineer the ALD film growth by changing the ALD growth parameters. ALD-grown TiO_2 with ozone instead of water can curb the overnucleation since there are no hydroxilic groups involved. Such pathway is actively explored in the research group, because ALD-grown films present the advantage of conformality and reproducibility, being an industrial standard.

The second option was to pursue a physical deposition technique such as reactive sputtering (Figure 3.5 for comparison). The ALD is a chemical based deposition technique where the surface chemistry plays an important role in the nucleation of the desired film on the substrate. Surface defects are hard to engineer since many different chemicals and processing steps are needed for device fabrication. Therefore it can be hard to control the right amount of nucleation during the ALD growth.

The physical deposition techniques are much less sensitive to surface chemistry than the ALD growth since no chemical reaction happens at the surface. The devices based on non-stoichiometric TiO_2 grown using reactive sputtering in-situ with the electrodes showed no presence of crystallites as expected and were selected for vertical stacking. The material is similar to the one developed by Hoskins as explained in detail in the supplementary of Nature paper by Prezioso et al. [2].

Figure 3.5. Comparison between chemically and physically grown TiO$_2$. The chemical method was ALD and the physical method was IBD (similar principle to reactive sputtering). Tests have been performed on (a) one layer film (30nm TiO2) and (b) two layer stacks (30nm x 2). The chemical deposition method shows heavy crystallite density while the physical deposition method shows smooth films for same thickness.
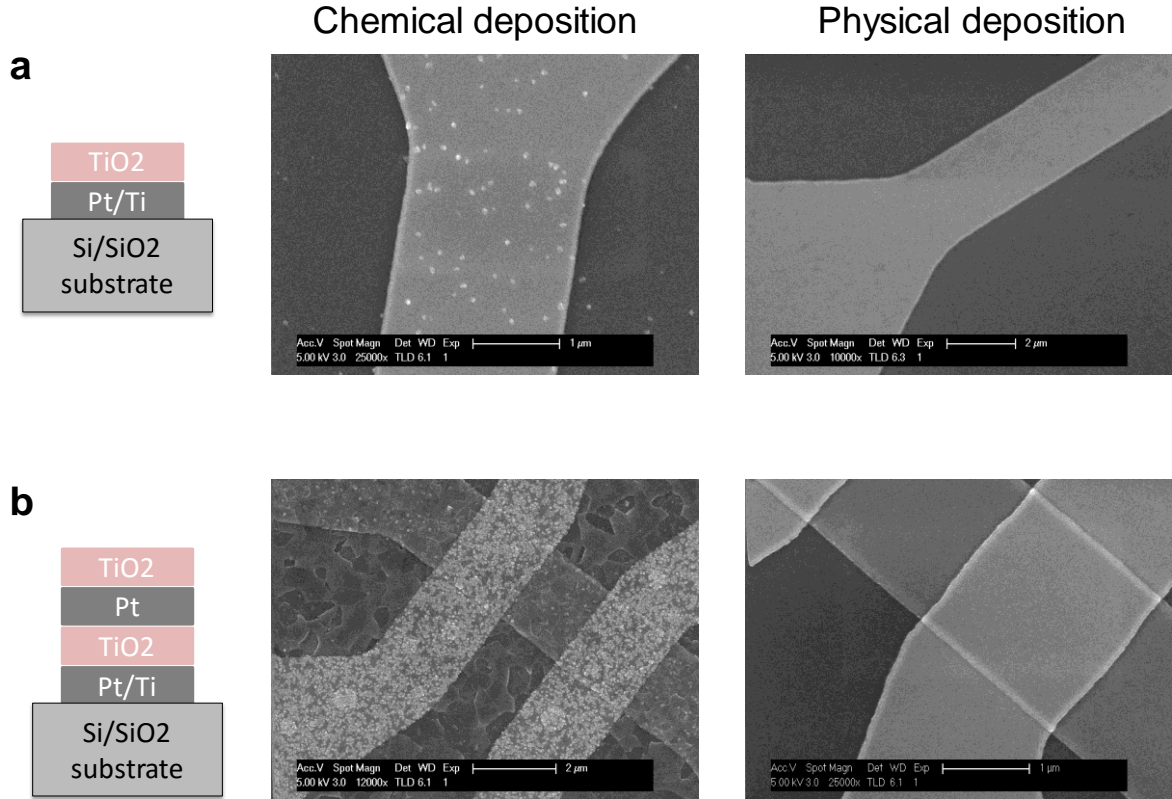
To minimize the sidewall redeposition on the walls of the photoresist undercut during sputtering of the middle electrode, which caused "bunny-ear" formation around the edges of middle electrode (Figure 3.6), both metals were deposited at 0.9 mTorr, the minimum pressure needed to maintain plasma in the sputtering chamber. Also, the thickness of the photoresist undercut layer was optimized to provide more shadowing by using a liftoff layer of LOL2000 (from Shipley Microposit, spin speed 3500rpm, bake 210ºC, thickness ~200nm) followed by the same DSK101/ UV210 stack mentioned above. Using the swabbing in isopropanol occasional lumps were reduced to the height of ~ 20-30 nm.

Figure 3.6. Middle electrode topology due to sidewall redeposition during sputtering (a) using standard process which results in > 200 nm lumps at the edges of the electrode and (b) after deposition optimization and swabbing method, which allows reduction of these features to 20-30 nm.
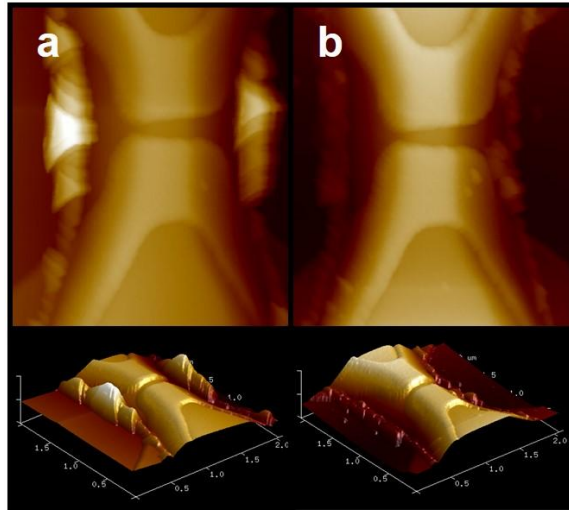
3. Bottom device planarization

Severe topography of the bottom layer devices (Fig. 3.7) may cause shorts and large variations in top layer devices. To overcome this potential problem, a planarization step was performed using chemical mechanical polishing and etch-back of 750 nm of sacrificial SiO_2.

SiO_2 was used with double purpose: as a sacrificial material for planarization and to provide insulation among devices. Different SiO_2 deposition temperatures were investigated (Figure 3.7). The SiO_2 deposited at lower temperatures seemed to not perform well as sacrificial layer during the chemical mechanical polishing step. At 250 ºC, the planarization was successful; however this temperature was too high for the devices to survive. Test devices annealed in an oxygen atmosphere at 200ºC for 25min (the deposition time of SiO_2 in PECVD system) were highly conductive in the virgin state and impossible to switch OFF.

Figure 3.7. CMP planarization of SiO2 deposited at different temperatures: (a) 50°C deposition using an ICP-PECVD system; (b) 100°C deposition using an ICP-PECVD system; (c) 250°C deposition using an standard PECVD system. A mixture of 400sccm of 2% SiH$_4$ and 1420 sccm of N$_2$O was used in both systems.
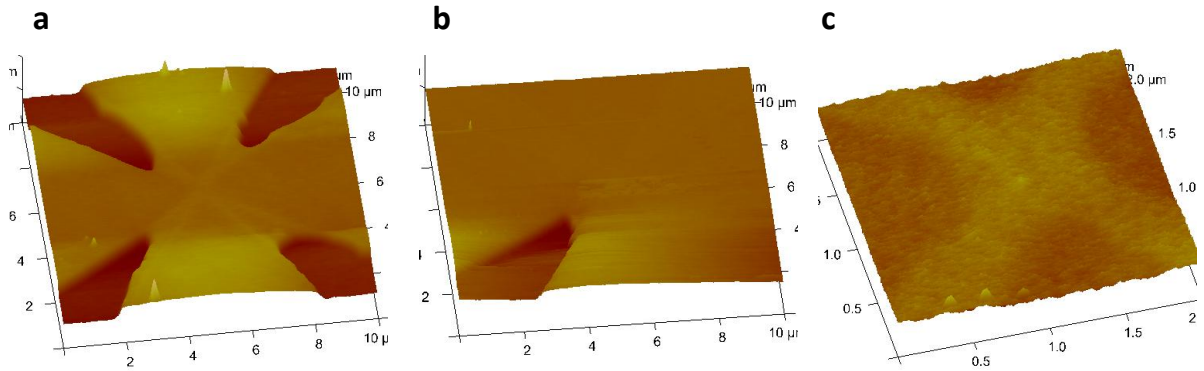
Figure 3.8. A top-view atomic force microscope images of the circuit during different stages of planarization, in particular showing: (a) bottom device before planarization; (b) after chemical-mechanical polishing of SiO$_2$ deposited over bottom device; (c) after etch #1 using CHF$_3$ for 1200 sec showing partially exposed 18-nm-high middle electrode; (d) after etch #2 using CHF$_3$ for 20 sec showing partially exposed 22-nm-high middle electrode; (e) after etch #3 using CHF$_3$ for 20 sec showing partially exposed 28-nm-high middle electrode. (f) Cross-sections profile taken across middle portion of the device (see marks on panel a) at the different etch-back stages
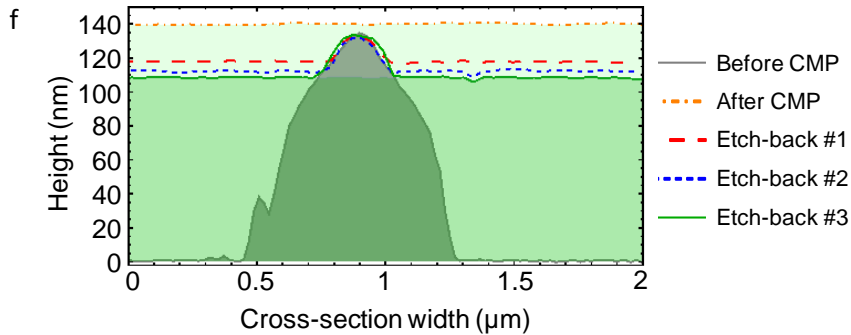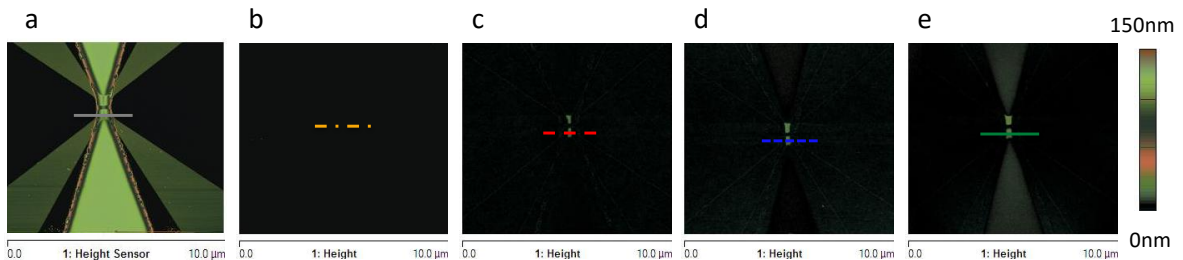
The most optimal planarization was achieved by depositing SiO_2 at 175ºC using PECVD method which did not impact the device performance. Following the deposition, 400 nm of SiO_2 were removed by chemical mechanical polishing for 3 min achieving surface roughness of less than 1 nm.

The last step in planarization procedure was to etch back ~ 250 nm of SiO_2 until the middle electrodes were exposed (Figure 3.8). Several chemistries were investigated with the best results achieved using CHF_3 at 50 W, which had an etch rate of 0.2 nm/s (Figure 3.9). In particular, the dry-etching with CHF_3 was done in steps to ensure < 5 nm roughness in the exposed middle electrode. AFM scans were performed after each etching step to check the thickness of the exposed electrode (Figure 3.8.f) and to confirm that the post-etch surface has no traces of bunny-ear formations.

Figure 3.9. Comparison of two etch back recipes for SiO_2. (a) SF6 with quadratic mean surface roughness $R_Q > 6$ nm and (b) CHF_3 with $R_Q < 1$ nm.
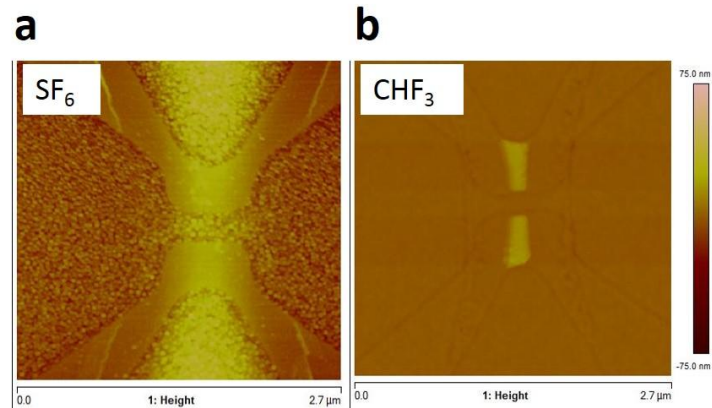
4. Device fabrication flow

Devices were fabricated on a Si wafer coated with 200 nm thermal SiO_2. Circuit fabrication involved four lithography steps performed using an ASML S500 / 300 DUV

stepper using a 248 nm laser. To prevent from misalignment of device layers, the bottom devices were made larger with an active area of 500 nm × 500 nm, as compared to 300 nm × 500 nm active area of top devices. In particular, in the first lithography step bottom electrode was patterned using a developable antireflective coating (DSK-101-307 from Brewer Science, spin speed 2500rpm, bake 185ºC, thickness ~50nm) and positive photoresist (UV210-0.3 from Dow, spin speed 2500rpm, bake 135ºC, thickness ~300nm). 5 nm / 20 nm of Ta / Pt was evaporated at 0.7 A/sec deposition rate in a thin film metal e-beam evaporator. After the liftoff, a "descum" by active oxygen dry etching at 200ºC for 5 minutes was performed to remove photoresist traces.

In the next lithography step, the middle electrode was patterned and the bottom device switching bi-layer 6 nm / 45 nm of Al_2O_3 / TiO_{2-x} and 15 nm / 38 nm of Ti / Pt metal were deposited using low temperature (< 300ºC) reactive sputtering in an AJA ATC 2200-V sputter system.

After planarization and partial middle electrode exposure using the technique explained in Section 3, the top layer devices were completed by in-situ reactive sputtering of the switching layer (4 nm /30 nm of Al_2O_3 /TiO_{2-x}) and the top electrode of Ti (15 nm) / Pt (25 nm) over patterned photoresist (DSK101/UV210). No oxygen descum was not performed before deposition in order to avoid potential oxidation of the bottom switching layer and maintain the controlled TiO_{2-x} stoichiometry.

Lastly, the pads of the bottom and middle electrodes were exposed through a CHF_3 etch of the sacrificial SiO_2 used for planarization. In all lithography steps, the photoresist was stripped in the 1165 solvent (from Shipley Microposit) for 24 h at 80ºC. Figure 3.10 shows AFM profiles taken after the main fabrication steps.

Figure 3.10. A top-view atomic force microscope images of the circuit during different stages of fabrication, in particular showing: (a) bottom electrode; (b) middle electrode; (c) middle electrode after planarization step; and (d) top electrode.
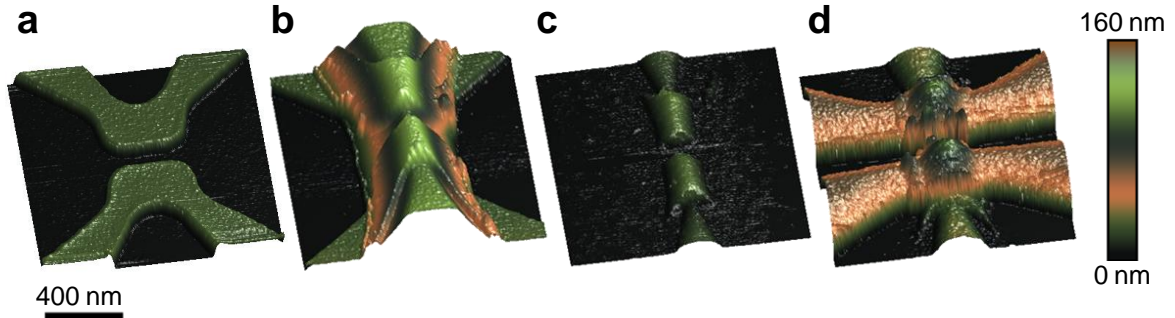
The device layer thicknesses and stoichiometry, which was precisely controlled by changing oxygen to nitrogen flow ratio during sputtering, were selected based on our earlier study [2] with the primary objective of lowering forming voltages. Thin Ti and Ta layers were deposited to improve electrode adhesion. Addition of Ti to the middle and top electrodes also ensured ohmic interfaces with titanium dioxide layer, which was important for device's asymmetry [3] Low forming voltages reduced electrical stress during electroforming [2] while in-situ contacts between titanium oxide and the metal electrodes fabricated without breaking the vacuum ensured high-quality interfaces [4] with both factors were essential for improving uniformity of memristor's switching characteristics. Furthermore, planarization reduced middle electrode roughness resulted from residual sidewall deposition and was critical for lowering variations in top-layer devices. The absence of annealing step, which is typically used for fine-tuning of the defect profile in metal oxide memristors [2,5], and low-temperature fabrication budget, with temperatures below 300ºC during the sputter deposition, simplifies three-dimensional integration and makes the fabrication process compatible with conventional semiconductor technologies. Figure 3.11 shows an SEM top view of the completed device structure.

Figure 3.11. A top-view scanning-electron-microscope image of the completed device structure. The red, blue, and purple colors were added to highlight the location of bottom and top devices, and their overlap, respectively.
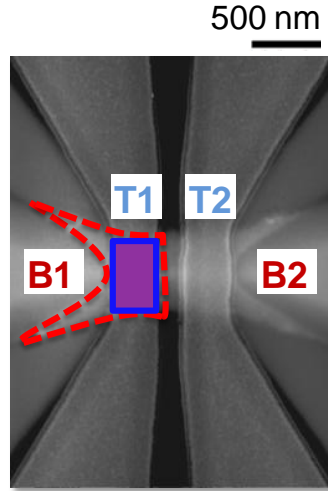
5. Device characterization

All electrical testing was performed with an Agilent B1500A Semiconductor Device Parameter Analyzer tool. The memristors were electroformed by grounding the device's bottom electrode and applying a current-controlled quasi DC ramp-up to the device's top electrode, while keeping all other circuit terminals floating. For all devices forming voltages were around ~2-3 V. To minimize current leakage during the forming process, each memristor was switched to the OFF state immediately after forming.

Figure 3.12 (a-d) shows typical memristor *I-V* characteristics obtained by applying positive and negative quasi-DC triangular voltage sweeps for 100 cycles per device. Switching polarities for all devices correspond to the bottom active interface, which is in agreement with the device's asymmetry. A slightly higher V_{SET} = 1.5 V for the bottom memristors, compared to V_{SET} = 1.2 V for the top ones (Fig. 10 e and f), is explained be somewhat thicker titanium dioxide layer for the former devices. The most severe are cycle-to-cycle variations in set switching threshold voltages, which ranges from 0.7 V to 1.6 V for the top layer devices, and from 1.1 V to 1.9 V for the bottom devices. For all devices, the

set switching is very sharp while the reset process is gradual. For example, for the bottom devices reset transition starts at ~ -1.5 V, however, to avoid partial switching voltages exceeding -2.5 V must be applied.

Figure 3.12. (a-d) *I-V* curves showing 100 cycles of switching for all devices and (e-f) the corresponding set threshold voltage statistics. Gray lines on panels (a-c) show typical current-controlled forming *I-V*s. (device T1 did not require forming). The dashed orange curve (b) is fitting used for numerical simulations in Chapter 2.
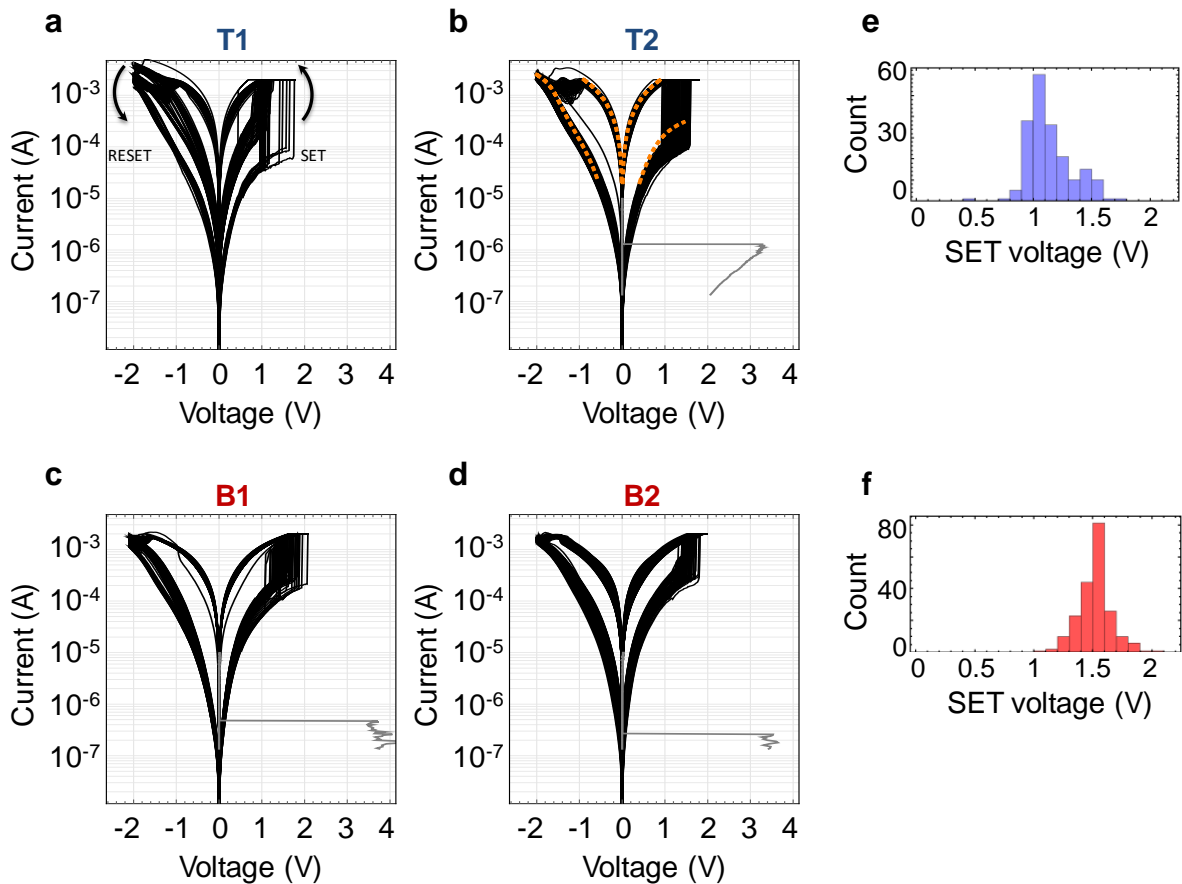
As Figures 3.13a and b show, repetitive switching between ON and OFF states of one device did not disturb the state of others, thus suggesting that thermal crosstalk is negligible. Ratio of currents measured at 0.1 V between the ON and OFF states were close to two orders

of magnitude. Other characteristics, such as endurance and retention, were close to those reported earlier for the similar devices [2].

Figure 3.13. (a) Conductance of the device B1 that was repeatedly switched 200 times and (b) those of the other three devices in the stack that were kept in the OFF states for the first 100 cycles, and then in the ON states for the remaining 100 cycles. The devices were switched by applying triangular voltage pulses.
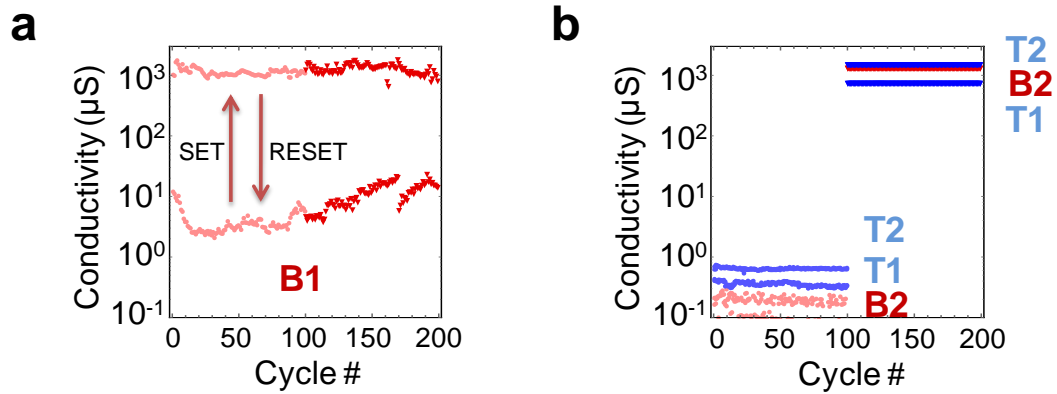
6. Disadvantages

The lift-off based fabrication flow presented has several important disadvantages that are worth mentioning. The metal features are prone to rabbit ear formations and it is unsuitable for aggressive down scaling because high aspect ratio features are hard to achieve. Moreover, the high energy beam used during deposition can damage the switching film decreasing the quality of the interface. Due to these reasons, it is not an industrial standard for CMOS processing anymore. In order to ensure smooth integration with CMOS and fast adoption of this technology, a new pattern transfer method of fabrication based on ion milling is presented in the next section.

B. Ion-milling-based fabrication

1. Advantages

This section presents the stacked memristor crossbars fabricated using ion-milling techniques. The improved manufacturability of this process is shown by presenting a larger system, two monolithically stacked memristor crossbars of 10x10 devices.

The ion milling-based patterning of metal lines is compatible with conventional, DUV, and, for ultra small features, the e-beam lithography. Moreover, the over-development of the Al_2O_3 hard mask allows sub diffraction limited features as shown in Figure 3.14b.

Figure 3.14. Reproducible continuous ~70nm wide metal lines fabricated using a 248nm DUV stepper and Ar ion milling (a) Schematics of process flow and (b) Example of memristive device showing a bottom electrode (horizontal) ~200nm wide fabricated at the limit of the DUV diffraction and a top electrode (vertical) ~70nm wide fabricated using a controlled over-development of the hard mask Al_2O_3.
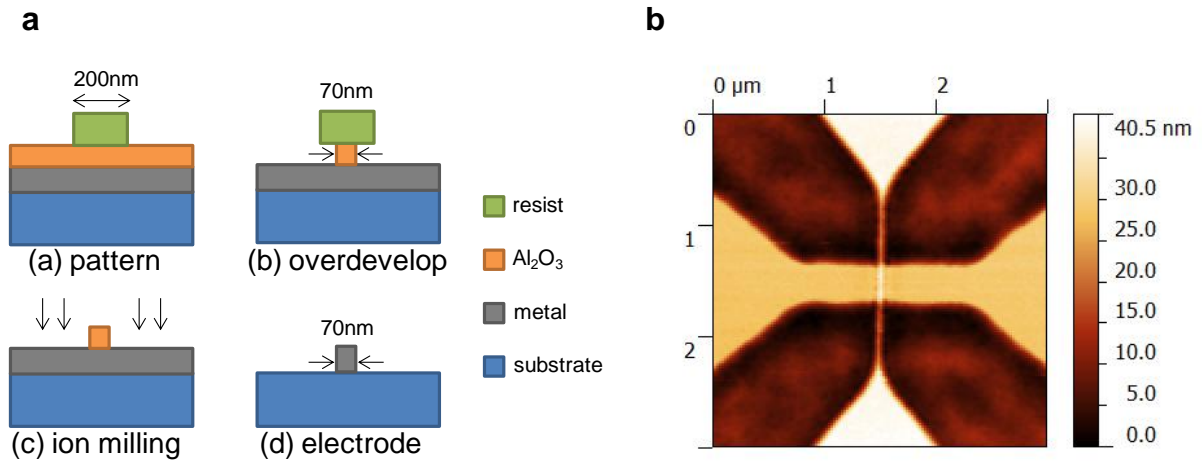
Enhanced control of metal feature shape is possible, by using highly selective hard masks such as Al_2O_3 and slightly tilted etching in order to eliminate the possibility of rabbit ear formations. Moreover, the elimination of evaporation and liftoff dramatically reduces time to manufacture electrodes, thus increasing the speed and efficiency in the R&D process.

2. Desired structure

The desired device structure for the ion-milling based crossbars of memristors is fairly similar to the one in Fig.3.2, with slight modifications to some film thicknesses.

The major steps involved in fabrication are (Fig. 3.15):

1) *Bottom electrode:* Deposit in blanket the adhesion layer (TiO_2 – 5nm) and metal (Pt - 30nm) using sputtering, followed by hard mask Al_2O_3 (30nm) by electron beam deposition. Pattern the Al_2O_3 layer using DUV lithography and developer and use it as a hard mask for the ion milling of the metal. Due to the fact that Al_2O_3 etch rate in developer varies with its deposition parameters, a more reliable processing is based on reactive-ion etching of this hard mask using an inductively coupled plasma (ICP) system using CHF_3 as etch gas.

2) *Switching layer and top electrode:* Deposit in blanket the switching layer (TiO_{2-x} – 30nm), the getter layer (Ti - 15nm) and metal (Pt – 15nm) using sputtering, followed by hard mask Al_2O_3 (30nm) by electron beam deposition. Pattern the Al_2O_3 layer using DUV lithography and etching and further use it as a hard mask for the ion milling of the metal and of the getter layer.

3) *Isolation:* Isolate the devices to reduce leakage and simultaneously expose the bonding pads for measurement by etching away the sacrificial layer and the switching material around the crossbar features.

Using the CMP planarization and controlled etch-back developed in previous section to planarize this fabricated crossbar, it is then possible to repeat these steps to stack additional crossbar layers as needed.

57

Figure 3.15. Schematics of process flow based on ion milling to pattern memristive devices using Al₂O₃ as hard mask. The angle of incidence between the ion beam and the sample plays an important role in the ion milling since it impacts the shape of the final metal line.
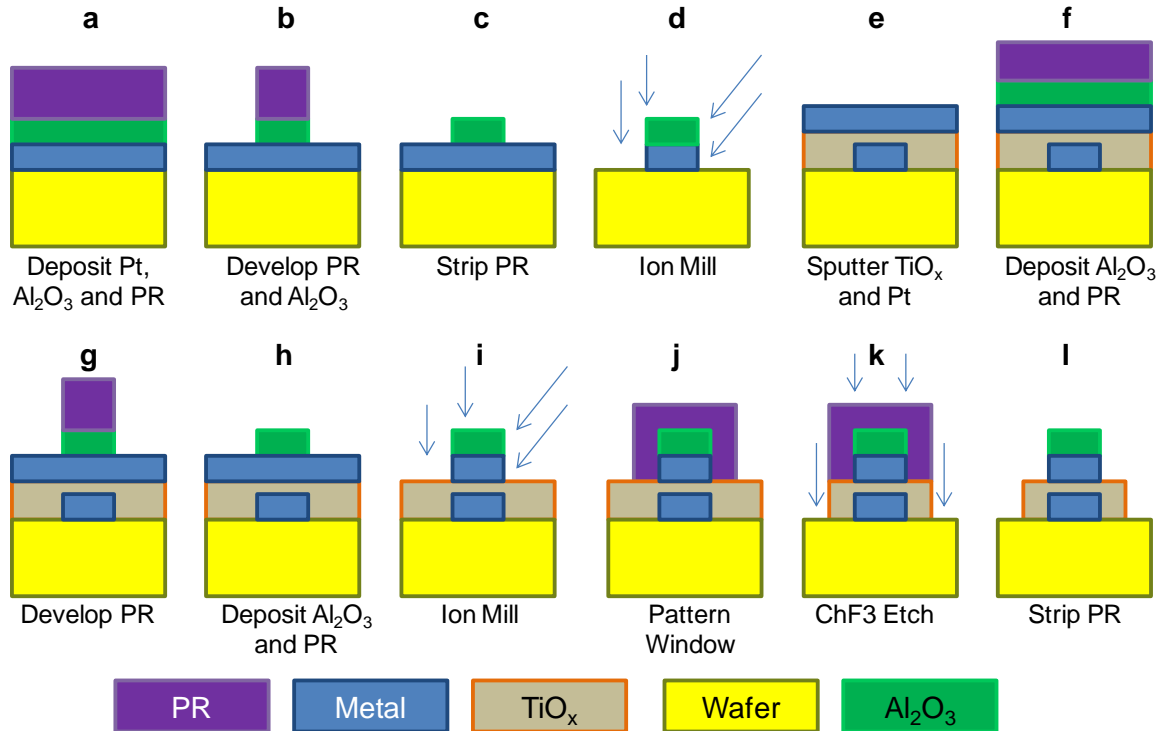
The sections focus on the most challenging step: the development of a reliable ion milling procedure in order to insure continuous metal lines with good step coverage.

3. Electrode patterning with ion milling

Firstly, individual devices were fabricated using an ion beam at normal incidence to the sample (0° tilt). The SEM figure 3.16 shows how the top electrodes are broken at the step with the bottom ones, creating very high resistance or not connected top metal lines.

Figure 3.16. Initial device structure fabricated using milling with an ion beam at normal incidence to the sample (no sample tilt).

200nm
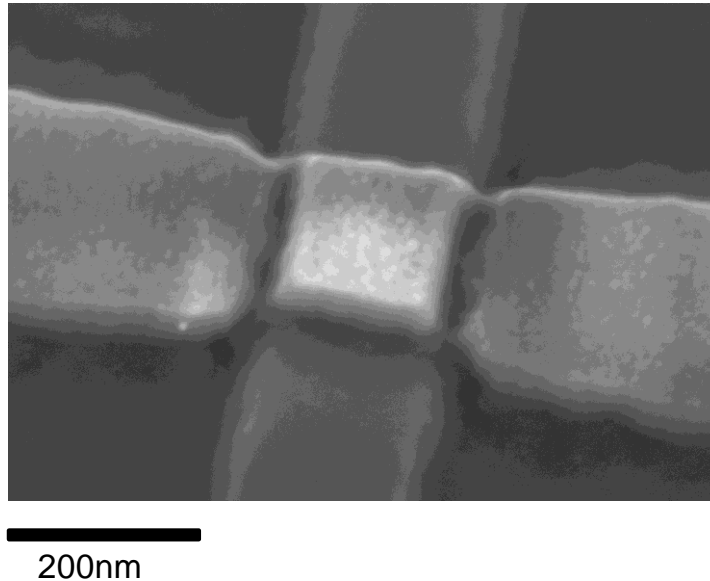
An investigation was carried to understand the influence of the tilt angle on the continuity of the metal lines. The conditions were investigated: (a) no tilt; (b) partial tilted; and (c) purely tilted. The results are summarized in Figure 3.17.

These conditions were created three very different shapes for the electrode (a-c, column 1). All the samples were then covered by blanket layers of metal (Pt) and hard mask (Al$_2$O$_3$) and milled in the no tilt condition. The Al$_2$O$_3$ blanket film should have protected the features, so no visible etching was expected in the ideal case. The electrode in the no tilt condition had an almost straight sidewall that contributed to poor step coverage and metal discontinuities. A partially tilted milling was chosen for the bottom electrode patterning. However, even if bottom electrode profile has no sharp edges, the step coverage of the top electrode can be very poor if the top electrode milling is at the 0° angle as shown by Figure 3.17a. This behavior is due to the classical sputtering curve, which shows that the sputtering yield increases with increasing angles. The feature most sensitive to discontinuity (in this case the step) should be at an normal angle with the incident ion beam, therefore the sample

59

has to be tilted accordingly. At the 50° angle condition, the electrodes show good continuity

and step coverage (Figure 3.18c).

Figure 3.17. Influence of the ion milling tilt on the bottom electrode shape and step coverage. (a) no tilt - 6 minutes milling at angle 0°; (b) partial tilt - 3 minutes no tilt milling and 2 minutes at 40° angle and (c) purely tilted - 4 minutes at 40° angle. (1) electrode milling; (2) after the deposition of a blanket metal film and hard mask layer and not patterned etch at angle 0°.
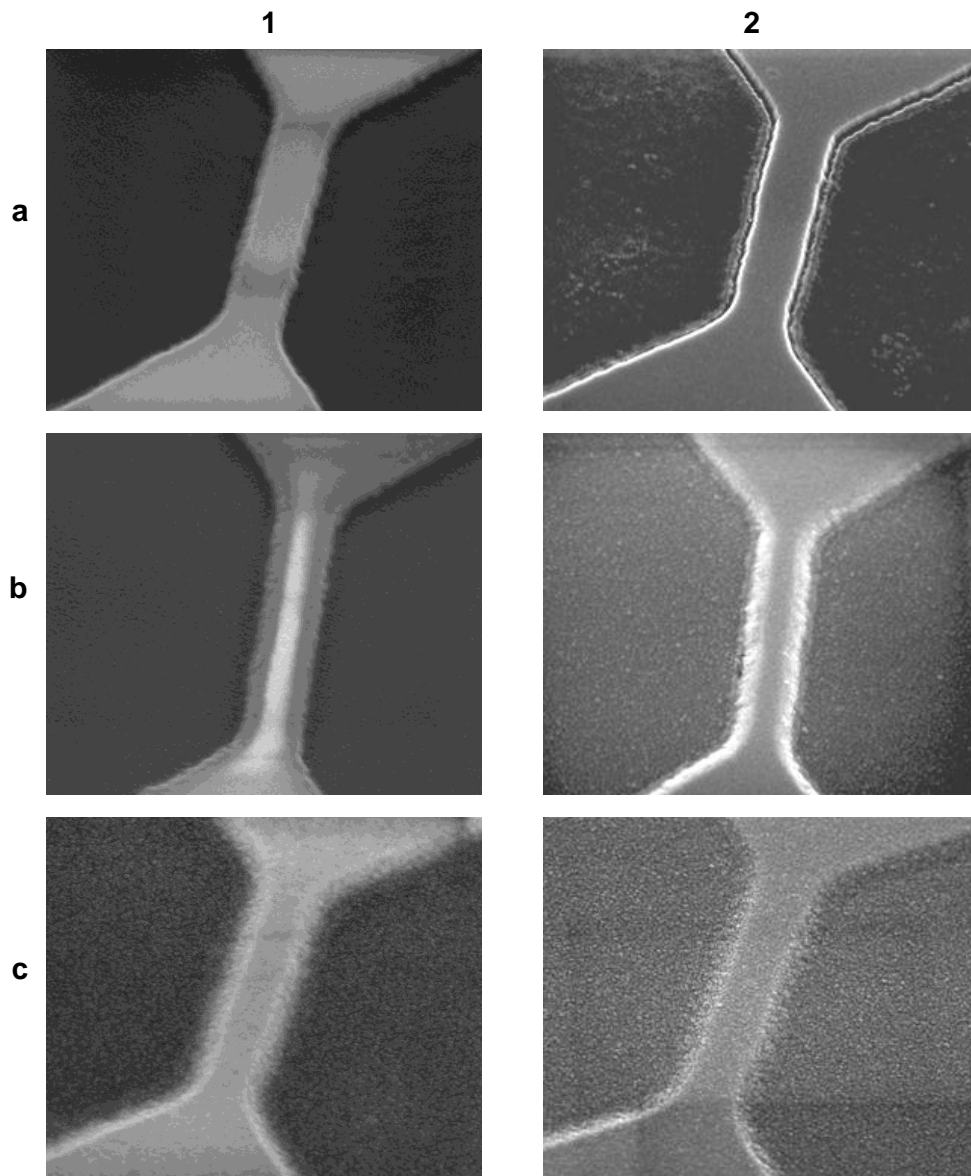
Figure 3.18. Influence of the ion milling tilt on the top electrode step coverage. (a) no tilt - 8 minutes milling at angle 0°; (b) 7 minutes at 30° angle and (c) 6 minutes at 40° angle.
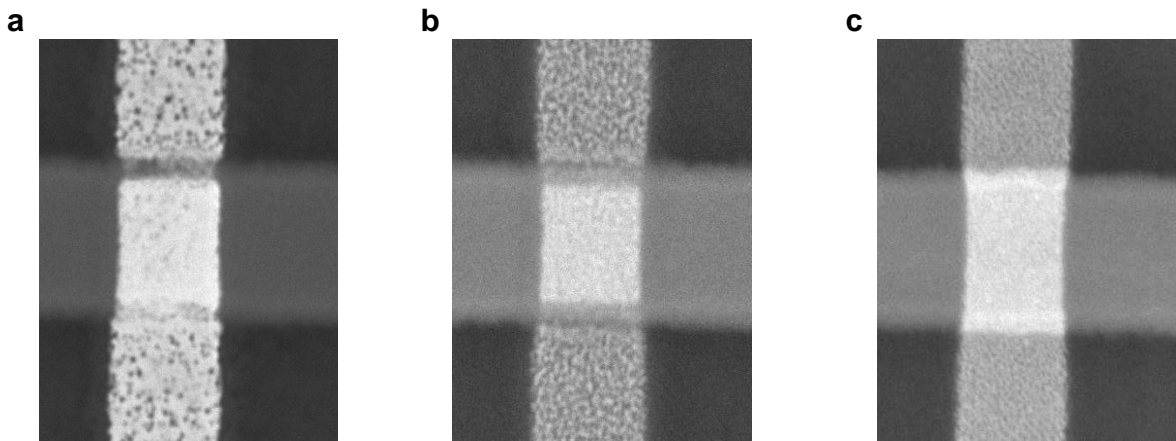
4. Device fabrication flow

Devices were fabricated on a Si wafer coated with 200 nm thermal SiO_2. The crossbar fabrication involved four lithography steps performed using an ASML S500 / 300 DUV stepper using a 248 nm laser. Both bottom and top devices had an active area of 500 nm × 500 nm. After a thorough cleaning in acetone, isopropanol and deionized water, the adhesion layer (TiO_2) and the metal (Pt) for the bottom electrode was deposited in blanket using low temperature (< 300ºC) reactive sputtering in an AJA ATC 2200-V sputter system. The Al_2O_3 hard mask (thickness ~30nm) to be used for ion milling was then deposited in blanket using an e-beam deposition system at very low deposition rates <5Å/sec. The first lithography step patterned the hard mask by using a developable antireflective coating (DSK-101-307 from Brewer Science, spin speed 2500rpm, bake 185ºC, thickness ~50nm) and positive photoresist (UV210-0.3 from Dow, spin speed 2500rpm, bake 135ºC, thickness ~300nm). Al_2O_3 is developable in most common photoresist developing agents (in this case AZ300MIF from Clariant was used for a total of 5min and 15sec). After the hard mask

patterning, all the photoresist was removed using O_2-based reactive ion etching (20sccm of O_2, 100W, 10mTorr pressure for 5min). The metal was then etched away using Ar ion milling (Oxford Flexal system, 30mA current source, 3min at 0° angle followed by 2 min at 40° angle). After etching, the remaining traces of Al_2O_3 were removed in AZ300MIF developer.

The switching layer (TiO_{2-x}, thickness ~30nm), the getter layer (Ti, thickness ~15nm) and the metal layer (Pt, thickness ~15nm) for the bottom devices were then deposited in blanket in the same vacuum using a reactive sputtering system. The switching layer thickness and stoichiometry, which was precisely controlled by changing oxygen to nitrogen flow ratio during sputtering, were selected based on an earlier study [2]. However, a slightly different reactive sputtering system was used than for the fabrication of the lift-off based devices. For the patterning, a similar recipe to the bottom electrode was followed to deposit and pattern the hard mask. The metal and the getter layer ware then etched away using Ar ion milling (30mA current source, 6min at 50° angle). After etching, the remaining traces of Al_2O_3 were removed in AZ300MIF developer.

Lastly, the pads of the bottom and were exposed through a CHF_3 etch of the TiO_{2-x} switching layer. Because the switching layer (TiO_{2-x}) is slightly conductive and deposited in blanket over the entire wafer, this step was also needed to isolate the large contact lines to prevent leakage. Figure 3.19a shows AFM profile taken after the main fabrication steps for a 10x10 crossbar and 3.19b. shows an SEM zoom-in photo of devices in the crossbar. No rabbit ears or broken contacts are presents.

Figure 3.19. Ion-milling based 10x10 memristor crossbar: (a) AFM view; (b) SEM showing device detail.
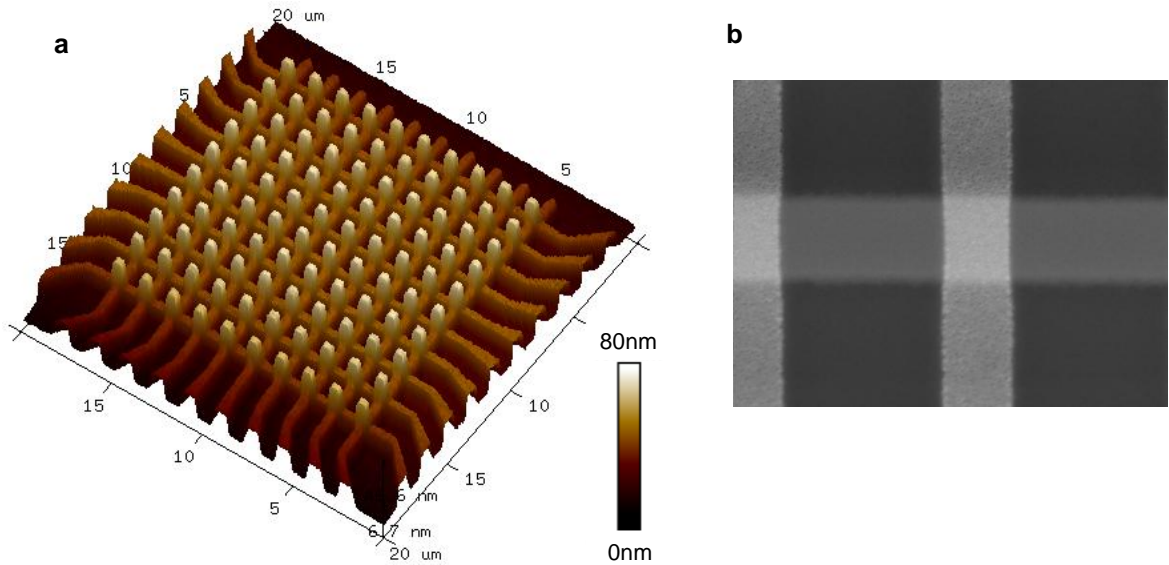
No annealing was needed, similarly to the lift-off based devices. The highest temperature needed during fabrication was 200ºC during the sputter deposition. This low temperature budget insures that the fabrication process is compatible with conventional semiconductor technologies for the purpose of monolithically integrating these stacked memristor layers on CMOS circuitry.

5. Device characterization

All electrical testing was performed with an Agilent B1500A Semiconductor Device Parameter Analyzer tool. The memristors were electroformed by grounding the device's bottom electrode and applying a current-controlled quasi DC ramp-up to the device's top electrode, while keeping all other circuit terminals floating.

Figure 3.20 shows a reliability comparison between lift-off and ion-milling based devices. 100 typical memristor *I-V* characteristics were obtained by applying positive and negative quasi-DC triangular voltage sweep for both lift-off and ion-milled devices using

similar device material stack. Ion-milled based devices showed improved reliability with

~32% less variation in the set switching voltages as compared to similar lift-off based ones.

Figure 3.20. Reliability comparison between lift-off and ion-milling based devices. *I-V* **curves showing 100 cycles of switching for a characteristic individual devices**
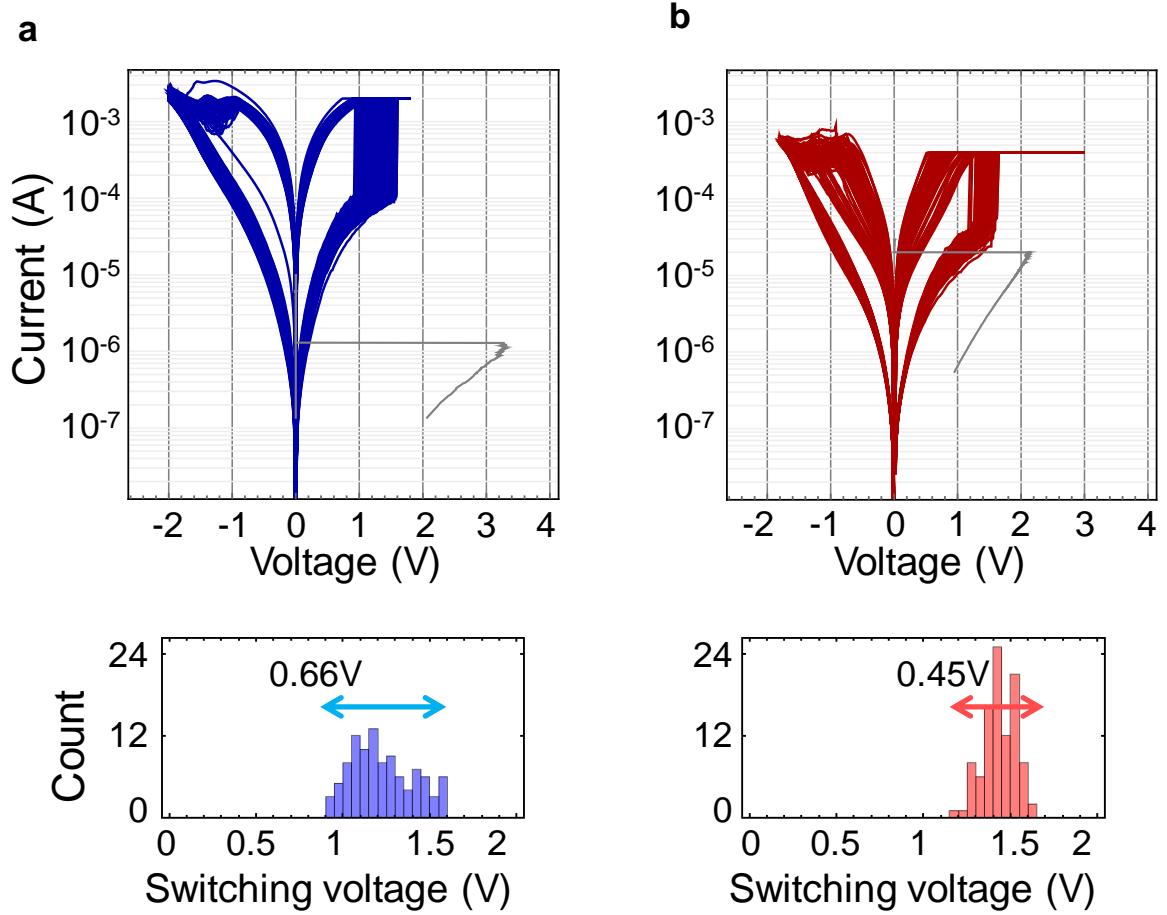
C. Summary

We have presented two different fabrication flows for building monolithically stacked

memristor structures. The first one, a lift-off based approach, was used to fabricate a small

2x2 array. The devices showed good reproducibility and no thermal crosstalk, but they had

high variations in the cycle-to-cycle switching. Moreover, this process was not

64

manufacturing-friendly, rabbit ears posing challenges throughout the processing. These challenges were a motivation to develop another process flow based on metal etching using ion-milling. Due to the thin hard mask Al_2O_3, the rabbit ear formations were not a problem in this process thus making it suitable for future multi-layer memristor crossbar stacking. Moreover, this process can be used to push the limits of optical lithography, by using the overdeveloping of the hard mask Al_2O_3 to reduce the feature size. Overall, the ion-milling process allowed for increased manufacturability and reduced fabrication time, so this ion-milling process will be more extensively used in future work for device development and applications.

The next chapter will use the fabricated stacked devices presented in this chapter to demonstrate stateful implication logic in three-dimensions. Reliable multi-cycle stateful operations are achieved using the approach presented in Chapter 2.

References for Chapter 3

1. Reiners, M., Xu, K., Aslam, N., Devi, A., Waser, R., & Hoffmann-Eifert, S. (2013). Growth and Crystallization of TiO_2 Thin Films by Atomic Layer Deposition Using a Novel Amido Guanidinate Titanium Source and Tetrakis-dimethylamido-titanium. *Chemistry of Materials*, *25*(15), 2934-2943 (2013).

2. Prezioso, M. *et al.* Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature* **521**, 61-64 (2015).

3. Yang, J. J. *et al.* Memristive switching mechanism for metal/oxide/metal nanodevices. *Nature Nanotechnology* **3**, 429-433 (2008).

4. Mikheev, E., Hoskins, B. D., Strukov, D. B. & Stemmer, S. Resistive switching and its suppression in Pt/Nb: $SrTiO_3$ junctions, *Nature Communications* **5**, 3990 (2014).

5. Govoreanu, B. *et al.* Vacancy-modulated conductive oxide resistive RAM (VMCO-RRAM). *IEDM Tech Dig.*, 10.2.1 - 10.2.4 (2013).

Chapter IV

3D Stateful Logic

This chapter describes 3D implication logic performed in monolithically stacked TiO_2-based memristive devices. The broad goal of this work is to enable intra-layer and inter-layer data manipulation since 3D stacked circuits allow for a much higher density as compared to the planar case. CMOS circuitry integrated with memristor multi-layers providing in memory computing capabilities can offer a viable solution to the von Newmann bottleneck.

In order to prove the stateful implication can be performed vertically between bottom and top layer memristors. The devices necessary were fabricated as a two-layer stack of metal-oxide memristors and their fabrication was presented in Chapter 3. All of the devices in the lift-off based memristor array shared a common middle electrode (Figure 4.1). At the end of the chapter, experimental results are included that show 3D implication logic performed in a ring sequence and an inter-layer NAND operation with reliable multi-cycle operation.

Figure 4.1. Schematics of stacked memristor structure. B1 and B2 denote bottom devices, while T1 and T2 the top ones. The desired operation to be proved experimentally is the bottom device IMPLY top device.
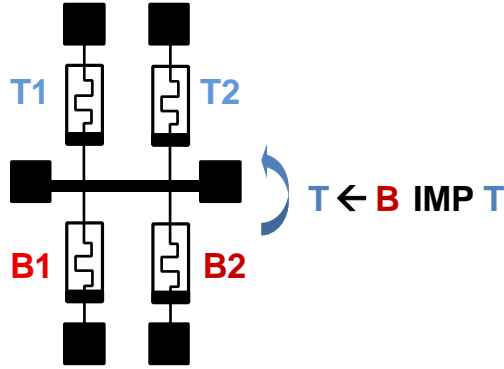
A. Measurement setup

All electrical testing was performed with an Agilent B1500A Semiconductor Device Analyzer that provided the source-measurement units (SMUs) for voltage and current measurements. In addition, Agilent 5250A low leakage switch matrix was used to re-configure connections as needed. A standard probe station from MicroXact equipped with W tips was used. The parameter analyzer and the switching matrix were controlled by a computer via a GPIB interface using a custom Visual C++ code used to automate the read, tuning and IMPLY operations.

The device state was read using a voltage pulse of small amplitude of 0.1V (Fig. 4.2). The conductance of the device was calculated from the readout as I_{read}/V_{read}. IMPLY logic was performed by grounding the switching device and DC biasing the conditional device to V_{cond} , then applying a current pulse to the common middle electrode (Fig. 4.3). After the IMPLY operation, each device was separately read by reconfiguring accordingly the switch matrix as described in Fig. 4.2.

Figure 4.2. The read operation of the T1 device state. (a) Schematics of the measurement setup used, including the connections inside the switch matrix that allow for the top electrode of the device to be connected to V_s and for the middle electrode to GND; (b) Schematics of the applied input voltage pulse of amplitude V_{read} and duration t_{read}; (c) Schematics of the read output pulse of amplitude I_{read} which can be used to determine the conductivity of the device at small bias ($=I_{read}/V_{read}$).
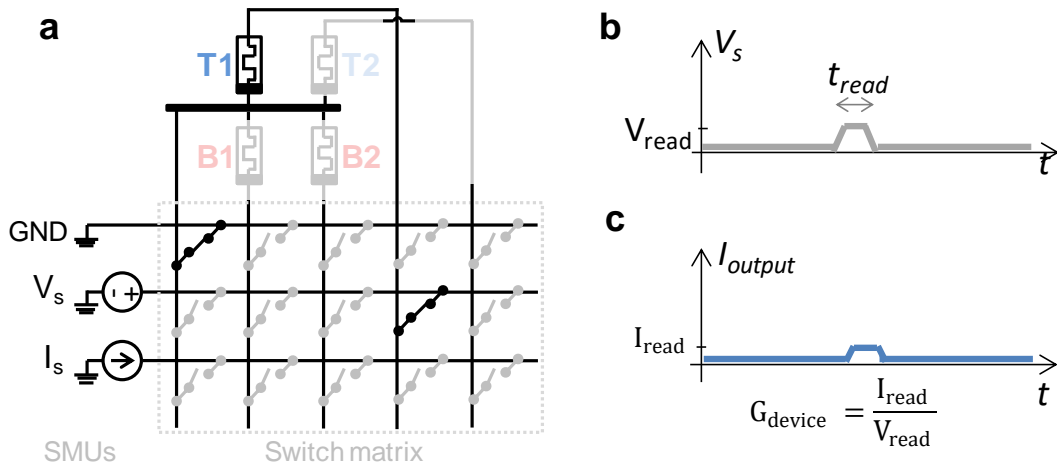
Figure 4.3. The IMPLY operation B1* ← T1 IMP B1. (a) Schematics of the measurement setup used, including the connections inside the switch matrix that allow for the bottom electrode of the B1 device to be connected to GND, for the top electrode of the T1 device to be connected to V_P and for the middle electrode to I_L; (b) V_{cond} is applied as a small DC voltage. I_L is applied as a current pulse of duration t_{IMPLY}. After the IMPLY operation, the switch matrix is reconfigured and the T1 and B1 device states read according to Fig. 4.2.
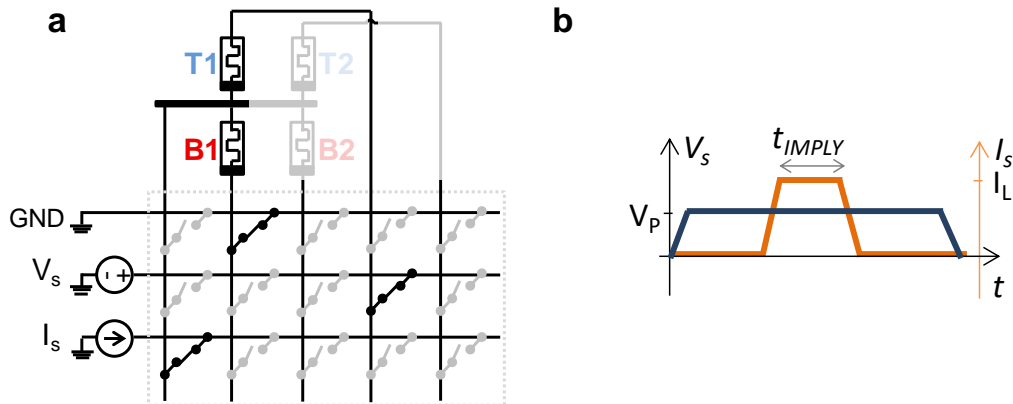
In accordance to the IMPLY truth table, the devices were programmed in the initial states using the state tuning algorithm [1] by applying voltage pulses with increasing amplitude (Fig. 4.4). Positive amplitude voltage pulses were used for the set operation and negative for the reset one.

Figure 4.4. The tuning operation of the B1 device state. (a) Schematics of the measurement setup used, including the connections inside the switch matrix that allow for the bottom electrode of the device to be connected to GND and for the middle electrode to V_s; (b) Train of current pulses used to set the device. Increasing amplitudes with $I_{set\text{-}step}$ are used until the maximum amplitude is reached, then that amplitude is maintained until the read pulse shows that the $G_{device} \geq$ desired G_{ON}; (c) Train of negative voltage pulses used to reset the device. Similar to the set operation with the stop condition that the read pulse shows that the $G_{device} \leq$ desired G_{OFF};
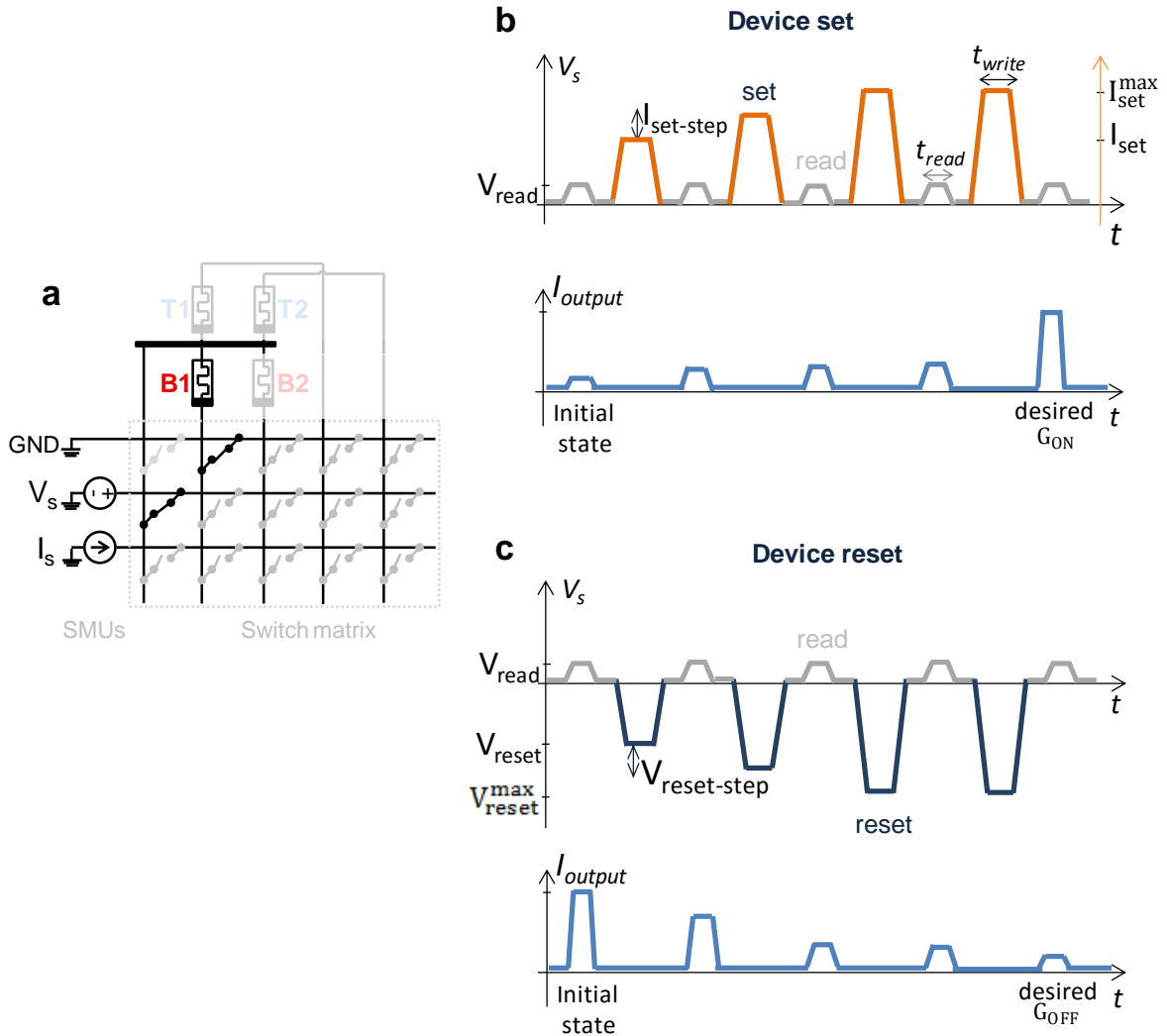
B. Implication ring

Significant set threshold voltage variations (Chapter 3, Figs. 3.12 e and f) is a major challenge for implementing material implication logic, so it is natural to choose circuit parameters to maximize voltage margins. Chapter 2 provides proof that the largest margins are for $G_L = 0$, which cannot be implemented with the original circuit, and requires replacement of the load resistance and voltage source with the current source. The transition from Borghetti's original circuit [2] with $G_{OFF}<G_L<G_{ON}$ to the modified one with optimal current source load I_L allowed increasing the largest permissible variations in set threshold voltage ($2\Delta V_{SET}$) by more than 20%, i.e. from ~ 0.7 V to ~ 0.9 V for the bottom layer devices, and from ~ 0.56 V to ~ 0.72 V for the top ones. Because our device variation in the set voltage was significant, such boost in defect tolerance was critical for our experiment allowing to cope with much of the experimentally observed variations. It should be noted that, in principle, material implication logic can be also implemented using memristor's reset transition, i.e. assuming that logic states "0" and "1" are represented by the ON and OFF states instead. However, this would not be helpful in our case, because the gradual reset transition imposes even sticker requirement on voltage margins. Using variation tolerant design with optimal values of I_L and V_P, which were obtained from accurate numerical simulations based on experimental (nonlinear) *I-V* curves (see Chapter 2), we successfully demonstrated material implication logic within fabricated memristor circuit.

In the first set of experiments, the goal was to demonstrate a series of material implication operations performed sequentially between different pairs of memristors from bottom and top layer (Fig. 4.5). The optimized circuit structure from Chapter 2 was used.

Figure 4.5. Circuit schematics for the 3D stateful logic ring. The optimized circuit structure with current load derived in Chapter 2 was used to provide maximum protection again device cycle-to-cycle variations.
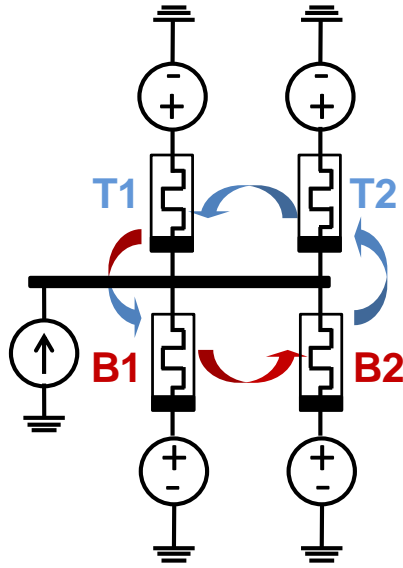
The stateful logic experimental results for all the four possible combinations between bottom and top layer devices are summarized in Figure 4.6. For all cases, V_P was selected as 0.25 V and the terminals of not participating devices were floated. A 10-ms pulse $I_L = 550$ µA was applied for the cases shown in panels a and b, i.e. when the result was written into the bottom device, and 10-ms pulse $I_L = 200$ µA was applied when the output was one of the top devices (panels c, d). In all experiments G_{ON} and G_{OFF} for the initial states measured at 0.1 V were always close to 115 µS, 115 µS, 125 µS, 120 µS and 10 µS, 10 µS, 5 µS, 8 µS for B1, B2, T1, and T2 devices.

The devices were programmed in the initial states using a modified state tuning algorithm (Fig. 4.4) by applying 1-ms pulses with increasing amplitude. For reset voltage pulses were used with amplitudes from 0.5 V to maximum 1.9 V and 0.1 V step, while for set current pulses from 50 µA to maximum amplitude 900 µA with 50 µA step.

Figure 4.6. 3D stateful logic experimental results showing device's conductances before and after IMP operation for different initial states and involving different pairs of memristors. Each panel shows the averaged conductances and their standard deviations for 20 experiments (details in text).
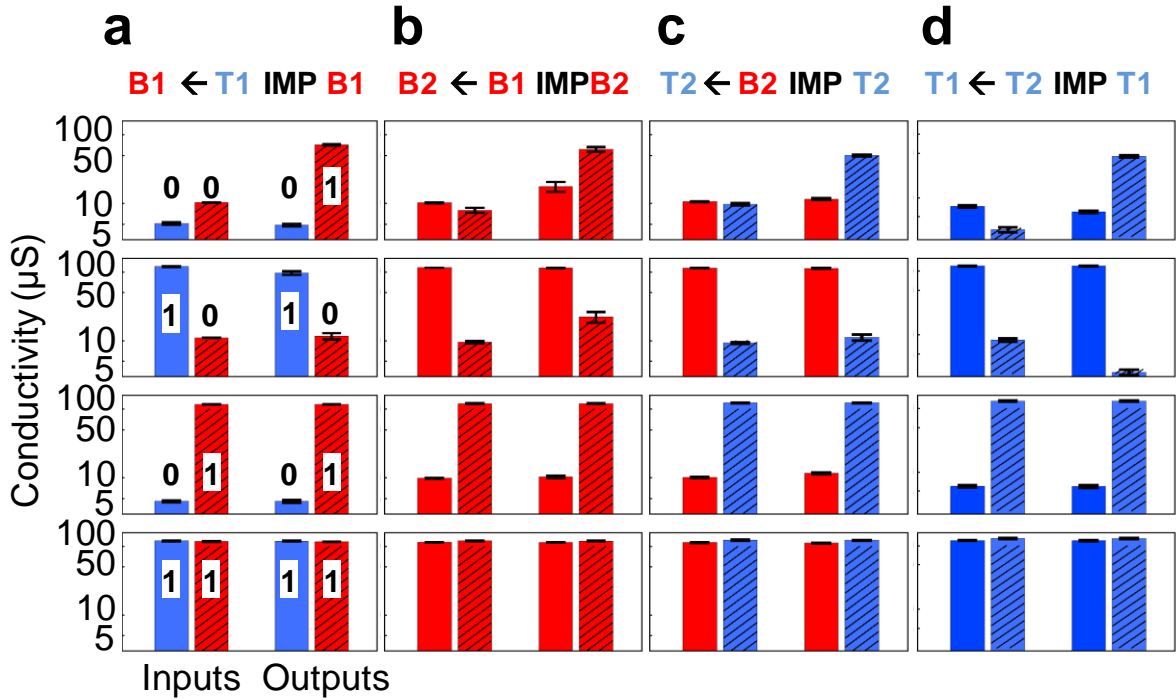
C. Inter-layer NAND gate

Before each logic operation in the implication ring (Figure 4.6), the devices were always written to the specified initial states, therefore this experiment is a proof of memory and logic functionality implemented within the same circuit. In most cases, output conductances are close to the extreme (initial) G_{ON} and G_{OFF} values, so that it should be possible to use the output of one stateful logic operations as an input for the next.

To confirm this, in the next series of experiments, we implemented NAND Boolean logic operation, for which inputs are the states of the bottom layer devices and the output is stored in one of the top layer memristors (Figures 4.7 and 4.8).

Figure 4.7. Schematics and truth table showing intermediate steps for the NAND Boolean operation via material implication logic.

a

b

B1	B2	T1	T1*	T1**
off	off	off	on	on
off	on	off	on	on
on	off	off	off	on
on	on	off	off	off

Step 1: T1 ← OFF   Step 2: T1* ← B2 IMP T1   Step 3: T1** ← B1 IMP T1*
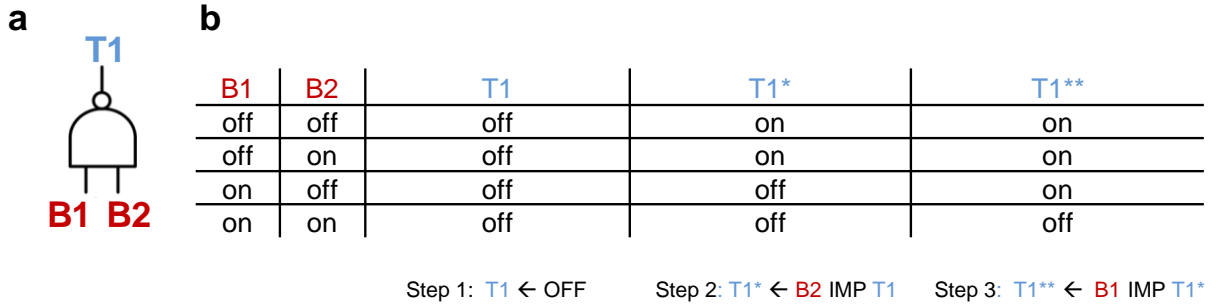
Figure 4.8. Experimental results for inter-layer NAND showing 80 cycles of operation with >93% yield for all four combinations of initial states.
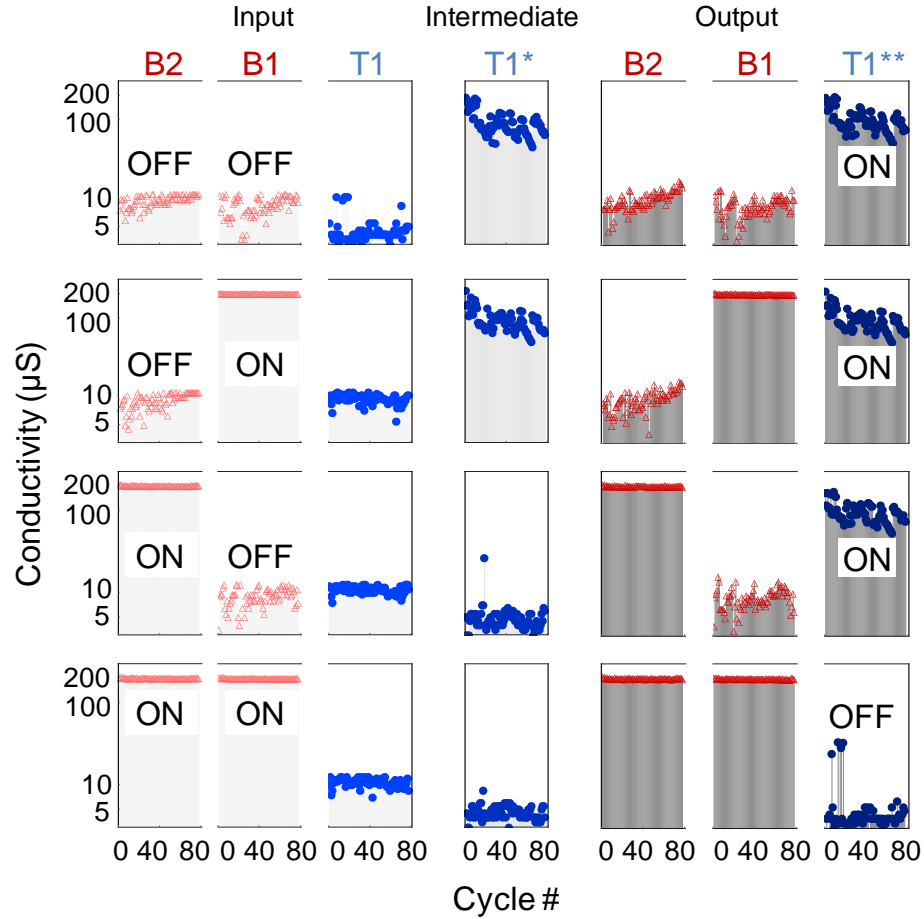
Figure 4.7 shows the NAND is realized in three steps - an unconditional reset, followed by two sequential IMP operations with the result of the first logic operation stored in top layer device, which is then used as one of the inputs to the second IMP. Figure 4.8

74

summarizes the results showing 80 cycles of reproducible operation with 93% yield. V_P was - 0.15 V and load current was a 10-ms pulse with I_L = -550 μA. Same tuning of initial devices was used as for the logic ring.

The optimal V_P and I_L were determined from numerical simulations with an additional constrain of using the same circuit parameters when the IMP logic output is in the bottom or top memristors. Such additional constrain is representative of more general case when parameters of biasing circuitry are not chosen based on switching characteristics of individual memristors. Using unoptimized values for V_P and I_L leads to incomplete ON switching of the switching device or wrong switching of the conditional device (Fig. 4.9).

Figure 4.9. Detailed information for 10 representative cycles for (a) T2*← B2 IMP T2 and (b) T1*← B1 IMP T1 that show incomplete switching due to poor choice of I_L and V_P.
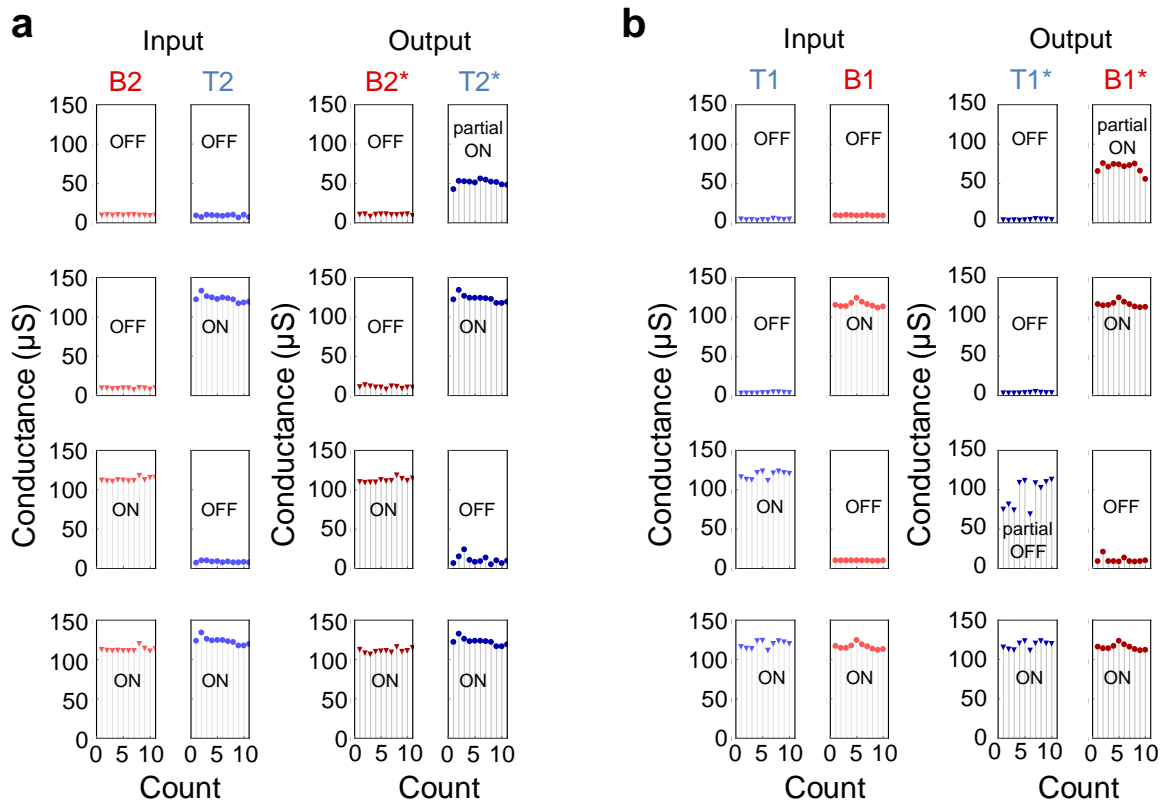
D. 1-bit Half Adder

The stack of four monolithically integrated memristor devices was further used to show the functionality of a 1-bit half adder. The Boolean variables a and b are added to calculate the sum s and the carry-out c_{out}. Firstly, a sequence of 4 NAND gates are used to calculate the sum. During this process, the devices that stored the input values need to be reused by resetting them to OFF state. The same stack of devices is then used to calculate the carry-out, while storing the previously determined value of s. Firstly, the input values have to be recopied, then the carry out can be calculated using a NAND gate and an IMP operation (Fig. 4.10). In total, 11 IMP and 6 RESET operations are needed for this implementation.

Figure 4.10. Half adder implementation (a) truth table and (b) steps needed to perform addition based on NAND gates using implication logic.

a

a	b	x_1	x_2	x_3	s	c_{out}
0	0	1	1	1	0	0
0	1	1	1	0	1	0
1	0	1	0	1	1	0
1	1	0	1	1	0	1

b

step	B1	B2	T1	T2	operation
1			a	b	write a, b
2	x_1		a	b	x_1=NAND (a, b)
3	x_1	x_2	a	b	x_2=NAND (x_1, a)
4	x_1	x_2	x_3	b	x_3=NAND (x_1, b)
5	x_1	x_2	x_3	s	s=NAND (x_2, x_3)
6		a	b	s	Rewrite a and b
7	x_1	a	b	s	x_1=NAND (a, b)
8	x_1	a	c_{out}	s	c_{out}=IMP$(x_1, 0)$

time

The half adder was experimentally implemented based on a NAND scheme as described in Fig. 4.10. Initial device conductances of ~500uS and of ~5uS were used as desired ON and OFF states respectively, since a higher ON/OFF ratio allows for better margin (Chapter 2, Figure 2.3b), The input bits a and b were copied into the devices T1 and T2 using the modified tuning algorithm presented above. The intermediate values x1 and x2 were calculated in devices B2 and B1 respectively. These two NAND gates were based on T

IMP B and B IMP B operations that required an I_L of 800uA and a V_P of 0.6V. The T1 and

T2 devices were then reused in the calculation of x3 and the sum. These two NAND gates

were based on B IMP T and T IMP T operations, which required an I_L of -375uA and a V_P of

-0.3V. The value of the sum was stored in T2 and the rest of the device stack was reused to

calculate the carry-out. Figure 4.11 summarizes the results.

Figure 4.11. Experimental results showing a 1-bit half adder implementation in a monolithically integrated system of 2x2 stacked memristors. The bit sum s is calculated and automatically stored statefully in one of the memristors while the bit carry-out c_{out} calculated.
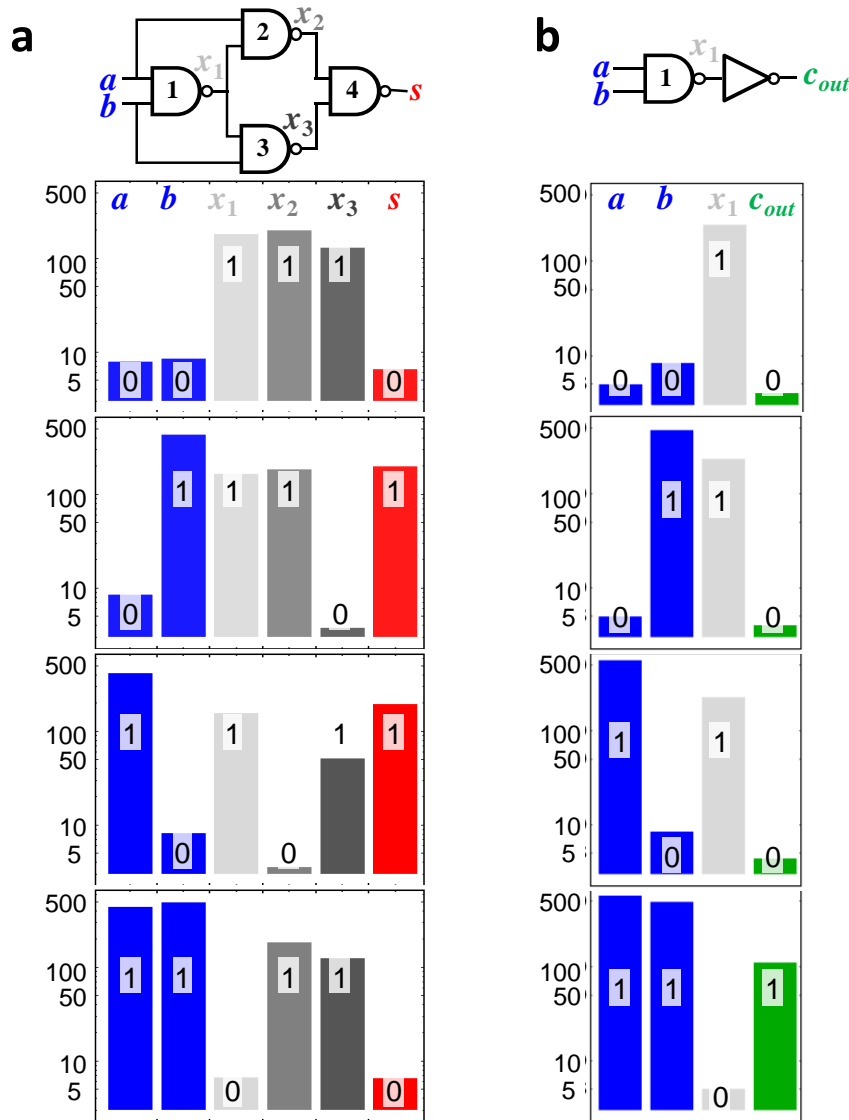
The conditional switching was close to the true "1" with an ON/OFF ratio of ~10-20, but a slight refresh was still implemented in order to make sure the maximum ON/OFF ratio between the calculation results (Fig. 4.12).

Figure 4.12. Experimental results showing conditional switching close to true "1" and slight refresh results.
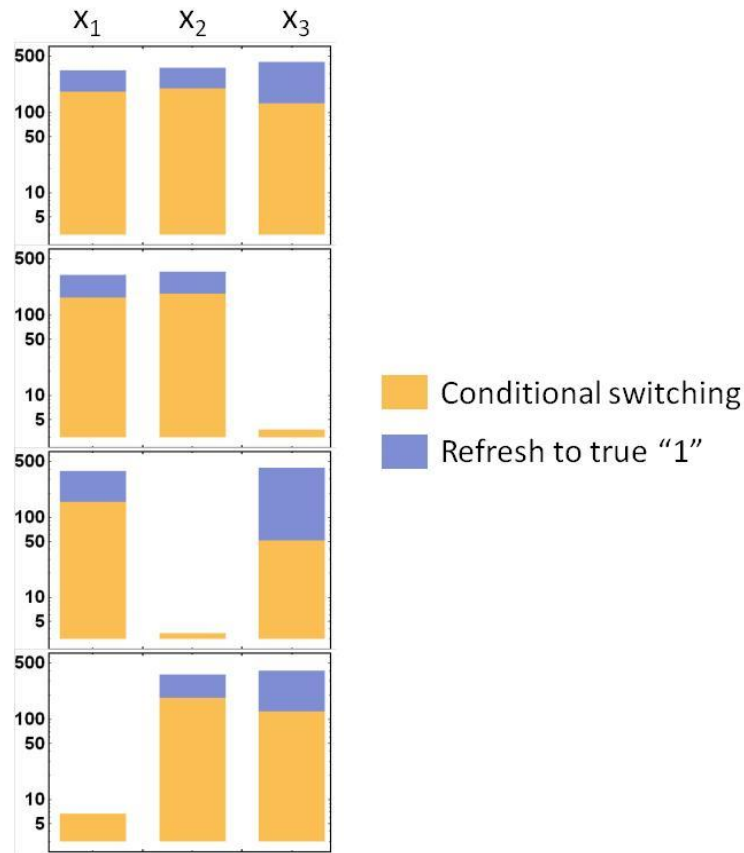
Very recently two implementations of adders using memristor-based implication logic were shown. Ballati and Ielmini [3] showed experimentally the a=0, b=1 and c_{in}=1 case for a 1-bit full adder implemented in a wire-bonded 1T1R system. Breuer and Waser [4] used complementary resistive switches (CRS) to show a 1-bit full adder based on a hybrid material implication approach. In comparison with the IMP gate suggested by Borghetti that uses 2 memristors, 1 load resistor and has 1 stateful input and 1 stateful output, the CRS

based implication logic implements a hybrid three-input gate with two non-stateful voltage inputs and one stateful (resistance) inputs. Different IMP operations are performed based on the value of the stateful input, RIMP for "1" and NIMP for "0". The output of this hybrid gate is stateful, however in order to be used for further computation, the state of the CRS device has to be destructively read, transformed into a voltage signal by using extensive external circuitry and then rewritten into the CRS device. These three-input IMP gates based on CRS devices reduce the number of devices required for computation and has the advantage of eliminating the sneak path problem, but the extra number of steps required due to the destructive read makes it impractical. A solution based on 1S1R devices was suggested by Siemon et al [5]. Since the applied voltages carry information, parallelism of such computations in a crossbar might be a problem. Due to required additional complex CMOS circuitry and challenges to parallel implementation, this hybrid solution is impractical for in-memory logic. Regarding the Feynman adder, all the external circuitry and the CRS devices both would have to be built in a 50x50x50nm cube, since the CMOS circuitry is an integral part of the logic processing, being required for transforming the resistance states into input voltages.

Our experimentally demonstrated adder is by comparison, implemented in a monolithically integrated stack of memristors. All the four cases of the truth table are shown for completeness. It uses a fully stateful IMP gate implemented with a current source to increase operational margin and maintain high switching speed and low energy consumption. The biasing can be viewed as a complicated clock that carries no information and therefore it can be shared between different computations happening in parallel across the crossbar. An improved device with an integrated selector would make such system

suitable for large array integration for in-memory logic. Moreover, this implementation allows for the realization of a Feynman adder in 50x50x50 nm [6], which will be explained in more detail in Chapter 5.

D. Summary

We have demonstrated logic-in-memory computing in three-dimensional monolithically integrated circuits. As the memristor technology continues its rapid progress (and will eventually become sufficiently advanced to allow sub-nanosecond and pico-Joule switching with $>10^{14}$ cycles of endurance, which have been demonstrated in discrete devices [7-8] in large-scale integrated memristive circuits), we expect that the presented approach will become attractive for high-throughput and memory-bound computing tasks suffering from memory bottleneck problems.

The presented approach can establish a pathway towards one of the Feynman's grand challenges-an 8-bit adder in 50x50x50nm. Resolving this challenge would require implementation of material implication logic in aggressively scaled crossbar circuits, which does not seem too taxing task given that metal-oxide memristors of the required dimensions and much larger passive memristive crossbar circuits have been already demonstrated [9-14].

The next chapter will take further steps in the realization of Feynman's proposed adder. The results from this chapter will be expanded to the case of stacked crossbars and more complex stateful-logic based 3D circuits, i.e. a 1-bit adder.

References for Chapter 4

1. Alibart, F., Gao, L., Hoskins, B. D., & Strukov, D. B. High precision tuning of state for memristive devices by adaptable variation-tolerant algorithm. *Nanotechnology*, 23(7), 075201. (2012).

2. Borghetti, J. *et al.* 'Memristive'switches enable 'stateful' logic operations via material implication. *Nature* **464**, 873-876 (2010).

3. Balatti, S., Ambrogio, S., & Ielmini, D. (2014). Normally-off Logic Based on Resistive Switches-Part II: Logic Circuits.

4. 4. Breuer, T., Siemon, A., Linn, E., Menzel, S., Waser, R., & Rana, V. (2015). A HfO2-Based Complementary Switching Crossbar Adder. *Advanced Electronic Materials*.

5. Siemon, A., Menzel, S., Chattopadhyay, A., Waser, R., & Linn, E. (2015, May). In-memory adder functionality in 1S1R arrays. In *Circuits and Systems (ISCAS), 2015 IEEE International Symposium on* (pp. 1338-1341). IEEE.

6. *Feynman Grand Prize*, full description available online at https://www.foresight.org/GrandPrize.1.html

7. Yang, J., Strukov, D. B. & Stewart, D. R. Memristive devices for computing. *Nature Nanotechnology* **8**, 13-24 (2013).

8. Wong, H.-S. P. & Salahuddin, S. Memory leads the way to better computing. *Nature Nanotechnology* **10**, 191-194 (2015).

9. Prezioso, M. *et al.* Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature* **521**, 61-64 (2015).

10. Govoreanu, B. *et al.* Vacancy-modulated conductive oxide resistive RAM (VMCO-RRAM). *IEDM Tech Dig.*, 10.2.1 - 10.2.4 (2013).

11. Yang, J., Strukov, D. B. & Stewart, D. R. Memristive devices for computing. *Nature Nanotechnology* **8**, 13-24 (2013).

12. Parkin, S. S. P., Hayashi, M., & Thomas, L. Magnetic domain-wall racetrack memory. *Science* **320**, 190-194 (2008).

13. Chevallier, C. J. *et al.* 0.13 μm 64 Mb multi-layered conductive metal-oxide memory. *ISSCC'10*, 260-261 (2010).

14. Yu, S. *et al.* HfO_x-based vertical resistive switching random access memory suitable for bit-cost-effective three-dimensional cross-point architecture. *ACS Nano* **7**, 2320-2325 (2013).

Chapter V

Adder designs for Feynmann challenge

Interestingly, three-dimensional IMP logic enables a practical solution for one of the Feynman Grand Challenges – the implementation of an 8-bit adder which fits in a cube no larger than 50 nanometres in any dimension. The specifications [1] for this adder are 1) accurate addition of any two 8-bit digital numbers (labelled A and B in this section) without overflow; 2) electrical or other types of input signals and 3) output readable from a pattern of raised nano-bumps using scanning probe microscopes or other appropriate equipment.

This chapter describes possible designs for this nanoscale adder. Since no precise specification is given regarding if the adder can have sequential or simultaneous operation, two adder designs are presented. The first one is a sequential input feed/output read 1-bit adder that has a lower number of devices, but larger number of steps. The second one is a

simultaneous input feed/output read 8-bit adder that needs higher number of devices but takes a reduced number of steps.

A. Sequential input feed/output read

The major building block – a full adder, which adds Boolean variables a, b, and c_{in} to calculate sum s, and carry-out cout, requires 6 memristors and consists of two monolithically stacked 2×2 crossbars sharing the middle electrodes (Fig. 5.1a). Two of the memristors in the crossbar are assumed to be either not formed or always kept in the OFF state (Fig. 5.1b), which eliminates leakage currents typical for crossbar circuits and makes IMP logic set margins similar to those of the demonstrated circuit.

Figure 5.1. Proposed memristor-based structure for the nano-adder. (a) Sketch of a structure and (b) its equivalent circuit.
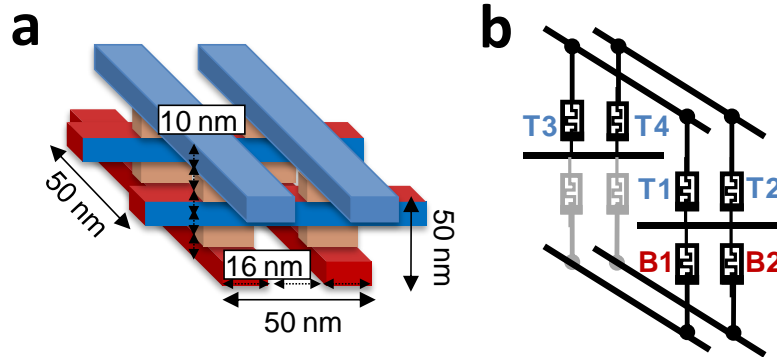
In particular, at the start of computation, a, b, and c_{in} are written to the specific locations in the circuit (Fig. 5.2b). A sequence of NAND operations, each consisting of one unconditional reset step and two IMPs, is then performed to compute c_{out} and s according to the particular implementation of Fig. Fig. 5.2a. An occasional NOT operation is implemented with one unconditional reset step and one IMP step and is used to move variables within the circuit. In total, this 1-bit full adder is implemented with 9 NAND gates

and 4 NOT gates (13 unconditional reset steps and 22 IMP steps). Finally, a full 8-bit adder
could be implemented in a ripple-carry style [2] by performing full adder operation 8 times.

**Figure 5.2. A full adder implementation with 3D IMP logic that can meet the size
requirements of the Feynmann nano-adder: (a, b) A sequence of steps and specific
mapping of logic variables to the circuit's memristors for a particular implementations
of full adder shown on panel d. The last step on panel b, in which c_{out} is placed in the
same location as c_{in}, is only required to ensure modular design, but might be omitted in
more optimal implementations.**

a

b

step	B1	B2	T1	T2	T3	T4	operation
1			a	b	c_{in}		write a, b, c_{in}
2	x_1		a	b	c_{in}		x_1=NAND (a, b)
3	x_1	x_2	a	b	c_{in}		x_2=NAND (x_1, a)
4	x_1	x_2	x_3	b	c_{in}		x_3=NAND (x_1, b)
5	x_1	x_2	x_3	x_4	c_{in}		x_4=NAND (x_2, x_3)
6	x_1	x_2	x_5	x_4	c_{in}		x_5=NAND (x_4, c_{in})
7	x_1	x_6	x_5	x_4	c_{in}		x_6=NAND (x_4, x_5)
8	x_1	x_6	x_5	c_{out}	c_{in}		c_{out}=NAND (x_1, x_5)
9	x_1	x_6	x_5	c_{out}	c_{in}	$\overline{c_{out}}$	$\overline{c_{out}}$=NOT (c_{out})
10	$\overline{x_5}$	x_6	x_5	c_{out}	c_{in}	$\overline{c_{out}}$	$\overline{x_5}$=NOT (x_5)
11	$\overline{x_5}$	x_6	x_5	x_5	c_{in}	$\overline{c_{out}}$	x_5=NOT $(\overline{x_5})$
12	$\overline{x_5}$	x_6	x_7	x_5	c_{in}	$\overline{c_{out}}$	x_7=NAND (x_5, c_{in})
13	$\overline{x_5}$	x_6	x_7	s	c_{in}	$\overline{c_{out}}$	s=NAND (x_6, x_7)
14	$\overline{x_5}$	x_6	x_7	s	$\textbf{\textit{c}}_{\textbf{out}}$	$\overline{c_{out}}$	c_{out}=NOT $(\overline{c_{out}})$

time

B. Simultaneous input feed/output read

The adder discussed in the previous section (Figs. 5.1 and 5.2) has sequential input/sequential output. By increasing the size of the two monolithically stacked crossbars to 4x4, all the 16 bits of input can be copied in the system at the beginning of the addition and the 8 bits of output can be read at the end from the top layer crossbar.

This section describes one possible set of sequences that accomplish a simultaneous input/simultaneous output. Fig. 5.3a shows a sketch of the computing nanodevice based on two stacked 4x4 memristor crossbars with 7nm half-pitch. All the devices in the crossbar are assumed to be in "0" state initially. The inputs are copied in pairs (A_i, B_i) in adjacent devices in crossbar, the first 4 bits of A and B on the bottom layer and the last 4 bits on the top layer as explained in Fig. 5.3b. The memristors B4, B8, B12 and B16 and T4, T8, T12 and T16 were selected to be used as working devices throughout computation because, having two stacked arrays of memristors as auxiliary reduces the number of moves since it increases the number of possible IMPLY operations intra- and inter-layer.

Figure 5.3. (a) Sketch of a memristor-based nanostructure that allows for simultaneous input/simultaneous output operation of an 8-bit adder in a cube of less than 50nm in any dimension (b) Device labeling on top and bottom crossbars and mapping of the input bits for A and B numbers.

86
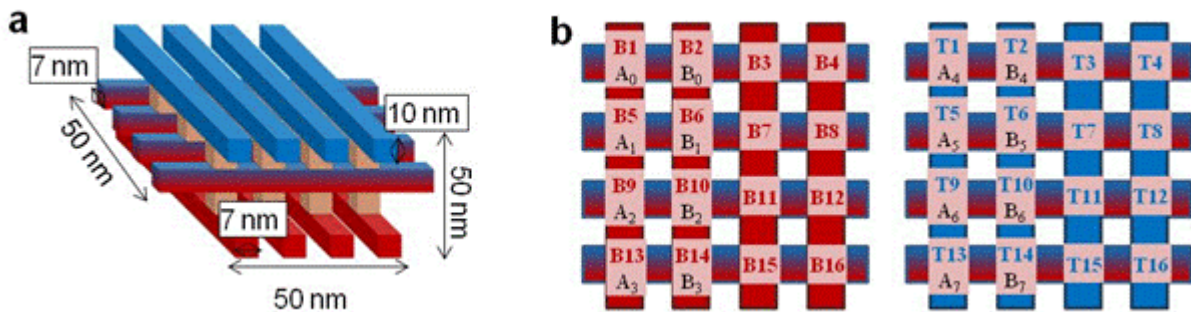
The circuit architecture selected for this exemplification is a ripple carry adder that has as building blocks 1-bit full adders based on NAND gates (Fig. 5.4a and b). Each 1-bit adder consists of 9 NAND gates labeled 1 to 9. Each NAND gate consists of a reset step and two implication logic steps. The conditional switching during implication logic is assumed to provide a value close to "1", so the output value can be reused in future computations. If this is not the case, buffers should be included to restore the value to a true "1".

Implication logic can be performed only by two devices sharing an electrode (either bottom, middle or top). A NAND gate is a succession of two implication logic operations using as switching device the output and as conditional devices the inputs. In the three-dimensional crossbar architecture, a NAND gate is implemented 1) in the same layer between three devices in the same row or column; or two input devices in different rows and columns if there is an output device sharing one electrode with each of them; and 2) in between layers with input device in one layer, output device positioned right under or above it in another layer and sharing a row or column with the other input device.

In Fig. 5.4c, a, A_i and B_i are the input bits, S is the output sum bit and Ci_i and Co_i are the carry in and carry out bits for each of the 1-bit adders. The specifications for the Feynman adder does not includes a carry in, so C_i for bit 0 is considered "0". The internal variables are labeled x1 to x7. Some variables are cleared when no longer needed in order to free up memristors for later use. The variables are reorganized at different steps to simplify computational sequence and enable read.

Figure 5.4. (a) One-bit full adder based on NAND gates (labeled 1 to 9) implemented with implication logic. The first three gates have inputs (A_i and B_i) and produce outputs ($x1_i$, $x2_i$ and $x3_i$) that change their device position in the crossbar based on the bit number i. In order to streamline the computation, these outputs are reorganized to have a fixed position in the crossbar independent of the bit number. After gate 9, the sum bit S and carry out bit C_o are moved for storage to a specified position based on the bit number. (b) 8-bit ripple carry implementation scheme. At the end of the computation, the sum bits that are stored on the bottom crossbar are moved to the top crossbar through a reorganization of the stored input and output bits. (c) Steps required for an 8-bit ripple carry adder implementation with the variables that can be cleared after use.
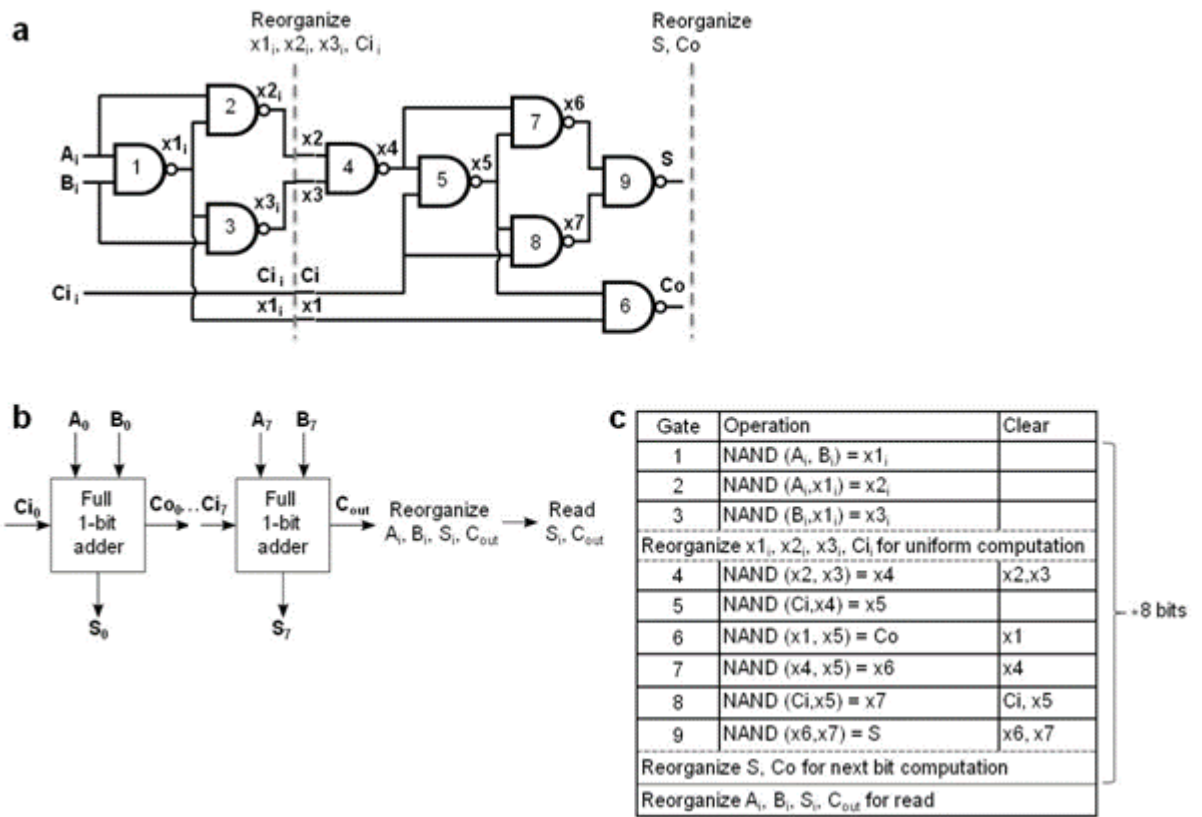
The first three gates of the 1-bit adder have one or both inputs one of the input bits A_i and B_i. These inputs are positioned differently in the crossbar based on the bit number i (Fig. 5.3b). Due to these differences and the limited space in the crossbar, the outputs of these gates ($x1_i$, $x2_i$ and $x3_i$) will be computed into different devices at each bit i (Fig. 5.5).

Because the crossbar is limited to 4x4 due to size constrains, variables have to be moved between steps to free up devices in strategic positions. In order to move the value stored in device P to device Q, a third auxiliary device R in the OFF state is needed. A NAND is performed between P and Q and its value (=P) is stored in Q. The device can be moved from P to Q only if an auxiliary device R is available and is positioned according to the requirements explained above.

Figure 5.5. Mapping of the variables in the first part of the 1-bit full adder to devices in the crossbar. This mapping is different for each bit number i because of the different positioning of the input bits A_i and B_i and the small number of crossbar memristors available. Only devices sharing an electrode can perform implication logic, so some variables are moved between steps (through a NAND operation with 0 as explained in legend) to make space for new computation.
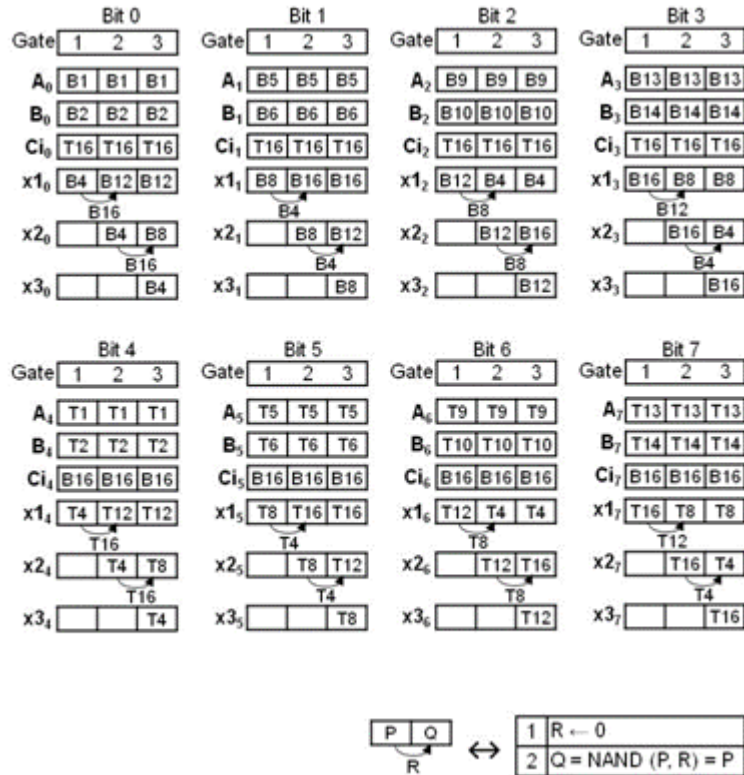
In order to streamline the computation, these outputs are reorganized to have a fixed position in the crossbar independent of the bit number (Fig. 5.6). Variables x1, x2, x3 and Ci are moved into devices T4, B4, B8 and T16 using a different sequence of moves for each bit.

Figure 5.6. Reorganization of variables from first part of the 1-bit adder. The variables are moved to specific positions in the crossbar (T4 for x1, B4 for x2 , B8 for x3 and T16 for Ci) that are the same across all bit numbers, which allows for remaining computation to be mapped to the same devices for all bits.

	Bit 0	Bit 1	Bit 2	Bit 3	Bit 4	Bit 5	Bit 6	Bit 7
$x1_i$	B12 T12 T4 T4 T8	B16 B4 T4 T4 T8	B4 T4 T8	B8 T8 T4 T12 T12	T12 T4 Move x3 Move x2 T8	T16 T4 T14	T4	T8 T4 T12
$x2_i$	B8 B12 B4 B16 Move x3 B16	B12 B4 B16	B16 B4 B8	B4	T8 T4 B4 B4 B12	T12 B12 B4 B4 B8	T16 T8 B8 B4 B8 B4 B12	T4 B4 B8
$x3_i$	B4 B8 B16	B8	B12 B8 B16	B16 B8 B12	T4 B4 B8 B8 B12	T8 B8 B12	T12 B12 B8 B8 T8	T16 T8 B8 B8 B12
Ci_i	T16	T16	T16	T16	B16 T16 T12	B16 T16 T12	B16 T16 T12	B16 T16 T12

Fig. 5.7 shows the second part of the 1-bit full adder that is mapped on the crossbars. This sequence is applicable for all bits. The results of the computation, sum bit S and carry out bit C_o are stored in T8 and B8 respectively and will have to be moved for storage and carry propagation.

After the 1-bit adder finished computing bit I, the sum bit S is stored in devices B8. In order to prevent its values to be overwritten in the next bit computation, it is moved for storage to a specified position based on the bit number (Fig. 5.8). The carry out bit Co stored in T8 is rippled to the next bit computation by moving it to the carry in position (either T16 or B16 depending on the bit number). For the last bit, Co represents the overflow and is moved later to position T4 for read.

Figure 5.7. Mapping of the variables in the second part of the 1-bit full adder to devices in the crossbar. This mapping is the same for all bit numbers. Variables x5, Co and x6 have to be moved into a new position a total of 4 times per bit in to facilitate computation. The sum and carry bits are stored in devices B8 and T8 respectively.
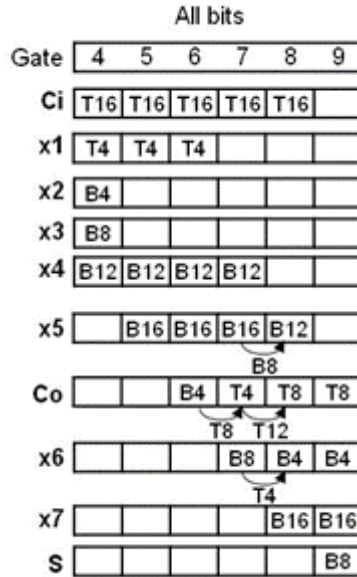
Figure 5.8. Reorganization of result variables of the 1-bit adder. For each bit, the sum variable is moved from device B8 to specific position in the crossbar for storage until final read. The carry out from T8 is moved to its carry in position for the next bit (either T16 or B16 depending on the bit number) except for the last bit.

At the end of the computation, the sum bits that are stored on the bottom crossbar are moved to the top crossbar to allow mechanical read (Fig. 5.8). Final mapping of variables to devices is shown in Fig. 5.9.

Figure 5.9. (a) Reorganization of variables after 8-bit adder to facilitate read. The sum bits 0 to 3 are moved from the bottom crossbar to the top crossbar in order to allow for their values to be mechanically read using scanning probe microscopes. Input bits B_4 to B_7 have to be moved to bottom crossbar to free devices in the top one. (b) Mapping of the input bits for A and B numbers and of the sum bits S. The sum bits are stored in the two middle columns of the top crossbar. The overflow carry Cout can be read from device T4. The values of the 16 input bits are preserved.
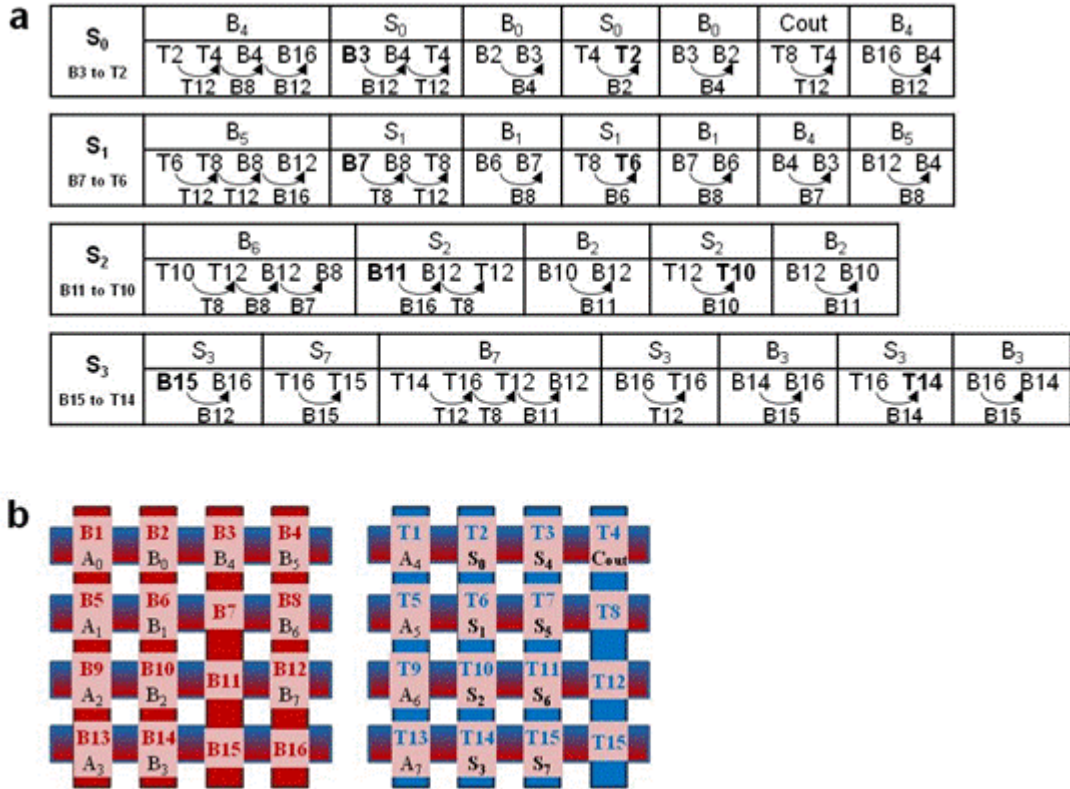
We have described the required steps to utilize three-dimensional material implication logic for the implementation of 8-bit adder which fits in a cube no larger than 50 nanometer in any dimension.[22] This adder stores all input and output signals as memristor values. The sum bits and overflow can be read mechanically from the top crossbar. The adder consists of 72 NAND gates, 150 moves and 64 clear steps (Table 5.1). A NAND gate consists of 1 unconditional RESET step and 2 IMPLY steps. A move step requires 2 RESET

steps (one for the input device and the output) and 2 IMPLY steps and a clear step is a

RESET step. In total, the adder requires 436 RESET and 444 IMPLY steps.

Table 5.1. Total number of NAND gates, variable moves and variable clear (reset) operations required for the implementation of an 8-bit adder in a 4x4x2 crossbar.

	NAND gates	Moves				Clear variables
		Adder (part 1)	Reorganize (part 1)	Adder (part 2)	Reorganize (part 2)	
Bit 0	9	2	5	4	3	8
Bit 1	9	2	3	4	2	8
Bit 2	9	2	3	4	3	8
Bit 3	9	2	3	4	3	8
Bit 4	9	2	6	4	5	8
Bit 5	9	2	5	4	6	8
Bit 6	9	2	6	4	5	8
Bit 7	9	2	5	4	2	8
Reorganize for read	-	S_0	S_1	S_2	S_3	-
		10	10	8	9	
		37				
Total	**72**	**150**				**64**

The simplest way to read an output of an adder is to measure electrically the state of memristors T2 and T3 (Fig. 6c). Alternatively, the output can be sensed thermally using scanning Joule expansion microscopy [3].

C. Summary

We have described the required steps to utilize three-dimensional material implication logic for the implementation of 8-bit adder which fits in a cube no larger than 50 nanometer in any dimension. This adder stores all input and output signals as memristor values. The sum bits and overflow can be read mechanically from the top crossbar. The sequential input feed/output read adder requires 6 memristors in a 2x2x2 crossbar

configuration (2 memristors unused in OFF state) and requires 104 RESET and 176 IMPLY steps. The simultaneous input feed/output read adder requires 32 memristor devices in a 4x4x3 crossbar configuration and needs 436 RESET and 444 IMPLY steps for operation. The large number of extra steps are required for copying the states between devices in order to preserve all the input and output states in the 50x50x50nm system.

Several major steps have to be made to practically realize the nano-computing device for the Feynman challenge. Self-evident is the task of scaling down the memristors to below 7nm features. Govoreanu has already shown 10nm x 10nm memristors [4]. The memristors utilized for this theses had >300nm feature sizes and showed no thermal crosstalk, but with scaling thermal crosstalk can become a problem [5,6]. Besides fabrication, there is also a challenge of implementing a circuit based on implication logic in a crossbar due to sneak paths as explained in Chapter 2, section C. However due to the small array size needed for the Feynman crossbar, the sneak paths might not be a problem. Through appropriate voltage schemes and with the help of selectors, this task might be achievable in a reproducible fashion.

References for Chapter 5

1. Foresight Institute https://www.foresight.org/GrandPrize.1.html#anchor183110 (Accessed on 10/18/2015)

2. Parhami, B. *Computer arithmetic: Algorithms and hardware designs.* (Oxford University Press, Inc., New York, NY, 2009).

3. Varesi, J. & Majumdar, A. Scanning Joule Expansion Microscopy at nanometer scales, Applied Physics Letters **72**, 37 (1998).

4. Govoreanu, B., Kar, G. S., Chen, Y. Y., Paraschiv, V., Kubicek, S., Fantini, A., ... & Jurczak, M. (2011, December). 10× 10nm 2 Hf/HfO x crossbar resistive RAM with excellent performance, reliability and low-energy operation. In *Electron Devices Meeting (IEDM), 2011 IEEE International* (pp. 31-6). IEEE.

5. Lohn, A. J., Mickel, P. R., & Marinella, M. J. (2014). Analytical estimations for thermal crosstalk, retention, and scaling limits in filamentary resistive memory. *Journal of Applied Physics*, *115*(23), 234507.

6. Sun, P., Lu, N., Li, L., Li, Y., Wang, H., Lv, H., ... & Liu, M. (2015). Thermal crosstalk in 3-dimensional RRAM crossbar array. *Scientific reports*, *5*.

Chapter VI

Conclusions and future work

The focus of this thesis was on monolithic stacking of memristive devices with a shared electrode for 3-dimensional computation, as a step towards a highly integrated and versatile CMOL architecture. The CMOL architecture is an integrated hybrid circuit between traditional CMOS technology and stacked layers of novel two-terminal active devices that offer unprecedented functionalities and scalability. Such CMOL systems could offer extremely high density of active devices ($10^{12}/\text{cm}^2$) that would allow for very high speed and throughput information processing, promising to make fast and cost effective hardware artificial neural networks a reality.

An example of a novel two-terminal device that can be easily implemented in a CMOL architecture is the memristor. The memristor is based on resistive switching, with its resistance capable of being modified in a non-volatile fashion using voltage or current. Due to their very small footprint, <5nm devices being reported, these devices are an excellent

candidate for terabit scale non-volatile memories. Due to the tight integration between CMOS and memristor capabilities, artificial neural networks could be easily implemented using CMOS-based neurons and memristor-based synapses.

This thesis was focused on another interesting application: performing logic using the passive two-terminal devices. The memristors can construct simple voltage dividers that allow for conditional switching, with the advantage that the output is instantaneous latched into a non-volatile memristive state. The natural performed logic operation is material implication. With such behavior, the memristive devices have a great potential for being a good candidate for large scale memories with in-memory computation capabilities. The in-memory computation would alleviate the burden posed by the von Neumann bottleneck, by dedicating some memory-specific computational tasks such as look-out tables, error-correcting, etc. to the memory itself, freeing the CMOS processor for other processing-heavy tasks. The instantaneous latching of the output into the memory, the so called "stateful" operation, is an interesting feature for energy scavenging devices, because memristive circuits can work even with intermittent power supply.

Although that the concept of memristive-based logic is attractive, there are many challenges before such technology can be adopted. Although significant improvements have been made in the past years in understanding the device physics and improve the device manufacturing, the existing memristive devices are still plagued by large device-to-device and cycle-to-cycle variations. This work provides an optimized circuit configuration that allows for reliable multi-cycle conditional switching using a memristive system with large voltage spread. The provided experimental results show for the first time hundred of cycles of implication logic between memristors in different layers. Moreover, this optimized circuit was also used to show for the first time a 1-bit half adder implementation in a monolithic

stack of memristors. This opens the road to implementations of more complex circuits, but more challenges have to be solved first. First of all, reliable multi-input/multi-output gates based on implication logic should be experimentally demonstrated which would allow for the Boolean design of the circuit to be greatly simplified. The crossbar architecture that the memristors are typically arranged in to maximize density, suffers from the intrinsic problem of electrical crosstalk due to leak paths. This prevents the output of implication logic to be written simultaneously in multiple memristors. Several theoretical solutions have been proposed, but more investigations are needed. The crosstalk also poses challenges for the parallel execution of multiple implication logic operations in the same crossbar, particularly if a large number of devices are in the low resistance state. Novel architectural designs should be investigated to indentify particular applications that in-memory memristor "stateful" logic would be the most beneficial for.

Two different fabrication flows were presented in this work, one based on lift-off techniques and one on ion milling. The ion milling-based processing shows a higher manufacturability and it allows for faster turn-over, which provides an essential advantage in the experimental investigation of the large space of variables that influence memristor design. Ion milling-based devices show tighter switching variations in comparison with the lift-off ones. Potential areas of further improvement for the ion milling-based device design could be reducing the bottom electrode slope and engineering the thin interfacial barrier for increased non-linearity in the ON state. Due to these advantages, the ion milling-based patterning greatly helps the CMOS/memristor integration in laboratory settings, where the CMOS circuitry is externally manufactured in foundries and available on small size chips. In order to monolithically integrate multiple layers of memristor layers on these CMOS chips, planarization is a requirement. Initial planarization using chemical mechanical polishing is

needed due to the high aspect ratio of the CMOS outer metal layer, but it is strenuous and, for such small chips, running the risk of damaging them. Subsequent planarizations are required after each memristor layer. Ion milling-based planarization could be developed to provide a safer and more controllable option. The ion-milled process will be further adapted for the CMOL architecture currently in development in the group, that would allow for the experimental demonstration of a variety of exciting applications.

A third topic of future work is in the understanding of device physics and fine tuning of device design in order to improve reliability, decrease variation and achieve operation voltage and currents in the desired range. Conditional switching is particularly sensitive to variation in the switching thresholds of the devices, so the devices should be as similarly as possible to each other and to their own past behavior over millions of switching cycles. Otherwise, some switching conditions will work for some devices but not for others. A very high ratio between the high and low resistance states or the integration with a selector device is also desirable in achieve a large scale working memristor crossbar, because it reduces the electrical crosstalk. The devices presented in this work showed no thermal crosstalk between the different layers. However, recent simulation work in the field has shown that with the down scaling of the devices, thermal crosstalk could become a problem, potentially posing additional constrains on the material choice and device designs. Device improvements will be possible only through thorough investigation of the physics of the resistive switching mechanism which is a topic of current and future work.

Appendix A

Process flow

In this appendix, the process flow details for lifted-off monolithically stacked memristive devices is described.

1. Wafer preparation

4" Si wafers, $525\pm25\mu m$ thick, covered with 2000Å of SiO_2 on both sides purchased from WRS Materials were used for the entire duration of this dissertation work. The wafers have a major and minor flat needed for the proper auto-alignment of ASML DUV stepper. Since the stepper only functions with full 4" wafers, no pieces were used.

2. Surface preparation for bottom contacts

➢ Standard solvent clean

    ➢ 3 mins ACE in ultrasonic bath at frequency 10 and intensity 10

    ➢ 3 mins ISO in ultrasonic bath at frequency 10 and intensity 10

    ➢ 3 mins DI rinse in ultrasonic bath at frequency 10 and intensity 10

    ➢ N2 blow dry

    ➢ Dehydration bake - 100◦C for 1 mins, 1 mins cool down

➤ Lithography for bottom contacts

- PR coat

- Spin DSK-101 @2500 rpm, 30s, recipe #5.

- Clean the backside of the wafer using ERB solvent for maximum flatness. Remove any particles on the backside.

- Bake at 180∘C, 60 secs on preheated aluminium top hotplate with the wafer placed in the middle of the hotplate for maximum uniformity.

- Cool down for 1min on metal plate on the bench.

- Spin UV210 @2000 rpm, 30s, recipe #4.

- Clean the backside of the wafer using ERB solvent for maximum flatness. Remove any particles on the backside.

- Bake at 135∘C, 60 secs on standard bench hotplate with the wafer placed in the middle of the hotplate for maximum uniformity.

- Cool down for 1min on metal plate on the bench.

- Expose the sample using ASML DUV stepper.

- PR development - 68 sec development in AZ-300MIF developer, slight continuous agitation for undercut, 3 min DI rinse and 2 min N2 blow dry.

- Inspect under DUV microscope to ensure that there is no scum, all the features have developed and the alignment marks have properly been exposed and developed.

- Descum – 30 sec descum in oxygen plasma (PE-II system) at 100 W, 300 mT.

➤ **Metal deposition for bottom contacts**

- Load private Ti and Pt sources in E-beam#1.

- Load wafer on rotating mount.

- Let chamber evacuate until $<3 * 10^{-6}$ Torr, the heat up the Ti source for 1min until it melts and any impurities evaporate.

- Deposit Ti/Pt contact – 50/200 Å thick. Deposit both metals at 0.5 A/sec with no sweep.

- Lift-off - Heat up the 1165 stripper at 80°C for 20-30 mins prior to immersing the sample. Use a sample holder to position the sample upside down, so pieces of lifted-off metal can fall off. Leave the sample in 1165 at 80°C for minimum 24h. Perform standard solvent clean. Check using AFM to make sure no rabbit ears are observed.

➤ **Lithography for switching layer and middle contacts**

- Spin LOL2000 @3500 rpm, 30s, recipe #6.

- Clean the backside of the wafer using ERB solvent for maximum flatness. Remove any particles on the backside.

- Bake at 210◦C, 90 secs on preheated aluminium top hotplate with the wafer placed in the middle of the hotplate for maximum uniformity.

- Spin DSK-101 @2500 rpm, 30s, recipe #5.

- Clean the backside of the wafer using ERB solvent for maximum flatness. Remove any particles on the backside.

- Bake at 180∘C, 60 secs on preheated aluminium top hotplate with the wafer placed in the middle of the hotplate for maximum uniformity.

- Cool down for 1min on metal plate on the bench.

- Spin UV210 @2000 rpm, 30s, recipe #4.

- Clean the backside of the wafer using ERB solvent for maximum flatness. Remove any particles on the backside.

- Bake at 135∘C, 60 secs on standard bench hotplate with the wafer placed in the middle of the hotplate for maximum uniformity.

- Cool down for 1min on metal plate on the bench.

- Expose the sample using ASML DUV stepper.

- PR development - 60 sec development in AZ-300MIF developer, slight continuous agitation for undercut, 3 min DI rinse and 2 min N2 blow dry.

- Inspect under DUV microscope to ensure that there is no scum, all the features have developed.

- No descum in oxygen plasma.

➢ **Deposition for switching layer and middle contacts**
- Use a reactive sputtering chamber (Sputter #3 or Sputter #4)

- Calibrate Al_2O_3, Ti and Pt deposition rates using test samples and elipsometry.

- Calibrate TiO_{2-x} stoichiometry.

- Calibrate TiO_{2-x} deposition rate using test samples and elipsometry.

- Pre-sputter in the empty chamber.

- Load the sample of interest and deposit Al2O3 (4nm), TiO_{2-x} (25nm), Ti (15nm), Pt (25nm) without breaking the vacuum.

- Lift-off - Heat up the 1165 stripper at 80°C for 20-30 mins prior to immersing the sample. Use a sample holder to position the sample upside down, so pieces of lifted-off metal can fall off. Lift-off the sample in an ultrasonic bath for 60min using the full swing setting and maximum power and intensity.

- In order to remove the rabbit ears, submerse the wafer in acetone and swab for 2-3 min focusing on the devices of interest. Use the standard solvent clean afterwards. Check using AFM to confirm that the rabbit ears are removed.

➢ **Planarization**

Sacrificial layer deposition

- Cool down the Advanced Vacuum PECVD #2 chamber from 300C to 175C until the temperature is stable (45-60min).

- Clean the chamber using standard clean recipe for 10 min

- Pre-coat the chamber using the standard SiO_2 recipe for 10min.

- Calibrate the deposition rate using a test sample and the elipsometer. The standard deposition rate is ~30nm/min.

- Load the sample of interest in the chamber and deposit SiO_2 at 175C for ~25min in order to achieve a thickness of ~750nm. Remove the sample as soon as the

deposition is done to prevent the over-annealing of the switching layer in oxygen atmosphere.

- Let the sample cool down.

Polishing

- Use a chemical mechanical polisher (CMP Logitech Orbis)

- Clean the hard polishing pad (if possible, have a new one installed right before to prevent potential scratching).

- Install a new 4" holding pad on the 4" head. Make sure no sticky pieces of glue are left from previous holding pads by wiping vigorously with acetone. Make sure no bubbles are formed between the holding pad and the head, which can cause non-uniformities.

- Install the wafer on the holding pad using one 50μm thick blue shimmy and water. Make sure the wafer holds well.

- Prepare the polishing pad by covering it in slurry for 2min at slow rotation (10).

- Polish the wafer fast with an 80 rotation and 50 slurry for 3 min.

- Flush with deionized water for 2 min

- Check using the elipsometer how much SiO_2 was removed. This recipe typically removes ~ 550nm. In order to make sure the polishing results are good, test wafers covered in blanket SiO_2 can be used before running the sample of interest.

- Rinse in deionized water by facing in a sample holder. In order to make sure all the slurry particles are removed, sonicate in ultrasonic bath for 10min.

- Check using AFM to make sure the surface is completely planar and smooth. The features beneath should be barely visible.

  Etch-back

- Change the gases to Ar and CHF_3 if needed in the ICP #1 system.

- Clean chamber for 10min using standard O_2 clean.

- Load the sample onto a 6" carrier wafer using oil. Make sure the entire surface of the sample is covered in a thin layer of oil in order to make good thermal contact with the carrier. No oil should be coming out at the edges of the sample once it is mounted on the carrier.

- Etch using recipe 187 using small time increments. After each etch, check using AFM to see if the middle electrodes are partially exposed. Repeat the etching adjusting the time accordingly and recheck using the AFM. Stop etching when ~ 15nm of middle electrode is visible. The color of the wafer should be very dark blue due to the SiO_2 remaining.

➢ **Lithography for switching layer and top contacts**
- Standard solvent clean the wafer.
- Spin DSK-101 @2500 rpm, 30s, recipe #5.

- Clean the backside of the wafer using ERB solvent for maximum flatness. Remove any particles on the backside.

- Bake at 180∘C, 60 secs on preheated aluminium top hotplate with the wafer placed in the middle of the hotplate for maximum uniformity.

- Cool down for 1min on metal plate on the bench.

- Spin UV210 @2000 rpm, 30s, recipe #4.

- Clean the backside of the wafer using ERB solvent for maximum flatness. Remove any particles on the backside.

- Bake at 135∘C, 60 secs on standard bench hotplate with the wafer placed in the middle of the hotplate for maximum uniformity.

- Cool down for 1min on metal plate on the bench.

- Expose the sample using ASML DUV stepper.

- PR development - 68 sec development in AZ-300MIF developer, slight continuous agitation for undercut, 3 min DI rinse and 2 min N2 blow dry.

- Inspect under DUV microscope to ensure that there is no scum, all the features have developed.

- No descum in oxygen plasma.

➢ **Deposition for switching layer and middle contacts**
- Use a reactive sputtering chamber (Sputter #3 or Sputter #4)

- Calibrate Al_2O_3, Ti and Pt deposition rates using test samples and elipsometry.

- Calibrate TiO_{2-x} stoichiometry.

- Calibrate TiO_{2-x} deposition rate using test samples and elipsometry.

- Pre-sputter in the empty chamber.

- Load the sample of interest and deposit Al_2O_3 (4nm), TiO_{2-x} (25nm), Ti (15nm), Pt (25nm) without breaking the vacuum.

- Lift-off - Heat up the 1165 stripper at 80°C for 20-30 mins prior to immersing the sample. Use a sample holder to position the sample upside down, so pieces of lifted-off metal can fall off. Lift-off the sample in a heated bath at 80°C for 24h.

- In order to remove the rabbit ears, submerse the wafer in acetone and swab for 2-3 min focusing on the devices of interest. Use the standard solvent clean afterwards. Check using AFM to confirm that the rabbit ears are removed.

➢ **Lithography for etching to expose measuring pads**

- Standard solvent clean the wafer.

- Spin UV6 @3500 rpm, 30s, recipe #6.

- Clean the backside of the wafer using ERB solvent for maximum flatness. Remove any particles on the backside.

- Bake at 135◦C, 60 secs on standard bench hotplate with the wafer placed in the middle of the hotplate for maximum uniformity.

- Cool down for 1min on metal plate on the bench.

- Expose the sample using ASML DUV stepper.

- PR development - 45 sec development in AZ-300MIF developer, slight continuous agitation for undercut, 3 min DI rinse and 2 min N2 blow dry.

- Inspect under microscope to ensure that all the features have developed.

- No descum in oxygen plasma.

➢ **Etching**

- Change the gases to Ar and CHF_3 if needed in the ICP #1 system.

- Clean chamber for 10min using standard O_2 clean.

- Load the sample onto a 6" carrier wafer using oil. Make sure the entire surface of the sample is covered in a thin layer of oil in order to make good thermal contact with the carrier. No oil should be coming out at the edges of the sample once it is mounted on the carrier.

- Etch using standard recipe 118 for 1'30". Check electrically that the pads are conductive.

- Lift-off - Heat up the 1165 stripper at 80°C for 20-30 mins prior to immersing the sample. Use a sample holder to position the sample upside down. Lift-off the sample for 2h. Perform standard solvent clean.

➢ **Sample measuring**

- Use Agilent B1500A Semiconductor Device Analyzer and attached probe station to measure the devices. Check for the lines to be connected and for the virgin state resistance before moving to form the device using current sweep.