S-type Negative Differential Resistance: Emerging Memory and Oscillators for Next-Generation Computation

A DISSERTATION
Submitted in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY
Electrical and Computer Engineering

by
Abhishek A. Sharma
M.S., Electrical & Computer Engineering, Carnegie Mellon University

CARNEGIE MELLON UNIVERSITY
Pittsburgh, Pennsylvania
December 2015

Abstract

Increasing data-centric nature of compute has motivated the need for overcoming the von Neumann memory-access bottleneck. Multi-functional beyond-CMOS have shown a great potential in uniquely complementing and augmenting the compute capability by utilizing emerging paradigms like on-chip memory and brain-inspired computing. In this work, we focus on first understanding the physics of devices that show S-type negative differential resistance behavior (S-NDR) and then engineering them for use in emerging memory (RRAM, selectors) and compute (oscillators for simulated annealing, neural networks, physically obfuscated keys) architectures.

To understand the electro-thermal dynamics of filament formation in metal- semiconductor/oxide -metal (MSM) stacks, we first develop a novel high-speed transient thermometry. This reveals a two-step current localization and nucleation process that is responsible for forming or threshold switching in these MSM stacks. This current localization event manifests as S-NDR in these devices, which we explore to variously understand threshold switching and oscillatory behavior. We also apply the developed nano-scale thermometry to resistive switching memory devices to extract the role of temperature in the switching process. After establishing self-consistency with microstructural changes under a TEM, we estimate the filament size and evolution with bias and current compliance.

In order to use these S-NDR devices as threshold switches and oscillators, we show for the first time, stack-engineering by changing the material composition, the electrode material and ballast-types to achieve > 500 MHz frequency, < 50 μW power, 0.6 V voltage-swing operation as a compact 1T1R oscillator, < 1 V operation, < 1 pA leakage current, ON-OFF ratio of > 10^6, and a $J_{ON,max}$ of > 1 MA/cm^2 for a threshold switch. Finally, using this engineered device, we show demonstrations of oscillator coupling and phase control, injection-locking and noise-reduction. On understanding the role of circuit parasitics on the oscillator behavior, we propose directly-connected simulated annealing for < 100 fJ/compute image feature extraction engine. The application of these oscillators in oscillatory neural networks (ONNs) and entropy sources for generating physically obfuscated keys (POKs) is also explored.

Acknowledgments

My sincere gratitude and deep respect are towards my advisors Prof. James A. Bain and Prof. Jeffrey A. Weldon for guiding me both technically and otherwise throughout my journey of doctoral studies. With the same breath, I must express my thanks to Prof. Marek Skowronski who also equally contributed to my mentorship. I'd also like to thank my committee members, Malgorzata Jurczak and Charles Kuo for their invaluable inputs throughout my time at Carnegie Mellon. Gratitude to SONIC,GRC,Intel-MSR, Smith Fellowship and others for financial support.

I'd also like to thank different professors and researchers that I worked with and learnt so much from – Andrzej Strojwas, Wojciech Maly, Diana Marculescu, David Greve, Yoosuf Picard, Elijah Karpov, Gilbert Dewey, Charles Augustine, Eriko Nurvitadhi, Lawrence Pileggi. Among peers and seniors, I'd like to thank Jonghan Kwon, Mohammad Noman, Ranga Kamaladasa, Gregory Slovin, Dasheng Li, Darshil Gala, Yunus Kesim, Keiko Nishikawa, Thomas Jackson, Kaustubha Neelathalli. Their contribution to my technical growth has been immense. It is almost impossible for me to think of how I could have accomplished as much as I could without their guidance and help.

If I list all of my close friends that helped to keep me always charged emotionally during my PhD career, it would make for a long list. Hence, I'd like to summarize them in form of the meanings of their names, as can best be translated from different languages to English – Light (x4), Chrysanthemum, Jewel-shard, Lotus (x2), Sun-beam, Foliage, Wisdom, the Ascendant (x2), bringer of thunder, Splendor (x2), Riverine (x3), the Truth that hides, Brilliant Red, Spirit, Rubia, Crescent of the Moon, the Singularity, the Multuplicity, and Rudra. They have all inspired me as much as their names and their nominal meanings are as essential a part of them to me, as they are. I'd also like to thank those who gifted me challenges and grief, for I do not think I could have grown intellectually and emotionally without them. I also thank my students.

Finally, I thank my family – my caring parents and nurturing grandparents, who were my first teachers and continue to sprinkle my life with that which is intellectually captivating and that which is emotionally pleasing. I hope I can unrelentingly rise in life along the routes I set for myself, consuming knowledge, bathing in logic, spirited with emotion and honing with truth.

Table of Contents

Chapter 1

Introduction

Increasing data-centric nature of computation has motivated the need for overcoming the memory-access bottleneck, also known as the von Neumann bottleneck. In massively parallel computational tasks, such as pattern recognition, conventional computing architectures have insufficient power efficiency for energy constrained environments. Thus, multi-functional beyond-CMOS have shown a great potential in uniquely complementing and augmenting the compute capability by utilizing emerging paradigms like on-chip memory and brain-inspired computing. By design, modern computer architecture assumes spatio-temporal locality of data and hence tries to maximize the data loaded in cache located on chip. If the desired data is not found in cache, it has to go off-chip to access the main memory (DRAM) and eventually storage (HDD/SSD). The latencies associated with data access range from ns for SRAMs and DRAM; and ms scale for HDDs. Most of this delay comes from transit delay associated with going off-chip or from mechanical access. This problem is exacerbated in data-intensive applications like image processing. Two possible methods of resolving this problem are: (1) Integrating a storage class memory at the backend of the line (BEOL) [1] or (2) Parallelizing compute using brain-inspired graphical methods [2]. As flash scaling is reaching its limits [3]-[5], newer memories like resistive random access memory (RRAM) and magnetic random access memory (MRAM) are being looked at, as a storage-class flash replacement and DRAM replacement [1] respectively. Similarly, some of the emerging devices like Carbon nanotubes (CNTs), RRAM and S-type negative differential resistance (S-NDR) oscillators can be utilized as a BEOL graphical processing engine for increased compute parallelism. These unique devices exhibit physical and electrical properties that make them uniquely suited to improve compute without the need for the expensive wafer stacking using interposers and through-Si-vias (TSVs). Moreover, because the operation of these devices involves a fundamentally different mechanism

than, for instance, transistors, they can enable novel methods of compute for which CMOS is comparatively inefficient.

Transistors, which form the basic building block of modern electronics, show tremendous versatility as multi-functional devices. In quest for another multi-functional device that can complement modern computation, we will focus this work on devices that exhibit S-type negative differential resistance or S-NDR. This work will attempt to interrogate these S-NDR devices to explore their potential as a transistor-like building block. Figure 1.1 shows a typical I-V schematic of a device with S-NDR characteristics.

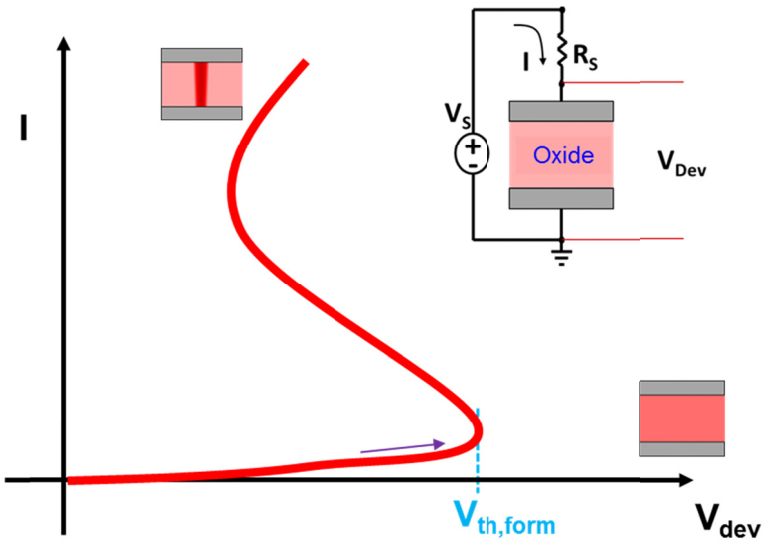


Fig. 1.1: Figure showing schematic I-V characteristics of an S-NDR device. The inset shows a schematic circuit diagram and stack that is used to generate the I-V.

The device shown in Fig. 1.1 is a leaky metal-insulator-metal stack. As fabricated, these devices have a very high resistance (also referred to as the pre-formed or OFF state), as shown in the I-V curve at low biases and currents. In this regime, the resistance (*V/R*) and the differential

resistance (*dV/dI*) are both positive. As the source bias is increased for this circuit beyond the labeled threshold or forming voltage, $V_{th,form}$, the device enters a negative differential resistance regime where the voltage across the device decreases as the current increases. If the bias is increased further, the device again reverts to a positive differential resistance, with a much lower resistance than the OFF state. This regime is known as the ON-state. Previous works [6] have proposed that the OFF-state corresponds to uniform conduction through the device whereas the ON-state corresponds to the device conducting locally. Associated with this localization event, the electrical properties of the device can thus be tuned temporarily or permanently. As the S-NDR element is a simple MIM/MSM stack, they can be integrated in large and dense cross-point arrays [11] with of minimum-pitch (4F2 footprint) that the lithography can provide. Moreover, as the conduction mechanism of these devices is filamentary in nature in the ON-state, they can be scaled to very small footprints of ~ 10 nm [9].

In this work, we focus on first understanding the physics of devices that show S-type negative differential resistance behavior (S-NDR) and then engineering them for use in emerging memory (RRAM, selectors) and compute (oscillators for simulated annealing, neural networks, physically obfuscated keys) architectures. The central idea is to integrate the working of S-NDR devices as memory units with their oscillatory response under the same framework.

1.1. Resistive Random Access Memory (RRAM)

RRAM devices generally have a very simple metal/oxide/metal heterostructure as shown in Figure 1.2, akin to the S-NDR devices explained in the previous section. The oxide thickness is usually in the range of 2-50 nm. Over the last few years, many different oxides have been found

to display resistance switching characteristics. TiO_x, HfO_x, TaO_x, NiO_x, AlO_x, SiO_x [7], etc. are just a few examples of this. Pt, Ti, TiN, and Al are some of the typical metals used for the electrodes. The resistance switching in these devices is typically represented by the current-voltage (*I-V*) curve shown in Fig. 1.2.

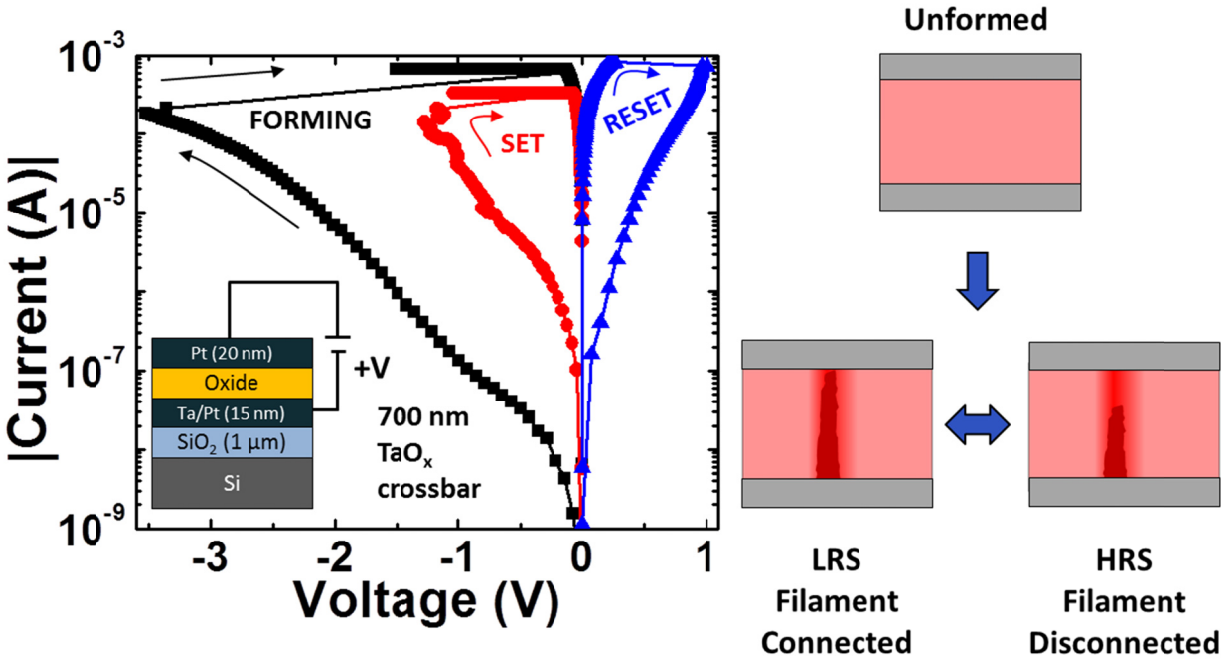


Fig. 1.2: A typical RRAM device (in this example TaO_x) undergoing forming (black), RESET (blue) and SET (red), alongside common visualization of the physical origin of resistance change in form of filament connection and disconnection.

These devices do not start off as being resistive switching memories; they have to go through a one-time programming process known as forming or electroforming. The forming process involves the application of a high voltage pulse that causes the oxide to breakdown and form a conductive filament that shunts the two metal electrodes, causing the resistance to decrease [8]. The LRS corresponds to the shunted conductive filament. This filament can be disconnected by

applying a voltage of the opposite polarity (referred to as RESET). Once the conductive filament is disconnected, the device resistance increases, and the device is said to be in the HRS. Applying a voltage of the same polarity as forming, causes the device to revert back to the LRS (referred to as SET). The device can now be cycled between the two states. The nature of the conductive filament is still a matter of debate. Typically, the conductive filament is assumed to be made up of a reduced oxide (sub-oxide) phase that is conductive in nature, often referred to as an agglomeration of oxygen vacancies (absence of oxygen atoms).

As memory (RRAM), these resistive switching devices have shown to offer great potential in terms of compactness & scalability [9], endurance [10], integration [11] and power efficiency [9]-[10]. Additionally, despite being an area of research since past 4-5 decades, it still continues to generate exciting questions regarding the fundamental nature of electronic conduction and ionic motion. While a huge corpus of research exists in dielectric breakdown phenomenon and resistive switching, one of the most significant challenges associated with commercialization of RRAM is the lack of depth in understanding the resistive switching physics, thus resulting in several often circumstantially degenerate models with limited predictive ability [12]-[15]. In this work, we will develop novel electro-thermal characterization techniques that can guide existing models. In this work, we will discuss how the forming process, which is central to the device operation as a memory (post-forming), is a natural extension of the S-NDR behavior that these devices exhibit pre-forming. As per the explanation of S-NDR behavior from the previous section, the devices that undergo forming process are biased in the ON-state of the S-NDR I-V curve. If the ON-state is maintained even after the bias has been removed, it implies that the devices has permanently changed, or formed or ‘memory switched’. If the ON-state is volatile, the device in ON-state is considered to be ‘threshold switched’. This will help us in addressing

the challenges associated with the physical understanding of the forming and subsequently switching process. Specifically, this work will attempt to examine the following questions: (1) What are the steps preceding the forming process? (2) What initiates vacancy migration? (3) What are the peak temperatures reached during the forming and switching process? (4) What is the role of temperature in the conductive filament formation and vacancy migration during forming and switching? (5) What controls the peak filament size during forming and switching? (6) What is the role of temperature in endurance failure of RRAM devices?

1.2. S-NDR Oscillators

Using the preliminary understanding of resistive switching obtained in the initial course of this study; we will also explore the possibility of using a similar device stack as an oscillator for brain-inspired oscillatory neural networks. These oscillatory elements utilize threshold switching phenomenon that exists in several sub-stoichiometric transition metal oxides, a phenomenon often associated with electronic localization effects that eventually lead to oxide breakdown [16]. Specifically, when the devices are biased in the negative differential resistance regime, they exhibit sustained relaxation oscillations, as shown in Figure 1.3 (intersection of the green dotted line with the I-V shows the bias-point corresponding to oscillatory behavior). Similarly, if the devices are biased to a higher current value in the ON-state, the device will undergo threshold switching i.e. settle to an ON-state. Depending on the material parameters (whether or not it undergoes forming or not), the device will remain in the ON-state till the bias is maintained, a phenomena referred to as threshold switching. Once the bias is removed, the device will revert back to the unformed OFF-state (violet dotted line represents the bias point required for threshold switching).

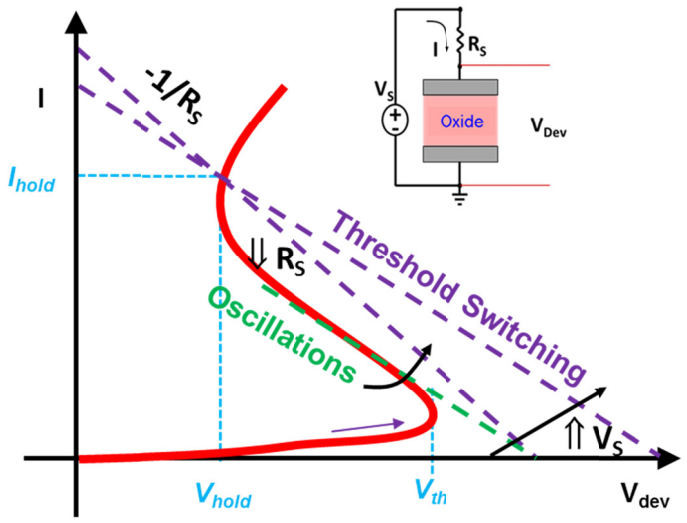


Fig. 1.3: I-V schematic of an S-NDR device oscillatory and threshold switching bias-points.

Figure 1.4 shows the coexistence of oscillatory (black) and threshold switching (red) response in these devices.

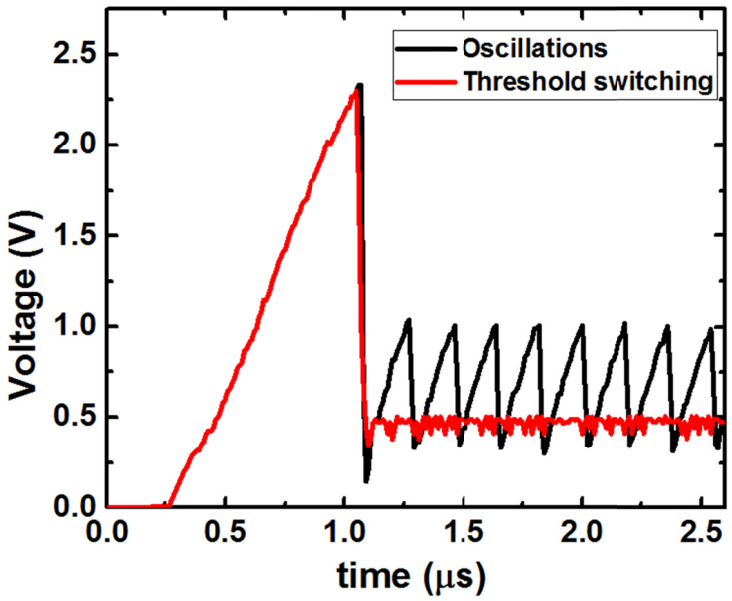


Fig. 1.4: Coexistence of oscillatory and threshold switching response in a typical S-NDR device.

As threshold switches, these device find applications as two-terminal threshold switch selectors for cross-point memories [17]. As oscillators, these devices thus serve as interesting building

blocks for large arrays, offering the possibility of coupling, compactness, and power efficiency in oscillator networks. However, the major challenges faced by this class of devices are: (1) Lack of physical understanding of the threshold-switching/oscillatory response, (2) High voltage and power needed to operate these devices, and (3) Poor understanding of co-design needed between device design and circuit topology connected to the device. This work will use TaO_x as a demonstration vehicle to show co-existence of memory switching, threshold switching and oscillations. Current attempts at engineering these devices for low-power and high-performance have met with limited success. In this thesis, we will generate an understanding of these oscillatory devices so as to engineer them for low-power and high-performance. Specifically, this thesis will explore the following questions: (1) What determines the transience of the ON-state? (2) How are threshold switches/oscillators different from memory switches? (3) What sets the performance metrics of the oscillator and how can it be modified?

1.3. Organization of thesis document

This document consists of a total of six chapters and a conclusion. The first chapter introduces the concept of S-NDR and its relationship with RRAM devices and oscillators. Detailed literature survey is intentionally left out to maintain high-level simplicity; literature will be surveyed at the beginning of each chapter for details. As the final goal is to attempt to shed light on the multi-functionality of the S-NDR element, we will first attempt at understanding the physics behind the device operation. This involves the understanding of the threshold switching/forming process. The forming process controls the filament diameter and its properties; however, it is a failure mechanism for threshold switches and oscillators that are created out of S-NDR devices. Chapter 2 will discuss the reversibility of the filament formation

during the formation process, thus clearly separating the difference in forming and threshold switching/oscillations. For this purpose, we develop a novel high-speed thermometry to understand the process of localization and the associated temperature excursions. In Chapter 3, we use the same thermometry to determine the filament properties (size, temperature, electrical properties etc.) of an RRAM device in the LRS. Thermometry in switched devices enables us to understand the role of temperature in switching, the microstructural changes that accompany these temperature excursions and failure mechanisms.

Chapter 4 revisits the S-NDR devices and discusses the origins of oscillations in TaO_x-based devices, clearly highlighting origin of frequency control, the role of parasitics and ballasts on the oscillation response. The devices explained in Chapter 4 cannot be used as is, for computation; they have to be engineered for low-power and high performance. Chapter 5 delves into the stack-engineering of TaO_x –based oscillators to yield the best-in-class oscillator (in terms of power and performance) and explores its use in oscillatory neural networks (ONNs). While compact oscillators provide unique advantage for dense array implementations, the CMOS circuitry around ONNs makes the implementation both power and area inefficient. We also discuss scalability in area and power, and variability in these oscillators. Thus Chapter 6 discusses how these oscillators can be directly coupled using unique physics that maintains the temporary ON-state in presence of coupling elements like capacitors. With this, we demonstrate for the first time, coupled oscillator pair with full phase coupling and control using transistors. To overcome some of the challenges discussed in Chapters 2, 4 and 5, we demonstrate injection-locking in these oscillators. This enables us to program the initial phase of these oscillators using a global clock and reduce the drift in frequency due to aging. In order to simulate dense parallel systems with these oscillators, we first develop a SPICE model of these relaxation oscillators using van

der Pol's formalism, which is a second order non-linear differential equation. Using this equation, we then construct a fully connected oscillator network and program it for robust image feature extraction engine. Using similar concepts, we finally demonstrate a full simulation of a stereo vision engine that utilizes the unique parallelism that is offered by the oscillator arrays to efficiently implement data and smoothness costs. A full scale foundry PDK and data driven SPICE simulation is described to understand the unique advantages that these networks offer for energy minimization-type problems.

Chapter 2

Forming in Binary Metal Oxide-based Resistive Switching Memory

2.1. Introduction to forming

Binary oxide-based resistive switching devices have shown a great potential as the next-generation non-volatile Resistive Random Access Memory (RRAM) elements [18]. These memory cells are based on changes in resistance thought to be due to the change in the stoichiometry of the oxide which typically becomes more conductive when deficient in oxygen [19]. It is now well established that the change of resistivity occurs only locally within a small diameter filament [7]. The switching between high and low resistance states (OFF and ON states, respectively) is thought as due to changes of the vertical extent of the filament [7],[20] or a formation of a lateral constriction within the filament [21]. Before a device can be switched, it has to undergo an initialization process referred to as the electroforming or simply the forming step. During this process, a voltage is applied to the device resulting in the creation of an metallic filament in the initially uniformly conducting device. Even though the forming establishes the active area of the device and determines many of its characteristics, very few reports have addressed the complex details of this process [22]-[25].

Recent experimental results on oxide-based devices indicate that the formation of a permanent conducting filament is not instantaneous. Instead, after application of a bias pulse, the device retains its high resistance for the time referred as the incubation time after which the resistance rapidly decreases. This process is associated with current localization, as the initially uniform current flow spontaneously localizes to a narrow filament. The incubation time is strongly dependent on applied voltage and temperature [26],[27]. It was also observed that the initial decrease of resistance is volatile i.e. if the voltage pulse is terminated soon after the drop of resistance (typically less than 1 μs), the device will return to its original highly resistive

unformed state. Only after the voltage is maintained for sufficiently long time, does the resistance change become permanent [24]. Similar phenomenon is well documented in chalcogenide-based phase change memory (PCM) where volatile resistance drop is referred to as threshold switching (with retention times of less than 1 μs) and is distinctly different from memory switching (with retention time of years) [28],[29].

The dynamics of the switching process is best described by a classical nucleation switching theory [31] applied in the past to interpret the results on chalcogenide [30] and oxide devices [31]-[34]. We note that the model is universal and has been applied to describe phase transitions of different physical nature, both involving the amorphous-crystalline transition in chalcogenides and those involving electronic changes, for example, Mott transitions in strongly correlated oxides [33]-[34]. However, it must be noted that the nucleation model does not specify the exact nature of the conductive phase. Thus, this phase could be either an electronic phase or a structural distortion. The model is briefly described here for the reader's convenience. It interprets the sudden drop in resistance as due to the nucleation of a conductive second phase inclusion in the functional layer under bias. This inclusion changes the free energy of the system, ΔG, by:

$$\Delta G = A\sigma + \mu\Omega + W_E \tag{2.1}$$

where the first term describes the energy of the interface between the two phases with A being the nucleus surface area and σ being the interface energy per unit area. The second term is a volumetric change, with μ being the free energy difference between the insulating and conducting phases per unit volume, and Ω is the nucleus volume. This term, if μ is negative, can lead to spontaneous nucleation of the conductive phase at elevated temperatures in absence of the

field. In the dielectric materials here, the nucleation model could describe the transition between the insulating disordered oxygen vacancies phase, and the small inclusion of conductive phase consisting of ordered vacancies. This inclusion can nucleate at any non-uniformity in the material, and will grow into a filament when the growth is thermodynamically favorable. The reduction of an energy barrier for a formation of a vacancy pair when two vacancies are close has been theoretically predicted for oxide materials, supporting that cluster and ordered vacancies filament formation is energetically favorable. Clustering was experimentally observed in TiO_2 single crystals with applied bias and ordered vacancy phases known as Magnéli phases had been observed in TiO_2-based RRAM devices [19]. Although the oxygen-vacancy model appears to be the leading interpretation of resistance switching in oxides, we do not want to exclude the possibility that the conducting phase is of different, as of yet unidentified, nature. One such possible phase transition would be an insulator-metal Mott transition in Ti and Ta suboxides. The nucleation of the inclusion of the critical size R_0 is delayed by the incubation time due to nucleation barrier (W_0 in Fig. 2.1) associated. The last term in Eq. (2.1) is the electrostatic energy associated with the conductive 'inclusion' (as shown in Figure 2.1 (b)) in the charged parallel plate capacitor. This term grows more negative with the applied electric field and causes the nucleation barrier to decrease (W_{eff} in Fig. 2.1). The barrier reduction lowers the incubation time. In parallel with decreasing the barrier, the electric field causes the size of the nucleus corresponding to the maximum of energy to shrink ($R_{eff} < R_0$). This has important consequences. For example, if the statistical fluctuation in the functional layer under bias produced a nucleus with size $R < R_{eff}$, such nucleus is expected to dissolve. Only the larger nuclei ($R > R_{eff}$) should be stable and grow in the presence of field. It is evident that the inclusion which, in the presence of field, is stable and starts to grow ($R_{eff} < R < R_0$), would become

subcritical and dissolve if the bias is removed. This, in the nucleation switching model, is the origin of the initial volatility of the ON state.

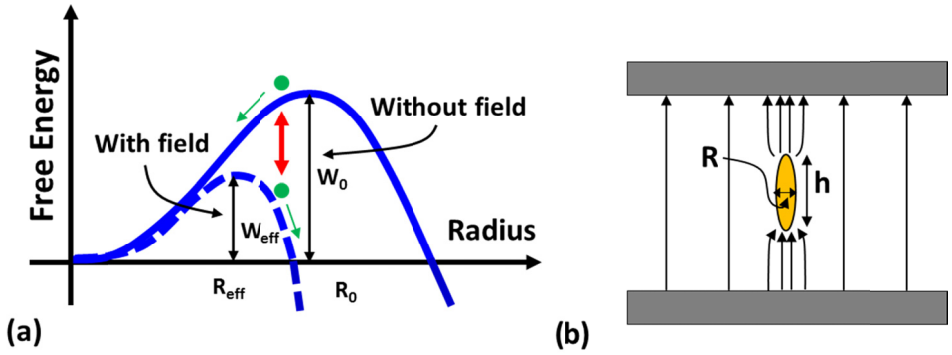


Fig. 2.1: (a) Free-energy curve showing the lowering of barrier in presence of field (from W_0 to W_{eff}). (b) A prolate ellipsoid inclusion assumed in the field-induced nucleation theory.

When a filament grows sufficiently large, it becomes non-volatile, i.e. stable after the external field is removed. The temporal dynamics of nucleation switching can be described as follows:

$$\tau = \tau_0 \exp\left(\frac{W_{eff}}{kT}\right) \tag{2.2}$$

Where, W_{eff} is a nucleation barrier, given by the maximum of the free energy ΔG in Eq. (2.1) along the nucleation trajectory. At high voltages, this expression can be approximated by the simple analytic expression:

$$\tau = \tau_0 \exp\left(\frac{W_0}{kT}\frac{V_0}{V}\right) \quad (V > V_0) \tag{2.3}$$

Here, V_0 is the voltage acceleration factor (a constant expressed in Volts and approximately independent of device voltage and temperature). The term W_0 represents the barrier height at zero applied voltage. Its value is determined by the volumetric (μ) and interfacial (σ) energy densities. T is the device temperature and τ_o is the attempt frequency.

In this chapter, we have measured the incubation time (τ) as a function of applied bias and stage temperature for TaO_x and TiO_x crossbar devices, extracted the activation energy for the incubation time as a function of applied voltage, and applied a nucleation model which was extended to include a self-heating effect in dielectric film to explain the data. We find that the model very well describes the complex functional dependence of $\tau(V, T)$. The extracted parameters of the nucleation switching model are compared with data available from other experiments. To understand the nature of the forming process better, we also conduct DC testing of RRAM devices in the presence of resistive ballast to understand the process of filamentation that precedes the forming process. The complex functional dependence of temperature and voltage are then utilized to estimate the temperature as the device undergoes filament formation process.

2.2. Experimental Techniques

The devices used in this study were TaO_x- and TiO_x-based metal-insulator-metal (MIM) crossbars designed specifically for high-speed pulse forming experiments. The vertical MIM stacks consisted of 15 nm Pt / 60 nm TaO_x / 5 nm Ta / 10 nm Pt and 15 nm Pt / 15 nm TiO_x /5 nm Ti/ 10 nm Pt, respectively. The top view of the device at two different magnifications is shown in Fig. 2.2 (a) and (b). The devices have been designed to be a matched 50 Ω air-coplanar waveguide in a Ground-Signal-Ground configuration. Moreover, 50 Ω characteristic impedance as seen from the source (pulse generator), the oscilloscope port, and the 2.92 mm cables (40 GHz bandwidth) ensured that the pulse consistency was maintained and there were no parasitic reflections. The switching dynamics was measured using a time domain transmissometry (TDT)

(in order to get better measurement accuracy than time domain reflectometry) method which allows for monitoring high-speed transients without interference of parasitic reflections [35].

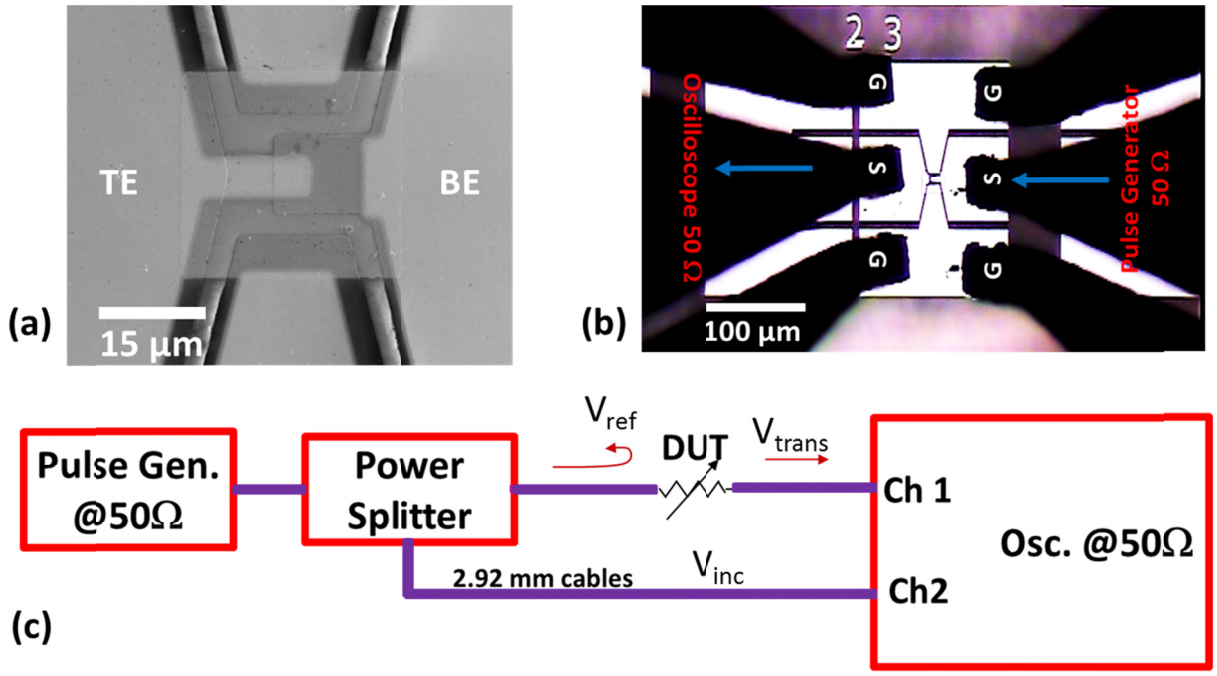


Fig. 2.2: (a) SEM image of the device. (b) Top view of the air-coplanar waveguide (ACPW) device with GSG RF probes with a pitch of 100 µm (23 in the image is the device ID). (c) Time Domain Transmissometry 50 Ω RF setup schematic.

Figure 2.2 (c) shows a schematic of the TDT setup. In TDT experiment, a pulse is launched from a pulse generator into one of the three ports of a resistive power splitter. Part of the signal is then propagated to one of the oscilloscope channels where it serves as the reference or incident signal. The third port of the power splitter is connected to the device through microwave probes and finally terminates at an oscilloscope port with 50 Ω impedance. Part of signal that is incident on the device is reflected back while the remaining part is transmitted to one of the channels of the oscilloscope. Thus, the oscilloscope can record both the incident and the transmitted waves. Knowing these, one can calculate the transmission coefficient (τ) and knowing the characteristic

impedance of the line (Z_0), one can extract the voltage across the device, the current flowing through the device and the resistance transient at the device. More details about the setup is explained in Appendix A.

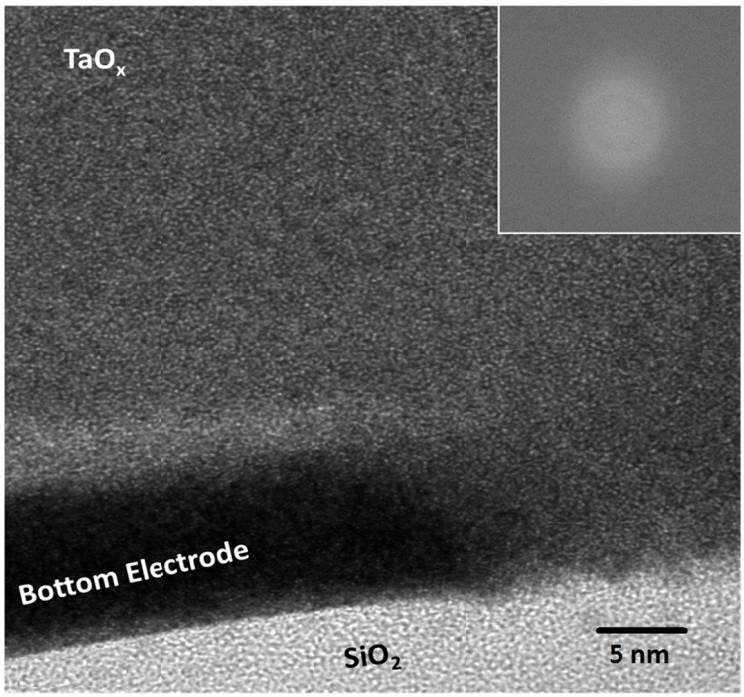


Figure 2.3: Bright-field TEM image of the sample with FFT pattern as an inset showing amorphous microstructure.

The functional films were deposited at room temperature and because of that should be amorphous. This was verified by a high-resolution TEM. Fig. 2.3 shows a HR TEM image of the structure while the inset shows a Fast Fourier Transform pattern of the functional layer image. The lack of any discernible directional intensity changes indicates the amorphous nature of the film.

2.3. DC Forming

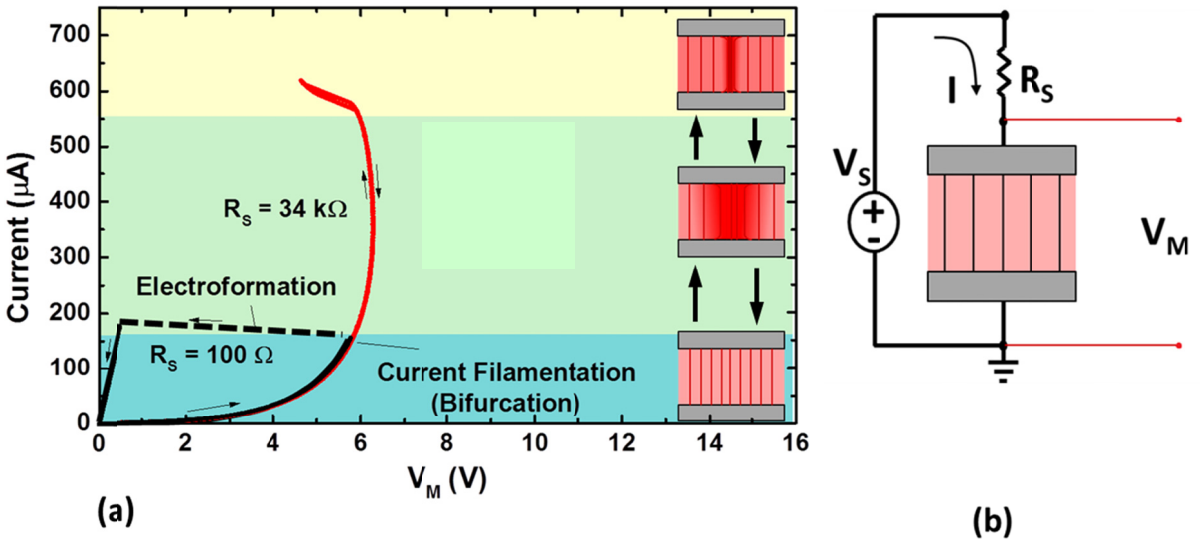


Fig. 2.4: Observation of negative differential resistance (NDR) in Ta_2O_{5-x}. (a) Electroformation with a 'snap' observed in samples with 100 Ω series resistance (black trace). This is prevented by using a large source resistance (red trace). Stable and reversible current filamentation causes the device *I-V* to show negative differential resistance (NDR), post-bifurcation. The three color-zones indicate uniform conduction (blue), thermal filamentation (green) and filament collapse (yellow). Inset on the right shows a schematic of the change in conducting area as the device is driven to filamentation. (b) Circuit schematic shows the experimental setup with V_M representing the device voltage.

Figure 2.4 shows four quasi-DC *I-V* characteristics obtained from a single crossbar-type Pt/Ta_2O_5/Pt device. The two curves were obtained using different values of series resistor, R_s (the circuit schematic is shown in Fig. 2.4(b)). The measurements were made by sweeping the source voltage while measuring the voltage across the device (V_m in the figure). Because increasing the series resistance in the circuit decreases the slope of the load-line (which is $-1/R_S$), a higher voltage from the source is needed to reach the same *I-V* point of the device. The black line marks the *I-V* obtained with $R_S = 100$ Ω. Starting at low voltages, the curve increases super-linearly and

then at 5.75 V snaps back (electroforms) leaving the device in the low resistance ON state. The decrease of resistance is permanent although the device can be switched between ON and OFF states repeatedly. The current value after the snap-back for the R_S = 100 Ω case, reaches about 180 μA due to the current compliance (with a likely overshoot during snap-back). The electroformation event is effectively instantaneous on the time scale of the source meter response with no intermediate states recorded. The red *I-V* curve which was collected beforehand, does not show a permanent change and can be retraced for decreasing source voltage. The red curve corresponds to R_S = 34 kΩ (standard ohmic resistors of 33.7 kΩ chosen for experimental convenience; ~300 Ω added extracted by calibrating for the resistance of pad and cross-bar traces). At low voltages, it follows the same path as the one for R_S = 100 Ω but extends to higher current values without forming. At about 6 V, the *I-V* trace gradually bends back forming a part of an S-type curve characteristic of current-controlled negative differential resistance (CC-NDR). The presence of CC-NDR usually indicates the presence of an instability that can lead to the spontaneous formation of localized high current density filaments within the device [6]. This phenomenon is characteristic of nonlinear dynamic systems and is often referred to as *bifurcation*. This also implies that the system will 'snap' to a low resistance state (in this case, the current runaway will result in permanent change of resistance i.e. forming) whenever the total differential resistance becomes negative (equation 2.4). By increasing the source resistance (R_S), one can limit the range of voltages corresponding to negative total differential resistance (dV_S/dI) and increase the range accessible to testing. This enables the device to support a stable filament and a corresponding negative differential resistance (dV_D/dI negative). At sufficiently large R_S, the total differential resistance becomes positive (due to dV_{RS}/dI positive) and the snap can be prevented altogether.

$$V_S = V_D + V_{R_S}$$
$$\frac{dV_S}{dI} = \frac{dV_D}{dI} + \frac{dV_{R_S}}{dI} \tag{2.4}$$

In other words, if the dV_{RS}/dI is not large enough (due to small R_S), the device will filament instantaneously, as it goes through a snap. The different types of *I-V* in Figure 1 are, therefore, fully consistent with each other and can be explained without invoking any changes of the device structure. In other words, the R_S enables us to stabilize the pre-forming *I-V* by stabilizing the electronic filament and defines the *I-V* path of the device post-bifurcation. This makes the post-bifurcation *I-V* distinct in each case.

While the data in Fig. 2.4 were obtained on Ta_2O_{5-x}-based devices, we have observed similar behavior in other switching oxides. Figure 2.5 and 2.6 present the results obtained on $Pt/TiO_2/Pt$ devices. Figure 2.5(a) shows a typical forming *I-V* with the source resistance of 700 Ω. Initially highly resistive device exhibits forming at 4.7 V with the device formed to high resistance OFF state. The dashed line in the figure corresponds to the load line. The electroformation of TiO_2 devices frequently leads to changes in the top electrode morphology with several groups reporting craters forming at the location of the filament [36],[37]. The changes in the device shown in (a) were imaged by Scanning Electron Microscopy (SEM) using a through-the-lens detector (Figure 2.5(b) and (c)). The low magnification image shows a horizontal stripe corresponding the to bottom electrode and much brighter the vertical top electrode stripe. The most visible feature is the dark dot with a bright edge located in the lower left corner of the active area. The dot is surrounded by a bright "halo". The high magnification image of this feature is shown in Figure 2.5(c). It is quite apparent that the dark dot corresponds to the area

where platinum film delaminated exposing the functional layer underneath. This could be due to either local melting of the metal or solid state diffusion at elevated temperatures. It is widely accepted that such features form at the location of the filament. The "halo" shows a contrast characteristic of polycrystalline grains in a metal. The center portion has grain size of about 100 nm (as measured by an SEM) gradually tapering down to 10 nm toward the edge. We interpret this contrast as the result of Joule heating in a small diameter filament and the resulting grain growth in the Pt electrode. By limiting the current overshoot during electroformation we can eliminate the delamination of Pt but all of out electroformed TiO_2 devices show signs of the halo.

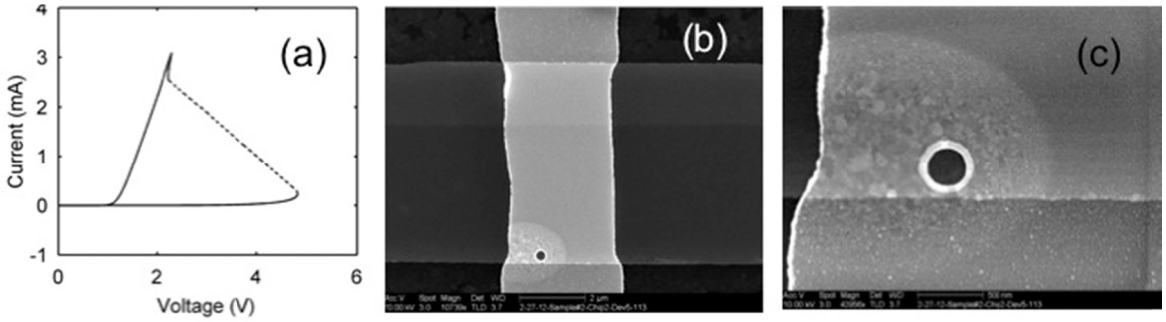


Fig. 2.5: (a) Electroforming *I-V* characteristics of the $Pt/TiO_2/Pt$ 5 μm cross-bar device with the source resistance of 700 Ω. (b) SEM image using through-lens detector of the device after electroforming and (c) a high magnification image of the lower left corner of the same device.

Figure 2.6 shows the I-V characteristics of a similar TiO_2 crossbar obtained with R_S = 4.7 kΩ. The overall I-V shape is similar to that obtained on Ta_2O_{5-x} samples with clear CC-NDR behavior and a distinct change of slope in the upper part of the curve at 2 V. In the later part to this report, we interpret such change as due to two different mechanisms contributing to the increase of conductivity with applied bias, namely thermal and electronic. SEM micrograph in

the inset shows the contrast of the top electrode after the test. The contrast across the device is uniform with the exception of a bright circular area with the diameter of about 0.7 μm located near the left edge of the device. The feature has all characteristics of the halo shown in Fig. 2.5 and is consistent with current constriction to a small diameter filament associated with NDR. The above observations indicate that the voltage and temperature non-linearity is characteristic of most resistive switching oxides.

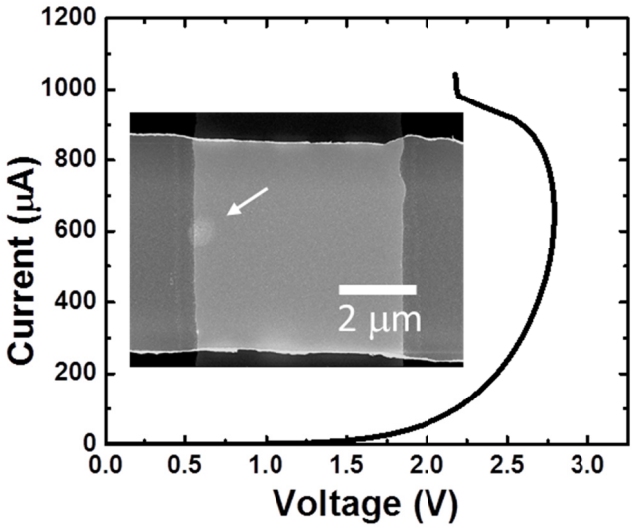


Fig. 2.6: Negative differential resistance observed in 20 nm thick TiO_{2-x} cross-bar with R_S = 700 Ω. SEM image (inset) shows the evidence of the current constriction in the CC-NDR voltage range.

2.4. Reversibility of Forming Process

Additional insight into the dynamics of the electroformation process and current filamentation was obtained using time domain transmissometry (TDT) (detailed description in the Appendix A). It must be noted that TDT is a technique that enables delivery of high-fidelity voltage pulses

to the device and reading off the transmitted voltage pulses and current through the device without any parasitic overshoots. For the pulse forming experiment, we use constant voltage pulses applied across the device and a 100 Ω resistance connected in series. TDT allows for monitoring the change of voltage across the device and current as a function of time. Identical devices can be formed at a wide range of voltages, with lower voltages requiring longer incubation time before the device undergoes the electroformation [25].

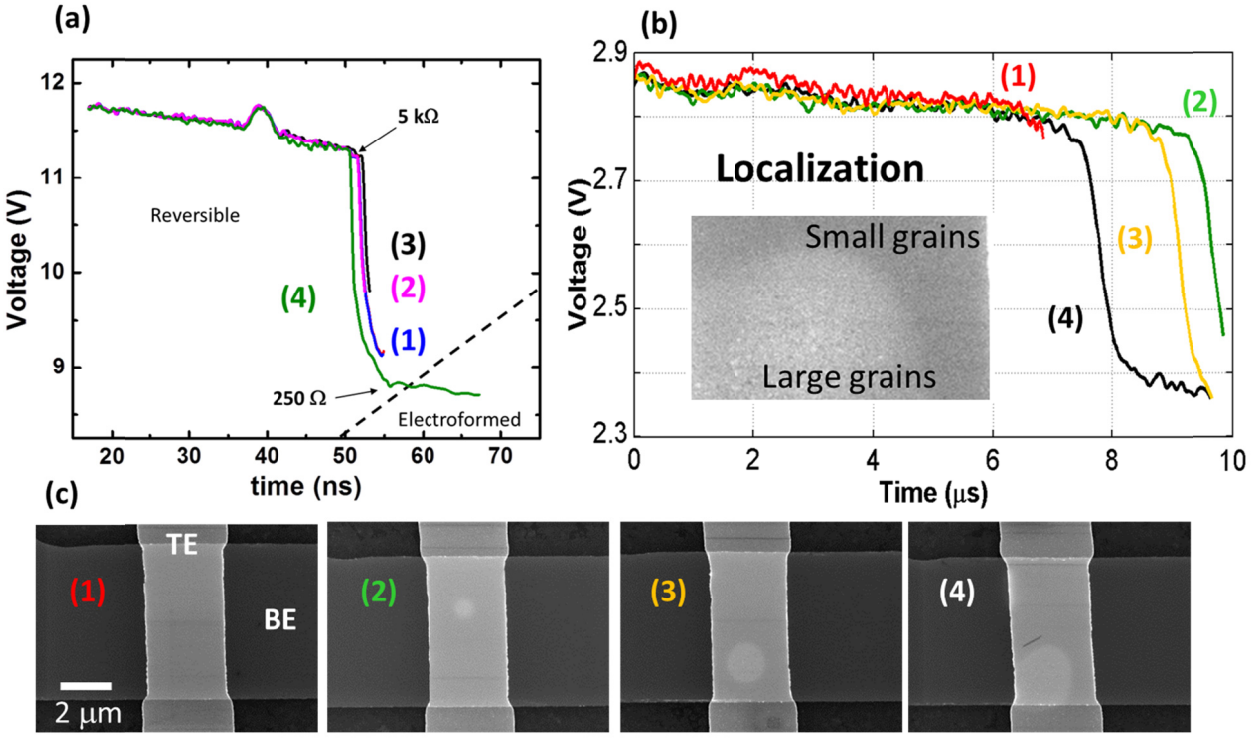


Fig. 2.7: Pulsed electroforming experiments. (a) Voltage dynamics obtained by pulsing the device repeatedly (1 through 4) with different pulse durations for the same Ta_2O_{5-x} device. Large resistance change observed without any permanent change until pulse 4 (b) Similar pulse reversibility experiments on TiO_{2-x} show morphological changes (shown in (c)) due to higher power dissipation compared to Ta_2O_{5-x}.

Fig. 2.7(a) shows the voltage across the device as a function of time. Each curve corresponds to the trace during a single voltage pulse applied to the same Ta_2O_{5-x} device. In this experiment, each pulse had the same amplitude but slightly different duration, allowing for the interruption of the process at different stages, a few nanoseconds prior to the completion of forming. As the pulses applied brought about reversible changes (explained below), we were able to use the same device for the entire experiment. The oscilloscope trace started at t = 0 with a fast rise if the voltage due to the leading edge of the applied pulse (not shown). This was followed by a gradual decrease of voltage (initial negative slope) which is associated with decrease of device resistance due to Joule heating. It must be noted here that, while the device has high resistance (~1 MΩ) at low biases, it shows a low resistance (~ 5 kΩ) at high biases (~ 11 V across the device), following a strong voltage non-linearity ($R \alpha V^{-2}$), as shown in Figure 2.7(b). This part of the transient can be accurately simulated using known materials parameters and assuming uniform current flow, [discussed in Appendix A for Ta_2O_{5-x}]. The bump at 38 ns is due to a parasitic pulse reflection in the system while the rapid drop between 45 and 55 ns corresponds to the beginning of the electroformation process (and we arrest the pulse at different points, in this range). We assert this based on the magnitude of the resistance change of the device. For example, the device resistance during pulse 4 (green curve in Fig. 2.7(a)) at the onset of the rapid drop is 5 kΩ and the resistance value at the point of pulse termination is 250 Ω. This change is too big to be explained by thermal effects in the uniform conduction regime. Moreover, while there was no permanent change of the device resistance after pulses 1-3 terminating before the completion of the process, the device was formed after the pulse 4. The resistance did not recover and remained at the 250 Ω level after the pulse. The conclusion here is that the rapid drop in resistance in Fig. 2.7(a) does correspond to the electroformation process. The initial part

of the sharp reduction of resistance is volatile and therefore has to be electronic in nature rather than one involving atomic motion.

The results of the similar experiment performed on TiO_{2-x} devices are presented in Figure 2.7(b). Each voltage transient and the corresponding SEM image were collected on a different but nominally identical device and each device was exposed to only one pulse. Different devices were used in this experiment for the ease of SEM imaging, while making sure that the incubation time before resistance change (reversible forming, akin to Ta_2O_{5-x}) is the same for all the devices under test. The images show the cross bar-type devices with the light grey vertical strip corresponding to the top electrode and the horizontal darker grey strip corresponding to the bottom electrode. The active area of the devices is the rectangle at the intersection of the electrodes. As in Figure 2.7(a), the pulses were interrupted at various stages of the electroformation process. Traces 1-3 correspond to devices that retained their original resistance after the pulse while trace 4 corresponds to the device on which the electroformation was completed with the permanent drop of resistance. The red trace (1) corresponds to the process interrupted during the uniform current flow stage. The SEM image obtained after this single pulse shows perfectly uniform contrast over the active part of the device. SEM image obtained on the device which experienced the first part of rapid decrease of voltage (trace 2 in Figure 2.7(b) and image 2 in (c)) shows a characteristic halo with diameter of 1.5 μm. The size of the halo on the device which experienced larger resistance decrease (trace 3) increased to 2 μm eventually attaining diameter of 3 μm on the device with permanent resistance change. Such morphological changes on the top electrode are not seen on Ta_2O_{5-x} devices apparently as a result of lower temperature excursions. This observation is a direct evidence of current filamentation

occurring before any permanent changes (such as vacancy accumulation) take place in the memristive devices. This instability is reversible and electronic in nature. Only at its later stages and after the core of the filament reaches high temperatures due to high current density, do the physical changes in the device structure take place. Thus, this experiment relates the electronic filamentation behaviour in the measured oxides to the transient threshold switching seen in chalcogenides (which involves volatile reduction in device resistance (lasting ns to µs), when the device is exposed to short pulses).

Figure 2.8 (a) shows the experimental set up for measuring the quasi-DC I-V characteristics of 60 nm TaO_x devices at room temperature. The devices were connected in series with the resistor limiting the current flow with the voltage used in Fig. 2.8 (b) and (c) measured across the device terminals. A clear thermal footprint in form of thermally grown grains of Pt top electrode representing a localized heat affected zone could be observed in these devices (not shown) and has been discussed earlier.

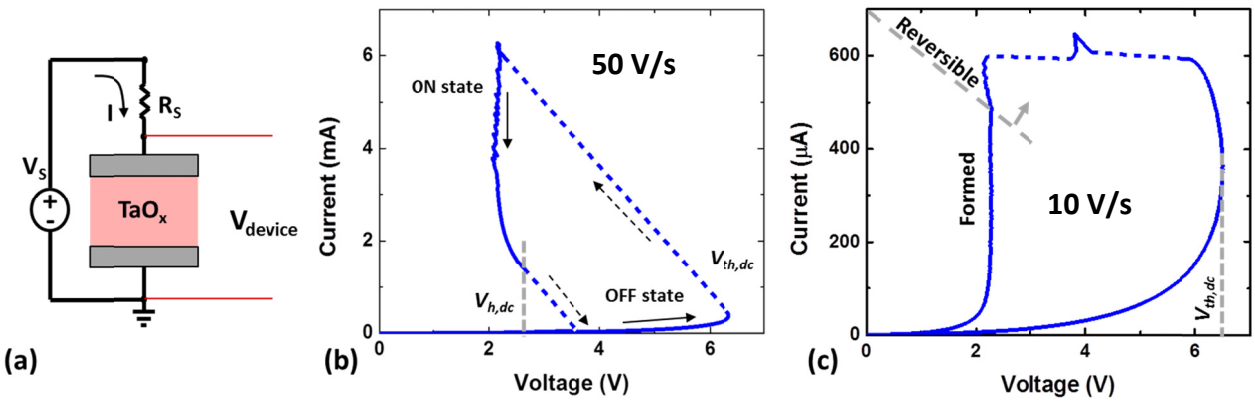


Fig. 2.8: (a) Setup used for DC testing of NDR and threshold switching. (b) DC sweep with a ramp rate of 50 V/s exhibiting completely reversible threshold switching in I_{device}-V_{device} characteristics in a 60 nm TaO_x device. (c) DC sweep with a ramp rate of 10 V/s showing permanent change in the device characteristics for the same device.

The nucleation model predicts that a prolate ellipsoid shape will produce the lowest energy barrier and will be the fastest to nucleate. The electric field will be locally enhanced at the tips of the spheroid what in turn should provide a positive feed-back for the nucleus growth. This will lead to fast elongation of the nucleus and shunting of the field across the dielectric layer. This behavior results in S-shaped I-V characteristic part of which is observed in Figure 2.8 (b). The device is initially in high resistance (OFF) state with current increasing super-linearly as the function of voltage. At about 6.3 V corresponding to threshold voltage, the device enters into the negative differential resistance portion of the characteristics due to filament nucleation. The I-V rapidly evolves along the load line (upper dashed line in Fig. 2.8 (b) corresponding to $R_S \sim 700$ Ω) to stabilize in a low resistance state at much lower voltage (due to voltage division). The voltage ramp rate in Fig. 2.8 (b) was 50 V/s and was fast enough that when the device arrived at the holding voltage (marked $V_{h,DC}$ in the figure) it was still in the volatile stage of the filament. At this point, the I-V snapped again along the load line to reach the OFF state at higher voltage. There were no permanent changes to the device and the procedure has been repeated many times. Figure 2.8 (c) shows the I-V characteristics of identical device with a slower ramp-rate (10 V/s). The nucleus has enough time to reach the critical size and the device permanently changes or 'locks-on' to a low resistance state, as indeed observed in Fig. 2.8 (c). One should note that the series resistance used in Fig. 2.8 (c) is much higher than the one used in 2.8 (b).

Whether the NDR-type characteristics are purely a result of heating or supplemented by field dependent conductivity is assessed in the following sections.

2.5. Voltage Dependent Transient Measurements

The incubation times in TiO_x and TaO_x devices were measured by applying a series of rectangular pulses to the device under test and monitoring voltage across the device as a function of time. The duration of pulses was kept constant while their amplitude was gradually increased. For low pulse amplitudes, the voltage across the device slightly drops with time due to decrease of device resistance associated with uniform Joule heating. At certain pulse amplitude, the device undergoes an abrupt reduction in resistance during the pulse, an event that is reflected in a rapid drop of the voltage across the device. This time is defined as the incubation time. The pulse width was determined by the temporal resolution of the oscilloscope. For example, pulse widths for high voltages (> 10 V) were kept constant at 100 ns and the amplitude was changed. After this, the incubation time (which is less than the pulse width) was recorded. Thus, the applied pulse widths were such that the incubation times were contained within the scope trace, without any loss of temporal resolution. The formed state was volatile similarly as happens in chalcogenide-based devices. Each device was used only once as the formation causes permanent change in characteristics.

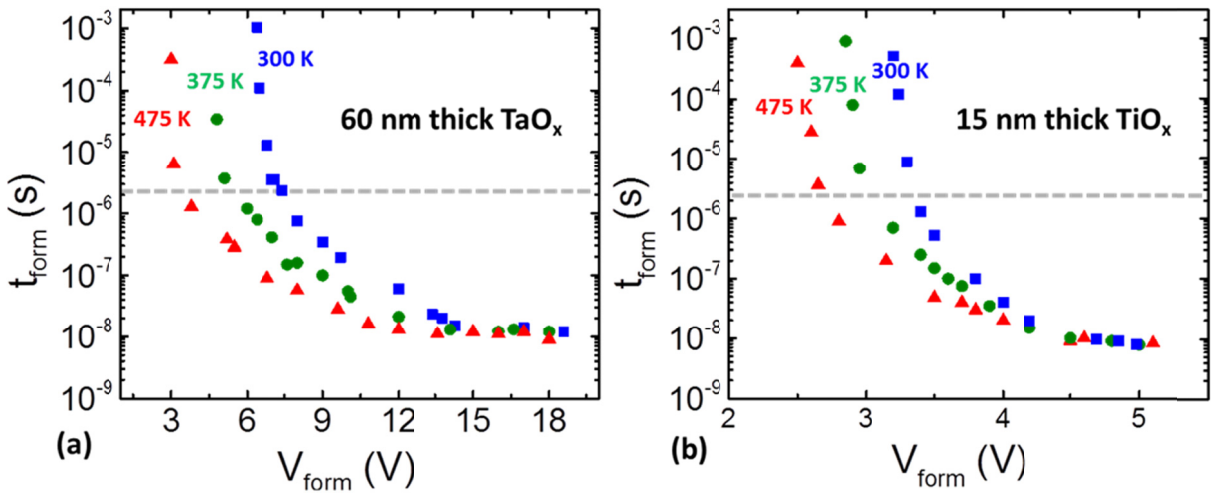


Fig. 2.9: Incubation time as a function of applied bias and stage temperature for (a) TaO_x and (b) TiO_x based crossbar devices. The horizontal dashed line denotes the value of the thermal time constant of devices with uniform current flow.

This experiment was repeated for pulses of different widths and various stage temperatures. Figures 2.9 (a) and (b) show plots of incubation time as a function of voltage across the device and stage temperature. Each data point represents the average of 5 devices that were subjected to a pulse of the same width and amplitude at the same stage temperature. Both materials exhibit three ranges with different functional dependence of incubation time on applied voltage. At high voltages (> 14 V for TaO_x and > 4.5 V for TiO_x), the incubation time is independent of voltage and stage temperature. At intermediate bias, the τ decreases with increasing bias and stage temperature. It should be noted here, that the stage temperature is not equal to device temperature as the devices experience considerable Joule heating.

2.6. Temperature, Filament-size and Activation Energy Estimation

Two procedures allowing for the estimates of the device temperatures have been described in our earlier publications [24],[26] and Appendix A. The first is purely experimental relying on acquiring the I-V characteristics at different stage temperatures using voltage pulses much shorter than the thermal time constant of the device. The accumulated data can be used as a look-up table to extract the temperature of the device with the uniform current flow. The second approach relies on finite element modeling of the temperature increase due to measured power dissipation during the voltage pulse.

For transient measurements, as we are interested in extracting the temperature as the device uniformly heats up (before filament formation takes place), the first method can be used to plot the device temperature as a function of time for every data point in Fig. 2.9 (a) and (b). It was observed that the temperature increases monotonically during the pulse and at the end of the incubation reaches 450 K and 600 K for TaO_x (60 nm) and TiO_x (15 nm) devices at the stage temperature of 300 K and highest voltages. Both types of devices reach the maximum of 700 K at the stage temperature of 475 K. At low voltages for which the incubation time exceeds the thermal time constant of our devices [24].[26] (marked with a horizontal dashed lines in Fig. 2.9 (a) and (b)), the dependence on the bias becomes very steep.

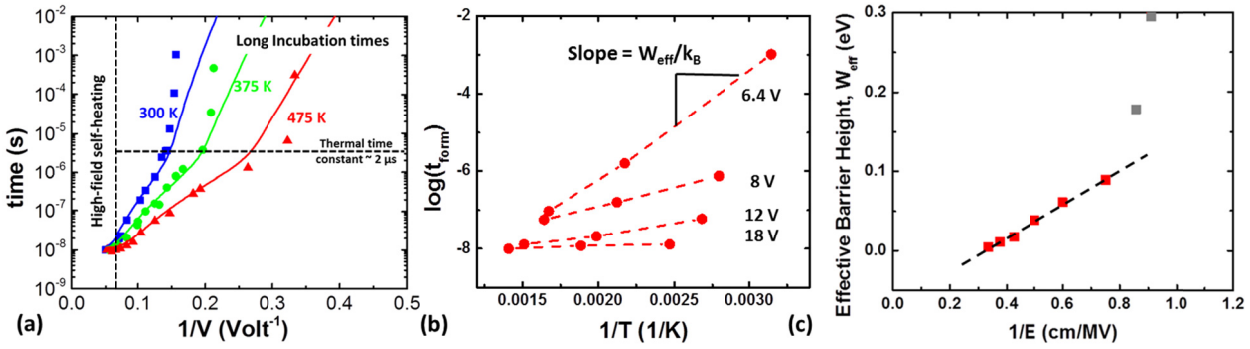


Fig. 2.10: (a) Incubation time as a function of *1/V* for different stage temperatures. The continuous lines are the results of the nucleation switching model extended to include self-heating effect in the dielectric film. Areas above red dashed line and to the left of blue dashed line delineate regimes of forming where self-heating is important. (b) The Arrhenius plots of $1/\tau$ versus *1/T*. The slopes correspond to effective activation energies at different fields. (c) Plot of activation energy for nucleation as a function of 1/electric field.

Figure 2.10 (a) shows the data from Figure 2.9 (a) plotted as a function of *1/V*. It is quite clear that in the intermediate voltage range, the incubation time changes as *exp(1/V)* in agreement with Eq. (2.3). The slopes of linearly dependent segments are different for different stage temperatures and converge for bias exceeding 14 V. The figure also clearly shows that at low voltages, the dependence on electric field becomes much faster than that predicted by Eq. (2.3). The continuous lines in the figure are the results of the modified nucleation switching model discussed below. Fig. 2.10 (b) presents the data obtained in the intermediate voltage range as a function of *1/T*. The values of the temperature *T* were actual device temperatures at the end of the incubation including the self-heating estimated using the experimental procedure outlined above. In Figure 2.10 (b) for stage temperatures 300 K, 375K, 475K the corresponding actual temperatures reached 325 K, 425 K, 573 K at 6.4V, 363 K, 450 K, 588 K at 8V, 375 K, 485 K, 635 K at 12V, 380 K, 525 K, 683 K at 18V for 60 nm TaO_x devices. The lines at constant bias are standard Arrhenius plots with slopes corresponding to the activation energy (energy of the

barrier in Fig. 2.1 (a)). It is apparent that the incubation time shows thermally activated behavior with barrier height changing with bias. Figure 2.10 (c) shows the extracted effective barrier height as a function of (1/electric field) and illustrates a strong dependence of nucleation barrier on the electric field. In the intermediate bias regime the dependence is linear in good agreement with Eq. (2.3) which gives the electric field dependent effective barrier height W_{eff} as W_0V_0/V.

It is apparent from Fig. 2.10(a) that, with the exception of the high field region, the incubation time depends strongly on both the applied voltage and the device temperature. In the absence of Joule heating and assuming that W_0 and V_0 do not depend on temperature and field, the ratios of slopes of $ln(\tau)$ *vs.* $1/V$ would correspond to inverse of the stage temperature (T_{stage}) ratios, which is not the case. To illustrate the origin of this discrepancy, let us consider the intermediate voltage regime at 475 K stage temperature where $ln(\tau)$ depends linearly on $1/V$. Decreasing the voltage between two arbitrary values should increase the height of the barrier and, therefore, increase the incubation time. If we consider the contribution from self-heating, the barrier change would remain the same but the device temperature in the denominator of eq. (2.3) would decrease due to lower dissipated power. It is this change of device temperature that changes the slopes from $1/T_{stage}$ dependence expected from Eq. (2.3) in Fig. 2.10(a).

The data in Fig. 2.10(a) can be explained fully by Eq. (2.2) if the voltage dependent internal device temperature instead of stage temperature is used. To calculate the temperature $T(V)$ at the nucleation event, we solve the thermal heat balance differential equation in a dielectric film under constant field E applied during incubation time τ:

$$c_V \frac{\delta T}{\delta t} = J\,E - \lambda\left(T - T_{\text{stage}}\right) \tag{2.5}$$

The nucleation event is assumed instantaneous. The temperature T of the nucleus is the same as the temperature of the surrounding dielectric which is higher than the stage temperature T_{stage} due to current J flowing in the film *before* the nucleation event. The equation (2.5) represents balance of the power components: heat absorption rate in the device (product of temperature increase rate and specific heat c_v) with Joule heating power and heat loss rate (product of temperature increase and thermal resistance characterized by the cooling coefficient λ). The thermal parameters c_v= 2.3x10^6 J/K/m^3 and λ= 2.2 x10^{12} J/s/m^3/K were calibrated by fitting the experimental thermal transient obtained for V = 3.76 V at 475 K stage temperature.

The Joule heating is generated due to the finite current density *J* flowing in the film. To model the current density *J* we fitted the pulsed I-V characteristics with the following function [38]:

$$I = JA_d = A_d E\sigma_o \exp\left(-\frac{\Delta E_A}{kT} + \frac{E}{a(T)kT}\right) \tag{2.6}$$

which is a generic formula used to describe field and temperature activated transport, which we attribute to the field and temperature assisted conduction through existing oxygen vacancies in the film. A_d is the device area, ΔE_A thermal activation energy, E is electric field, and $a(T)$ describes the field dependence of conductance. We found $a(T)$ to be weakly temperature dependent: $a(T)$= 1.25x10^9 V/m/J – (T-300) K x 2.5 x10^6 V/m/J/K**.** ΔE_A= 0.249 eV and the film conductivity σ_0 = 20.9 S/m. For TiO_x, while the temperature dependence is the same, the parameters are numerically different.

We solved above equations self-consistently to calculate the electroforming time. We used parameters W_0 = 0.65 eV, R_0 = 3 nm and the high aspect ratio approximation for the oblate nucleus to calculate the effective barrier W_{eff}. Note that W_0 and R_0 were obtained by fitting the

model using Eq. (2.4) to the electroforming data in the intermediate voltage regime at 300 K stage temperature. The prolate nucleus radius is found to shrink from its zero-field value to $R_{eff} \sim$ 1 nm. This is in good agreement with the experimental results in our earlier work [24],[26]. The attempt frequency $\tau_0 \sim 10$ ns is read off the data in the high field saturation regime in Figure 2.10 (a).

The nucleation switching model with self-heating describes all experimental results in three different regimes and at 3 different ambient temperatures (Fig. 2.10 (a)). At high fields, all three curves level off as the barrier height approaches zero. At intermediate values of bias, the model replicates *1/V* dependence with contributions from both the field dependent barrier height and self-hearing. At low voltages, the incubation times exceed the thermal time constant in the film, and the device has enough time to self-heat. In this regime *1/V* voltage dependence is therefore strongly enhanced by the field dependence of the internal device temperature.

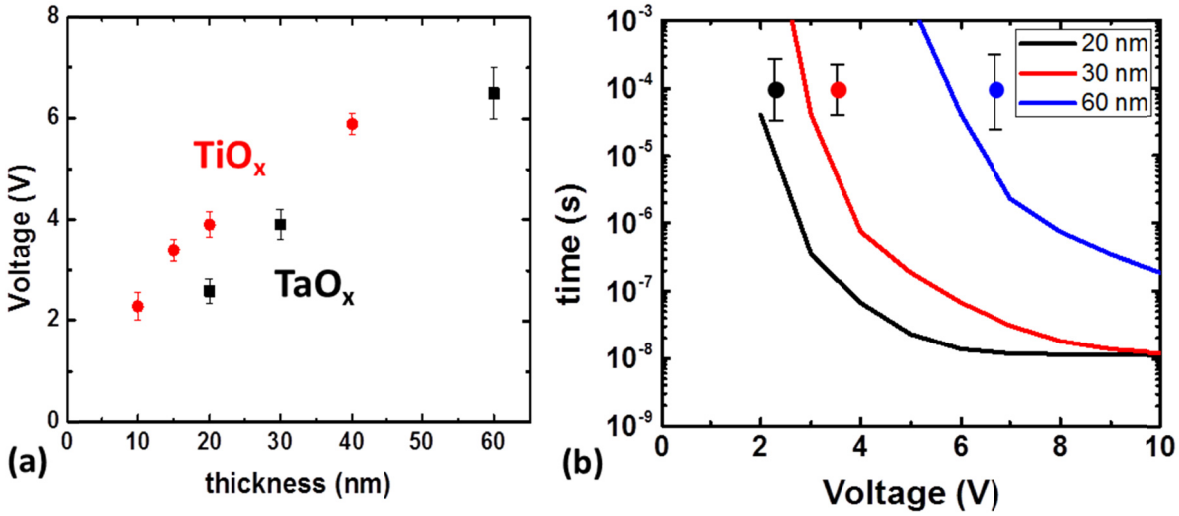


Fig. 2.11: (a) Linear increase in the DC /forming voltage as a function of oxide thickness indicating a field-driven forming process. Forming time as a function of applied bias at 300 K stage temperature for 60, 30, and 20 nm TaO_x films (lines: model, circles: data) in (b). The error bars show distribution over 10 devices.

The nucleation switching model with self-heating was also used to explain another characteristic of the devices, namely the change of the threshold voltage as a function of oxide layer thickness. The experimental data for the threshold voltage values obtained by a quasi-DC sweep for TaO_x- and TiO_x-based devices are shown in Fig. 2.11 (a). The voltage scales linearly with the functional layer thickness indicating that a unique field inside the functional layer is needed to form the device. Thus, while there may be some interface effects associated with the forming voltage, the electric field has the most significant effect on the forming. Figure 2.11 (b) shows the calculated incubation times as a function of applied voltage at three values of the oxide layer thickness: 60, 30, and 20 nm. The dots correspond to experimental values in good agreement with the calculated ones. Fig. 2.11 (a) and (b) illustrates that the nucleation switching model with self-heating reproduces the scaling of forming voltage with the film thickness within the data error bars.

However, as the devices undergo NDR, the temperature extraction methodology used in the transient experiments fails. This is because the device no longer conducts uniformly (Appendix A) and needs a self-consistent methodology to estimate the local temperature during the forming process.

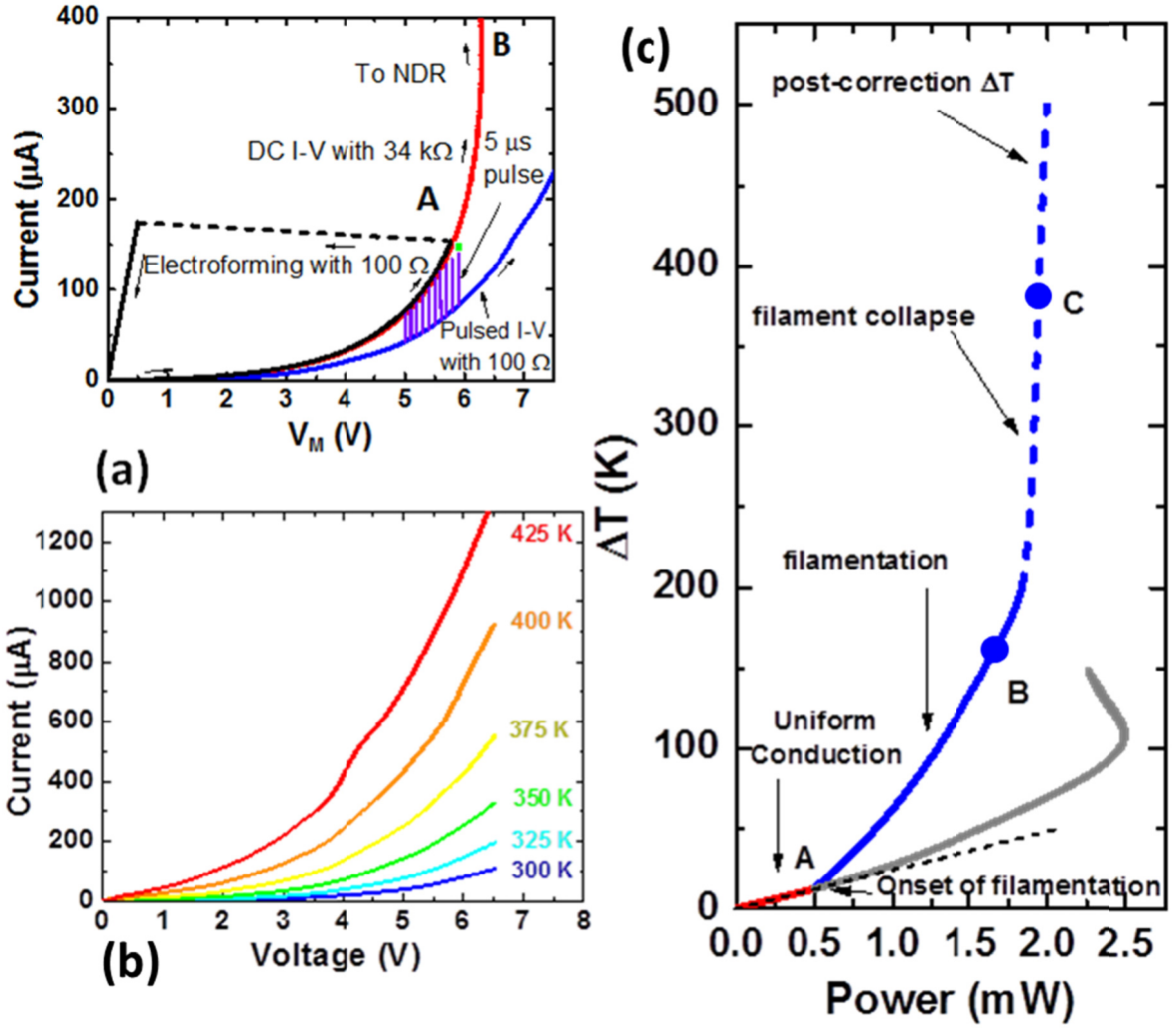


Figure 2.12: Temperature and voltage non-linearities as the origin of filamentation. **(a)** Adiabatic pulsed *I-V* shown in blue taken when the device did not undergo any self-heating. The violet *I-V* dynamic curves connect the initial and the final I vs. T and V vs. T points at 1 ns and 5 μs. Electroformation occurs in the 5 μs long pulses at the same V, I combination (green spot) as the DC case (black) **(b)** Adiabatic pulsed *I-V* curve taken as a function of temperature displaying non-linearities in both temperature and voltage. (c) ΔT vs. P obtained by applying thermometry to the DC *I-V* taken with a 34 kΩ resistor in series (red, gray curve). Blue curve shows the corrected local temperature. Non-uniform conduction sets in at point A due to thermal non-linearity. Region B represents the thermal onset of filamentation followed by a sharp collapse of the filament due to field/current density dependent instability (Region C).

We will now try to experimentally assess the role of Joule heating in NDR and filament formation. Figure 2.12(a) shows three *I-V* characteristics: the black and red traces are DC *I-V* with R_S = 100 Ω and 34 kΩ, respectively, as shown in Fig. 2.4(a). Since the sweep rate is low, the devices reached thermal steady state at every point (i.e. they heated up). The points on the blue curve were obtained in a pulsed experiment where the current and voltage across the device was measured 1 ns after the pulse leading edge. Since the thermal time constant of our devices was experimentally measured in our earlier work [24][26] to be about 2.5 μs, one can assume that the temperature of the device remained at the stage temperature. It is apparent that excluding the self-heating (blue curve) extended the range of voltages that could be reached without device forming and reduced the current at any given voltage eliminating the CC-NDR. In other words, a device not undergoing self-heating (pulsed *I-V*), would have a much higher breakdown voltage and current compared to ones that undergo DC forming. Next, we try to experimentally establish the relation between the pulsed *I-V* and the DC *I-V*. In addition to three *I-V* curves, Figure 2.12(a) also shows the results of another series of pulsing experiments represented by almost vertical violet lines. The pulse duration in this experiment was always 5 μs and the violet line corresponds to voltage and current evolution across this 5 μs pulse due to Joule heating. During the pulse, the temperature of the device evolves (Appendix A), approaching the steady state at 5 μs. The coincidence of the end points in this experiment and the black trace (DC with the same load resistor) confirm that filamentation and CC-NDR in the low voltage/current range is purely a thermal phenomenon. In other words NDR appears because the device becomes more conductive as it carries more current because of self-heating creating a positive-feedback. Accordingly, the origin of NDR in this part of the I-V characteristics appears to be thermal.

While we have argued that CC-NDR indicates filamentation, it is not clear, at which specific voltage the filament forms or the temperature the local electronic filament reaches before breakdown. We can estimate this by analysis of the dependence of device temperature on dissipated power. For the sake of discussion, let us assume that the current flow is uniform for the entire *I-V* curve obtained with $R_S = 34\ \text{k}\Omega$ (Fig. 2.4(a)). The device temperature at every voltage was extracted from the pulsed *I-V* measurement calibration as a function of stage temperature, shown in Figure 2.12(b). This data maps the non-linear dependence of current on voltage and temperature and allows the device resistance itself to serve as a thermometer. It must be noted that the pulsed-I-V measurements were taken up to 18 V (not shown here) as the device does not electroform for 5 ns pulses [41]. In Figure 2.12(c), we use this thermometry technique to plot the rise in steady-state temperature due to Joule heating during the DC voltage sweep as a function of power dissipated in the device (red trace). The expected rise in temperature should depend linearly on power:

$$\Delta T = R_{th} P \tag{2.7}$$

where, ΔT is the rise in temperature in Kelvin, R_{th} is the thermal resistance seen by the source of heat, and P is the power dissipated at the heat source. The thermal resistance, in turn, can be expressed as:

$$R_{th} = \frac{1}{k_{th}} \frac{t}{A} \tag{2.8}$$

where, k_{th} is the thermal conductivity of the materials leading to the thermal ground, t is the distance the heat travels to thermal ground and A is the area getting heated up (in steady state) and depends on the filament size. At low voltages, the current flow is uniform and A corresponds to the area of the device. This gives the constant slope of 0.025 K/μW in Fig. 5(c)). This slope corresponds to an R_{th} which is consistent with the thermal resistance felt by a uniformly conducting 5 μm square device deposited on 1 μm thick SiO_2. At 600 μW, the $\Delta T(P)$ slope increases indicating the onset of current constriction (point A in the DC *I-V* curves). The subsequent section of the curve is gray (at powers above 600 μW) to indicate that the temperature calibration is no longer correct when filamentation sets in; the gray section can only be taken to be a lower bound on the device temperature. As the bias increases further past the onset of filamentation, the slope continues to increase indicating gradual reduction of the filament diameter. The deviation from the initial slope up to 2 mW is attributed to thermally-induced CC-NDR. Following this, a steep change of slope occurs at higher powers indicating that the mechanism of the non-linearity is completely different from the one that occurs at point A. This region corresponds to the collapse of the filament resulting into a very localized current flow, which we refer to as the electronic filamentation (making a clear distinction in the non-linearities).

After filamentation onset, the temperature reached is a strong function of filament diameter, with greater current localization leading to higher temperature. Rather than simply postulating a filament size and then estimating the temperature based on that assumption, we have attempted to extract a filament size self-consistently from our data by reconciling temperature rise as estimated from thermal modeling and temperature rise estimated from conductivity change. We

assume, in this case, that the conduction through the filament dissipates power uniformly across the filament. The true temperature is calculated with a self-consistent solution for a filament size under the constraints of simultaneously satisfying the adiabatic *I-V-T* relationship (Figure 2.12(b) with extrapolation as necessary) and the R_{th} experienced by the same filament size. The adiabatic *I-V-T* relationship measured with pulsed *I-V* technique indicates how much power is generated in the device at a given temperature for a combination of current density, *J,* assuming a filament size. The R_{th} is a measure of how effective can the generated power be dissipated and is unique to a given filament size. As R_{th} represents the thermal resistance that is connected between the filament as the heat source and the thermal ground, it can be easily calculated from material properties and a steady state finite element simulation. We use Comsol Multiphysics finite element method solver for the calculation of the R_{th} as the ratio of the rise in temperature experienced with unit increase in the power dissipated in the filament, at steady state. Figure 2.13 (a) shows the simulation setup used for the solver. The results of the simulation are summarized in Fig. 2.13 (b). From this figure, then, it is possible to estimate the temperature given a filament radius and the measured power dissipated in the device after the onset of the filamentation.

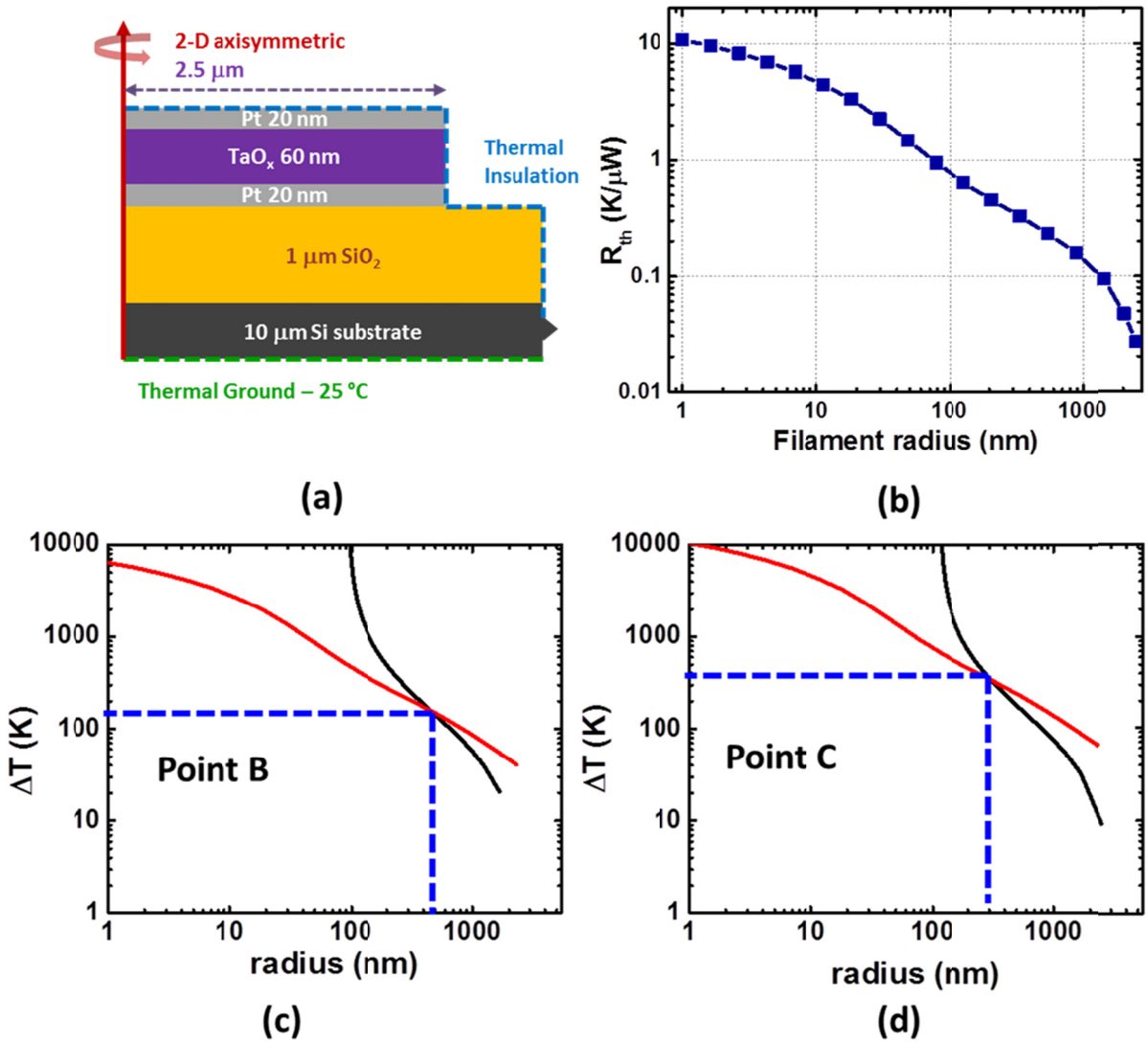


Fig. 2.13: True filament temperature extraction. **(a)** Schematic of the Ta_2O_{5-x} device used in Comsol Multiphysics electrothermal simulations. **(b)** Simulated thermal resistance (R_{th}) as a function of filament radius. **(c)** & **(d)** ΔT extracted by parameterizing different radii with (b) – red curve; and using pulsed I-V (black curve) shown at different powers corresponding to points B and C in Figure 2.12(c).

We also know that when we estimate ΔT from the pair of *I-V* coordinates, we assume that the current flows through a filament of radius 2.5 μm, i.e. uniform conduction. Thus, for a device undergoing filamentation, we will always underestimate the local filament temperature. Scaling

the current axis in the *I-V-T* thermometer (Figure 2.12(b)) by the ratio of the uniform device area (conduction radius r = 2.5 μm) and the filament radius, r we get a new range of temperatures for an effectively higher current density. Thus, the corrected curve so constructed is a SECOND, independent figure we can consult to extract a temperature from an assumed filament radius and known I and V measurements. Figures 2.13 (c) and (d) consist of two curves each. The red curves represent the ΔT obtained by multiplying the power at points B and C in Figure 2.12(c) with the R_{th} and parameterizing the radius. The black curves represent the temperature rise predicted by the *pulsed I-V* with parameterized current density for different filament radii. The unique intersection of the two curves is used at each power level to calculate the true filament temperature and radius as the device undergoes electronic constriction of current. Thus, these two constraints are simultaneously applied to the DC measurement to yield the true localized temperature of the device as it undergoes thermal and electronic filamentation. This gives the blue trace (full details of the extraction procedure are explained in the Appendix A). The solid line represents the extraction of temperature made within the experimental limits of the adiabatic *I-V-T* measurement. The dotted lines indicate the non-linear extrapolation of the measurement data. The non-linear extrapolation was made by assuming a space-charge limited conduction (to which the data matches). It must be noted that the entire NDR curve (up to point C) is reversible and hence we assert the nature of this localization to be electro-thermal and preceding the motion of atoms or vacancies.

Moreover, such deviation is also seen in dynamics measurements in TiO_{2-x}, (more details in Appendix A for Ta_2O_{5-x}). The device temperature can be both measured and simulated till just prior to resistance drop. Temperature rises ranging from 10-150 K have been measured in

devices prior to filamentation in pulse experiments (range refers to the pulse voltage used to form the device, like our previous work [24][26]).

We have presented the experimental evidence of electronic instability in oxide materials commonly used in fabrication of memristive devices. During the approach to the instability, the temperature gradually increases linearly with power to about 320 K. At this point the current flow constricts and the actual temperature of the filament increases faster than uniform current flow calibration. For the Pt/Ta_2O_{5-x}/Pt device discussed in this work, the steady temperature reaches ~500 K as the extracted filament diameter reduces from 5 μm (uniform conduction) to 1 μm due to thermal NDR. Additional power produces further filament collapse, and temperature increases rapidly, reasonably estimated as high as 1000 K as the filament diameter collapses to ~ 10 nm (same as reported by Kwon *et al*. [19]). The filament diameter estimation has been explained in the supplementary document. These temperatures provide sufficient activation energy to change the oxide in the different ways reported in literature – cause oxygen vacancy creation, crystallization, secondary phase formation[30] and/or melting of the top electrode [19]. The proposed mechanism explains these changes and removes the inconsistency in explaining reduction of the oxide at low temperature. It parallels the mechanism that has been widely accepted in chalcogenide glasses and referred to as "threshold switching" [37],[39]-[41].

The CC-NDR in metal/oxide/metal structures have been reported number of times in the literature [39]-[41]. Most observations have been made on electroformed devices that already contained a permanent conducting filament. This clearly is only remotely relevant to the discussion of the electroformation process presented here. As noted in the introduction, the model presented in the paper and its experimental evidence agrees with the simulation of the *I-V*

characteristics reported by Alexandrov *et al*. [42] The CC-NDR is caused by increase of the conductivity with temperature and electric field. One could, therefore, pose the question if the observations reported here are consistent with what is known about conductivity mechanisms in oxide thin films. The as-deposited oxide films in memristive devices are typically n-type but highly resistive indicating that the Fermi level is located significantly below the conduction band. The mechanisms most frequently identified as responsible for conductivity in such layers are thermionic emission, hopping between defects sites, Poole-Frenkel effect, or small polaron hopping [43]-[48]. All of these mechanisms result in the conductivity increase at higher temperatures and can lead to thermally induced current constriction. At high electric fields, Poole-Frenkel and hopping-based models could lead to sudden mobility increase by transfer of electrons from localized states to extended band states with high mobility. On the other hand, the traditional field-induced nucleation model described above assumes existence of two distinct phases of the material to account for threshold switching that precedes forming: the stable insulating phase and a metastable (at low temperatures) metallic one. The structure of the functional film in our devices has been assessed by transmission electron microscopy (TEM). The image in Fig. 2.3 shows a characteristic mottled contrast of amorphous material for TaO_x and SiO_2 layers in the device structure. The Fourier transform of the image (inset) has a perfect radial symmetry in agreement with this assertion. Since the sputtering was done in argon (no oxygen present), the oxygen content of the film is likely significantly below Ta_2O_5 fully oxidized tantalum. The exact composition is very difficult to assess. The initial state of the threshold switch is the amorphous tantalum sub-oxide. The difference between the two phases in the well-understood threshold switch materials, namely chalcogenides such as $GeTe_x$ [33] and suboxides such as VO_2 [34] and NbO_2 (from the same column of periodic table as Ta), is the change of

atomic arrangement without change of composition. The Ge atoms change their coordination from tetrahedral to octahedral in GST while in transition metal oxides the metal ions form pairs in the low temperature low symmetry phase. This characteristic allows for fast switching observed in all of these systems. This re-bonding is the underlying reason for metal-insulator transition and the conductivity change of both types of structures making it is conceivable that a similar transformation is responsible for the threshold switching in TaO_x.

2.7. Summary of S-NDR during forming in RRAM devices

In summary, we argue that electronic current localization precedes permanent filament formation during electroformation in oxide-based resistive switches. The presence of negative differential resistivity in the material causes the device to go into a negative differential resistance regime which causes current constriction. Unless prevented by the circuit load, this process frequently occurs in the form of an uncontrolled runaway. We support these claims by analysis of the steady-state DC behavior and the dynamics of the instability. Both DC and dynamic measurements indicate the presence of an instability that is reversible and, hence, electronic in nature as distinct from structural. The initiation of the constriction is temperature dependent and higher temperature is shown to cause the point of bifurcation to appear at a lower voltage. Hence, we propose the following mechanism of electroforming - with increasing bias, the device conducts uniformly throughout its area. At a well-defined point depending on source voltage, series resistance, temperature and time, the device enters into the *I-V* range of negative differential resistance which results in the electronic current filamentation. This current filamentation starts off with being thermally induced (due to the thermal non-linearities) before the effects of voltage non-linearity set in. This final stage of current filamentation causes the

device to change resistance to a value close to the post-forming value. We estimate the localized temperature in the current filament using a self-consistent electro-thermal measurements and simulation. Temperature excursions that exceed 500 K (over the ambient) were estimated in a localized sub-20 nm region on the onset of forming. This localized temperature excursion then triggers the physical changes in the structure to form the permanent filament. The constriction can be controlled with the use the external circuit loading thus affecting the permanent filament structure. In order to corroborate the results with temporal dynamics of filamentation, we observed and explained the three regimes of electroforming time dependencies on forming voltage. The observed *1/E* field dependence of forming times is consistent with field-induced nucleation model from which we extracted material properties such as the nucleation barrier height at zero bias ($W_0 \sim 0.65$ eV) and voltage acceleration factor V_0 of ~ 2.8 V (for 60 nm TaO_x film) at intermediate voltages. While the nucleation model provides a robust framework that can be populated with details of the reversible conductive phase (electronic or structural), It is agnostic to the exact mechanism that results in the reversible nature of threshold switching that precedes forming. Also consistent with nucleation model, a clear difference in the temporal dynamics was identified for low voltages with corresponding forming times longer than the thermal time constant and at high-fields where the film self-heating is important. Moreover, we were able to detect, and study the volatile filament that precedes formation of the non-volatile filament. The forming process was analyzed in the framework of nucleation model which was extended to include the self-heating effects. This yielded an estimate of the critical nucleation radius ($R \sim 1$ nm) below which filament is always volatile. This implies that the filament-based RRAM technology can thus be scaled to a physical limit dictated by the critical nucleus size which could be as low as 1 nm in size. We also demonstrated that forming is a field accelerated

phenomenon and that the forming can be sped-up by nearly 6 orders of magnitude compared to DC forming typically used for RRAM. The thermometry is fairly general in its applicability and hence can be applied to switched RRAM devices in LRS. The next chapter (Chapter 3) will deal with how the temperature excursions affect the switching behavior, filament size and endurance failure modes. The understanding of S-NDR in oxides acquired as a part of this chapter will be later used in Chapter 4 to explore the applicability of these devices as oscillators and threshold switches.

Chapter 3

Switching Thermometry and Modeling in RRAM

3.1. Introduction

Although there has been a huge corpus of research conducted to understand the underlying resistive switching mechanism and improve device performance [7],[36], the fundamental nature of the conductive path and basic switching mechanisms are still under debate. This is largely due to a lack of microstructural proof of the suggested switching and failure mechanisms. Indeed, the microstructural proof (e.g. size, location, distribution, and configuration of conductive filament in oxide functional layer during resistive switching), associated with the switching models is essential for its eventual commercialization because it is believed that the non-volatile nature and device reliability issues (i.e. thermal stability and endurance) are directly related to the irreversible structural transformations in the device [19].

It is well known that the temperature plays a key role in microelectronic device performance and reliability, and thus device temperature estimation during switching has also long been an active research topic. Chapter 2 discussed the evolution of temperature in RRAM devices as they underwent the forming process. However, the role of temperature on switching is also regarded as being central to achieve resistive switching. This is primarily attributed to the fact that the motion of oxygen vacancies that participate in resistive switching is a thermally activated process [7, 50].

The temperature estimation so far, however, has been largely relied on simulations that make several assumptions about both the geometry and microstructure of the filament [49]-[51]. However, to understand the physics of resistive switching in RRAM devices, it is important to first characterize the temperature excursions that filamentary RRAM devices experience. In spite of the importance of thermal effects in VCM RRAM devices, experimental evaluations of the

filament temperature are scarce [52]. Experimental evaluation of the temperature in a confined conduction area is critical for the device technology development because an accurate modeling of the complex electro-thermal behaviors at the nanoscale is not possible yet: heat dissipation mechanisms as well as the thermal properties of resistive switching materials e.g. the thermal boundary resistance values are not well known.

Direct measurement of the local filament temperature is extremely challenging for the following reasons. The typical device is composed of a vertical metal-insulator-metal (MIM) structure, where the insulator is a thin resistive switching film. The temperature increase occurs locally in the insulator or insulator-electrode interface on a scale of few nano-meters [49]-[52]. Several studies on the local filament temperature in RRAM materials were previously reported. Janousch et al. were able to show qualitatively local heating at the anode side of a large lateral device using IR thermal imaging [53]. Other studies compared electro-thermal simulations with 2-terminal electrical measurements to model the temperature distribution in the RRAM cell [49-51]. In order to obtain the actual temperature profile in the device using the aforementioned method the following inputs are required: filament geometry, dimensions, heat generation/dissipation mechanism, thermal/electrical properties of the filament and its surrounding, and in particular the thermal and electrical interface resistance which become dominant at the nano-scale.

In this chapter, we will discuss the applicability of pulsed thermometry discussed in chapter 2, for switched RRAM devices in the low resistance state. Unlike the thermometry applied to the forming process where the conducting area evolves as a function of applied bias (due to current localization), the area of the conductive filament remains constant in RRAM devices in the low

resistance state (LRS). Thus the concept simply relies on using pulses in order to determine I-V-T characteristics of the device in the absence of self-heating, without the need for maintaining self-consistency needed in Chapter 2. This can be accomplished if the pulses used are short enough that significant temperature increases will not occur. Once collected, the I-V-T data serve as a look-up table for extracting the actual temperature of the device (in the presence of self-heating). If the device has not yet undergone forming process, the conduction is uniform. Thus, the area of conduction is fixed to the device area at low biases. At high biases, the current constricts into a narrow filament (i.e. conduction area changes), with temperature to be sensed by the thermometry (Chapter 2). We will also validate our physical picture by providing a direct evidence of microstructural change (local crystallization) associated with the local temperature excursion and dynamics during switching by using a high-resolution transmission electron microscopy (HRTEM). The TEM analysis is done by Jonghan Kwon to supplement and validate the results of thermometry. Thus, we use a combination of these two techniques to determine the filament size evolution with cycling, shedding light on the endurance failure mechanisms in RRAM devices. Crossbar patterned resistive switching device stacks, TiN (40 nm) /$HfAlO_x$ (5 nm) /Hf (10 nm) /TiN (30 nm), have been prepared on Si substrates as shown in Figure 1 by IMEC (Leuven, Belgium). The functional $HfAlO_x$ layer was deposited via atomic layer deposition (ALD), and the bottom TiN and the top Hf/TiN electrodes were sputter deposited. The sputter deposited Hf-cap acts as an oxygen getter layer that allows formation of the oxygen deficient, off-stoichiometric $HfAlO_x$ functional layer, facilitating conductive filament creation under applied electrical bias. The devices are defined by the region where the bottom electrode (BE) and top electrode (TE) overlap. These devices were encapsulated by a SiO_2 passivation

layer. The device sizes used for the current study were 85×85 nm^2 unless specified. Fig. 3.1 shows the device schematic and an SEM micrograph.

The pulsed I-V characteristics of the device were obtained by pulsing the device with Tektronix 80060A pulse generator using time domain transmissometry [35], using the setup shown in Chapter 2. Pulses of widths < 85 ps were delivered to the device with microwave RF probes, with a rise time of < 55 ps.

3.2. Thermometry

Figure 1 shows the DC forming (black) and switching I-V characteristics of the TiN(TE)/Hf/$HfAlO_x$/TiN(BE) stack for compliances of 50 µA, 100 µA and 200 µA. In all of these cases, the current compliance was applied by the means of an on-chip ballast element.

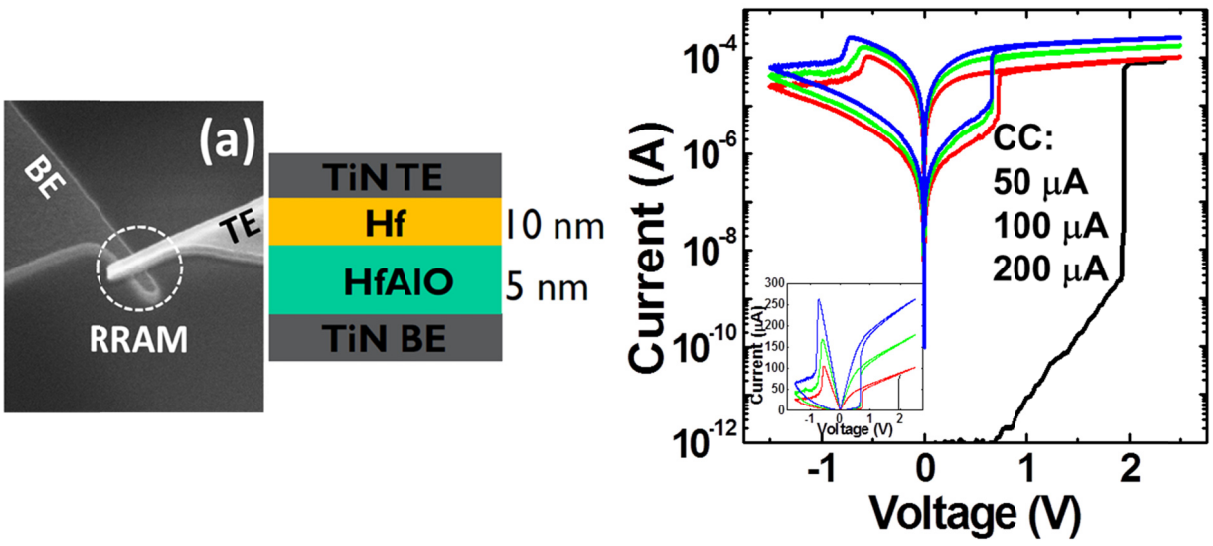


Fig. 3.1: DC forming and switching I-V characteristics of the devices at different compliance current (CC).

We use the pulsed I-V characteristics of the device (see Methods section) to extract the temperature excursions in the device. Fig. 3.2 show the pulsed I-V characteristics of a device in LRS.

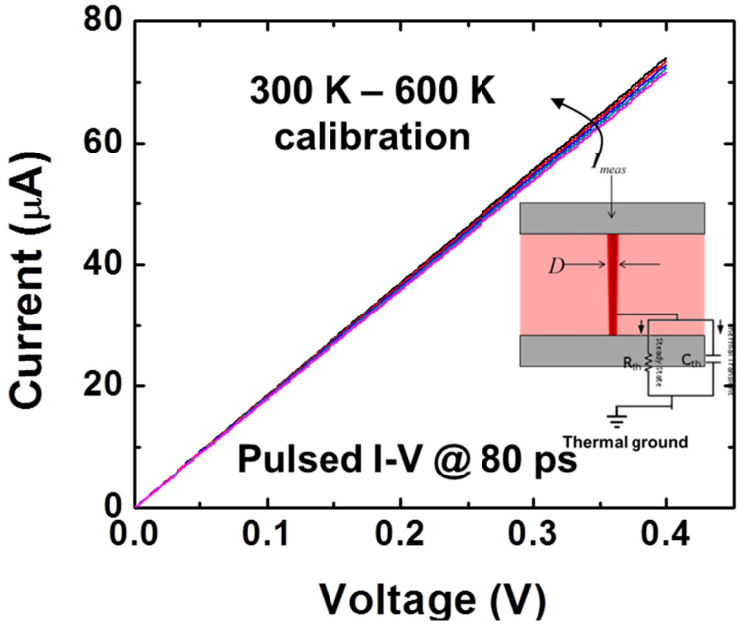


Fig. 3.2: Pulsed I-V characteristics as a function of stage temperature for RRAM devices in LRS.

Once the temperature is extracted (using a similar pulsed thermometry as described in Chapter 2), we plot the temperature excursion (ΔT) as a function of the applied power, P, (as obtained from the switching I-V) in the low resistance state (LRS), as shown in Fig. 3.3.

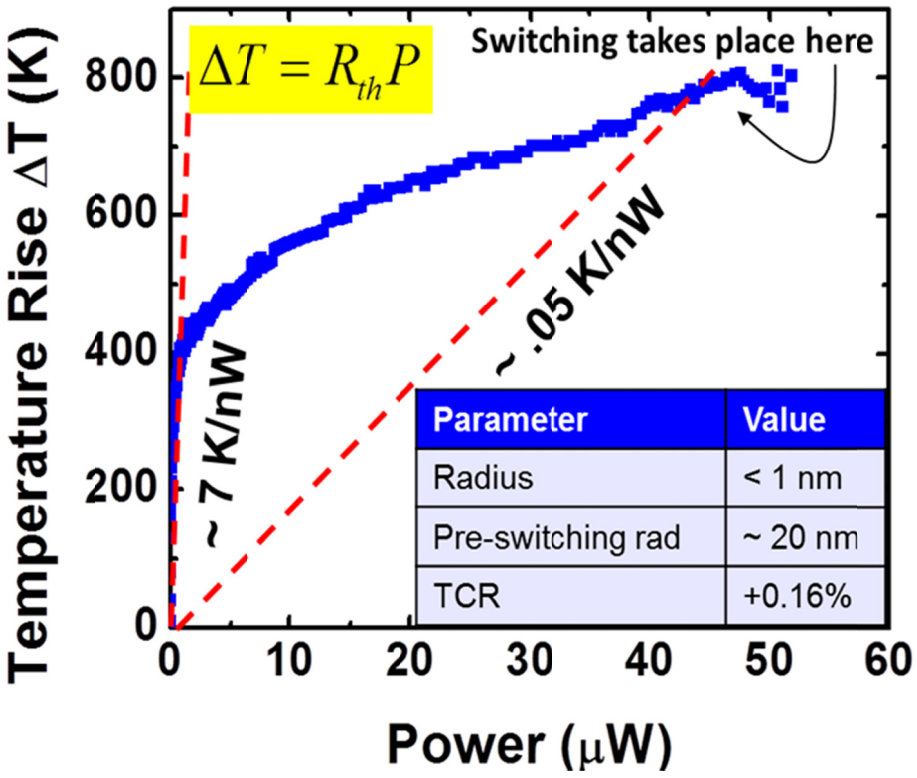


Fig. 3.3: ΔT vs. power in the low-resistance state.

The slope of the ΔT versus Power curve represents the thermal resistance (R_{th}) that is experienced by the heat-source. This can be represented with the following equation in the steady-state:

$$\Delta T = R_{th} \times P \tag{3.1}$$

Here, R_{th} depends on the thermal conductivity of the material surrounding the filament, the length over which the heat is dissipated (thickness of the device) and the area of the heat source. Out of these parameters, the area of the filament is the only variable that could represent a change in R_{th} (thermal conductivity only has a weak dependence [54] on temperature). Thus, the change in the slope (R_{th}) that is seen in Fig. 3.3 can be attributed to the change in the heat, as will

be explore in the discussion section. Near the origin, the slope of the ΔT vs Power plot exhibits a slope of ~ 7 K/nW. As the power through the device is increased, the slope decreases to ~ 0.05 K/nW before switching.

To interpret the slope in terms of radius of conduction, we use a finite element simulation to understand the change in R_{th} experienced by the filament. Figure 3.4 shows a COMSOL simulation of the device, plotting the thermal resistance for filaments of various sizes, akin to chapter 2. However, owing to the differences in the thermal environment of these devices compared to the devices discussed in Chapter 2.

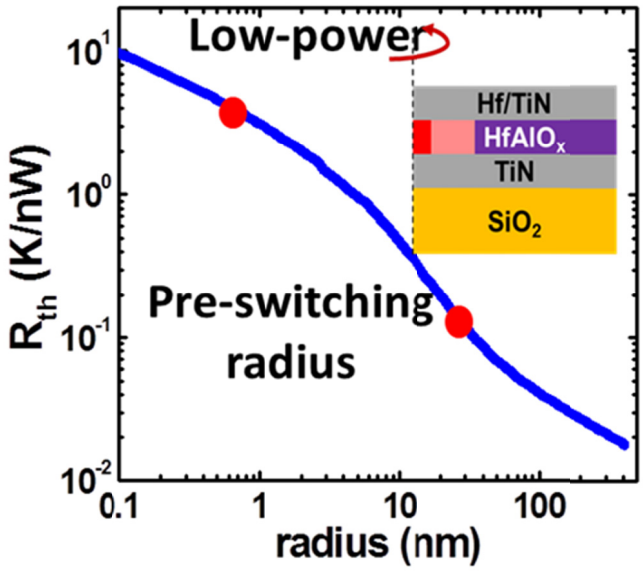


Fig. 3.4: Thermal resistance for varying filament size

More details about the electro-thermal simulation are mentioned in Appendix A. This plot is used as a lookup table to extract the radius of the heat source (filament), for a known R_{th}. The radius corresponding to a R_{th} of 7 K/nW is < 1 nm implying that the heat source at low-powers

can be identified with a nano-scale filament. At high voltages, the power density in the filament is very high, causing large temperature excursions. These temperature excursions can be seen in the simulation shown in Fig. 3.5.

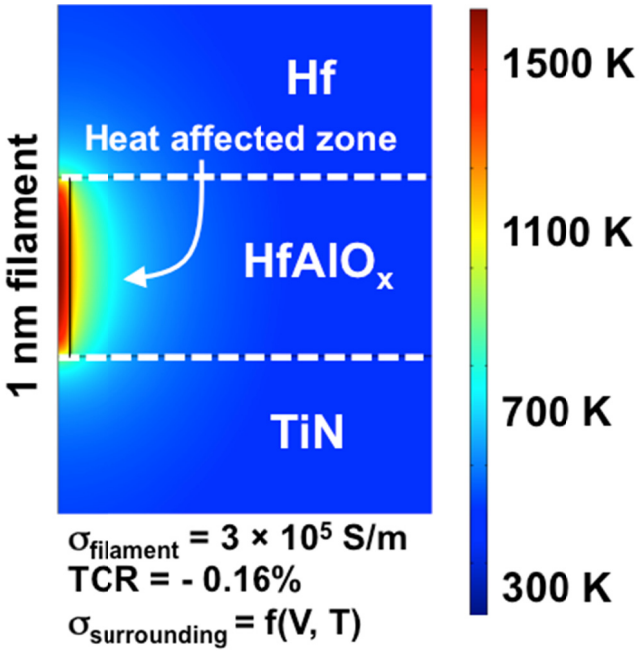


Fig. 3.5: Electro-thermal simulation to estimate temperature excursion at high voltage.

However, it must also be noted that due to heat diffusion, the oxide region surrounding the filament gets heated up. This causes the conductivity of the surrounding region to increase [54], eventually leading to current spreading. This causes the apparent slope of the heat-source to reduce. Thus, at high-powers, the slope of the ΔT vs power plot reduces to 0.05 K/nW. This corresponds to a radius of ~ 20 nm. One must note that this radius is averaged by the conductivity and the temperature is also similarly averaged. The true local temperature can be obtained from the same electro-thermal simulation (Figure 3.5) alongside the heat affected zone

around it. The peak local temperature suggested by the simulation is ~ 1600 K at the core of the filament with the pre-switching heat affected zone experiencing temperature in excess of 800 K over a 5 nm radius.

It must be noted that these values are sensitive to the filament size, the oxide matrix and the testing methodology (DC versus AC). Our thermometry is capable of determining the size of the filament agnostic to all of the aforementioned parameters as long as the pulsed-IV calibration is done for each device state. Fig. 3.6 shows the filament size extracted using the pulsed thermometry technique for different current compliance values. It is clear that the filament size seems to increase with increasing compliance, consistent with previous works.

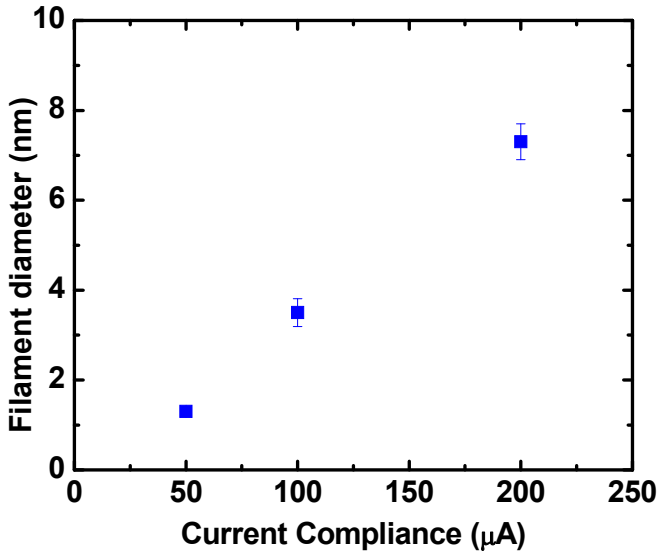


Fig. 3.6: Increasing filament diameter as a function of current compliance.

3.3. Microstructure Analysis

In order to confirm the temperature excursions and dynamics during switching and the simulated temperature profile, we conducted cross-sectional HRTEM analyses of the device stacks. Figure 3.7 (a) is a cross-sectional view of the pristine device showing TiN(TE)/Hf/$HfAlO_x$/TiN(BE) stack and Figure 3.7 (b,c) represent magnified views at left and right side of the device, respectively.

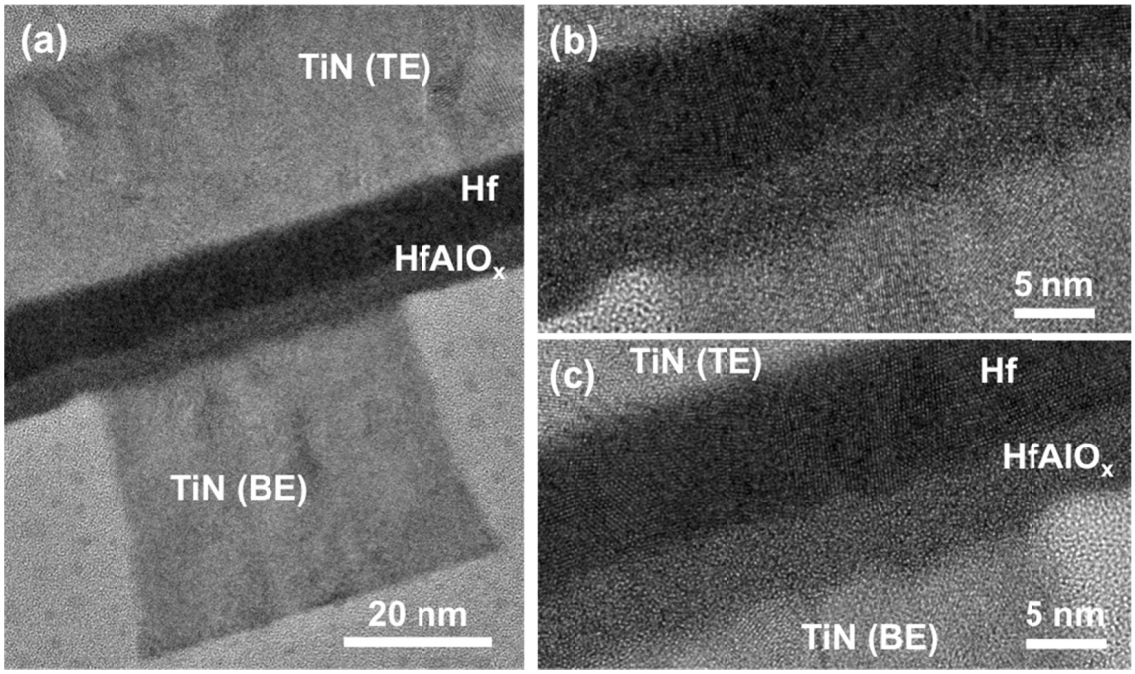


Fig. 3.7: Cross-sectional TEM image showing the device stacks. (a) Low magnification and high magnification views of the left (b) and right (c) corner of the device.

Since the focused ion beam (FIB) prepared device was cut along the TiN (TE) of the crossbar device, the TE appears much longer in width than the TiN (BE). The two TiN electrodes are polycrystalline and they exhibit a microstructural morphology consistent with columnar growth.

The $HfAlO_x$ functional layer exhibits slightly lighter contrast than the Hf oxygen getter layer because of the atomic mass difference. The lightest contrast surrounding the TiN BE is the SiO_2 passivation layer. Recesses are seen at the corners of the device that is presumably due to etching process during device fabrication. It is noted that the recesses provide landmarks to confirm that the magnified images were recorded at the same location. Speckle contrast is seen in the SiO_2 layer which is indicative of deposited Ga^+ ions from the FIB specimen preparation process. Note that the speckle contrast is not seen in the device area and does not affect phase contrast imaging of the device layers. The initial microstructure of the $HfAlO_x$ in as-fabricated device is amorphous. No crystallinity was observed in the $HfAlO_x$ functional layer, even when performing a through focusing series and FFT analysis. The result has been further confirmed by examining a much larger volume of $HfAlO_x$ layer in another larger device (1×1 μm^2).

A cross-sectional HRTEM analysis has been conducted on devices operated at different compliance currents and local crystallization (lattice fringe) of $HfAlO_x$ layer is observed as seen in Figure 3.8. The micrographs are magnified views of right corner of the devices (the recess is seen at the bottom-right corner) showing microstructure of the programmed devices at 10 μA (Figure 3.8a), 50 μA (3.8b), and 200 μA CC (3.8c).

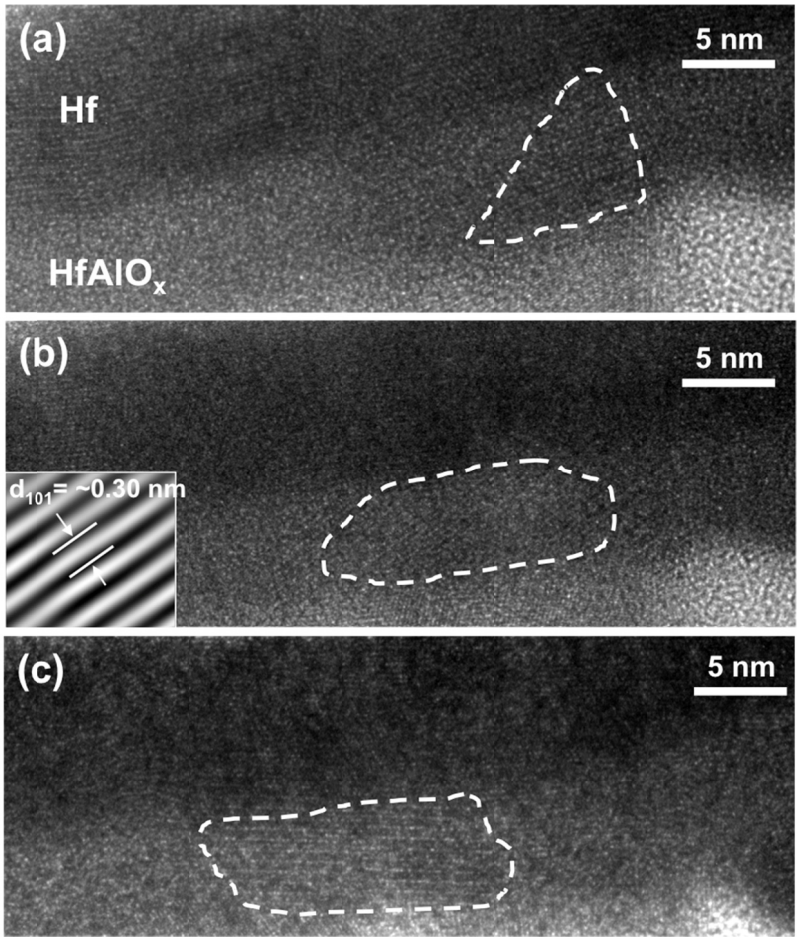


Fig. 3.8: HRTEM micrographs showing a localized crystallite embedded in amorphous matrix programmed at 10 μA (a), 50 μA (b), and 200 μA CC (c).

Imaging through focus series results in observation of a single crystalline region embedded in the amorphous $HfAlO_x$ layer for each device and it is delineated by dashed line. The d-spacing (~0.3 nm) was measured by using Fast Fourier transformation (FFT) analysis and it is likely d_{101} of tetragonal $HfAlO_x$. The crystallite sizes are ~10 nm (diameter) at 10 μA CC and ~20 nm at both the 50 μA and 200 μA CC. Note that the observed crystallite sizes are similar to the one extracted from the transient thermometry in the previous chapter. The changes in compliance current represent the change in dissipated power inside of the device. Since the temperature rise

is proportional to the dissipated power (Equation 3.1), local Joule heating during programming event presumably causes the crystallization.

Crystallization temperature of $HfAlO_x$ has been estimated by combining rapid thermal annealing (RTA) and cross-sectional HRTEM analysis. 3×3 μm^2 crossbar type devices were used for annealing experiment and ~4 μm wide TEM lift-out specimens were prepared. 5 different locations were randomly selected and through focus imaging has been performed. The devices were annealed at fixed temperatures for 2 s in a forming gas. Figure 3.9 shows cross-sectional views (one of the through focus series) of the devices annealed at (a) 820 K, (b) 870 K, (c) 920 K, (d) 970 K.

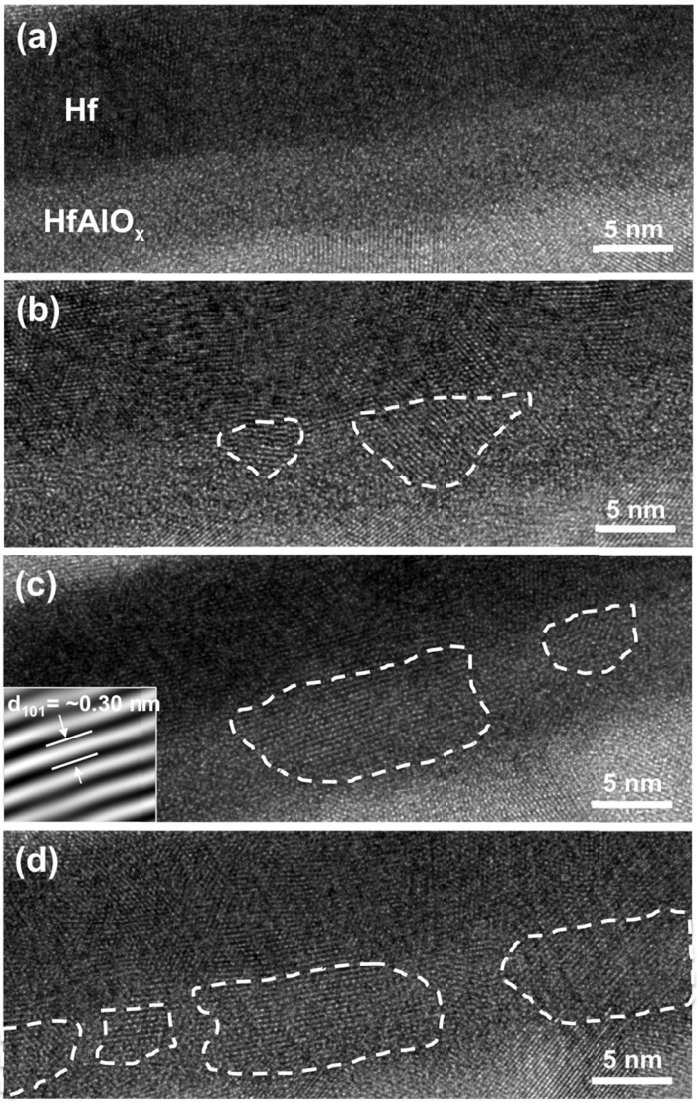


Fig. 3.9: HRTEM micrographs of RTA specimen at different temperatures: 820 K (a), 870 K (b), 920 K (c), 970 K (d).

From the top, Hf, HfAlO, and TiN layers are seen with the different contrasts. There is no crystallite found in 820 K RTA specimen and it is just like the as-fabricated specimen. Above the 870 K, crystallites are seen in the $HfAlO_x$ layer surrounded by amorphous matrix and they are more pronounce as the annealing temperature increases. The crystallites in $HfAlO_x$ layer are outlined by white dashed lines. The observed d-spacing is the same with the one spotted in the

programmed devices as seen in Fig. 3.8. Figure 3.9 represents this data across a range of temperatures plotting the % transformed material for a 2 s RTA at increasing temperatures.

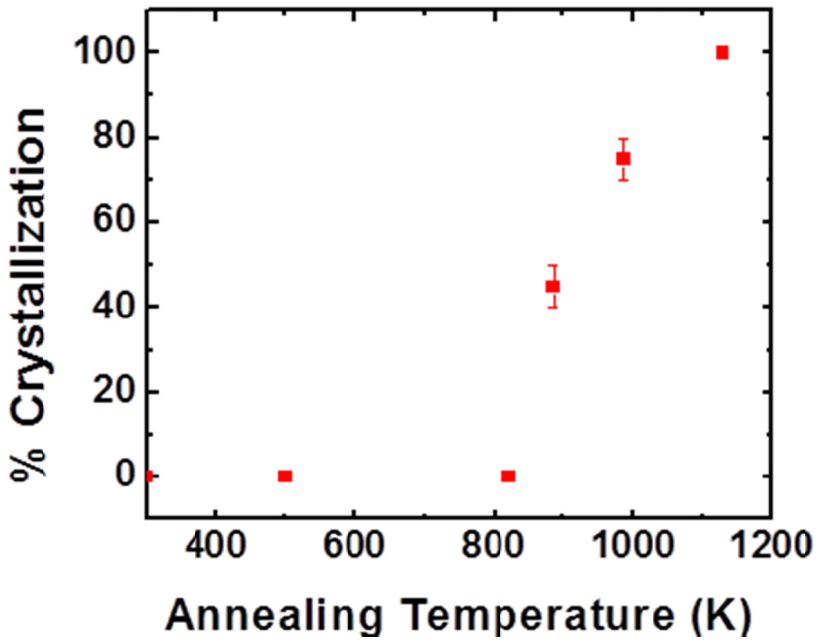


Fig. 3.9: % of crystallization as a function of different RTA temperatures showing a clear transition occurring at temperatures > 900 K.

The results indicate the switching temperature is above 870 K, which is close to the value estimated by the transient thermometry and the numerical simulation in the previous chapter. It is important to note that the resistivity of the device has not been changed before and after the annealing, denoting the crystallization per se may not be the conducting phase but rather a byproduct of the local Joule heating.

3.4. Discussion of Pulsed-Thermometry Methodology

For the simulations of the pulsed thermometry devices we assume that the current at low bias is flowing only through a uniform metallic filament with small diameter. The size of the filament

does not change in LRS or HRS except during SET and RESET processes at the end of *I-V* plot. The power dissipation causes temperature increase of the filament and through thermal conduction of the surrounding oxide. Since the conductivity of the oxide is increasing with temperature, at higher power levels the surrounding heated oxide starts to contribute to total current flow. In other words, the volume where the power is dissipated expands. The consequence of this should be the reduction of the thermal resistance associates with the heat flow through the stack and the substrate to the thermal ground.

In DC measurements, the temperature rise is approximately proportional to the dissipated power with a proportionality constant of the thermal resistance (R_{th}). As stated above, at low bias the current is flowing through the filament with fixed diameter. Plotting the measured temperature rise as a function of power (Fig. 3.3) and finding the slope at low bias allowed for an estimate of the filament diameter and as a consequence its electrical conductivity. Also, it is clear that the slope of the temperature increase versus power (i.e. the thermal resistance) decreases with increasing bias indicating power dissipation in the oxide. One should be able to reproduce this effect in the electro-thermal simulations. It will be shown in the following sections that the heating taking place in the filament during the first 100 ps of the pulse is significantly lower than the peak temperature that is reached at longer times.

In order to estimate the electro-thermal transients in the device, we simulated heat and current flow assuming the cylindrical geometry of the filament. The electrical conductivity and TCR (temperature coefficient of resistance, positive for the metallic filament) of the filament can be extracted once the area of the filament and the resistance measured during 100 ps are known. Our simulation involves surrounding the filament with an oxide which has an electrical

resistivity exponentially decreasing with temperature (non-linear negative TCR). The thermal properties of the materials assumed in our simulation are listed in Appendix A.

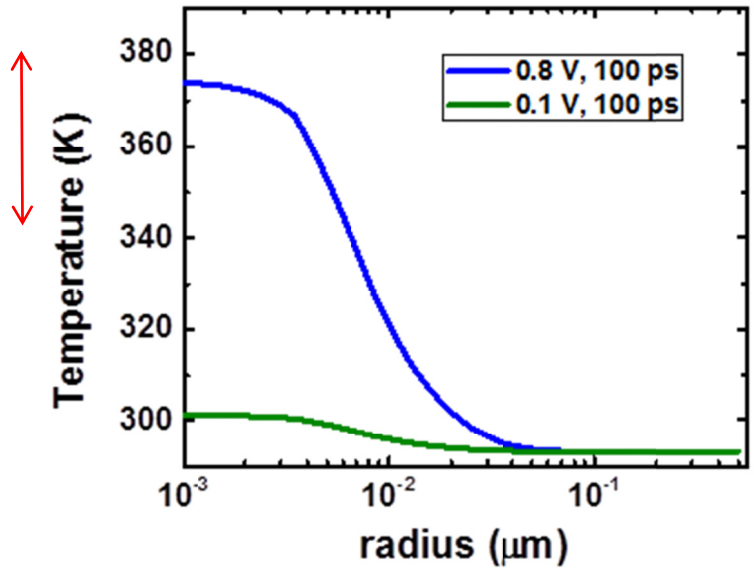


Fig. 3.10: Thermal transients for 0.1 V (green) and 0.8 V (blue) biases across the device, at the end of 100 ps. X-axis represents distance from the center of the filament.

This error is introduced due to self-heating in the material during the pulsed-IV calibration, at high-power levels.

3.5. Endurance Cycling

Using the self-consistent evidence obtained from the thermometry and TEM analysis above, we use the same methodology to determine the evolution of the heated zone near the filament. Fig. 3.11 shows the endurance cycling data for RRAM devices that were cycled with a current compliance of 50 μA with a V_{SET} of 1.5 V and V_{RESET} of -1.75 V, showing an endurance failure

after 10^8 cycles. A recalibration of the filament characteristics using pulsed I-V yields a new low-power slope, as shown in Fig. 3.11(a).

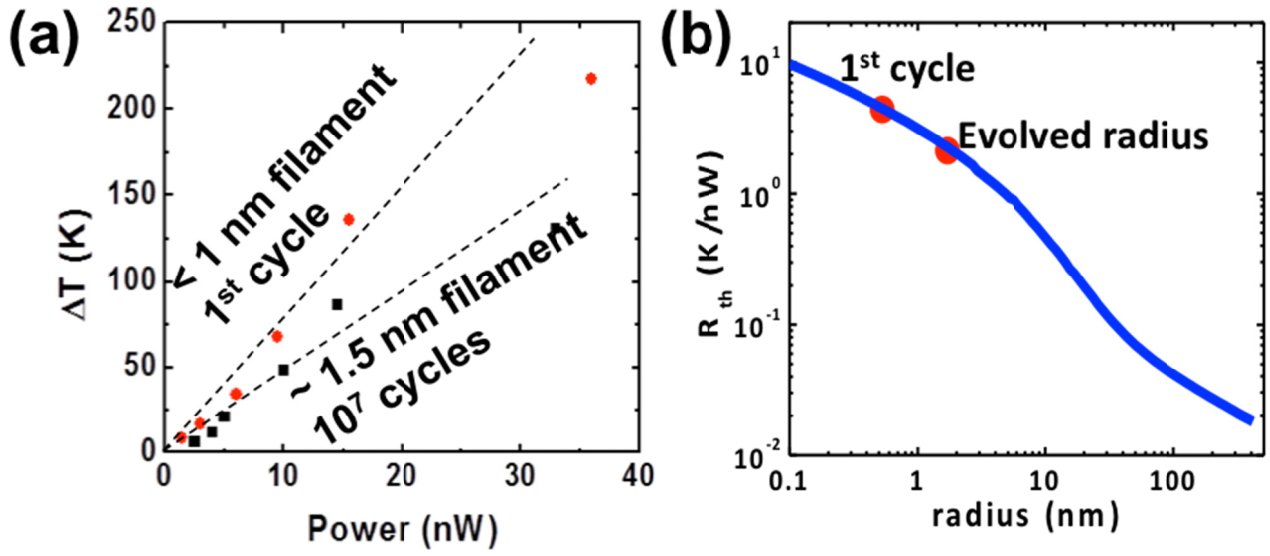
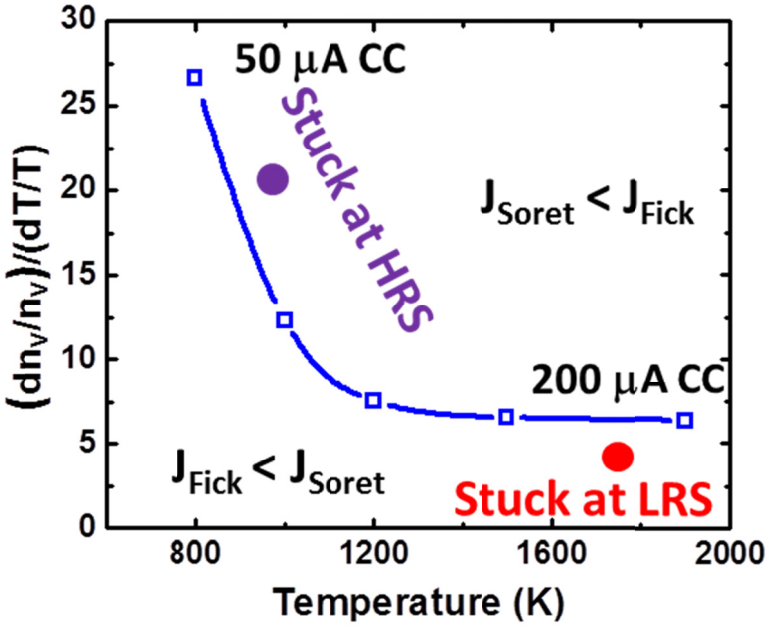


Fig. 3.11: (a) ΔT vs. power at low-power after endurance failure. (b) R_{th} vs. radius profile showing filament size evolution upon endurance cycle.

The red symbols indicate the original filament radius and the black symbols indicate the ΔT vs power extraction after the endurance cycling. The post cycling R_{th} was observed to indicate a slope corresponding to a filament of 1.5 nm radius (Fig. 3.11(b)). This yields a wider heat affected zone due to higher power levels needed to switch the device.

3.6. Discussion of endurance failures

Based on the thermometry carried out in the previous sections, one can estimate the role of temperature in the switching mechanism. We use the measured pre-switching temperatures (from five devices operating at different CC) and their gradients to estimate a Fick-Soret vacancy radial flux balance (J_{Fick}/J_{Soret}) [55], as shown in Fig. 3.12.



$$J_{\mathrm{Fick}} \approx -D_{\mathrm{V}}\mathrm{d}n_{\mathrm{V}}/\mathrm{d}x$$
$$J_{\mathrm{Soret}} \approx -D_{\mathrm{V}}S_{\mathrm{V}}n_{\mathrm{V}}\mathrm{d}T/\mathrm{d}x$$

Fig. 3.12: Plot showing balance of Fick and Soret effects for 5 switching samples (symbols) and samples that incurred endurance failure due to SET (stuck at HRS) and RESET (stuck at LRS) failures.

At each operation point, it is assumed that there is a balance between concentration driven out-diffusion of vacancies and vacancy in-diffusion driven by the Soret effect [55]. Devices that undergo RESET failure (stuck at LRS) are shown as red symbols and they clearly show a dominant Soret contribution producing an excessive in-diffusion of vacancies. Devices that undergo SET failure (stuck at HRS) are shown as violet symbols and indicate dominant Fick flux producing an excessive out-diffusion of vacancies. Numerous observations are consistent with this flux balance model in which crystallization is a second order consideration and not a central participant. These include the observations that increasing endurance is possible with fast pulses to cold-switch the device [56], that devices undergoing switching at lower temperatures show no

post-forming microstructural change (i.e. $HfAlO_x$ remains amorphous) until the endurance failure. Moreover, as discussed in the previous section, after ~ 5×10^5 cycles, thermometry indicates that the CF diameter (at 50 μA CC) grows by a factor of ~2, and a higher voltage is needed to RESET, causing temperatures to locally exceed 1500 K and leading to the temperature excursion predicted by thermometry. Recently, SET failures have been attributed to narrowing and eventual disconnection of the filament [57]. This creates a lower temperature but high concentration gradients consistent with the relative position of SET and RESET failures in the flux-balance plot (Fig. 3.12-13).

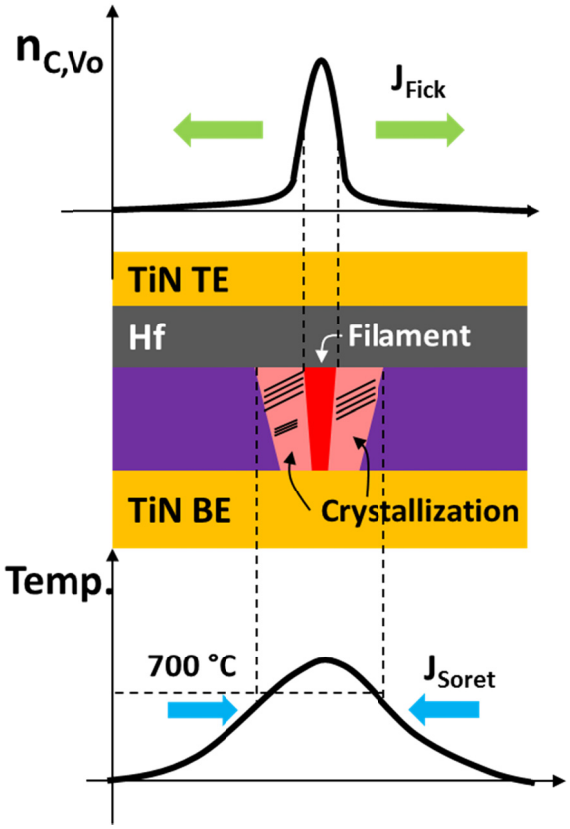


Fig. 3.13: Schematic of the device cross-section clearly indicating the filament (red), crystallized region (hatched light-red) that gets heated up and causes the apparent ΔT *vs* Power slope (R_{th}) to reduce.

The thermometry used in this chapter indicates that the device experiences temperatures in excess of 1000 K which eventually causes the surrounding oxide matrix to become conductive, thus acting as a self-limiting process to reduce the power density and the peak temperature. Apart from changing the area of conduction, the switching process also creates a change in the lateral temperature gradient. This change in gradient could affect the lateral out-diffusion of vacancies, eventually causing SET and RESET failures. While the primary actor in endurance failure is excess vacancies being created due to thermal cycling of the device and vertical temperature and concentration gradients, the lateral gradients play an important role to ensure that the filament does not get disconnected when the filament self-heats. More experiments are needed to understand the role of lateral temperature gradients on the switching mechanism. This is also partly because the vertical temperature gradients may be a strong function of electrode material properties (most notable thermal properties).

Chapter 4

S-type Negative Differential Resistance for Compact Oscillators

4.1. Introduction to S-NDR Oscillators

Brain-inspired neurocomputing is considered as an emerging alternative to computing based on traditional techniques due to its massive parallelism. A neurocomputer attempts to mimic the human brain via a network of coupled artificial neurons that process information in parallel. Each brain neuron represents a computational unit in a neural network and a connection between two such neurons represents is known as a synapse. The strength of this connection, the synapse is in form of a synaptic weight in a neural network that relates one artificial neuron to another. Traditional computing schemes (that utilize the von Neumann architecture) run a software algorithm for a specific application by sequentially executing each line in the instruction code. Even though each execution might take a very short time the overall computation efficiency is not that high due to the serial execution of instructions [58]. Instead, a neural network performs pattern recognition via associative memory in a massively-parallel manner. It maps a set of input patterns to a set of output patterns via synaptic weights, whereby an output pattern can be retrieved for a given initial pattern. Graphical applications would otherwise require numerous memory fetch operations and a processor that is executing a list of commands for optimization.

Oscillatory neural networks (ONN) are one such example of phase-based neurocomputing, in which the state variable is represented by the *phase of an oscillator*. For this, frequency-tunable oscillators are used to implement circuit blocks. Specifications of these circuits typically have stringent power and area constraints. One of the implementations of ONNs requires oscillators with frequency as the processing state-variable [59],[60]. These oscillators utilize frequency shift keying (FSK) to implement states, and thus, must be frequency tunable. The other implementations of ONN require phase coupling and control (Phase Shift Keying, PSK). One of

the challenges in implementing such oscillators in CMOS (Voltage Controlled Oscillators, VCOs) is the relatively large footprint needed to realize them along with the need for a large area phase-locked loop (PLL) [60] for ONNs, and the consequent power consumption. This challenge becomes exacerbated in highly parallel oscillatory neural networks that need dense oscillator arrays for associativity [59]. If ring oscillators are used, small areas can be achieved but providing a wide tuning range is very difficult. Furthermore, coupling of oscillators requires significant additional circuitry adding to the overall size and power. This has increased interest in oscillatory behavior exhibited by emerging nanoscale devices based on chalcogenides, oxides and spin-torque oscillators [59]. Moreover, these oscillators can be easily coupled directly [61],[62], thus naturally lending themselves to ONNs. In this chapter, we will revisit the concepts of S-NDR devices developed in Chapter 2, but under the context of how the same devices are used as oscillators due to their unique characteristics.

Oscillators based on devices that exhibit S-shaped negative differential resistance (NDR) have been explored for several decades [63-65]. These devices consisted of metal-semiconductor - metal structures in which the functional layers were either chalcogenides or oxides. Recently, parallel efforts by Parihar et al [62] have shown phase coupling of two such resistance ballasted VO_2 oscillators. However, these devices have to be fabricated as lateral structures on rutile single-crystal TiO_2 substrates, severely limiting their prospects for CMOS integration in the BEOL. Moreover, VO_2 has a very low transition temperature of ~85 ºC associated with insulator-metal transition (IMT), rendering it impossible to expose it to typical CMOS operating temperatures. Thus, we will focus our attention to materials that have high glass transition temperatures at which the ON state can be thermally activated.

This indicates that there is a clear need for vertical device stacks with > 500 K transition temperatures for CMOS compatibility and area scaling. In order to address this, we use TaO_x as the functional layer. As we noted in Chapter 2, TaO_x does not change its state at least up to 500 K and shows characteristics that are desirable for oscillators. Thus, this work serves as the first report of TaO_x as a material that exhibits an ON state that that can be stabilized and used for oscillators. While frequency tunability has been shown in both oxide and chalcogenide-based devices, it usually involved changing the ballast resistance. In addition, frequencies higher than 10 MHz have never been reported which heretofore has represented a serious limitation.

In this chapter, we demonstrate precise frequency control over four decades (20 kHz - 250 MHz) using an emerging class of oscillators based on metal-insulator-metal (MIM) structures, where the insulator is an oxide thin film. We exploit unique properties of these materials and devices which result in filamentary relaxation oscillations. Thus, the oscillator is a single MIM device in series with a ballast, capable of displaying a large resistance change (> 100x) while being both CMOS compatible and scalable. Additionally, we show the oscillator operation in two distinct regimes in which the frequency tunability is dictated by two variables – ballast and the source voltage. We explore how the oscillations are affected by using a linear resistor-ballast and a non-linear transistor-ballast. Moreover, the oscillation characteristics reported here shed light on the dynamics of the resistive memory switching devices during forming.

4.2 S-type Negative Differential Resistance and Oscillations

To implement these oscillators, we use two MIM crossbars to prove scalability of oscillators and generality of the oscillation phenomenon. The first structure consists of a 5 μm crossbar in form

of a Pt (20 nm)-TaO_x (RF sputtered, 60 nm)-Ta(5 nm)/Pt (15 nm) stacks; the second is a scaled 700 nm TaO_x-based crossbar. Transition metal oxides (in this case, TaO_x) have a unique property which enables the current flowing uniformly through the device to spontaneously and reversibly collapse into a narrow electronic filament – a property known in dynamical systems modeling as a 'bifurcation' phenomenon [6], discussed in detail in Chapter 2.

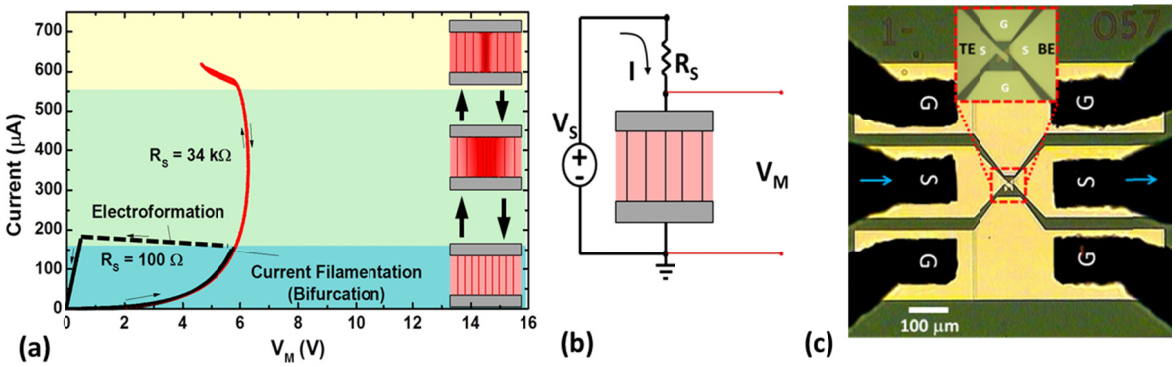


Fig. 4.1: (a) S-type NDR seen in the I-V characteristics in devices as they undergo volatile filament formation and dissolution. Line contours represent the current density. (b) The circuit schematic used to ballast the device to observe NDR. (c) Optical image of the device with the probe setup. The dotted red region represents the crossbar device.

The formation of narrow electronic filaments in oxide and phase change switches typically leads to current runaway and permanent changes in the device [13]. They can, however, be stabilized by adding a series resistance in the circuit path. This limits the total current that flows through the device in the filamented state and prevents atomic motion (details in Chapter 2; Fig. 4.1 repeated for reference). As shown in Fig. 4.2, if the forming step in RRAM devices is prevented from reaching completion (by means of using a current-limiting ballast), one can stabilize an electronic filament. One can also say that post resistance change after an incubation time

(defined in Chapter 2) has passed, the device in the ON-state will maintain the volatility of the ON-state till permanent atomic motion occurs. This time is referred to as the locking time – good memory switches have a small locking time and hence form immediately; good threshold switches/oscillators have theoretically infinite locking time, which makes the filament always transient. Repeated localization and delocalization of current in form of a transient filament results in oscillations, which will be investigated in this work.

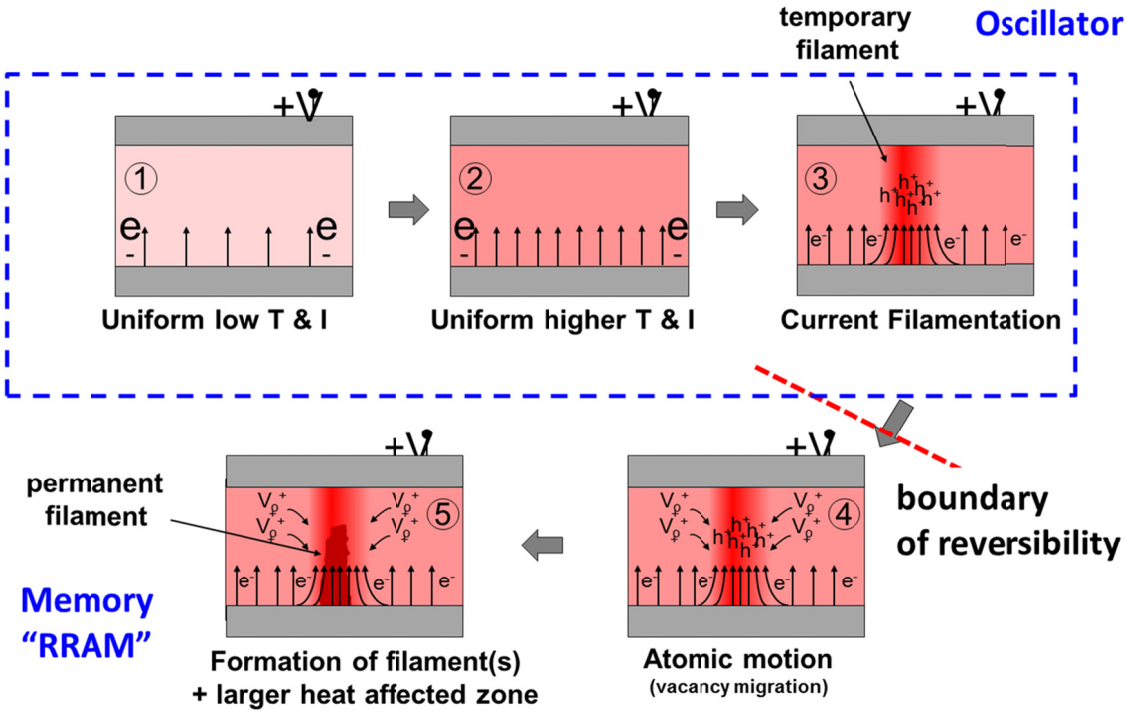


Fig. 4.2: Device undergoing forming process with a clear boundary of reversibility.

As the NDR cannot be stabilized without a series resistor, we use it to guide the device from the high resistance state to low resistance state. Unlike oxide based resistive memories (RRAM), the resistance changes are not permanent i.e. the high and the low resistance states are thought to be purely volatile, which prevents any drift in resistance. We use this bi-stable electronic nature of

the material in order to obtain self-sustained oscillations, with no external components other than the test setup.

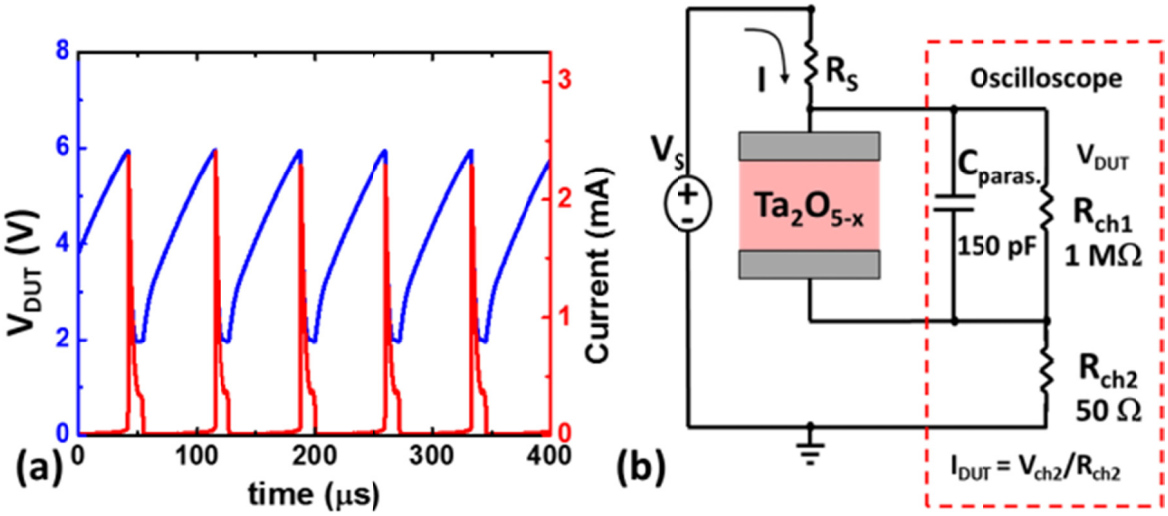


Fig. 4.3: (a) Self-sustained voltage and current oscillations at 12 kHz with a 1.2 kΩ resistor in series and the complete circuit with measured parasitics used to measure oscillations (b). Schematic of the oscillator with a source ballast and parasitics.

Figure 4.3(a) shows the voltage and current oscillations produced in the 5 μm device; Fig. 4.3(b) shows the equivalent circuit containing all of the measured parasitics loading the device. Here the subscript 'ch' refers to the channel of an oscilloscope.

4.3 Effect of Resistor and Transistor Ballast

We will first present a study that elucidates the effects of large R_S values in frequency tuning of these oscillators. The devices used in this study will be 5 µm TaO_x devices.

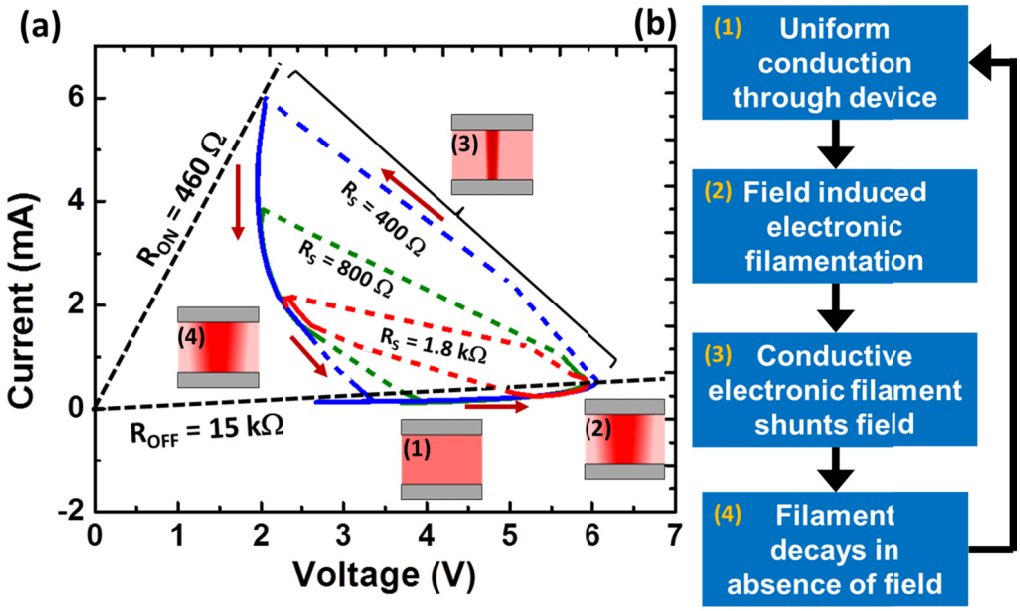


Fig. 4.4: (a) I-V phase portrait of devices during oscillations and (b) the behavioral block diagram.

In order to precisely understand the device behavior and the dynamics associated with these oscillations, we recast the data as a phase portrait in the I-V plane – Fig. 4.4(a). This represents the I-V trajectory of the device (with different series resistors) as it undergoes oscillations i.e. repeated filament formation and dissolution. Initially, the device is resistive at low biases and the current conduction in the device is uniform in state (1). As the device enters NDR, the current conduction becomes localized as it goes through state (2). At this point, the device resistance suddenly drops to a low value (state 3). This transition is abrupt, as seen in the falling edge of the

voltage waveform (rising edge of the current waveform) in Fig 4.3(a). This transition is thus markedly different from the thermally induced slow-NDR process, obtained through DC I-V measurement in the same device, as shown in Fig. 4.1(a). Thus, the same devices can either undergo thermally-induced NDR (Fig. 4.1(a)) or abrupt threshold switching process (Fig. 4.4(a)) depending on the applied V_S and R_S in the circuit. Note that the peak current levels exhibited under these two NDR conditions differ by an order of magnitude (600 μA in Fig. 4.1(a) vs 6 mA in Fig. 4.4(a)). The difference between the slow thermally-induced NDR and the abrupt threshold switching induced NDR is discussed in Chapter 2. This causes the voltage across the device to drop significantly as the voltage division with the series resistor evolves to a new operating point. The reduced voltage drop across the device, in turn, causes the electronic filament to become unstable and eventually dissolve through state (4) (Fig. 4.4(b)).

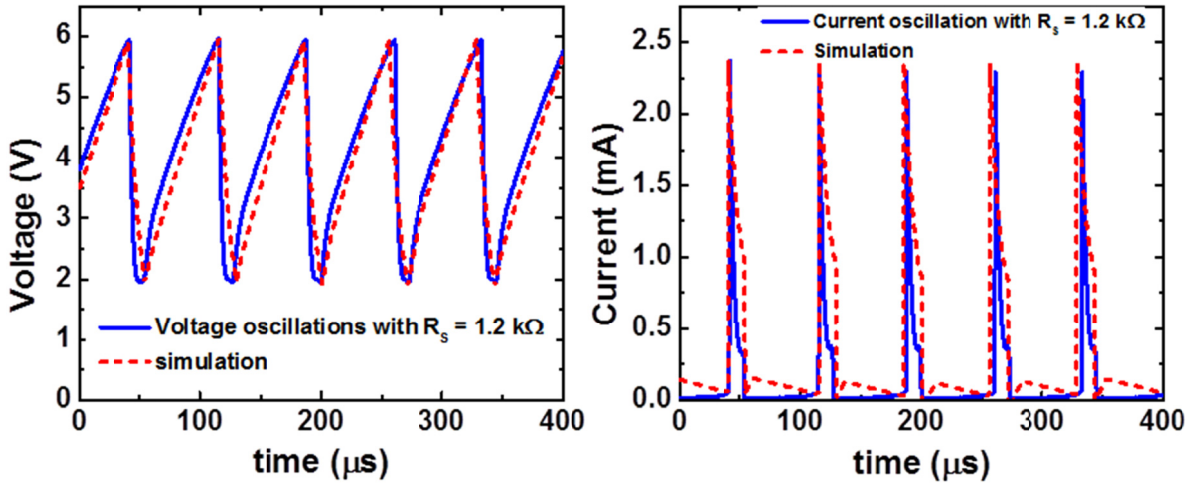


Fig. 4.5: Overlay of measured data (blue, solid) from voltage and current oscillations with a 1.2 kΩ resistor and simulation (red, dashed).

We used a reduced order model to represent this bi-stable behavior – using two linear resistors to approximate the high resistance (R_{OFF}) and low resistance (R_{ON}) states. The filamentation

dynamics that set the transition between these two states was modeled as triggered voltage thresholds having a sigmoidal transition between resistance states, with two distinct parameterized transition times. Figure 4.5 shows an overlay of the measured voltage and current oscillations and the simulation results. Chapter 5 will discuss a more detailed model. In the next section, we will describe how the devices operate under two distinct regimes: (1) Accelerated filament dynamics due to overdrive voltage when a large resistor ballast is used; (2) Parasitic dominated regime when a transistor is used as a ballast element.

When a ballast resistance is used in series with the device, the frequency can be tuned from 3 kHz to 500 kHz with a single series resistor as shown in Fig. 4.6(a). With increasing ballast resistance, a larger source voltage is needed to raise the potential of the device to threshold voltage and induce oscillations. In such cases, the frequency is found to increase as the series resistance is increased indicating that the frequency is not only controlled by the electrical parasitics, but also the acceleration of filament dynamics in the presence of a high overdrive voltage, as has been reported in other material systems [63],[66]. Overdrive voltage can be defined as the voltage that device experiences beyond the V_{th} (i.e. $|V_{app} - V_{th}|$). Sakai et al. [66] have presented a comprehensive summary of how large R_S results in biasing the R_S-device combination with a larger supply voltage, V_S, which in turn increases the frequency of oscillations. Thus, it appears that as the frequency increases with increased R_S. This may appear to be counterintuitive but it should be noted that the supply voltage needed to initiate oscillations ($V_{device} > V_{th}$) also increases with increased R_S. Conversely, if the V_{app} is held at a constant value and R_S is gradually increased, the frequency decreases, self-consistent with parasitic [63] and overdrive-dominated [66] oscillations.

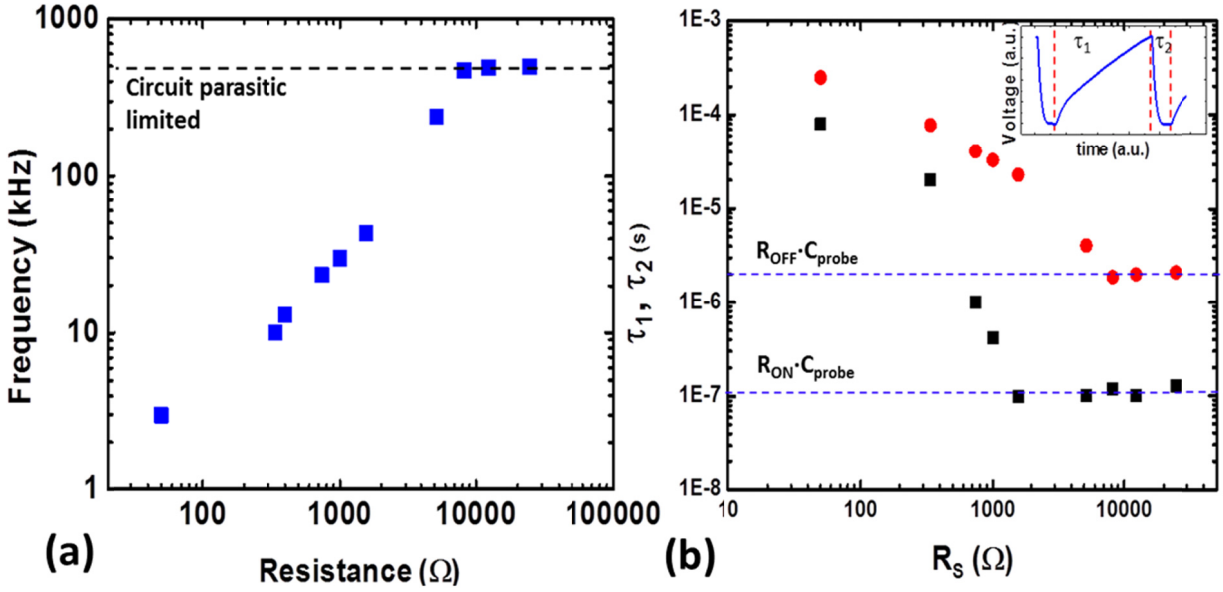


Fig. 4.6: (a) Frequency tuning using series resistors indicating an increase in frequency. (b) Both of the transition times saturate with circuit parasitics at high frequency. At low-frequency, charging times are dictated by filamentation dynamics. τ_1, τ_2 represent the rise and fall times of the oscillations.

The *maximum* frequency observed in our experiment is set by parasitics in the test setup. This indicates that the filamentation dynamics in presence of an overdrive voltage dictate the low frequencies obtained with smaller series resistance values. Figure 4.6(b) shows both time constants (rise time, τ_1; and fall time τ_2) of the device saturating to a constant value due to the circuit parasitics (as previously reported [63-65]). Interestingly, it is the filament formation dynamics in presence of appropriate R_S and V_S which gives rise to τ_1, and occur under conditions of uniform current conduction, rather than the filament collapse (τ_2) which set the frequency. The variation with series resistance suggests that the filament undergoes a more complete collapse (larger gap) and takes longer to reform under low series resistance, while only partially collapses and takes less time to reform under high series resistance, evidenced by the device returning to OFF state at higher voltage (indicated in Fig. 4.4 (a)).

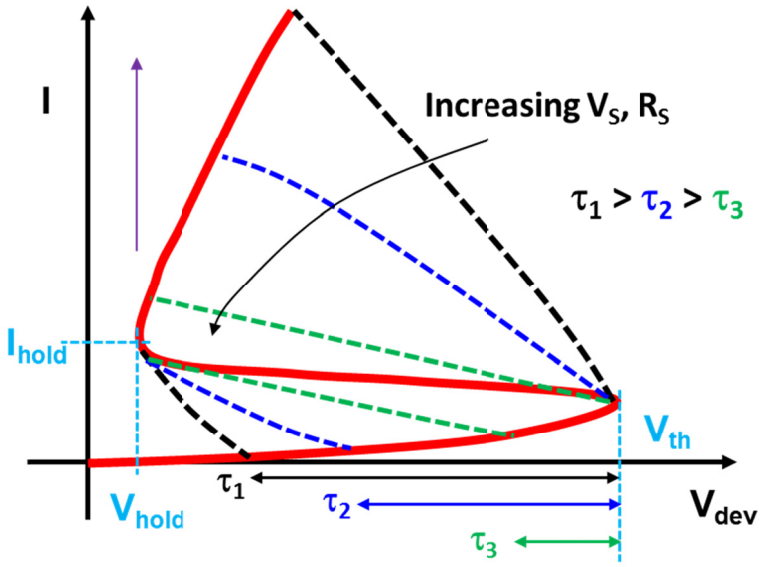


Fig. 4.7: Schematic showing the decrease in the rate limiting time constant as the R_S and consequently V_S is increased.

This is further clarified in Fig. 4.7, which has been schematically used to represent the data in Fig. 4.4(a). It indicates the change in the oscillation phase portrait as the resistance and the overdrive voltage increase. As the device goes from ON state to OFF state, it arrives at higher voltage on the OFF state branch. This effectively reduces the rate limiting charging time (RC) of the device as it approaches the threshold voltage (V_{th}). Moreover, for a higher R_S, the loop in the phase portrait nearly closes (Fig. 4.4(a), Fig. 4.7), resulting in a *saturation* of frequency (Fig. 4.6(b)) instead of a decrease with increasing R_S. In this regime, the frequency stops changing (becomes independent of R_S) for $R_S > 10$ kΩ. It is likely that the dynamics of filament formation and dissolution are being dominated by capacitance and parasitic resistance, at this stage. This indicates that the electro-thermal device dynamics are faster; and that overcoming this

shortcoming would lead to faster and more energy efficient oscillators. The above argument indicates that, if the parasitic capacitance due to the leads in our set up is eliminated, the frequency is modulated by the overdrive voltage which changes the device dynamics (in addition to being affected by the parasitics), which can unlock higher frequency and low-power operation at the same time. This is tied to previous reports that when the device is in the ON state, the maximum current determines the radius of the filament [6,67]. An increase in the overdrive voltage is estimated to cause the filament size to increase. Subsequently, the recovery time associated with a larger R_S (and consequently higher current and larger radius) is expected to be longer [6,67]. Thus, in the low-current, high-frequency regime, the oscillations would result in very narrow filaments, making the device highly scalable when used at high frequencies.

In order to look at the effect of scaling and non-linear ballast, we present oscillation dynamics in 700 nm TaO_x devices that are dominated by circuit parasitics, similar to previous works [63-65]. For this case, we use a transistor as a ballast which results in a decrease in frequency as the channel/output resistance is reduced. The transistor used is a PMOS and enables lower resistor values that can unlock higher frequencies without any additional need for overdrive voltage. Figure 4.8 shows a schematic of the transistor load-line intersecting with the device I-V.

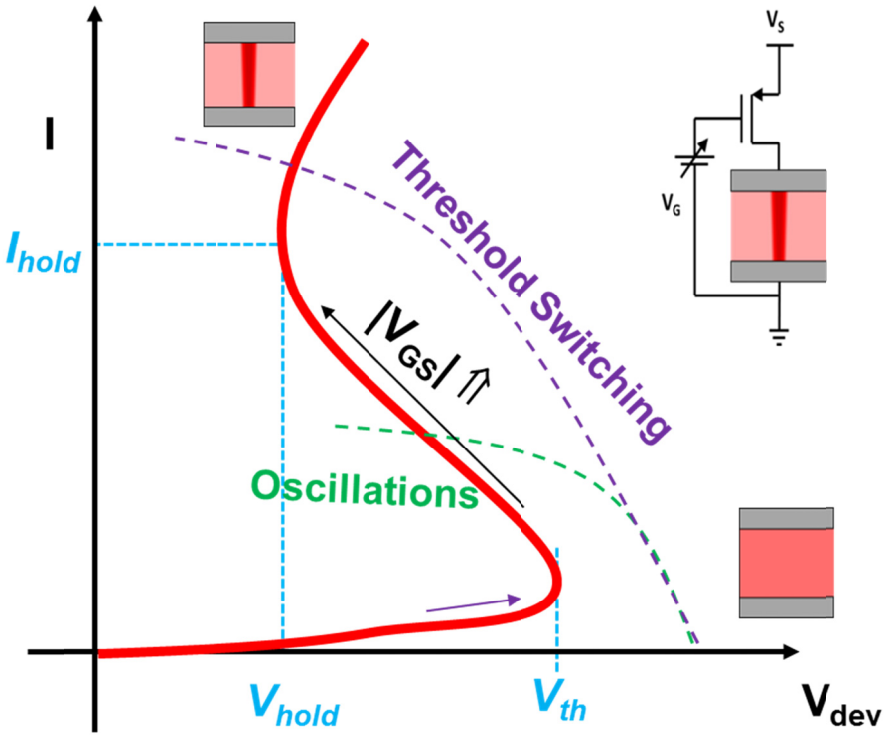


Fig. 4.8: Load-line of the transistor (green and violet), intersecting with the device I-V (red).

Figure 4.9 shows the circuit schematic of the PMOS-ballasted oscillator alongside *I-V* phase portrait of a device. It must be noted that unlike Fig. 6.4(a), the OFF state is completely traversed, irrespective of the gate voltage used. This is primarily because the transistor offers significantly lower output resistance than the series resistors. Thus, the effect of overdrive voltage (Sakai et al. [66]) is minimal compared to the effects of the parasitics. Thus, the filament is completely formed and dissolved for all V_G's considered.

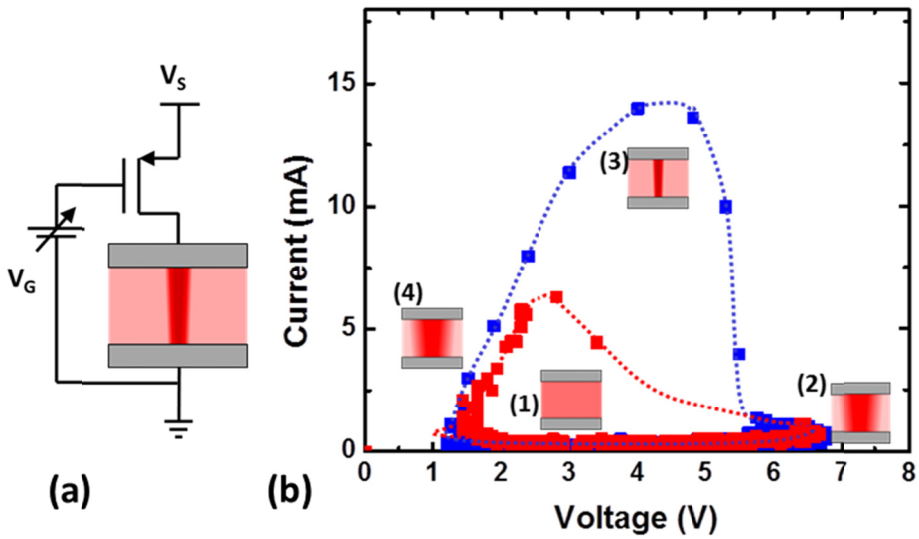


Fig. 4.9: (a) Circuit schematic and (b) *I-V* Phase portrait of oscillations in a 700 nm device with a PMOS ballast. Note that the OFF state is fully traversed for both cases, V_G = 4.2 V (red) and V_G = 4.5 V (blue).

In this configuration, an increase in the V_G (bias applied at gate) corresponds to a reduction in the V_{GS} (gate to source voltage) because the source is pulled up to the V_S rail in the PMOS. This causes the large-signal output resistance of the transistor to increase, which reduces the frequency of oscillations. Using PMOS ballast scheme, we are able to tune the frequency from 250 kHz to 250 MHz. Figure 4.10 shows the frequency control of the oscillator as a function of gate voltage. This enables the FSK scheme and potentially phase shift keying (PSK) scheme for ONNs where there is a need for compact coupled phase-locked loops (PLLs).

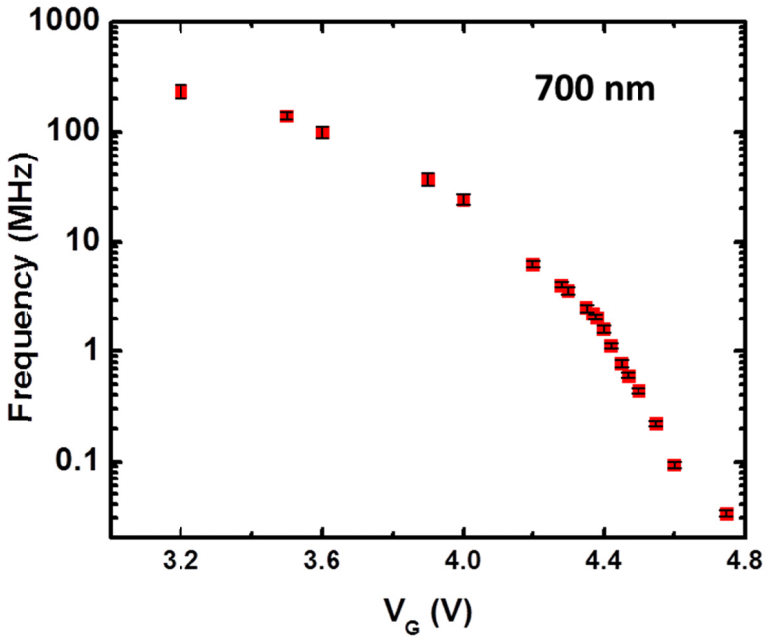


Fig. 4.10: Frequency tuning with a change in the gate voltage, V_G. Highest reported frequency of ~250 MHz reported with the 1T1R configuration. Error bars superimposed on the data points represent the spread for 15 devices.

One must note that while increasing the source voltage, V_S and the subsequent overdrive [66] was the cause of increasing frequency in the resistor ballast scheme, decreasing the V_G (increasing V_{GS}, absolute value) indeed causes an increase in frequency. This is because, increasing the V_{GS}, causes a larger current to flow through the circuit, charging the output node faster. This causes the oscillation frequency to increase.

4.4. Discussion on scalability, variability, failure modes and performance metrics

One of the most interesting observations has been the low ON-currents seen in the device when driven to high-frequencies – a behavior unique to these devices irrespective of the type of ballast used in series with the device. It has been long noted that the peak current in the ON state of the device dictates the size of the transient filament/nucleus ($I_{ON} \alpha r^2$) [6,63,64,68]. This implies that

higher frequencies require the device to operate such that the filament formed during the oscillations is small. This is very amenable to scaled implementations of the device as both frequency up-scaling and power downscaling are compatible. This is especially important in dense array-type implementations. The next chapter will shed significant light on methods of reducing power of this class of oscillators down to < 100 μW. Being relaxation oscillators, the peak power is dissipated in these oscillators during the capacitive discharging event that leads to a current spike. It must be noted that while the peak power in these oscillators ranges from 4 mW to 60 mW, the average power is significantly lower, ranging from 300μW to 3 mW. This is due to the asymmetric nature of oscillations – most of the oscillation period is spent in the device being in the OFF state with low power dissipation (Fig. 4.3(a)).

While oscillations have been stable, without permanently changing the resistance state of our device (i.e. forming), permanent filament formation has been previously reported as a failure mechanism for these type of oscillators [64,65]. However our oscillators show a different source of failure - abrupt ceasing of oscillations leading to a threshold switching event, without a permanent forming event. This usually occurs when the overdrive voltage is very high and can be mitigated by adjusting the source voltage. Detailed study of the failure modes and the persistence of ON state are needed to understand different failure modes. In all of our oscillations, we see the device repeatedly recover its original, pre-formed resistance state each time the device oscillates back to the OFF state. This suggests that the filamentation occurs without significant migration of oxygen vacancies that the forming process usually involves. Because these oscillations are not based on vacancy motion, the randomness in the initial distribution of vacancies due to process variations is likely to have minimal effect on the

oscillator, which is a positive for durability. Moreover, we chose TaO_x which is known to have very narrow distribution of forming and switching voltages with high endurance [69], to ensure minimum pre-forming variability in threshold-switching. For these oscillators, the oscillations last for > 10^6 cycles when operating at 3 mW average power (> 50 mW peak power), improving to ~10^7 cycles when the average power is 300 μW (peak power < 10 mW). This implies that significant improvement can be achieved with this class of oscillators when the peak current overshoot is minimized (typically by increasing the ballast). Prior to failure (by forming), the oscillators could undergo a change in frequency, especially for the high power modes of operation, as has been reported earlier [64]. Chapter 5 and 6 will show that presence of different load ballasts or drive elements can be used to provide a degenerative feedback mechanism that prevents a change in frequency, effectively locking it.

The primary reason for the oscillatory nature of these oxide stacks is negative differential resistance, which occurs due to a strong dependence of conductivity on temperature and current density. Mott transitions [62] are one of the examples of how this steep dependence of conductivity can be observed. While specific mechanisms of conductivity increase can vary, the overall behavior could be universal [6,29,67]. More experiments are needed to establish the universality of oscillatory behavior in other TMOs. While reliability is still a valid concern, circuit techniques could potentially mitigate failures arising from eventual forming by using feedback. More extensive investigation of such techniques is needed to validate this hypothesis. Moreover, it must be noted that if the same device is biased in the positive differential resistance regime of the ON-state, it behaves as a threshold switch. This understanding is also important from the point of view of the use of threshold switches in RRAM memory arrays. If an S-NDR device is biased inappropriately, it may start oscillating. This challenge becomes especially

exacerbated due to the distributed trace resistance of the word-lines and bit-lines of RRAM array. Thus the engineering requirements of an oscillator and threshold switch are fundamentally different. The holding voltage must be low for both oscillatory and threshold switching applications. However, threshold switches require the holding current to be as low as possible. This is to enable threshold switching instead of oscillations. On the other hand, a high holding current for oscillators implies the device has a wider selection of bias points for phase-frequency tunability.

Chapter 5

Engineering S-NDR Oscillators

5.1. Introduction to S-NDR Performance Metrics and Challenges

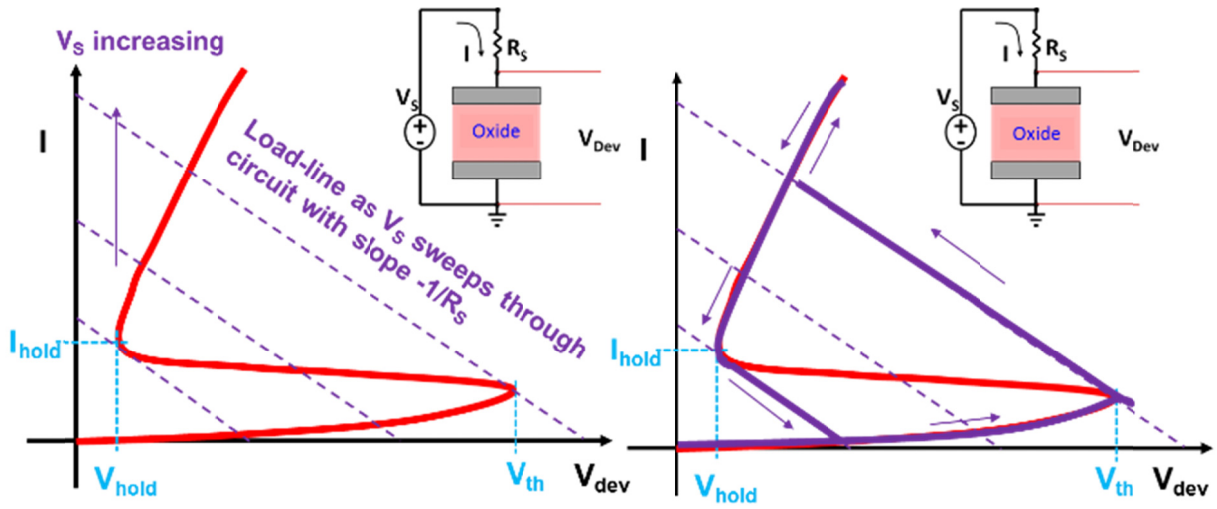


Fig. 5.1: *I-V* portrait of a threshold switch – intrinsic & in presence of load-line.

In Chapter 2, which studied the forming process in RRAM cells, we observed negative differential resistance (NDR) as a unique feature of the oxide devices. Any element that shows NDR can be appropriately ballasted and biased to give oscillations just like cross-coupled transistors. As discussed earlier, NDR gives the device a possibility to exhibit two stable states depending on how the load-line affects the characteristics. Let us take an example of an I-V plot shown in red, in Fig. 5.1. Let us assume that this plot with S-NDR (in red) represents the I-V characteristics of the device based on its material properties. One must note that these characteristics represent an I-V that consists of the device showing more of the device in the ON-state (as opposed to Fig. 4.1) i.e. akin to the observations made in Fig. 2.8(b). As this phenomenon has been explored for several decades (discussed in Chapters 2 and 4), there is a rich family of literature that tries to study oscillatory behaviour using a non-linear dynamical systems formalism [6]. Despite such detailed works, most of it is rarely cited due to the multi-

disciplinary nature of this area and lack of applications that needed such highly compact oscillatory units. It is now well-known that the circuit and the NDR element interact with one another to result in a self-sustained limit cycle that races a loop in the I-V trajectory. Several examples of such oscillators using chalcogenides, VO_x, NiO_x and NbO_x exist in literature. The frequency of these oscillators was found to decrease with increasing value of ballast resistor, implying that the frequency is limited by the charging time associated with charging the node shared by the ballast and the oscillator.

One of the challenges that these oscillators face include poor parameterization of physics associated with oscillations. For example, the threshold voltage (akin to forming voltage) has been stated to linearly depend on the thickness of the material and yet the holding voltage has been found to be relatively independent of thickness [70]. This has been attributed to three factors viz. (1) Electrical contact resistance, (2) Schottky barrier height at the electrode-semiconductor (or insulator) interface, (3) Resistance of the filament [6]. It must be noted that other effects associated with the filament field and temperature [72] also play a very important role in deciding the holding voltage and current values. Sustainable scaling of threshold voltage has to be accompanied by scaling of holding voltage to retain the large voltage swing as being the advantage of these oscillators. Thus careful engineering of the device is crucial.

This chapter builds upon the understanding from Chapter 4 and examines TaO_x-based compact oscillators (RRAM-type) as neuronal elements of an ONN compute block and experimentally demonstrates: (1) First ever 1T1R integrated structure, (2) Maximum frequency of ~500 MHz, > 2 orders of magnitude higher than reported in this class of oscillators, (3) Lowest reported power down to < 200 μW, one order of magnitude lower than best reported, (4) Full-system simulation

of an ONN-based associative memory (design and simulations carried out by Thomas C. Jackson).

5.2. Device Structure and Transistor Integration

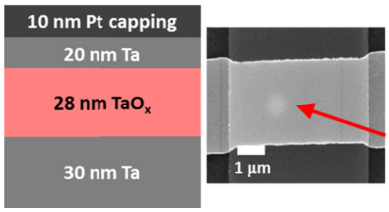


Fig. 5.2: Device & thermal footprint of a filament heated-zone in a typical TaO_x device

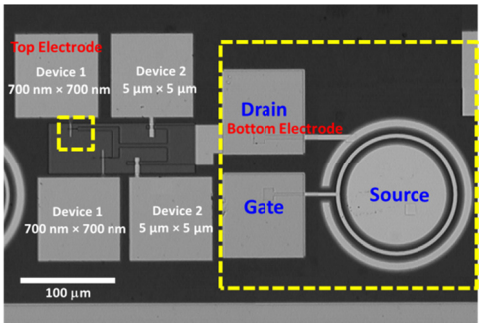


Fig. 5.3: Image of 1T1R structure

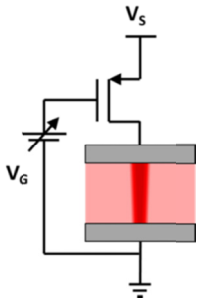


Fig. 5.4: Schematic of 1T1R

Fig. 5.2 shows a typical oscillator structure with an SEM of a device that has undergone filamentation. The structure consists of a MIM structure with Ta electrodes (choice of which is described in the following section). The SEM image, akin to Chapter 2, shows a heat affected zone on the top electrode that serves as a signature of local heating that would have occurred in the device during the oscillation process. These structures have been integrated on PMOS transistors as shown in Fig. 5.3, with a schematic in Fig. 5.4. It must be noted that the transistors have a circular geometry with a circular gate structure that is used to control the frequency. The source of the PMOS device is connected to a DC power source and the drain is connected to the MIM device. This enables the PMOS device to starve the oscillator of the minimum holding current. Moreover, being an integrated implementation, there is a lower displacement current (discharging from the node between the MIM and the transistor, unlike the configuration described in Chapter 4) which further makes the oscillations stable.

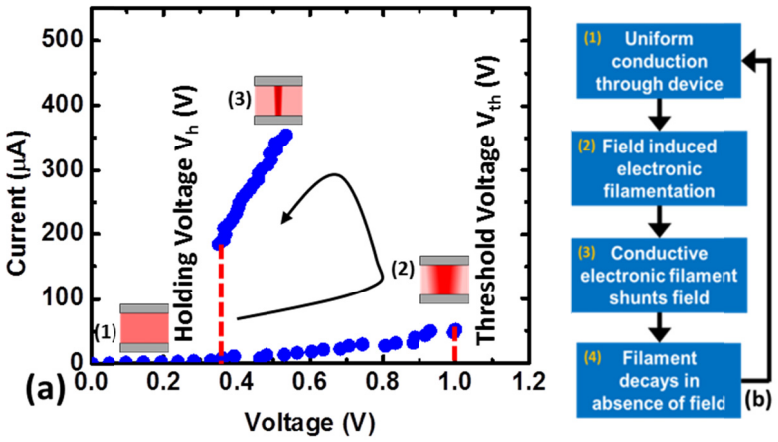


Fig. 5.5: Characteristics of devices in I-V plane as it undergoes filament formation and dissolution

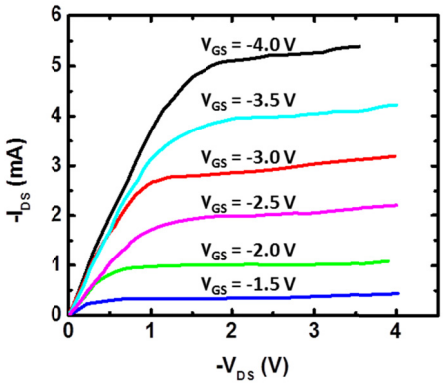


Fig. 5.6: PMOS I_D-V_{DS} characteristics. Overdesigned for high-drive and low-power

Fig. 5.5(a) shows the experimental *I-V* trajectory of a device, indicating the threshold (V_{th}) and holding voltages (V_h), which define the voltage swing. When the power is supplied to the oscillators, the voltage across them increases (while they are in OFF state) eventually crossing threshold at 1V which causes the nucleation of a temporary filament and the resistance and voltage to drop. The drop in voltage across the device causes the unstable filament to dissolve causing the device to go back to the OFF state, as shown in Fig. 5.5(b). Fig. 5.6 shows the I_{DS}-V_{DS} characteristics of the PMOS with a W/L of 22 designed for low power and high fan-out oscillators. The transistors were fabricated at Stanford University by Max Shulaker with the device integration of the threshold switches/oscillators completed at Carnegie Mellon.

5.3. Engineering Oscillators for Low-Power and High Performance

The key parameters that need to be engineered in these S-NDR oscillators are: (1) Threshold voltage, (2) Holding voltage and current, (3) Frequency of oscillations and (4) Scalability in area and power. As discussed in Chapter 2 and 4, the threshold switching is an electric field – mediated event and hence scaling down the thickness of the device should scale down the device threshold voltage. It must be noted that these devices do not show a first-fire threshold voltage that is greater than the oscillatory threshold voltage. Engineering of the oxide thickness and electrode work function enables the low-power operation of the device which results in low-power, large-voltage swing oscillations, as shown in Fig. 5.7.

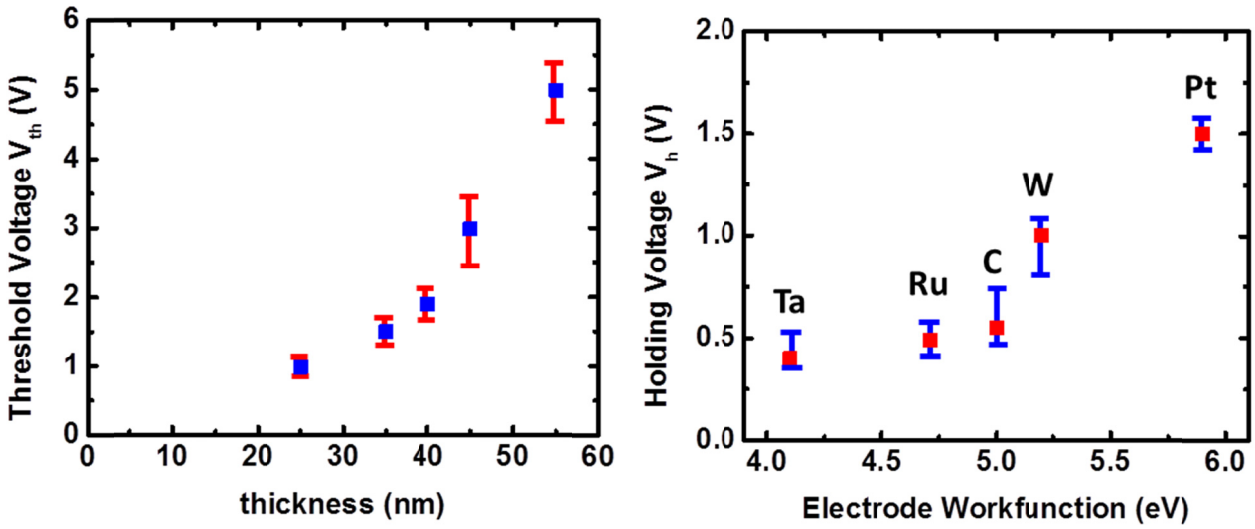


Fig. 5.7: (a) Nearly linear increase in the threshold voltage with increasing thickness, indicating a field effect. (b) The holding voltage also increases with increasing workfunction of the electrode material. The symbols represent a median value out of 10 measurements of different samples.

The work-function of the electrode material dictates the height of the potential barrier (Schottky) formed at the metal-oxide interface. This potential barrier induces a built-in voltage at the interface. This depends on the difference in the workfunction of the metal and the Fermi level of the oxide (Ta_2O_5 has a Fermi level at ~ 4.3 eV). This potential barrier will reduce the voltage at which the carriers stop conducting and the device reverts back to the OFF state – holding voltage.

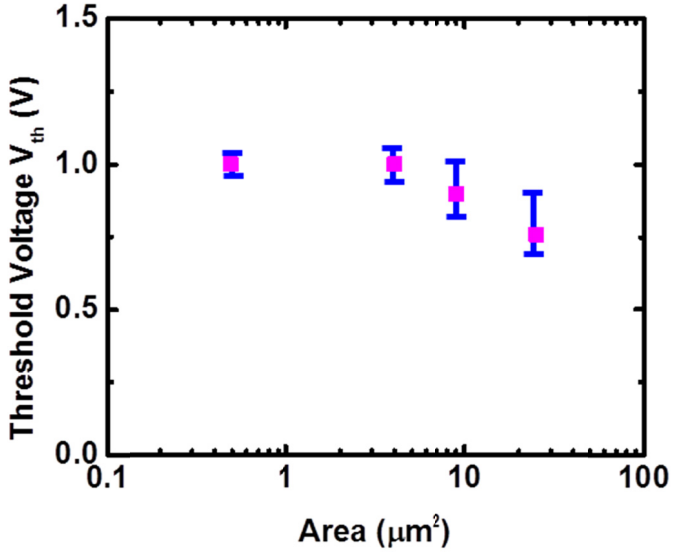


Fig. 5.8: Threshold voltage of the devices plotted as a function of lateral dimension scaling. The threshold voltage becomes nearly independent of lateral dimensions at small sizes.

As the microstructure of TaO_x was found to be amorphous (as discussed in Chapter 1), the nucleation sites (in form of defects) are distributed randomly (without any preferred locations such as grain-boundaries). Thus, the breakdown process in these materials, which perhaps is mediated by pre-existing conductive defects is retarded if the area is scaled below a certain size. This makes the threshold voltage nearly area independent, indicating a great potential for scaling in the sub-100 nm scale.

Based on this stack engineering, 700 nm × 700 nm crossbars made of 30 nm Ta as the bottom electrode with 28 nm of TaO_x as the bistable oxide with 20 nm of Ta as the top electrode, capped with 10 nm of Pt (Fig. 5.2) are used for performance benchmarking.

The oscillation frequency can be tuned up to 500 MHz by changing the channel conductance on the PMOS, as shown in Fig. 5.9.

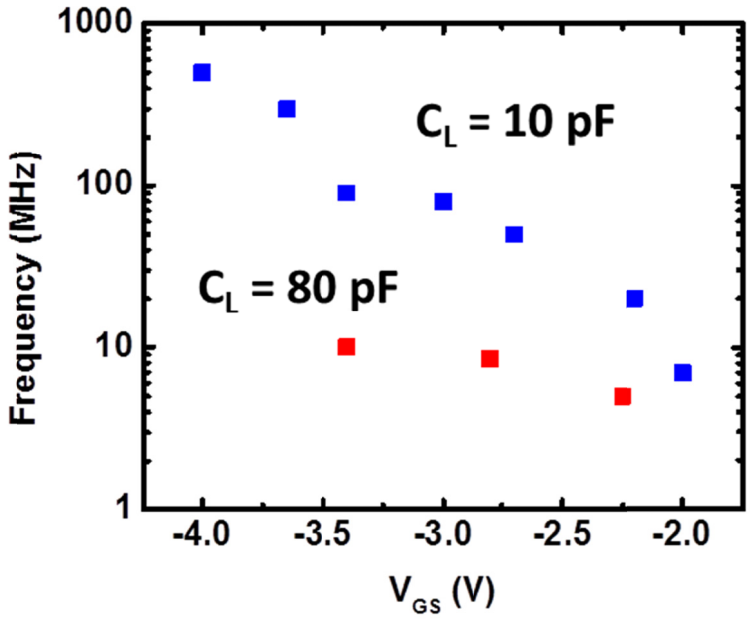


Fig. 5.9: Frequency tuning in 1T1R oscillators with (80 pF) and without (10 pF) external capacitive ballast.

The frequency is limited by the parasitics of the 1T1R configuration (as discussed in the previous chapters); an effect obvious in comparison between the oscillatory node loaded by < 10 pF (as fabricated) and 80 pF (after loading with a capacitor externally). This is consistent with the observations made in Chapter 4 which relate the frequency to the ballast and the parasitics. Thus, frequency in expected to increase with device scaling. The temperature coefficient of frequency

of these oscillators is < 1%/°C over 25 °C to 150 °C making them insensitive to on-chip temperatures.

5.4. Discussion

It has been observed in previous works on chalcogenides that the non-linear dynamics of the oscillators are very tightly coupled with the circuit parasitics that they are loaded with (Chapter 4 and [73]). Thus device-circuit co-optimization is essential to ensure that the devices are scalable alongside CMOS scaling and that they do not show fundamental limitations in being deployed as nano-primitives for ONNs. In this section, we will look at the effect of scaling and circuit parasitics on the oscillator dynamics.

It has also been observed that the threshold voltage associated with these oscillators is dependent on the number of nucleation sites existing in the insulator matrix for the temporary filamentation to occur, the thickness of the oxide/semiconductor film and the temperature at which these films operate as oscillators [74]-[76]. As the device area is scaled down, a weak increase in threshold voltage is often observed due to the reduction in the nucleation sites needed for the temporary filament to form. However, previous works in threshold switches [6],[75]-[76] have shown that scaled devices can be formed into a permanent secondary high-resistance state consisting of conductive phases, which in turn act as nucleation centers for the temporary filament during oscillations. It has also been shown that semiconducting glasses (both chalcogenides and oxides) have an almost linear dependence of threshold voltage on the thickness of the film. For threshold switches, sub-1 V switches have been developed with both oxides and chalcogenides. This implies that these devices can be scaled down to the minimum nucleation size (~3 nm) [77].

Thus, even integrated implementations of these oscillatory elements would be limited in area by the ballast and coupling transistor pitch. However, it must be noted that as the oscillators scale, they would have lower threshold voltages and peak currents. This would be consistent with the scaled transistors that would only be able handle limited power resulting in the temporary ON state of the device to be highly unstable (as I_{DSAT} would be lower than holding current for the device) and hence the device would oscillate at a higher frequency. Figure 5.10 shows frequency tuning in 100 nm TiO_x based crossbars (same electrodes, 20 nm TiO_x). It is noteworthy that the frequency of oscillations is of the same order of magnitude as TaO_x devices because of the oscillatory process being parasitic dominated.

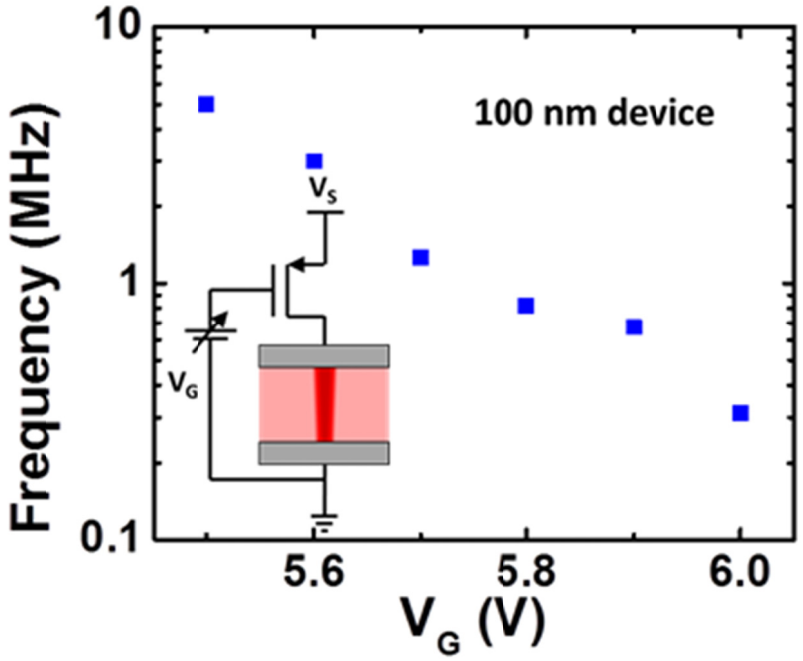


Fig. 5.10: Frequency tunings in a scaled 100 nm TiO_x oscillator.

Because the frequency is almost always limited by the device dynamics in the OFF state, the role of the parasitic capacitance charging at the oscillating node is crucial to ensure reliable oscillation phase portraits.

Shaw et al. [74] have shown that the maximum current after the device transitions to the ON state is governed by the displacement current associated with the parasitic node capacitance as well as the inductance of the system. Ideally, if the system has minimal capacitance, the oscillating nodes would be limited by the filament dynamics. In such cases, it is observed [74] that the oscillations have much higher frequencies which implies lower holding currents. Typically, holding current is a DC quantity that refers to a critical nucleus size that the filament has to reach, below which, the filament dissolves. However, transient ON Characteristics [71] show that the ON state would have much lower holding currents if the duration for which the nucleus exists is shorter. Thus higher frequencies which are possible with low parasitics would eventually correspond to a smaller duration for which the subcritical nucleus persists and this results in a lower holding current.

The nature of filament formation is considered to be an electronic or nucleation switching process process. In Appendix B, we will present some results showing NDR at cryogenic temperatures which indicates that the process should not involve motion of atoms, making the cycle to cycle variation minimal. Thus the same defect site serves as the initial nucleation site for the oscillations to initiate. However, the parasitic heating in presence of oxygen getters or presence of point defects with low activation energies [79] is the main cause of oscillation failure. In most cases, this problem is exacerbated due to the large discharge current that accompanies the onset of ON state. This discharge current is directly proportional to the

parasitics that load the oscillator and the change in the device voltage during the switching event. Thus, scaling down the device laterally (lower capacitance) and in power (lower dV/dt) should help reducing the variability and failure.

When oscillators are coupled, their coupling is expected to be far more robust as the synchronization is mainly dominated by the coupling element (as we will discuss in the next chapter). Hence, small changes in the individual frequencies can be compensated for as the frequency to which the system is locked to, is set by the coupling element.

As a mark of contribution of this work, Fig. 5.11 shows the improvement in performance and power achieved over the development time of this class of oscillators.

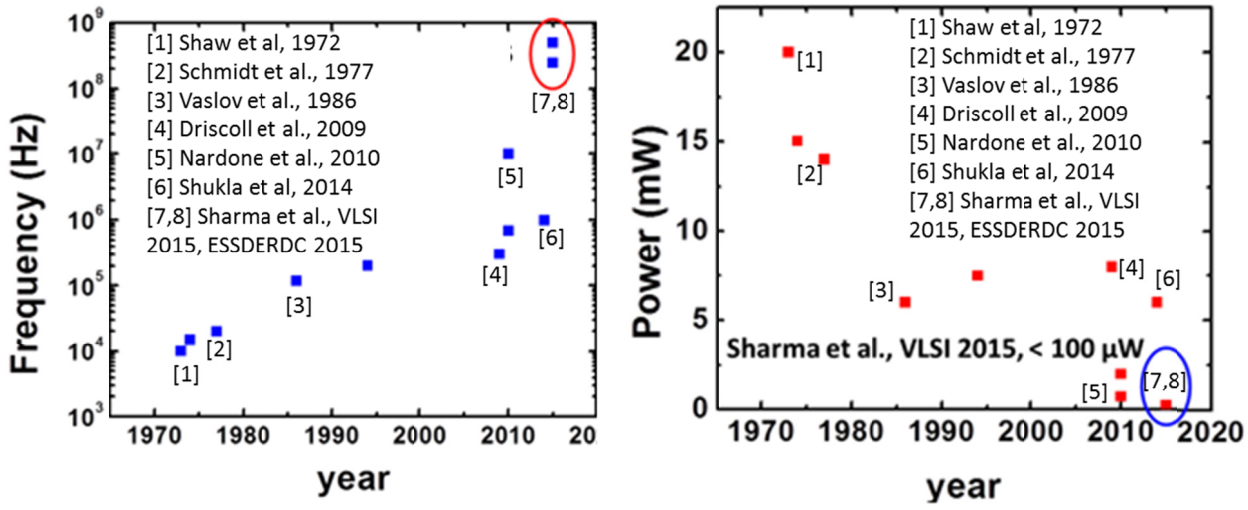


Fig. 5.11: Development of S-NDR based oscillators in terms of performance and power over time.

5.5. Oscillatory Neural Networks (ONNs)

Brain-inspired neurocomputing is considered as an emerging alternative to computing based on traditional techniques due to its massive parallelism. A neurocomputer attempts to mimic the human brain via a network of coupled artificial neurons that process information in parallel. Each brain neuron represents a computational unit in a neural network and a connection between two such neurons represents is known as a synapse. The strength of this connection, the synapse is in form of a synaptic weight in a neural network that relates one artificial neuron to another. Traditional computing schemes (that utilize the von Neumann architecture) run a software algorithm for a specific application by sequentially executing each line in the instruction code. Even though each execution might take a very short time the overall computation efficiency is not that high due to the serial execution of instructions. Instead, a neural network performs pattern recognition via associative memory in a massively-parallel manner. It maps a set of input patterns to a set of output patterns via synaptic weights, whereby an output pattern can be retrieved for a given initial pattern. Graphical applications would otherwise require numerous memory fetch operations and a processor that is executing a list of commands for optimization. Oscillatory neural networks (ONN) are one such example of phase-based neurocomputing, in which the state variable is represented by the *phase of an oscillator*. But each neuron requires a voltage or current controlled oscillator and a means of programming the phase relationship among all neurons to represent the stored information. CMOS voltage-controlled oscillators (VCOs) are theoretically viable to represent oscillations, but completely impractical from an energy standpoint. Moreover, implementing the phase relationships among the VCOs is even more inefficient. This implies that there is a need to explore directly coupled ONNs as proposed

by Nikonov et al [59] as they offer a more compact and energy efficient implementation. In this section, we will utilize NDR exhibited by S-NDR devices in order to implement oscillatory units.

Oscillatory neural networks (ONNs) composed of coupled phase-locked loops (PLLs) were proposed by F. Hoppensteadt and E. Izhikevich in [60]. In an this style of ONN, a PLL acts as the "neuron," integrating and storing the state of the system as its phase while connections act as the "synapses," or the weighted influence of one neuron on another. It has been shown that in this network, the neurons all synchronize to the same frequency and that their relative phases settle to a pattern stored in the network.

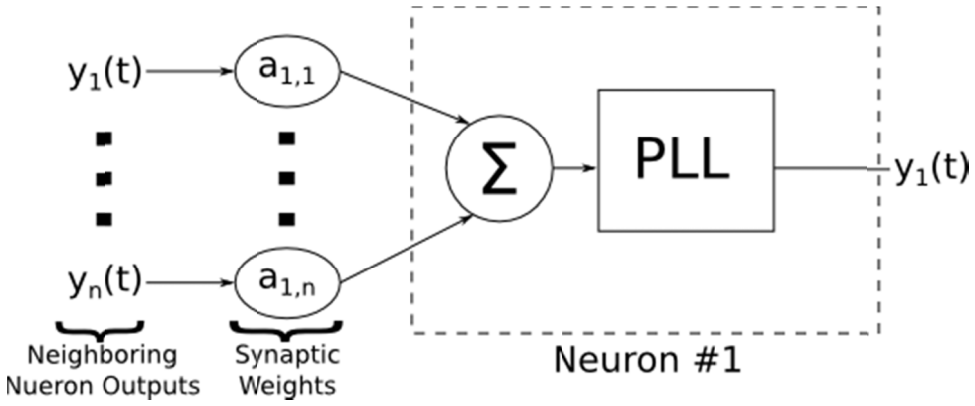


Fig. 5.12: A conceptual diagram of one cell of an Oscillatory Neural Network (ONN). The information in the system is stored as the phase of the output signals of each of the PLLs.

Many ONNs have been proposed and simulated, but a combination of architecture and technology that is scalable and reprogrammable has not yet been proposed. For example, in a system using lasers as neurons and holographic interconnect as synapses is proposed and

simulated in [78], but significant implementation challenges remain before such a system would be feasible, including the fabrication of holographic interconnect.

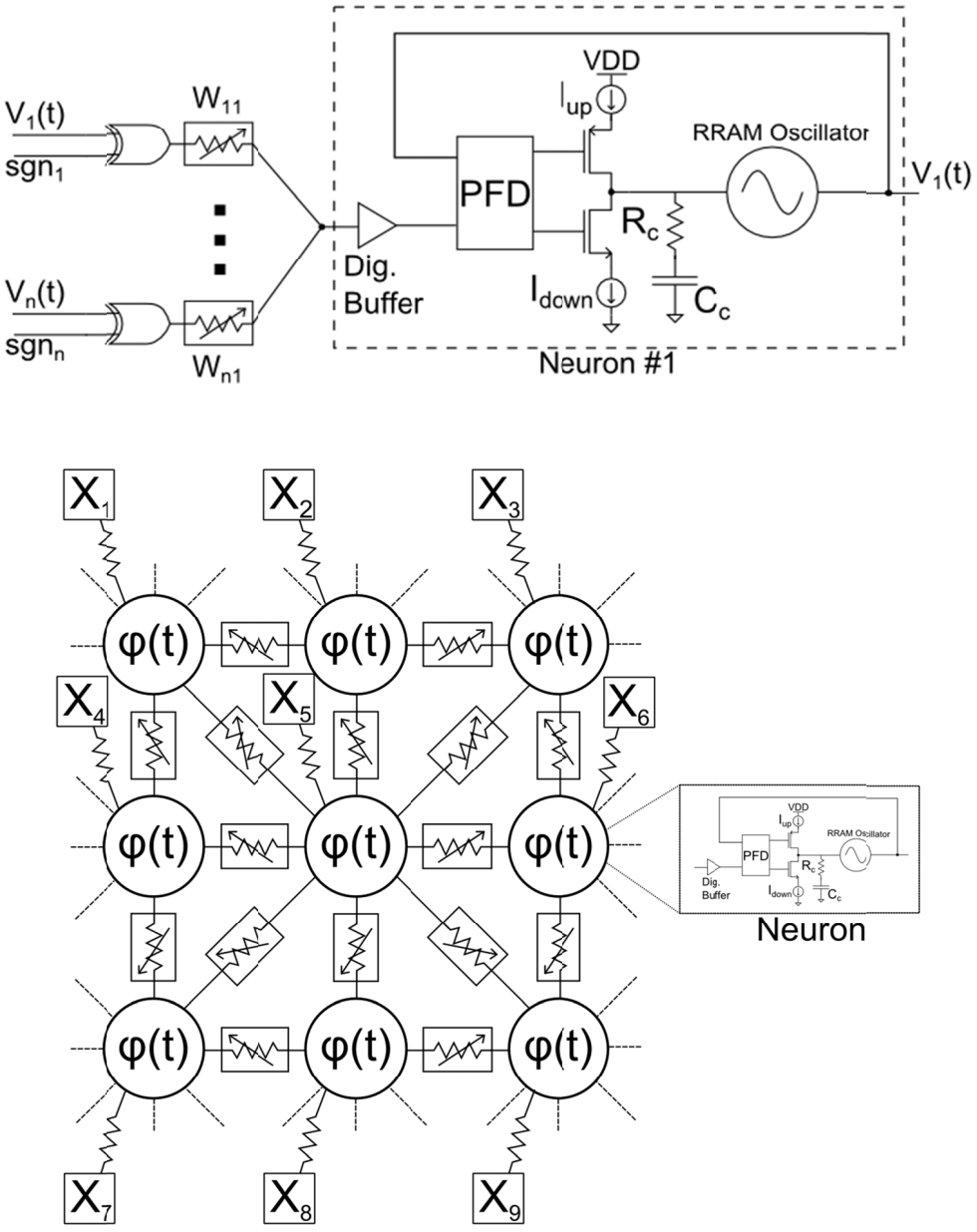
A system with a more feasible implementation was proposed and simulated in [80]. This system uses MEMS-based oscillators as neurons and variable electronic connections between them. Although the technology exists to build these systems, there is no strong evidence that such a system will be able to scale in area and power due to limitations in the mechanical oscillators.

An implementation using more traditional circuit elements was proposed and built in [81]. The neurons in this network are built around van der Pol oscillators, and the connectivity is achieved through a variable impedance. The example system in this paper is constructed from discrete components, and translating it to an efficient deeply-scaled CMOS circuit would be particularly challenging due to the inclusion of inductors in each oscillator and the need for active analog circuitry to provide the needed negative impedances.

ONNs based around spin torque oscillators (STOs) have recently been proposed and built in [61], and these systems are capable of scaling well in terms of area and power. The only demonstrated systems thus far have not been reprogrammable, as the coupling strength is based on the physical distance between the oscillators. Additionally, STOs have a voltage swing in the mV range which makes the design of interface circuitry very difficult.

The architecture and technology proposed in this paper is scalable and programmable. The neurons are represented by RRAM-based oscillators that can scale deeply in area and power while still operating at voltages easy to use with CMOS circuits. The synaptic connections use the same physical RRAM component, therefore they scale similarly in area and power. A single

neuron and its synaptic connections are shown in Fig. 5.13(a), while multiple neurons connected in two different configurations are shown in Fig. 5.13(b) and (c).



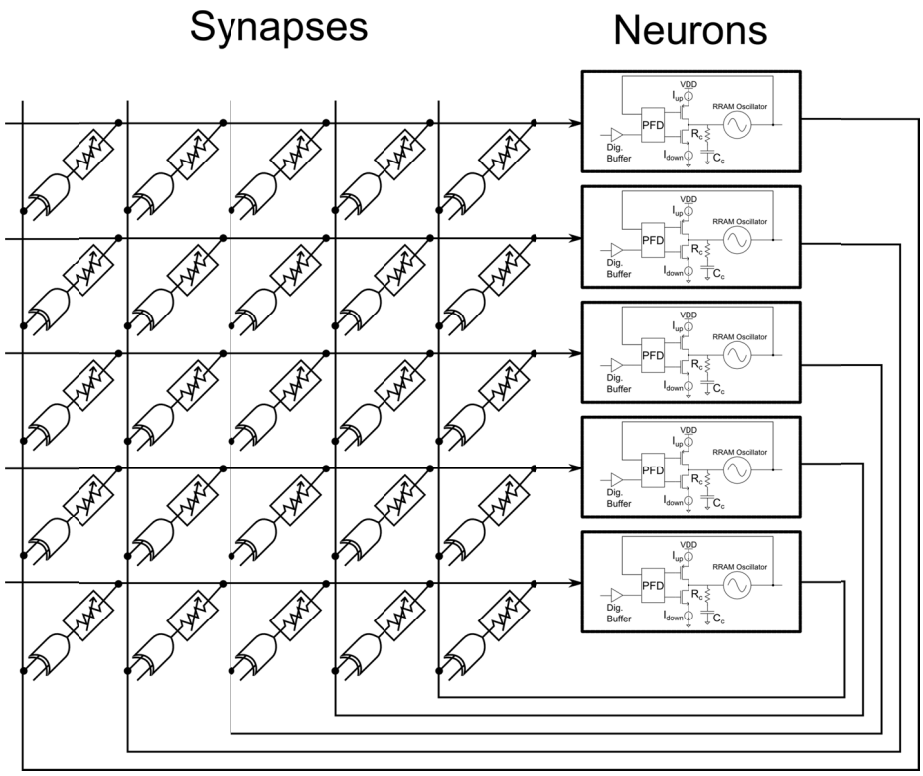


Fig. 5.13: The RRAM-based ONN. (a) A single neuron with its synaptic connections. The phase of $V_1(t)$ is the state of the network, and the resistance values of W_{11} to W_{n1} are the values of the synaptic weights. (b) An example configuration of neurons in a cellular connectivity pattern. The details of the neurons and synapses are abstracted away to show connectivity. X_1 to X_9 represent inputs to the system. (c) An example configuration of the neurons in a fully-connected pattern. A cross-bar array is used to achieve high synapse density.

The synaptic connections in an ONN require analog programmability, and therefore in a pure CMOS implementation a DAC would be required for each unique weight, which is infeasible in terms of power and area scaling. A VCO in CMOS designed to compete with the power and area scaling of the RRAM oscillators would be at a frequency so high that higher performance (and therefore power) digital circuitry would be needed. Unlike with a CMOS-only implementation,

in this architecture, the analog components are implemented by RRAM devices so the remaining CMOS circuitry is primarily digital and therefore able to take full advantage of process scaling.

The oscillators fabricated in this chapter were used to create a resistance switching VerilogA model which was used to simulate the operation of an ONN constructed with these 1T1R voltage controlled oscillators (VCOs). The behavior of relaxation oscillations can be modeled within the framework defined by Balthasar van der Pol [82]. Here, any S-NDR device can exhibit self-sustained relaxation oscillations due to the nonlinear damping once it is integrated into the second order circuit given in Fig. 5.14 (a) [83]. This model was implemented for these our oscillators in collaboration with Yunus Kesim.

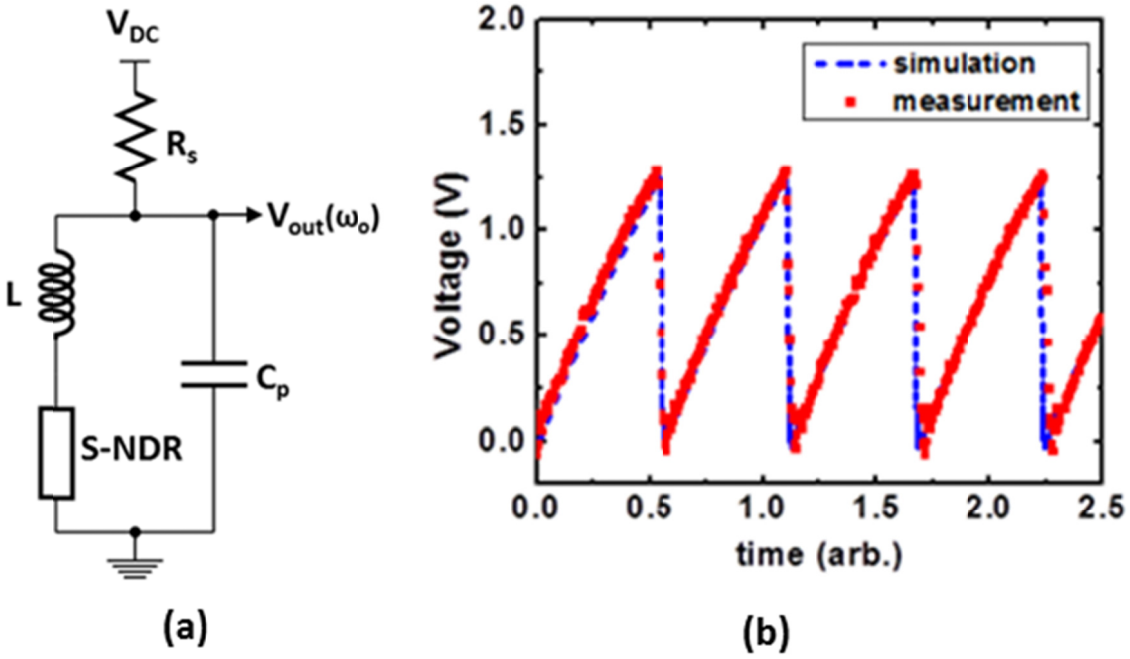


Fig. 5.14: (a) van der Pol circuit model and (b) overlay of simulation (blue) and measurement waveform of voltage oscillations.

The capacitance (C_p) represents the parasitic capacitance and the inductor (L) represents the persistent behavior of the filament. S-NDR block represents the DC behavior of the stack. For this system to oscillate, the load line defined by the DC voltage and the ballast resistor should cross the I-V graph in the NDR region [83], discussed in the previous chapter. Here, the origin of C_p can be considered to be parasitic. However, the inductance (L) simply represents the sweep-dependent transient nature of filament. This is better explained in Figure 5.15(a) which shows the circuit schematic of a TaO_x device connected in series with a resistor.

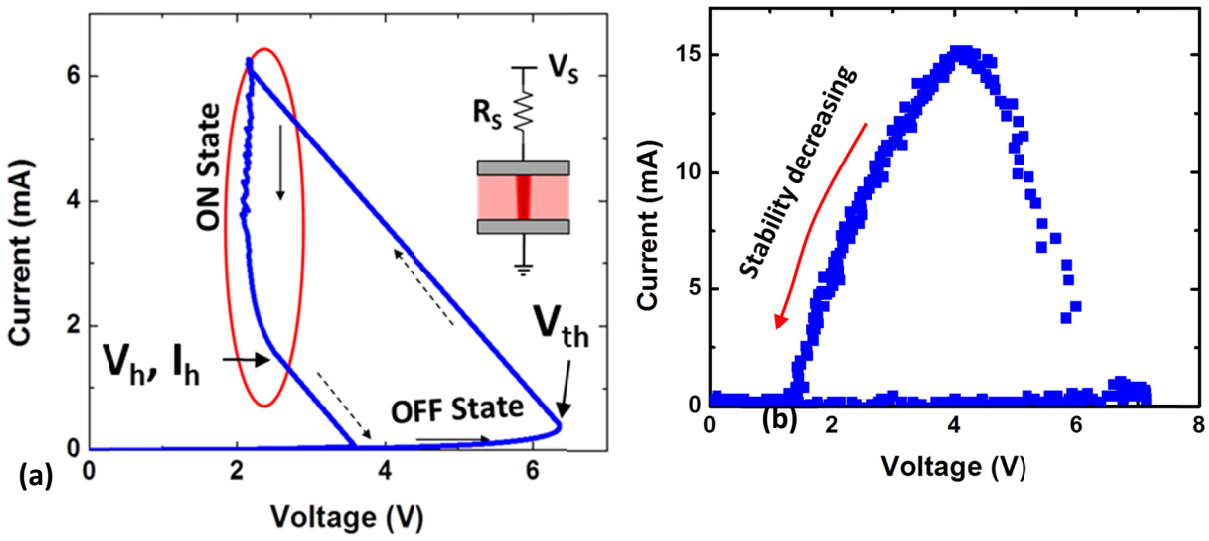


Fig. 5.15: *I-V* characteristics of an oxide device showing metastable ON state when stressed with a triangular pulse with a ramp-rate of 0.1 V/ms (a) and 1 V/µs (b).

As the bias across the device-resistance pair is slowly increased (0.1 V/ms, triangular pulse), the current through the device increases and eventually, at a threshold voltage, the device enters into the negative differential resistance regime. This implies that the device forms a conductive filament as it enters NDR and this abrupt reduction in resistance is responsible for the differential resistance becoming negative. Depending on the overdrive-voltage (differential voltage beyond

the threshold voltage) applied to the device, the device may settle down to various low resistance states, or ON states. The ON state is completely volatile (corresponding to a volatile filament) and the device will revert back to the OFF state (filament dissolved) once the voltage is removed. The voltage and current associated with this reversal is designated as the "holding voltage" (V_h) and the "holding current" (I_h). However, if the triangular pulse is significantly faster (1 V/μs), a device in ON state does not return back to the OFF state. Figure 5.15(b) shows the persistence of ON state, as previously seen in transient ON-Characteristics (TONC) [71]. This behavior is a representative of an inductor as the device cannot change its resistance fast enough if the voltage ramp is faster than the relaxation time.

For simulation purposes, the circuitry given on Fig. 5.14 (a) is created in SPICE. In order to import the DC I-V relationship into the simulation and create the S-NDR block, a Verilog-A model is developed in which the piece-wise linear fit is used. This simulation setup can effectively recreate the simulation behavior of S-NDR type oscillators when the correct component parameters are chosen. The ballast resistor and the DC supply voltage are already known. The parasitic capacitance, C_p, can be fitted based on the frequency of the oscillations. The inductor, L can be adjusted such that the amplitude of the oscillations matches the experimental data, as it is an indicative of the nature of TONC that the material exhibits. Fig. 5.14 (b) shows an overlay of the oscillation waveforms of the simulation and the experimental results.

The frequency of oscillations is predominantly determined by the parasitic capacitance, C_p. In previous chapters, oscillations at the frequency of 600 MHz were demonstrated when the ballast device was integrated on chip for reduced parasitics. In our simulation the values of C_p are used

from the experimentally measured value of ~10 pF. The remaining chapter will describe the network behavior of the proposed ONN. For more details, the full-paper by T. C. Jackson, A. A. Sharma [84] must be referred to.

The system was tested as an associative memory, and an 8 neuron network was tested. The test patterns and results are shown in Fig. 5.16 and 5.17.

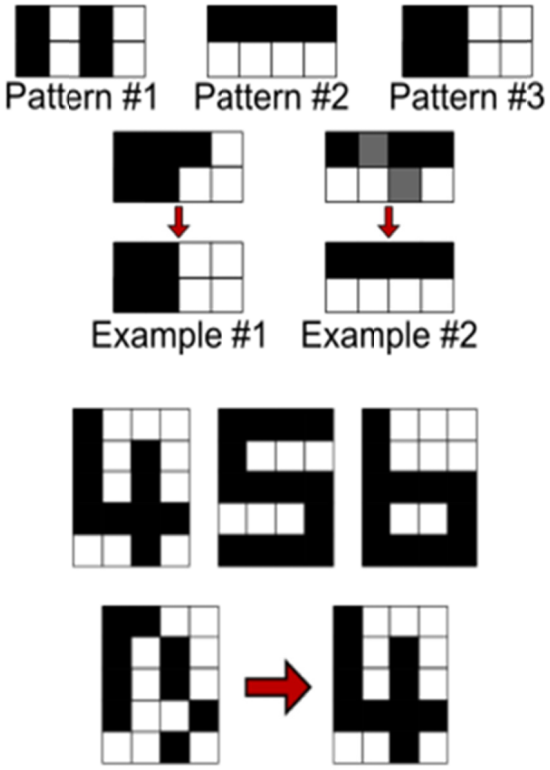


Fig. 5.16: Pattern recognition by associative memory

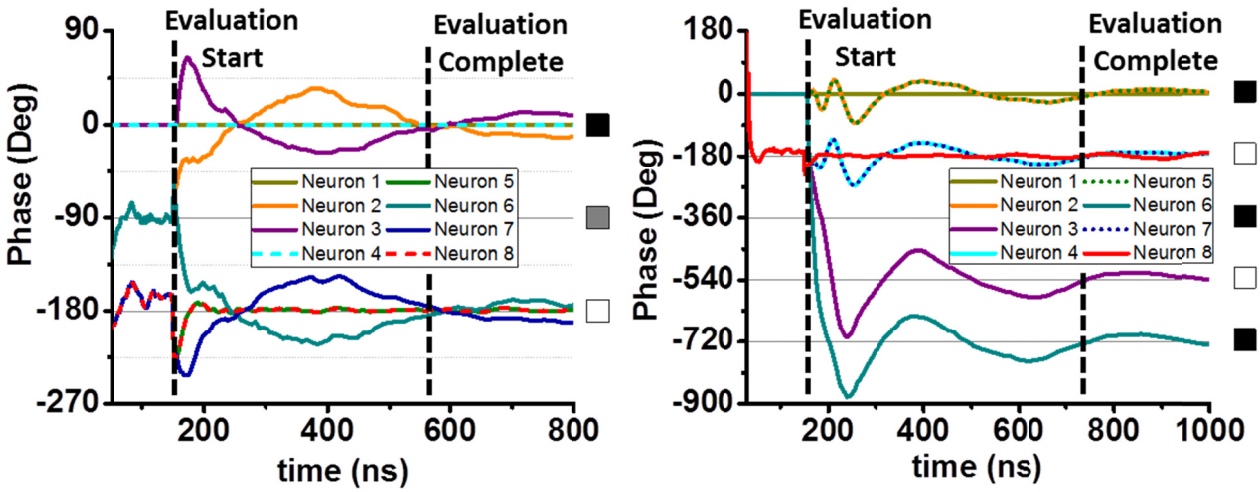


Fig. 5.17: Settling times for two patterns in an 8 neuron oscillatory neural network simulated with a VerilogA model of the measured device.

Due to the high frequency operation of these oscillators, the network is able to converge with a smaller settling time than published devices. However, it must be noted that the CMOS circuitry around the VCO could be prohibitive in terms of both power and area.

In this chapter, we discussed how these 1T1R oscillators can be tuned in order to be CMOS compatible. Following this, we demonstrated an application of these oscillators for oscillatory neural networks. To address the challenges posed in the previous section, next chapter will discuss applications of these 1T1R oscillators that are amenable to dense 4F2 array implementations for highly-parallelized directly coupled network-based computation.

Chapter 6

S-NDR Oscillators: Network Applications

As traditional CMOS device scaling for von Neumann architectures is nearing limits with device behavior becoming increasingly stochastic, continued increases in the computational ability per unit area and power cannot be sustained. Hence researchers have started looking for alternative computational schemes. Neuromorphic computing, often simply rendered as neurocomputing is one such paradigm that tries to mimic the same level of superior error-resilient compute, power and area efficiency as the human brain. This is specifically targeted towards a class of problems associated with pattern recognition that existing architectures are inefficient in handing with the same level of speed and power as the brain. Moreover, CMOS implementations of such neuromorphic systems have failed to achieve level of parallelism and energy efficiency exhibited by the brain [85],[86]. Hence, there is an increasing interest in looking at emerging devices that can enable some of the analog functionality needed for neuromorphic systems. Several implementations of neural networks have been proposed such as cellular neural networks (CNNs) which use multi-level resistance states exhibited by resistive random access memory (RRAM) cells or phase change memory (PCM) cells, as the state variable [87]. Likewise, oscillatory neural networks (ONNs) use oscillator arrays that consist of coupled oscillators in order to implement the parallelism, with each node representing a vector element that will be coupled with their neighbors [88]-[89].

Nikonov et al. [59] have reviewed and proposed several oscillatory neural network topologies that are broadly classified by the quantity that they use as a state variable namely, frequency and phase. The architectures that use frequency rely on frequency shifts or frequency shift keying (FSK) to match patterns. In FSK, each element oscillated at a frequency that is directly proportional to the vector distance between individual elements of the test vector and the memorized vector. All of the oscillators have the same coupling and an averaging block is used

to detect the closest match by averaging the distances between each vector element. In FSK-based systems, there is a clear need for frequency tuning over a large spectrum. Prior work on some of these emerging oscillators has shown limited tuning range of ~10 kHz to ~2 MHz. This not only limits the resolution of image processing (due to small margins between distinguishable frequencies), but also performance limitations due to its low-frequency operation (2 MHz being the maximum reported [16]).

A more common approach to implement ONNs is using *phase* as the state variable. In such coupled oscillator arrays [59], *phase* is the state variable that carries out the computation, encoding the information by phase shift keying (PSK). Furthermore, two distinct topologies of PSK-based systems have been proposed in prior work [59] – (1) an indirectly coupled ONN in which each oscillating node (neuron) is coupled to its neighbor using indirect coupling units (synapses); (2) star-coupled ONN in which each node is directly coupled to its neighbor through a coupling element with variable strength to control the phase between two oscillatory elements. Thus, a test vector has to compare the phase with all memorized vectors each with different set of coupling coefficients. Again, an average eventually activates the correct recognition of pattern. The star-coupled or directly coupled ONN offers distinctive advantages in terms of compactness, ease of integration in the back-end of the line (BEOL), and reduced circuit complexity. Moreover, CMOS-based indirectly-coupled ONNs use Phase-locked Loops to build oscillator arrays and tend to be power and area inefficient. This has led to investigation of compact oscillatory systems which can be used in ONNs. This makes the star-coupled ONNs a significantly more attractive implementation compared to indirectly coupled ONNs.

In the previous chapters, we developed an understanding to tune the oscillator parameters, making it amenable to scaled implementations. In this chapter, we will focus on: (1) Phase coupling and control of oscillator pair. This will help us gain a better understanding of the oscillator dynamics when loaded with another oscillator. (2) Injection locking of oscillators to enable variability tolerant phase initialization. (3) Application of S-NDR oscillators for edge detection (in collaboration with Yunus E. Kesim). (4) Stereo Vision using directly coupled networks.

6.1. Phase Coupling and Control of S-NDR Oscillators

To date, phase manipulation of compact electronic oscillators has proven to be a challenging problem. While gyromagnetic spin torque oscillators and VO_2-based oscillators have shown promise, their applicability has been limited to a binary phase contrast when coupled. Most image processing applications rely on gray-scale processing (and by extension, color) [62], motivating a need for fine grain phase-control (i.e. gray-scale levels). To date, fine-grain frequency and phase control have not been achieved, motivating our work to explore this emerging class of oscillators with a goal of implementing them in the star-type directly coupled configuration.

The desirable features of oscillators for oscillatory networks include: (1) phase and frequency coupling, (2) fine-grain control over these state variables, and (3) potential for dense, scalable arrays. In order to understand the mechanism of coupling, we first explore the coupling of two oscillators. The physics of coupling is expected to give us deep insights into a multi-oscillator coupled network. Moreover, such fine-grain control over both frequency and phase serve as the

desirable characteristics of oscillators in directly coupled FSK and PSK-based neuromorphic systems.

If we start oscillations in two different oscillatory elements (A & B), at nominally the same frequency, they oscillate with arbitrary phase due to variability in the time it takes to start oscillations. Moreover, their actual frequencies will differ due to variation in either the oscillator or the transconductance of the MOSFET in series. Fig. 6.1 shows voltage oscillations of two such uncoupled devices. In contrast, if the oscillating nodes are shorted, as shown in Fig. 6.2, the frequency and phase are both completely locked.

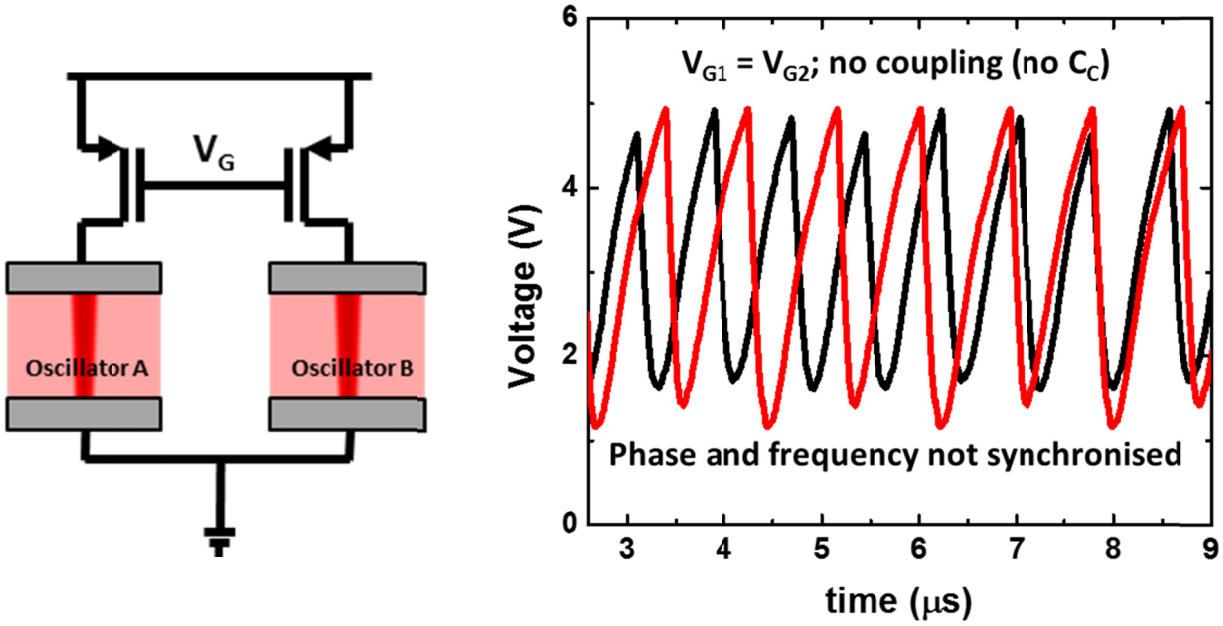


Fig. 6.1: Circuit schematic and oscillation waveforms of two uncoupled oscillators with same ballast V_G. Frequency and phase uncoupled because of drift that is observed due to device to device variations.

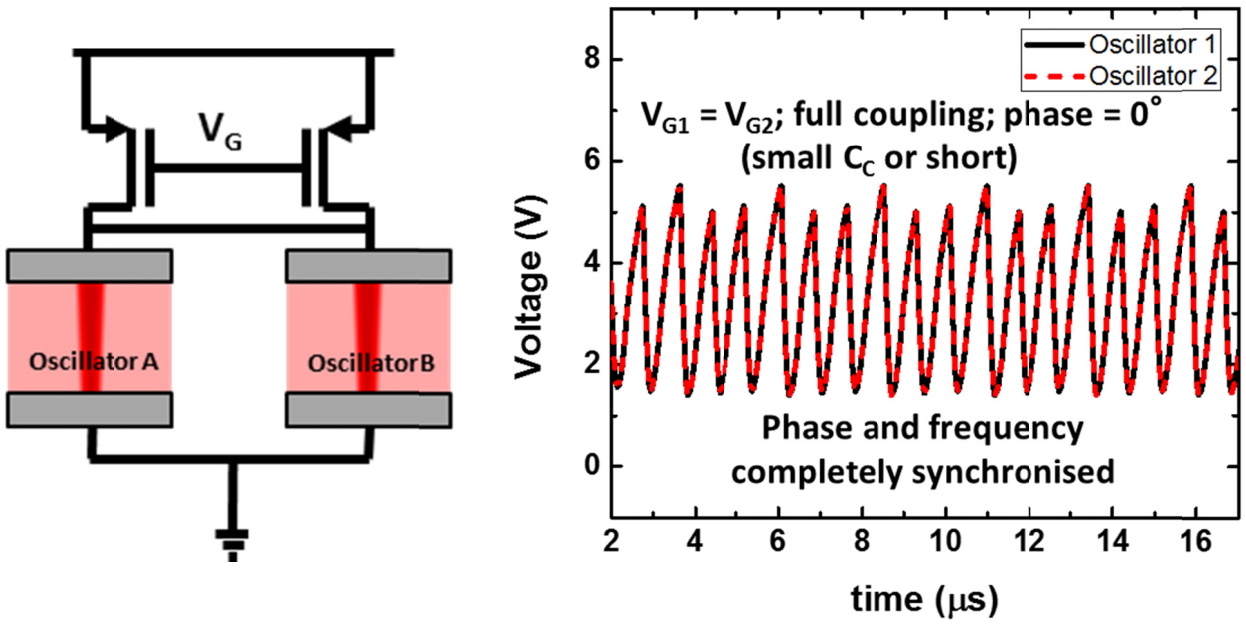


Fig. 6.2: Circuit schematic and oscillations in a fully-coupled oscillator showing complete phase and frequency locking.

In order to controllably phase-couple two free-running oscillators, we use a simple capacitor between their oscillating nodes. A circuit schematic of this configuration is shown in Figure 6.3(a). The capacitive coupling element acts as a high-pass filter and thus, a current would flow through the coupling element every time a transition takes place. Figure 6.3(b) shows 180º phase difference between the voltage transients on the coupled pair of oscillators (black and red respectively).

We know that for a single element (uncoupled), when the device transitions to the ON state, the filament formed is unstable i.e. the current in the ON state is lower than the holding current (I_h). Thus, the filament will dissolve and the voltage across the device will start increasing (OFF state).

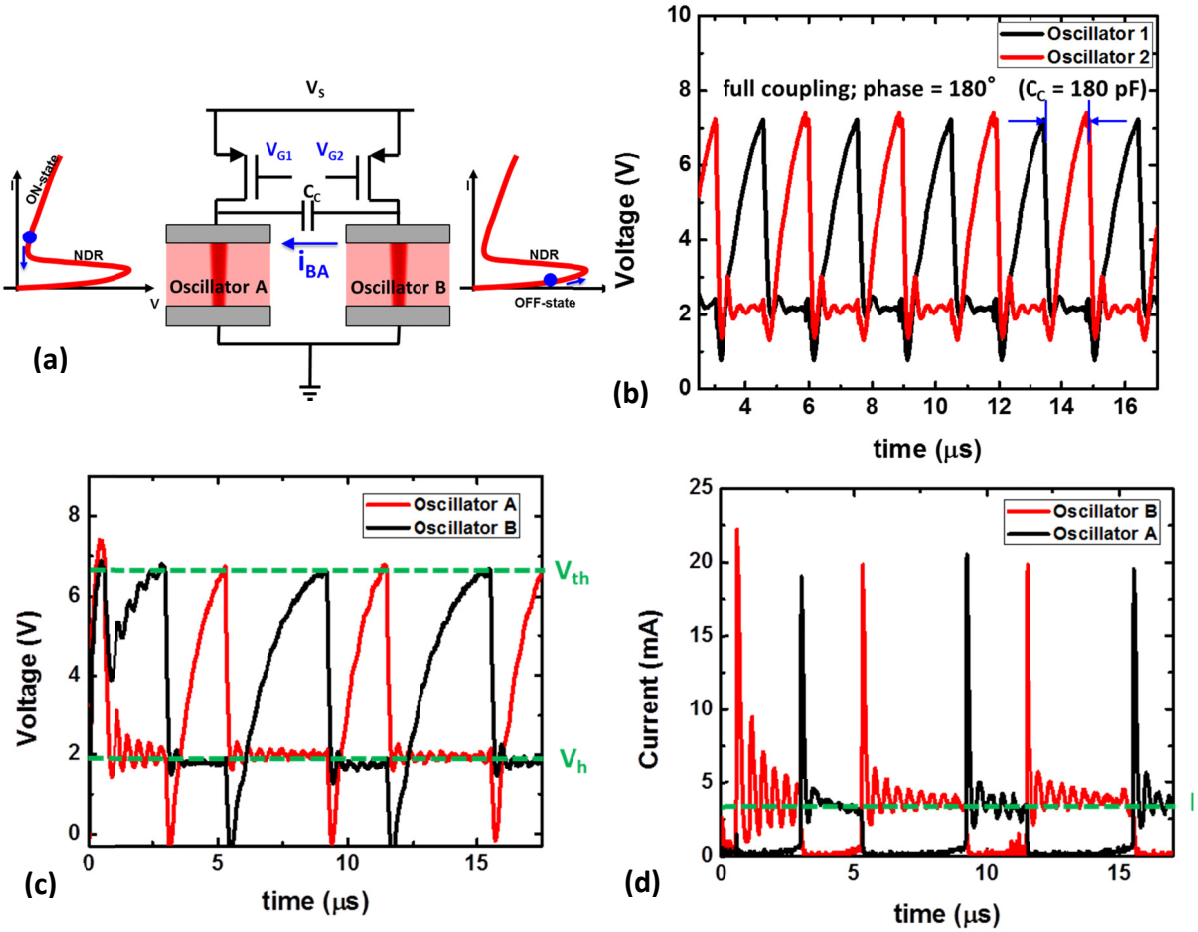


Fig. 6.3: (a) Circuit schematic for variable phase coupling with schematic NDR *I-V*'s shown to depict the oscillators. (b) 180° phase coupling using a coupling capacitance. (c) Voltage across oscillators and (d) current through the oscillators showing a clear stabilization of ON-state due to the displacement current.

To understand the electrical dynamics associated with the coupling, assume that at $t = t_0$, oscillator A is in the ON state while oscillator B is in the OFF state. Thus, at t_0, a large current is flowing through A (thus the voltage across A is low) while the voltage across B is gradually rising. The rising B causes the high-pass filter to shunt a substantial amount of current into A. While in uncoupled oscillators, any ON state would have been unstable but the supply of this extra current to A from B, results in the net current in A being the sum of currents from the supply and the coupling element. This results in the total current through A to be above the

holding current and thus the device ON state is stabilized. This is clearly shown in Fig. 6.3 (c) and (d). The voltage and current waveforms of the coupled pair with dissimilar voltages indicates the stabilization of one of the oscillators in the ON state at V_h, I_h while the other oscillator charges. At $t = t_1$, B switches to the ON state (as the node voltage exceeds threshold voltage), the displacement current (i_{BA}) through the capacitive branch reverses direction and reduces the current through A. This causes the stabilized ON state of A to become unstable and the device resistance reverts back to OFF state. As this process is occurring, the displacement current from the coupling stabilizes the ON state of B till A reaches its threshold voltage. This process keeps repeating itself, resulting in a full 180º out of phase coupling.

It is thus predictable that the bandwidth of the coupling element plays an important role in stabilizing the filament during the coupling process. If the displacement currents do not last for long enough, we could get coupling ranging from 180º out of phase down to fully coupled (0º phase) oscillators. This is achieved by changing the frequency of one oscillator w.r.t. another. Once coupled, the oscillators would settle down to a unique frequency. This results in the oscillator pair to achieve a phase difference w.r.t. each other. If the frequency difference between the two oscillators is too large, they seem to not couple but rather just both threshold switch to an ON state. The coupling capacitance that would enable such coupling is a very strong function of the frequency at which it operates. We will refer to this method of phase control as *differential gating*. Such coupling of oscillators is ideal for ultra-low power ONNs during the decision stage when each of these units is programmed to be at a certain frequency and depending on the frequency of the oscillations and the coupling coefficient; the device can give settle down to a phase that represents the vector distance. In implementations involving star-like network of PSK'ed oscillators, as discussed by Nikonov et al., differential gating serves as a method of

programming the coupling depending on which memorized pattern the test vector is being compared with. If the vector distance is too large, the oscillations stop and the system settles to the ON state.

Figure 6.4 represents the voltage waveforms of the coupled oscillators in time domain at ΔV_{GS} = 0.15, to give phase coupling at 120º phase. Figure 6.5 shows the fine-grain phase control obtained by using differential gating. The figure shows nearly continuous linear change in phase from 180º to 120º as the ΔV_{GS} changes between 0 and 0.15 V, for 10 samples.

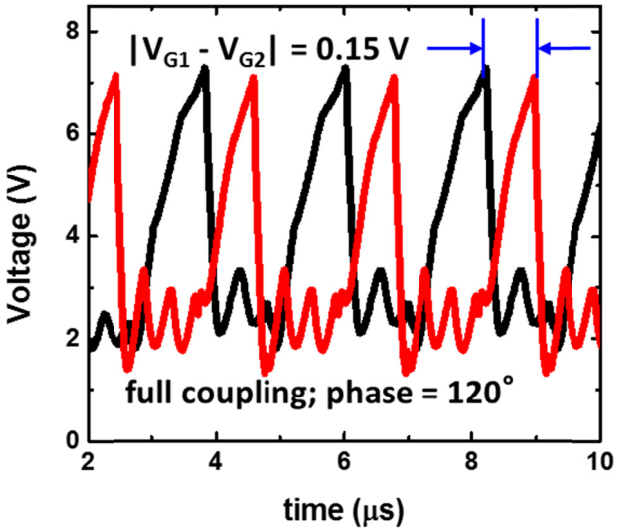


Fig. 6.4: An example of 120° phase coupling.

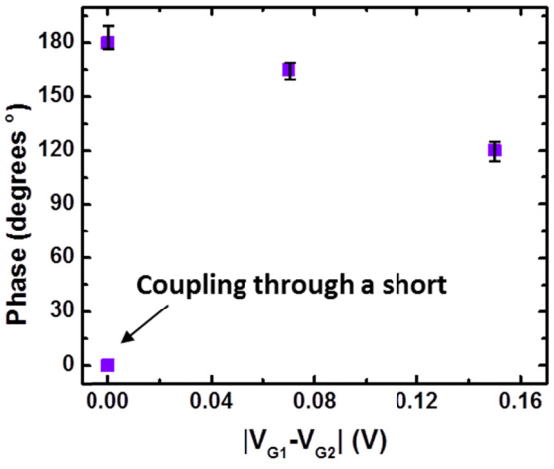


Fig. 6.5: Fine-grain phase control for dissimilar V_G, for 10 devices (symbols represent median).

In the implementation of ONN discussed in Chapter 5, there is a need to program different phases at the beginning of the recognition step, before coupling is turned on; and the oscillators settle to different phases once the coupling is enforced. To enable such primitives, we use a coupling NMOS transistor between oscillators, as shown in the circuit schematic in Fig. 6.6 (a).

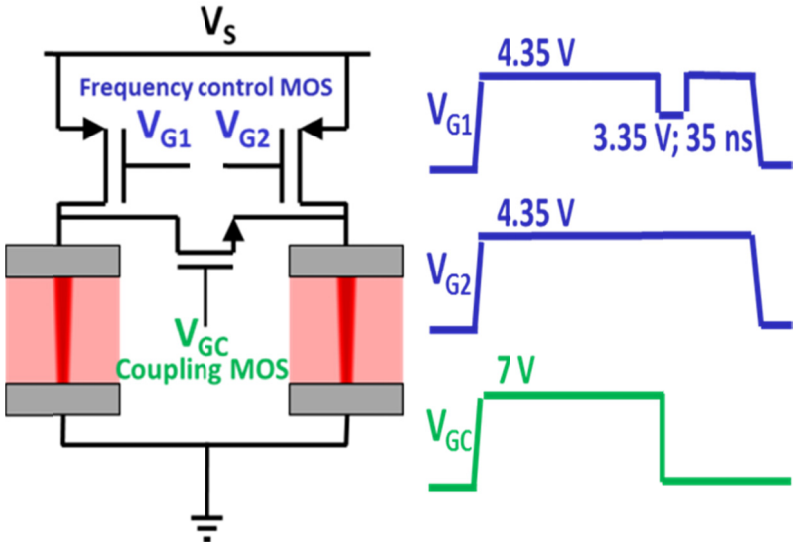


Fig. 6.6: Circuit schematic and timing diagram used to introduce local phase offset that propagates by decoupling.

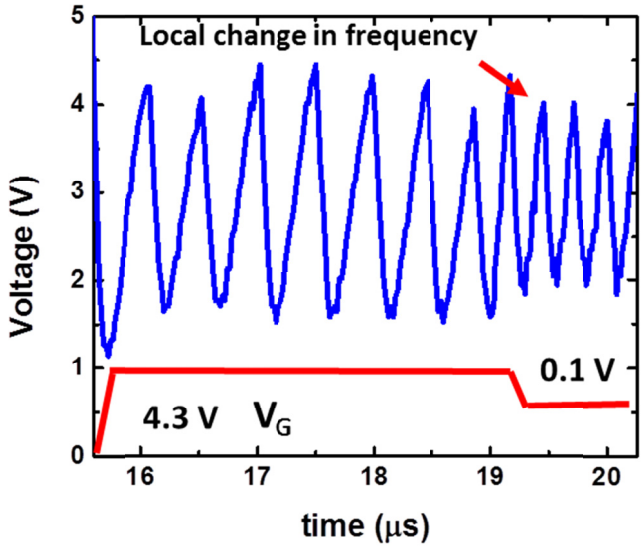


Fig. 6.7: Demonstration of local change in frequency, following Fig. 6.6.

Figure 6.7 shows a temporally localized frequency change when the gate voltage of a single oscillator is altered. This temporary change in frequency shifts the phase of the original signal. Figure 6.6 (b) shows a timing diagram of the voltages at the gates of the two ballast PMOS devices and the coupling NMOS device. In order to program a phase difference between the two

oscillators, V_G of one PMOS device is lowered for a short time, thus temporarily increasing $|V_{GS}|$ of one PMOS. Simultaneously, using this as a trigger, the V_G on the coupling NMOS, is reduced to 0, effectively coupling them only through MOS capacitance. This results in an abrupt increase in frequency and ultimately a phase-shift of the signal. In this method, the control variable is the magnitude of the pulse of a constant duration (i.e the amount of frequency speed-up); Figure 6.8 shows time domain waveforms of PSK-based phase programming in the oscillator pair, in which V_{G1} was pulsed 2 V below V_{G2} (4.53 V).

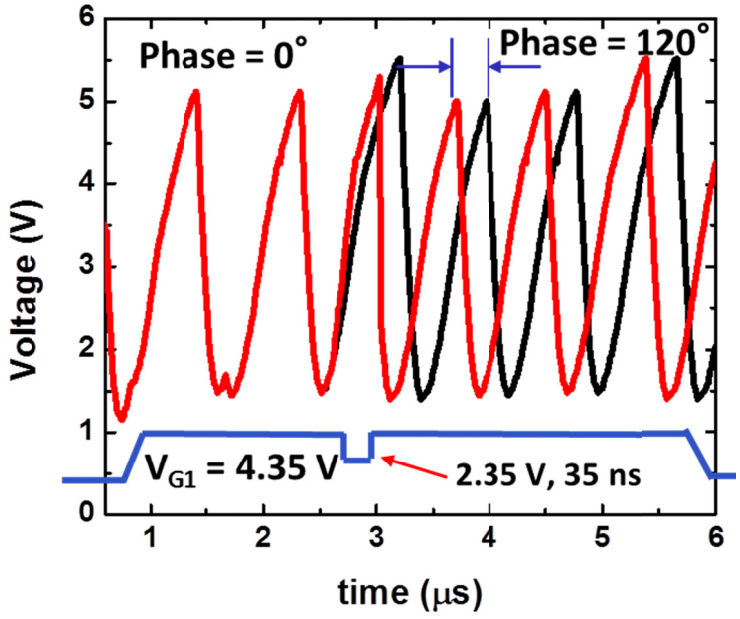


Fig. 6.8: Oscillations showing phase control by introducing a temporary increase in the frequency

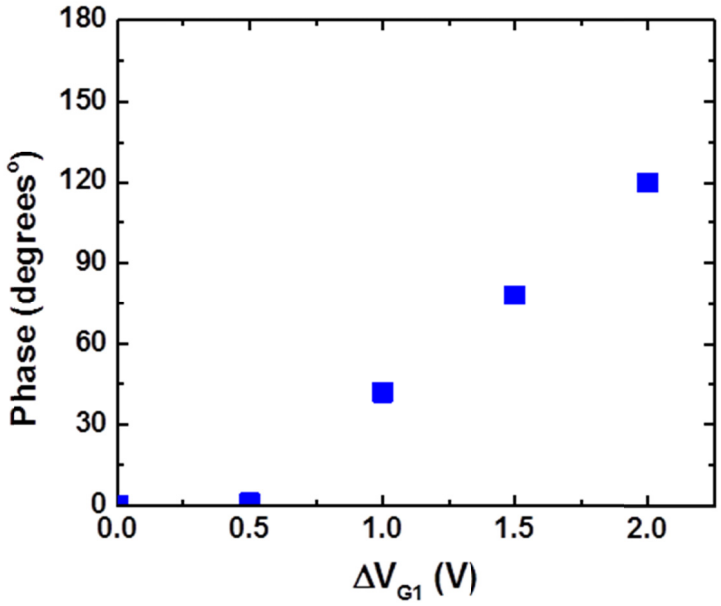


Fig. 6.9: Fine-grain phase control obtained by introducing phase offset.

The primary difference between variable phase coupling and variable phase programming lies in the ability of the coupling NMOS to decouple the two oscillators once the phase shift is introduced. If the coupling is maintained, the two oscillators would be expected to reach a steady state condition with a certain phase difference. Because decoupling stops the oscillators from interacting between each other, the introduced phase does not display any period doubling effects seen in variable phase coupling. Figure 6.9 shows the fine grain control on the phase coupling obtained by the use of variable pulse widths used to introduce phase delays.

In both the cases of coupling and control discussed in the previous sections, the oscillators were free-running i.e. they were connected to an external DC source. This may insert an initial phase different due to incubation time variation, as discussed in Chapter 2. Thus, we discuss more deterministic excitation of oscillators in the next section to reduce the inherent variability in the oscillator start-up time i.e. initial phase.

6.2. Injection Locking

Injection locking (frequency entrainment) is the phase and frequency locking of an oscillator to an external AC signal. This phenomenon arises due to the nonlinear nature of the oscillator and is observed only if the frequency of the external signal is close to the natural oscillation frequency of the oscillator (ω_o). Such synchronization is widely observed in the field of biology, such as circadian rhythm of 24 hours, synchronizing fireflies etc [91].

When the frequency of the externally applied signals is higher, the oscillator can lock to the exact integer sub-multiples of the external frequency. For example, as shown in Fig 6.10 (a); if the external frequency (2ω) is around $2\omega_o$, than the oscillation frequency becomes ω i.e. half of the frequency of the externally applied signal. This phenomenon is referred to as sub-harmonic injection locking (SHIL) and used in frequency synthesizers [92, 93]. In this section, we experimentally demonstrated SHIL for S-NDR type oscillators. Figure 6.10 (b) shows that when the oscillator is driven using a pump signal at 28 MHz, the output frequency becomes 14 MHz.

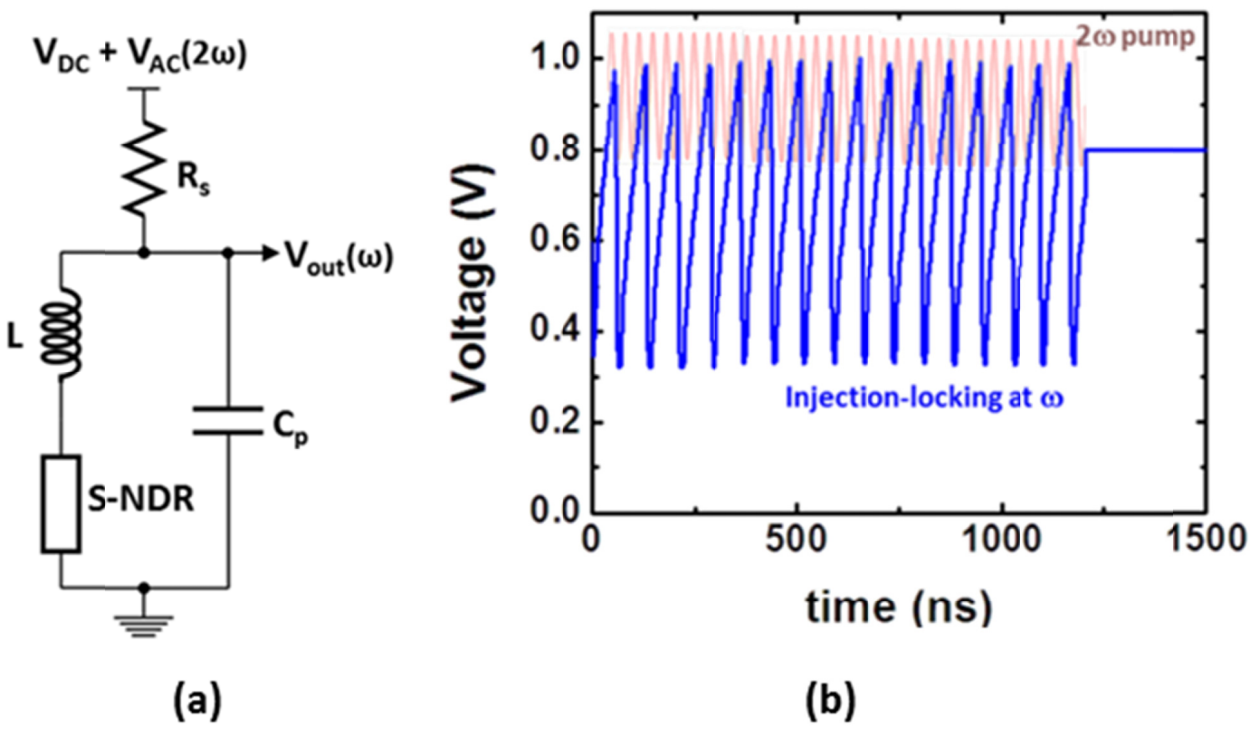


Fig. 6.10: (a) Circuit schematic representing SHIL. (b) Experimental demonstration of sub-harmonic injection locking

When the sub-harmonic injection locking occurs, the phase of the oscillator can assume two distinct values (0° and 180°) with respect to each other. The oscillator "prefers" to lock to the one that is closer to its initial phase. For the oscillators we are working on, the initial phase can be set by pre-charging the output node of the oscillator. Fig. 6.11 shows that any initial voltage of this node will result in locking to one of the stable phases. When the initial voltage is 0.1 or 0.4 volts the oscillator will lock the one phase, and it will lock to the other allowed phase if the initial voltage is 0.7 or 1 V. The phases can be named 0° and 180°, respectively. The frequency of the pump signal is 1.16 GHz and the resulting oscillation frequency is 580 MHz.

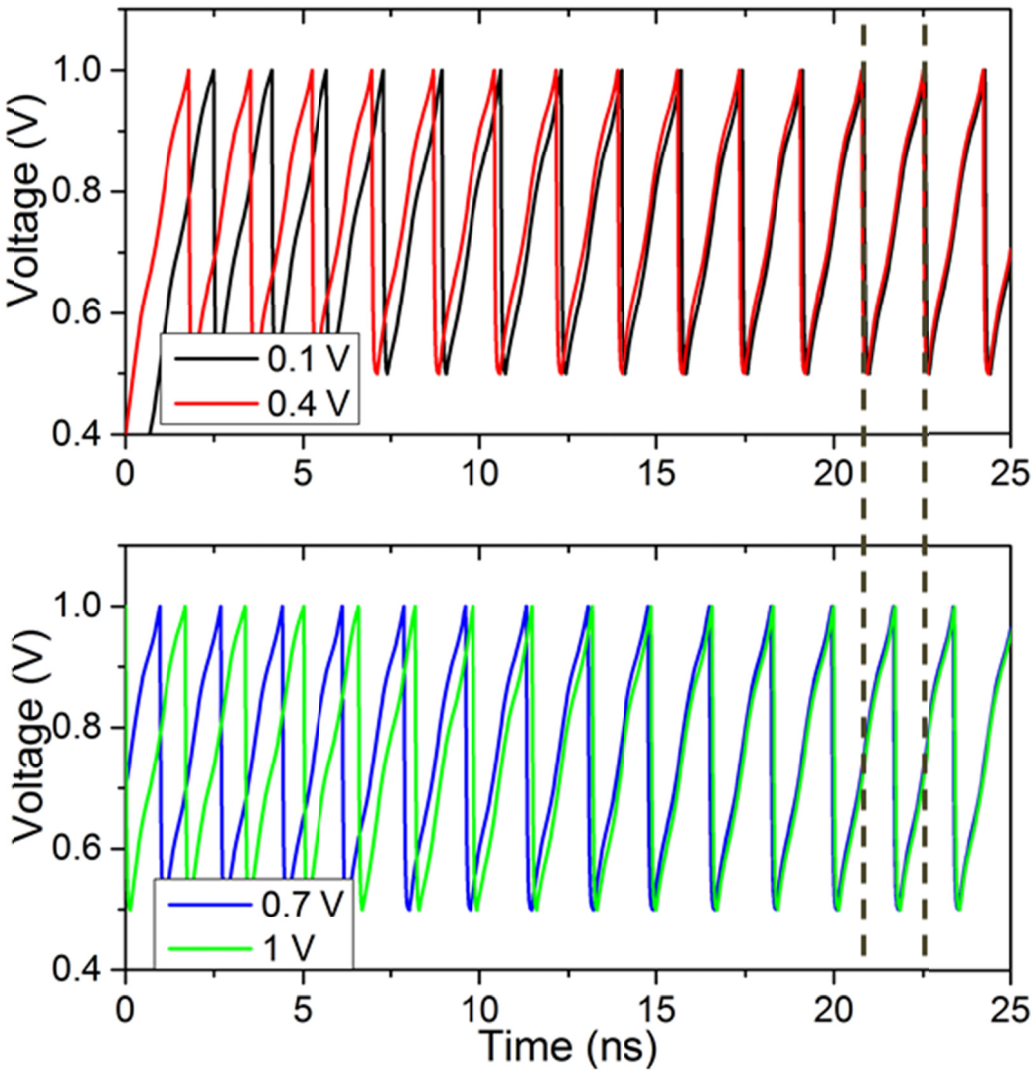


Fig. 6.11: The SPICE simulations based on van der Pol equation discussed in Chapter 5, confirm that SHIL results in two different phases at the output depending on the initial (pre-charged) voltage of the output node. The legends show these initial voltages.

SHIL not only eliminates the phase and frequency drift that is inevitable due to fabrication variations, but it also allows us to encode information in phase. In the following section, it will be used to encode the brightness of the pixels of the image to be processed.

6.3. Edge Detection using Directly Coupled Networks

With increasing exchange and processing of graphical data such as images, compression and classification of its features becomes very important. Thus, hardware implementation of these operations can significantly improve the overall performance. Directly coupled oscillator networks are known to exhibit the level of parallelism and efficiency that can significantly accelerate image segmentation through feature extraction. Typically, CMOS implementations of these systems are [84] inefficient in terms of power, area and performance (due to memory access), thus creating a need for dense networks that can be directly coupled for efficient feature extraction. In this work we explore the unique attributes of S-type negative differential resistance (S-NDR) based nano-oscillator network to enable efficient edge-detection, offering high performance and scalability beyond CMOS. Edge-detection is of immense importance for applications like character recognition, content-based image retrieval, 3D vision etc. These graphical applications require massive parallelism for efficient implementation, and directly-coupled oscillator arrays lend themselves naturally for such parallel computation. Edge detection abilities of capacitively coupled single-electron tunneling-junction oscillators have been reported [94,95]. In this section, we will discuss the network settling properties of an edge detection network. The network has been designed and simulated in collaboration with Yunus Kesim.

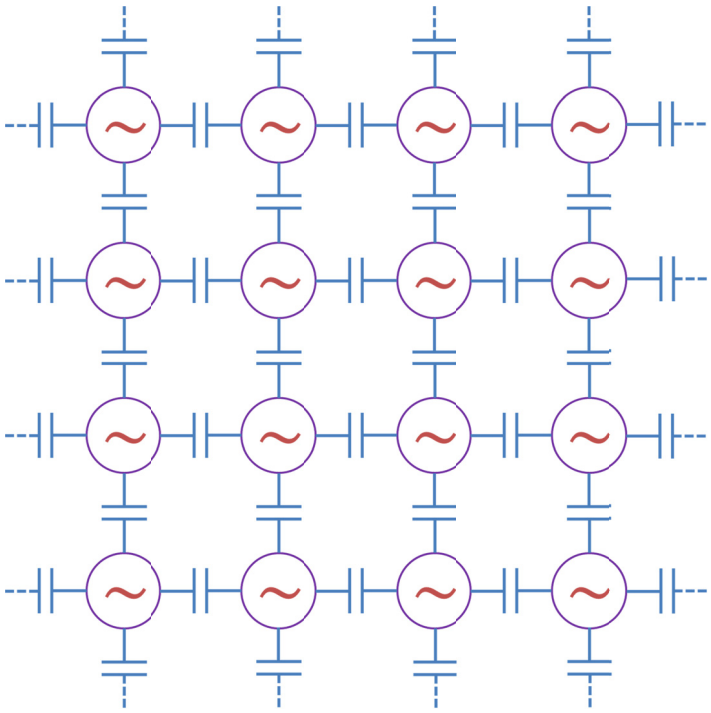


Fig. 1.12: Circuit schematic of the coupled oscillator network

A capacitively coupled oscillator network that is similar to the one shown on Fig. 6.12 is simulated in SPICE. In this network, each oscillator corresponds to a single pixel of the image to be processed and the pixel brightness needs to be represented by the phase of the corresponding oscillator. For simplicity, we have chosen a binary image as the input and set its initial phase using different initial voltages on the capacitor. For the white pixels, the initial voltage of the oscillators are set to 1 V, and for the black pixels, the initial voltage is 0.1 volts. After the image is encoded into the network, a transient simulation is conducted. Before each oscillator in the network settles to one of the two stable phases (0° and 180°), the oscillators corresponding to the edge pixels are clearly differentiable in terms of their phase from the rest of the circuit.

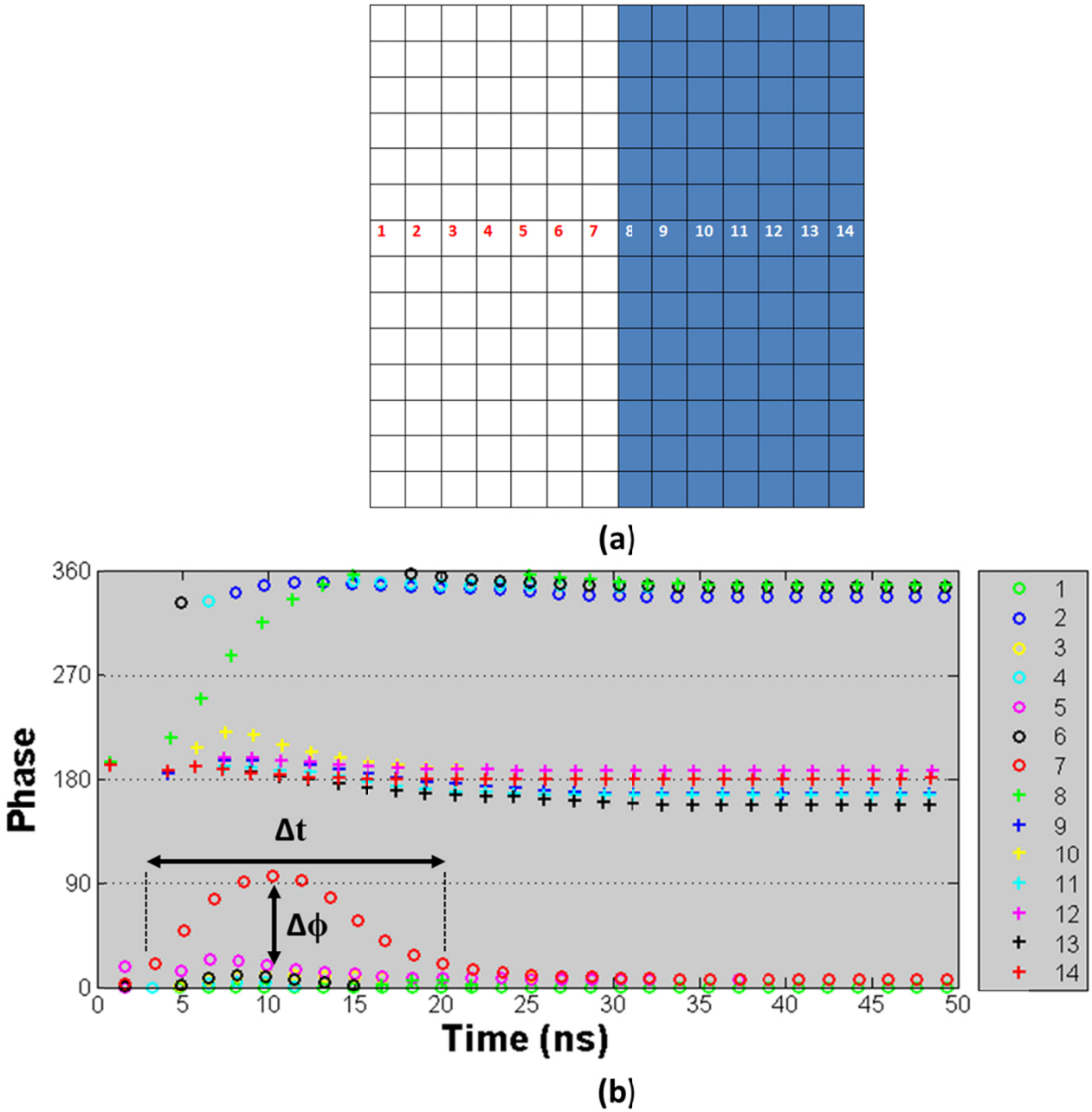


Fig. 6.13: (a) 14×14 network used for the simulations and the location of the pixels 1-14. (b) Time evolution of the phases of oscillators in a network.

If the initial phases of two capacitively coupled oscillators are the same, the displacement current passing through the coupling capacitor (C_c) is limited and the two oscillators merely interact with each other. However, if they have different initial phases, when one oscillator charging, the other one will be discharging which will result in a large dV/dt on the coupling capacitor and a significant interaction between the two oscillators. This approach can be generalized to more than two oscillators as in the network discussed here. If an oscillator is at the same phase with its neighbors initially (i.e. non-edge cell), it will not interact with its neighbors and therefore it will not experience a loading effect from its neighbors. However, if it has out-of-phase neighbors, it will see a loading effect. In this case, since some of the current supplied by the source will be drained by the coupling capacitor, the oscillation will momentarily slow-down and the oscillator will experience a lag while settling to the stable output phases. Due to this lag, the edge cells become distinguishable. Fig. 6.13 (b) compares the phases of 14 pixels along a line of a 14×14 network that is given on Fig 6.13 (a). Pixels numbered 1-7 (denoted with o symbols) and 8-14 belongs to different segments of the image. Pixel 7 and 8 are the neighbors at the edge. Oscillators in each segment settle to their final phases of 0 and 180, yet, this process is slower for the edge pixels (7,8), which makes them distinguishable.

To reconstruct the image based on the phase, the information needs to be mapped from phase space to binary space. This can be done using a thresholding function as follows:

$$f(\phi) = \begin{cases} 0, & -45 < \phi < 45 \text{ or } 135 < \phi < 225 \\ 1, & 45 < \phi < 135 \text{ or } 225 < \phi < 315 \end{cases} \tag{6.1}$$

which will convert the edge cells to binary 1 and non-edge cells to 0.

In the coupled oscillator based edge detection application, the synchronization of individual oscillators spreads over the network such that the regions containing similar features share the same phase. The resulting computation resembles finding cuts in a graph [19], the common method in computer vision. Here, the computation is achieved by simply letting a network of coupled oscillators relax, as opposed to carrying out numerical procedures to find eigenvectors of the graph Laplacian. A summarizing flowchart of the edge detection procedure using directly coupled network is given in Fig. 8.

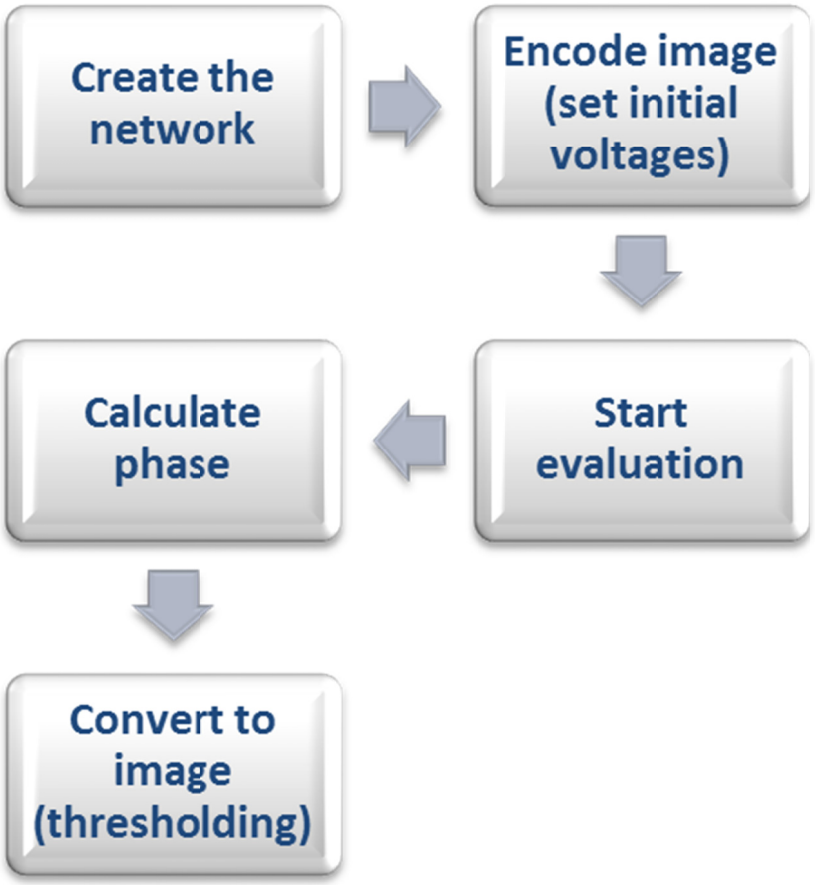


Fig. 6.14: Edge detection flowchart using coupled oscillator network

Fig. 6.15 shows a 50×50 binary input image and the corresponding output image where the edges are detected. Note that, this edge detection method can be generalized to grayscale images by using different initial voltages and it can be generalized to color images since it can be applied to the each component of an RGB image separately.

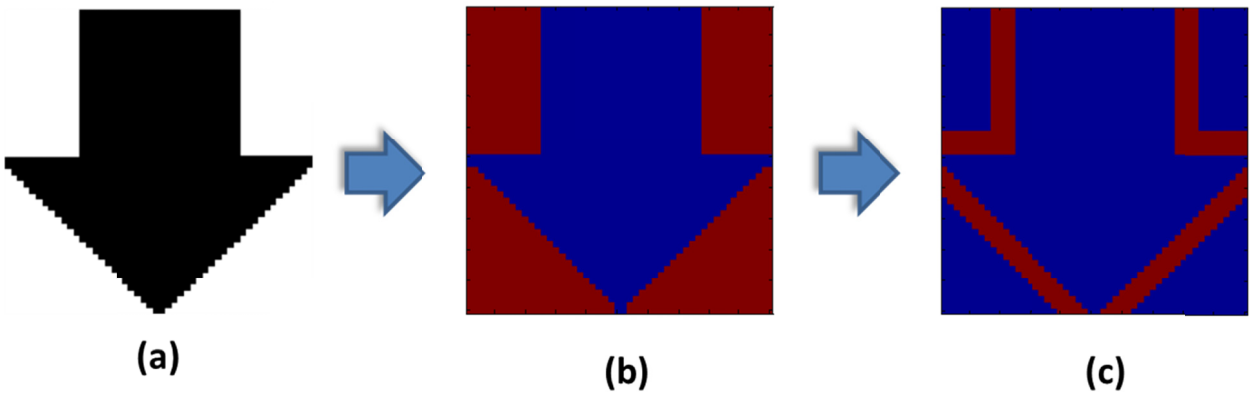


Fig. 6.15: (a) Input binary image (b) image is encoded in initial phases and (c) after the evaluation, the edges of the feature are detected.

The coupling element, C_c, plays a crucial role in the synchronization process of the oscillators. Firstly, when C_c is very small, the displacement current is negligible even if the neighboring oscillators are out-of-phase. If C_c is very large, the coupling becomes too strong and the oscillation frequency of oscillators reduces significantly and they cannot lock to the driving AC signal anymore. Moreover, since C_c determines the level of interaction between the oscillators, it directly affects how much the edge cells are separated from the others in terms of phase ($\Delta\phi$). Secondly, the coupling element determines how quickly the oscillators synchronize and this puts a restriction on the allowed time for evaluation (Δt). For example, for the case pictured in Fig.

6.15, the evaluation should take place before 20 ns, with the optimal window being the 5-15 ns range.

6.4. Stereo Vision using Coupled Oscillator Networks

Many computer vision applications, such as 3D stereo estimation, rely on inference computation on Markov Random Fields (MRFs) formulated as graphs, where computation is done through message passing over the nodes of the graph. These applications promise disruptive new capabilities for embedded systems. For example, smart phones, security cameras, and even glasses [108] that can view the world in 3D would open up new usage scenarios and market opportunities.

Many computer vision applications involve assigning labels optimally to nodes in a graph that represent an image. For example, in stereo estimation, the nodes represent pixels from a pair of stereo images, and the labels denote the 3D depth inferred from the image pair. The optimal labeling problem is typically formulated as an energy minimization on Markov Random Field [109]. Among various MRF solving methods, TRW-S [113][114] is known to provide better convergence and energy than others [109].

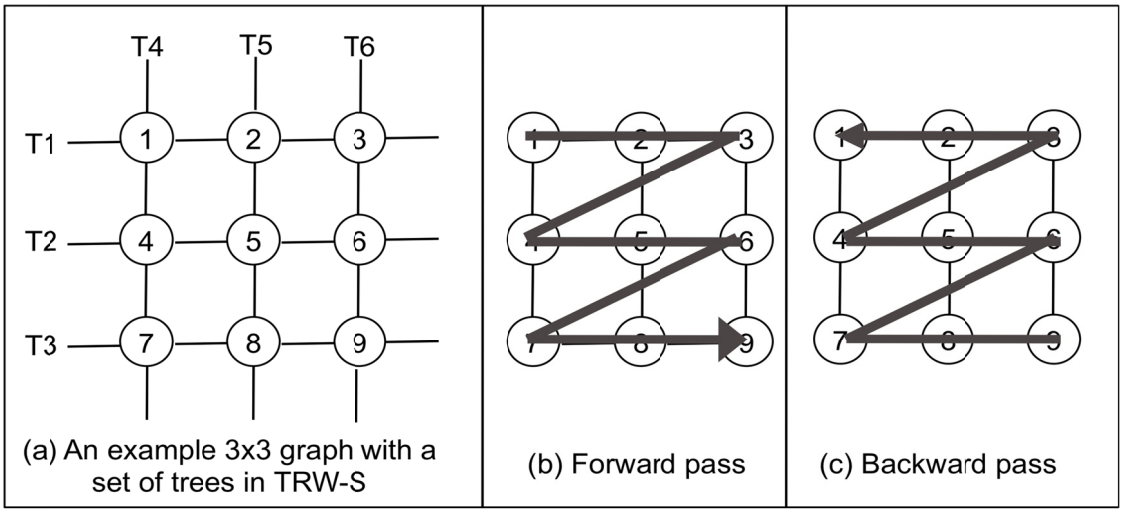


Fig. 6.16: TRW-S go over a set of trees (T1 to T6) in a monotonically increasing order, through forward/backward passes.

In TRW-S, the energy minimization problem is cast as a set of minimization problems on trees that cover the graph. Figure 6.15(a) shows an example MRF graph with the associated set of trees, T1 to T6, representing a 3x3 pixel image. The TRW-S computation follows a sequence of update functions applied to the nodes in the trees in a monotonically increasing order. This "sequential" update order impacts convergence as it avoids oscillating energy value. As a specific example, T1 tree in Figure 6.16 would update node 1 and sends the output message to node 2, then from node 2 to 3. For T4 tree, the order is node 1, 4, and 7. To achieve monotonically increasing order for all the trees in the graph, the typical approach is to iterate over the graph by passing messages from top-left to bottom-right (forward pass), and then in the opposite direction (backward pass), as illustrated in Figure 6.16(b) and (c).

The convergence of this algorithm is determined by two factors: (1) Data cost – that determines how important the value of any pixel is to itself; and (2) Smoothness cost – how coupled are the

neighborhood pixels. As the nodes keep getting updated, new data and smoothness costs are calculated and messages are generated, as shown in Fig. 6.17.

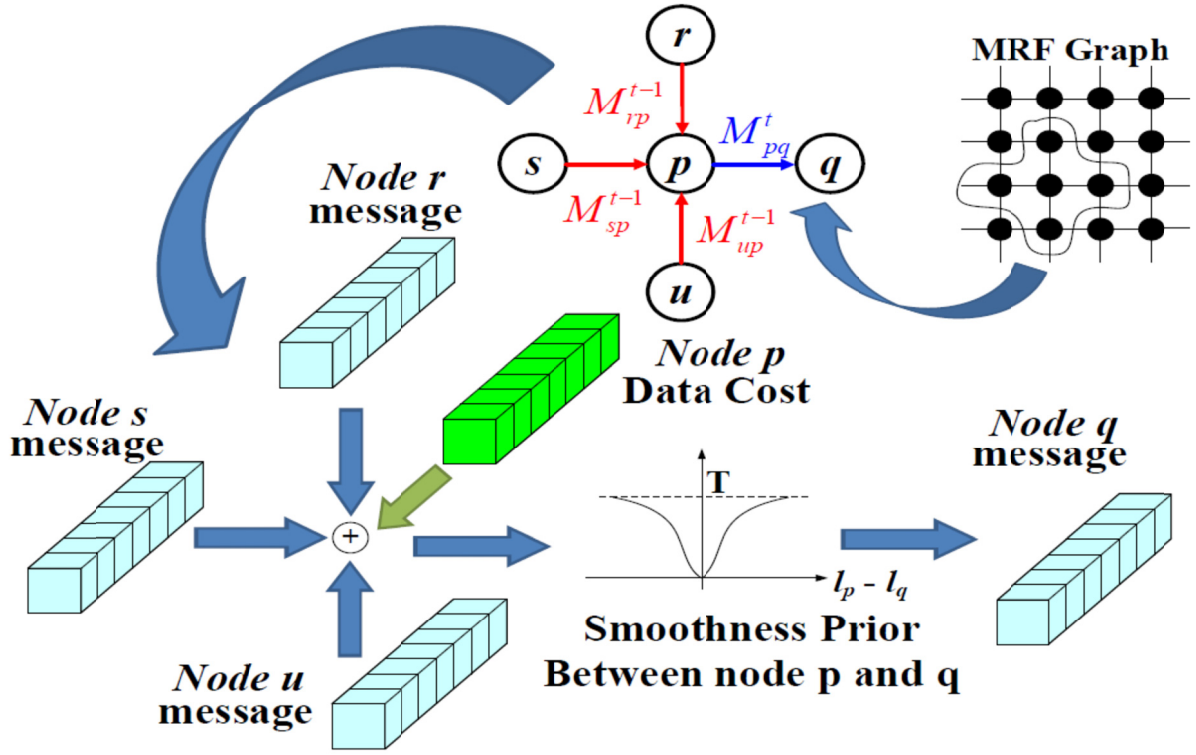


Fig. 6.16: Message updating process in message-passing based inferencing algorithms.

However, the biggest challenge in the implementation of these algorithms is the wide memory bandwidth and constant latencies introduced by memory access and raster scanning the pixels for message computation.

Our previous work on one such graphical software implementation [115] concluded that most of the computation (message generation and passing) is impeded because of raster movement of messages. As shown in Fig. 6.18, we create the same system using an oscillator network (this

section has been entirely executed by Abhishek A. Sharma), in which a disparity image is first created from images as seen by left and right camera (similar to human eye) - this is simply the color difference between left and right images.

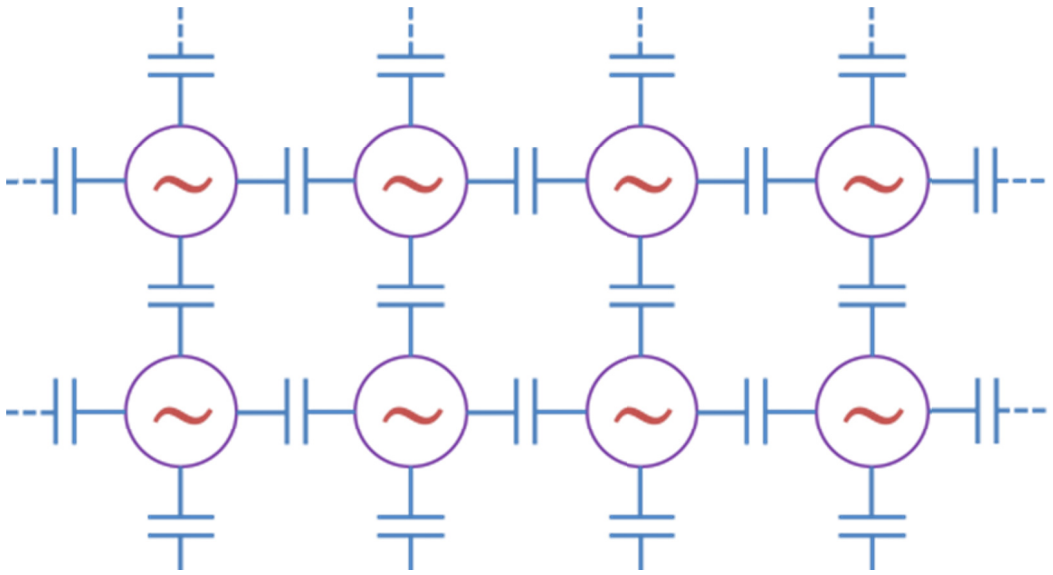


Fig. 6.18: Coupled oscillator network array.

It is then encoded onto the gate voltage of each 1T1R oscillator (mapping 0 – 255 over a V_G of 0 - 1 V in Fig. 6.18). Thus, images with larger color difference (objects closer to the camera) will translate to a higher V_G and thus, a higher frequency. Thus, all of the nodes which are a part of the object that is closer to the camera operate at a higher frequency and are locked to the same 180° phase, whereas objects far away operate at a lower frequency, and the edges of these objects too, lock to a 180° phase, as discussed in our previous work [115]. This phase locking provides frequency stability and prevents drift. Because of variable phase locking, the edges automatically lock at disparate phases. In order to design large system implementation, we develop a data-driven Van der Pol model [16] for the oscillator array simulation. The presence of

an inductor-like circuit element relates filament size to the oscillation frequency, which changes depending on the current that flows through it. Thus, the data cost (value of depth) is represented by the frequency and the smoothness cost (edge transitions) is controlled by coupling capacitance. We connect nearest 4 neighbors together using 10 fF – 50 fF capacitors (for the device capacitance of 20 fF – 50 fF). The oscillators were made to operate at 1 GHz. We use a 128 x 128 pixel oscillator network for a 128 x 128 pixel image. Interestingly, we can also 'tile' an image if our oscillator network size is limited (example: our oscillator network is 128x128 but the image is 1k x 1k i.e. 1MP image) - we can send one part of image at a time, if network is smaller. The benchmark tested here is the standard Middlebury benchmark – Tsukuba. Fig. 6.19 shows the convergence of a disparity image to form a stereo image once the oscillators have all coupled. Low-power devices enable large arrays that can overcome aliasing (limited memory access) and the array connectivity contribute to the co-design essential for stereo vision.

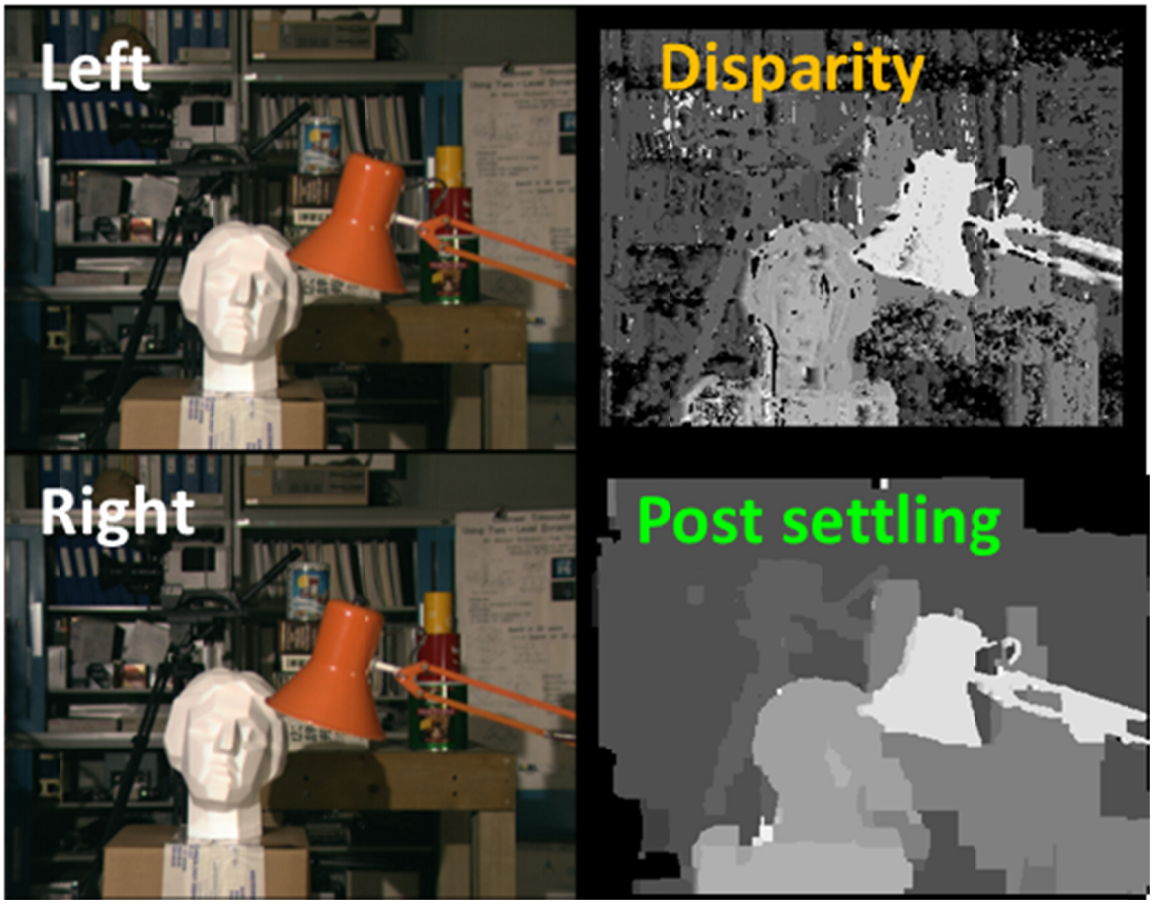


Fig. 6.19: Post-settling image from Tsukuba benchmark.

The results of this benchmark show an image misprediction of < 22% when compared to ground truth. This can be optimized by using injection-locking techniques so as to reduce the variability in the initial phase. The variability causes a delay in the coupling stabilization and eventually errors. Compared to our previous work on TRW-S, this method offers a compute speed-up of 16x with a power reduction of up to 100 times.

In conclusion, we demonstrated phase coupling and control in 1T1R S-NDR oscillators and experimentally demonstrated for the first time – injection locking. This enables us to use these oscillators in robust simulated annealing type-problems. As an example, we have simulated edge

detection using such directly-coupled, sub-harmonic injection-locked S-NDR oscillator networks. The models and SHIL are based on experimental results. Moreover, we analyzed the effect of the coupling element in the network dynamics and performance. We have shown that oscillation frequency can go as high as a few GHz and which reduces the time and power expense of the edge detection operation to a few ns and pJ level, respectively. Finally, we have also demonstrated a more focused implementation of stereo vision using coupled oscillator network that is uniquely positioned to solve energy minimization type algorithms.

Chapter 7

Conclusion

Several emerging technologies have shown promise in augmenting modern computation. The key feature that these technologies must possess is multi-functionality and reconfigurability for the same footprint. In this work, we have focused on devices that exhibit S-type negative differential resistance (S-NDR). The fundamental operating principle of these devices participates during the forming process in RRAM devices (as also perhaps SET process) as well as in its operation as a compact relaxation oscillator. In this work, an attempt has been made to first understand the physics of the device and then use the tunable the parameters to engineer these devices for use as memory and oscillators. Finally we show system-level demonstrations of beyond-CMOS architectures that can be uniquely realized with these S-NDR devices. In this concluding chapter, we will revisit the primary learning from each chapter to generate a more holistic picture of S-NDR devices.

While forming may be a one-time process needed to initialize MIM stack to function as memory cells, it is by far the most important procedure that the device needs for stable and reliable switching. Prior works have debated the physical origin and nature of forming process in RRAM devices in detail. The current understanding has been that this process is initiated by vacancy migration due to the presence of imperfections at one of the electrode-oxide interfaces. In our work, we have demonstrated that the forming process is a two-step process. This manifests itself in the form of negative differential resistivity in the material causes the device to go into a negative differential resistance regime which causes current constriction, prior to forming. Unless prevented by the circuit load, this process frequently occurs in the form of an uncontrolled runaway. We support these claims by analysis of the steady-state DC behavior and the dynamics of the instability. Both DC and dynamic measurements indicate the presence of an instability that is reversible and, hence, transient in nature as distinct from vacancy migration

initiated. The initiation of the constriction is temperature dependent and higher temperature is shown to cause the point of bifurcation to appear at a lower voltage. Hence, we propose the following mechanism of electroforming - with increasing bias, the device conducts uniformly throughout its area. At a well-defined point depending on source voltage, series resistance, temperature and time, the device enters into the *I-V* range of negative differential resistance which results in the electronic current filamentation. This current filamentation starts off with being thermally induced (due to the thermal non-linearities) before the effects of voltage non-linearity set in. This final stage of current filamentation causes the device to change resistance to a value close to the post-forming value. We develop a novel pulsed thermometry to estimate the localized temperature in the current filament using a self-consistent electro-thermal measurements and simulation. Temperature excursions that exceed 500 K (over the ambient) were estimated in a localized sub-20 nm region on the onset of forming. This localized temperature excursion then triggers the physical changes in the structure to form the permanent filament. The constriction can be controlled with the use the external circuit loading thus affecting the permanent filament structure. In order to corroborate the results with temporal dynamics of filamentation, we observed and explained the three regimes of electroforming time dependencies on forming voltage. The observed *1/E* field dependence of forming times is consistent with field-induced nucleation model from which we extracted material properties such as the nucleation barrier height at zero bias ($W_0 \sim 0.65$ eV) and voltage acceleration factor V_0 of ~ 2.8 V (for 60 nm TaO_x film) at intermediate voltages. A clear difference in the temporal dynamics was identified for low voltages with corresponding forming times longer than the thermal time constant and at high-fields where the film self-heating is important. Moreover, we were able to detect, and study the volatile filament that precedes formation of the non-volatile

filament. The forming process was analyzed in the framework of nucleation model which was extended to include the self-heating effects. This yielded an estimate of the critical nucleation radius ($R \sim 1$ nm) below which filament is always volatile, in line with the prediction from the thermometry. This implies that the filament-based RRAM technology can thus be scaled to a physical limit dictated by the critical nucleus size which could be as low as 1 nm in size. We also demonstrated that forming is a field accelerated phenomenon and that the forming can be sped-up by nearly 6 orders of magnitude compared to DC forming typically used for RRAM. The thermometry is fairly general in its applicability and hence can be applied to switched RRAM devices in LRS.

RRAM filament thermometry developed in this work has become a key for understanding the physics of RRAM devices due to the crucial role of the temperature in these devices. When the thermometry was applied to scaled 85 nm x 85 nm devices, we found that the devices can reach temperature in excess of 1000 K during the switching event. This was confirmed using TEM experiments that confirmed the temperatures due to the crystallization (indicating temperatures > 850 K) of the initially amorphous oxide matrix. At low-biases, the filament size was extracted to be ~ 1.2 nm (for a compliance current of 50 μA) and increased if the compliance current was increased. This is a direct continuation from the current constriction size that was found during the forming process. At high biases, the filament size appeared to increase. This was interpreted as an increase in the conducting volume due to the oxide in the proximity of the filament getting heated up and conducting. This was found to be consistent with a significantly wider crystallization zone when observed under a TEM. Such current spreading may be responsible for reducing the power-density in filamentary switching RRAMs, thus limiting the peak temperature and hence failure. Moreover, this was also used as a tool to understand the role of lateral thermo-

diffusion due to the presence of strong thermal gradients and it was found that devices with larger filament diameters (for the same switching power) failed due to the presence of a gentler thermal gradient. More detailed studies can now be pursued to track the role of lateral and vertical thermal and concentration gradients on resistive switching process. Thus, the thermometry technique offers unique insights into: (1) Peak temperature in the filament, (2) Geometry of the filament (3) Filament growth. (4) Dynamics of the heated zone. One must point out that these thermometry tools can be applied, in principle, to any thin film having filamentary conduction. This experimental technique can serve as a tool to design materials and devices with optimized thermal and electrical characteristics.

With the information about filament formation and the role of bias, temperature and microstructure clear from the thermometry, we revisited the S-NDR devices and found that certain compositions of TaO_x and chalcogenides do not permanently change their resistance during the forming process. In other words, they have an indefinitely long lock-on time post resistance change. These materials have been referred to in the literature as threshold switches. We demonstrated a novel ultra-compact oscillator that is based on the same material, TaO_x, displaying precise frequency control over more than four decades of frequency (20 kHz - 250 MHz) with the potential for an even larger frequency range. This range was obtained by using two different ballasting techniques – (1) a linear resistor and (2) a PMOS transistor. We have presented evidence that depending on the operation mode, these oscillations can be regulated by controlling the dynamics of the current filamentation internal to the MIM structure or by the external parasitics. We have also shown that high-frequencies are obtained by lowering the peak current in the ON-state in regimes where the frequency is not RC dominated. In addition to its

CMOS compatibility and scalability, this oscillator provides large-signal oscillations and can be used for dense oscillator arrays.

While the TaO_x, oscillators had a performance metric that was significantly better than prior works, the voltages and power consumption were still unreasonably high. Thus, we engineered these devices for CMOS compatibility, based on the learning from the forming process. By engineering the electrode material, oxide thickness, minimization of parasitics loading the oscillator, this thesis demonstrated: (1) First ever 1T1R integrated structure, (2) Maximum frequency of ~500 MHz, > 2 orders of magnitude higher than any report, in this class of oscillators, (3) Lowest reported power down to < 200 μW, one order of magnitude lower than best reported, (4) Full-system simulation of an ONN-based associative memory. Thus, we can build a basic instruction set for engineering these oscillators: (1) Thin films result in a lower threshold voltage or forming voltage, as it is a field-mediated process, (2) Electrode work-function creates a band-offset at the metal-oxide interface. Depending on the electron affinity and bandgap of the material, a high or lower work-function metal may be needed to ensure Ohmic conduction and hence a lower holding voltage. (3) Holding current determines the frequency tuning window and the peak power consumed during the oscillator operation and hence it must be minimized while keeping it above threshold current. (4) Presence of parasitics causes the devices to reach an ON-state that is more conductive due to the capacitive overshoot. To prevent this, a ballast device must be placed adjacent to the MIM stack to minimize capacitance.

Based on these recommendations, we explored the applicability of these oscillators for oscillatory neural networks, using these oscillators as oscillatory units. For this purpose, we used the van der Pol model to model these oscillators so that their properties can be studied in large network simulations. While the networks showed reasonable network characteristics as an associative memory but it was found that the CMOS surrounding the 1T1R oscillators (for connectivity) may be severely area inefficient. This problem could be exacerbated for very dense oscillator array implementations. Thus, we moved our focus to directly coupled oscillators for networks that require less CMOS related to coupling and to bring about system evolution. These networks relied on the circuit behavior at the device-circuit level, making it easy to engineer the network as a whole.

To explore this application, we demonstrated as the first, coupling between two oscillators using capacitors and transistors. By coupling two oscillators using a capacitor, we were able to demonstrate phase coupling from 0° to 180°, by the means of differential gating of ballasts. Similarly, we were also able to show initialization of phase by coupling oscillators through transistors. In such configurations, we were able to achieve a phase control from 0° to 120° by the means to local-FSK in coupled oscillators. It was observed that these oscillators exhibit an initial phase related to the incubation time. This causes the oscillators to show slight variability in phase due to device to device process variation; moreover, such oscillators show a frequency drift (increase) before failure due to partial permanent crystallization/forming. To overcome these phenomena, we used injection locking and showed that the oscillators can be locked to a carrier at half the excitation frequency (this is also referred to as sub-harmonic injection locking). This made the response of the oscillators significantly more deterministic and immune to failure modes.

Such demonstrations of oscillatory response tuning uniquely positioned us to explore large directly coupled oscillator networks as feature extraction engines like robust simulated annealing type-problems. As an example, we have simulated edge detection using such directly-coupled, sub-harmonic injection-locked S-NDR oscillator networks. The models and SHIL are based on experimental results. Moreover, we analyzed the effect of the coupling element in the network dynamics and performance of image feature extraction. Finally, we have shown that oscillation frequency can go as high as a few GHz and which reduces the time and power expense of the edge detection operation to a few ns and pJ level, respectively.

This thesis attempts to bridge the gap between the attractive properties of an emerging technology and its applicability in real-world problems by understanding the circuit-device and system-device interactions. While S-NDR devices have been a subject of research for several decades, no attempts were made to understand this phenomenon in the light of its applications – physical processes in RRAM and oscillators. This work endeavors to create headway and open up new research area that can make these devices accessible and useful for augmenting next-generation compute.

References

[1] Burr, Geoffrey W., et al. "Overview of candidate device technologies for storage-class memory." IBM Journal of Research and Development 52.4.5 (2008): 449-464

[2] Takefuji, Yoshiyasu. Neural network parallel computing. Springer, 1992.

[3] Bez, Roberto, and Agostino Pirovano. "Non-volatile memory technologies: emerging concepts and new materials." Materials Science in Semiconductor Processing 7.4 (2004): 349-355.

[4] Ielmini, D., et al. "Statistical modeling of reliability and scaling projections for flash memories." Electron Devices Meeting, 2001. IEDM'01. Technical Digest. International. IEEE, 2001.

[5] Hsiao, Yi-Hsuan, et al. "A critical examination of 3D stackable NAND flash memory architectures by simulation study of the scaling capability." Memory Workshop (IMW), 2010 IEEE International. IEEE, 2010.

[6] Schöll, Eckehard. *Nonlinear spatio-temporal dynamics and chaos in semiconductors*. Vol. 10. Cambridge University Press, 2001.

[7] Wong, H-S. Philip, et al. "Metal–oxide RRAM." *Proceedings of the IEEE* 100.6 (2012): 1951-1970.

[8] Rohde, Christina, et al. "Identification of a determining parameter for resistive switching of TiO 2 thin films." Applied Physics Letters 86.26 (2005): 262907-262907.

[9] Govoreanu, B., et al. "10× 10nm 2 Hf/HfO x crossbar resistive RAM with excellent performance, reliability and low-energy operation." Electron Devices Meeting (IEDM), 2011 IEEE International. IEEE, 2011.

[10] Lee, H. Y., et al. "Low power and high speed bipolar switching with a thin reactive Ti buffer layer in robust HfO2 based RRAM." Electron Devices Meeting, 2008. IEDM 2008. IEEE International. IEEE, 2008.

[11] Tz-yi Liu et al., "A 130.7mm2 2-layer 32Gb ReRAM memory device in 24nm technology," Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2013 IEEE International , vol., no., pp.210,211, 17-21 Feb. 2013

[12] Ielmini, Daniele. "Modeling the universal set/reset characteristics of bipolar RRAM by field- and temperature-driven filament growth." Electron Devices, IEEE Transactions on 58.12 (2011): 4309-4317.

[13] Yu, Shimeng, and H-SP Wong. "A phenomenological model for the reset mechanism of metal oxide RRAM." Electron Device Letters, IEEE 31.12 (2010): 1455-1457.

[14] Degraeve, Robin, et al. "Dynamic 'hour glass' model for SET and RESET in HfO 2 RRAM." VLSI Technology (VLSIT), 2012 Symposium on. IEEE, 2012.

[15] Gao, Bin, et al. "Unified physical model of bipolar oxide-based resistive switching memory." Electron Device Letters, IEEE 30.12 (2009): 1326-1328.

[16] Pickett, Matthew D., and R. Stanley Williams. "Sub-100 fJ and sub-nanosecond thermally driven threshold switching in niobium oxide crosspoint nanodevices." Nanotechnology 23.21 (2012): 215202.

[17] Lee, Jong Ho, et al. "Threshold switching in Si-As-Te thin film for the selector device of crossbar resistive memory." *Applied Physics Letters* 100.12 (2012): 123505.

[18] Nishi, Y., "Challenges and opportunities for future non-volatile memory technology." Current Applied Physics 11 (2011): e101-e103.

[19] Kwon, D. H., Kim, K. M., Jang, J. H., Jeon, J. M., Lee, M. H., et al., "Atomic structure of conducting nanofilaments in TiO2 resistive switching memory." Nature nanotechnology 5 (2010): 148-153.

[20] Miao, F., Yi, W., Goldfarb, I., Yang, J. J., Zhang, M. X., Pickett, M. D., Strachan, J.P., Rebeiro, G.M. & Williams, R. S. "Continuous electrical tuning of the chemical composition of TaO_x-based memristors." ACS nano 6 (2012): 2312-2318.

[21] Goux, L., Fantini, A., Degraeve, R., Raghavan, N., Nigon, R., Strangio, S., et al. "Understanding of the intrinsic characteristics and memory trade-offs of sub-μA filamentary RRAM operation." VLSI Technology (VLSIT), 2013 Symposium on. IEEE, 2013.

[22] Gilmer, D. C., Bersuker, G., Park, H. Y., Park, C., Butcher, B., Wang, W., et al. "Effects of RRAM stack configuration on forming voltage and current overshoot." Integrated Memory Workshop (IMW), IEEE Proceedings on, 2011.

[23] Ielmini, D., C. Cagli, and F. Nardi. "Resistance transition in metal oxides induced by electronic threshold switching." Applied Physics Letters 94.6 (2009): 063511.

[24] Sharma, A. A., Noman, M., Abdelmoula, M., Skowronski, M., Bain, J.A., "Electronic Instabilities Leading to Electroformation of Binary Metal Oxide-based Resistive Switches." Advanced Functional Materials 24 (2014): 5522-5529.

[25] Yang, J. J., Miao, F., Pickett, M. D., Ohlberg, D. A., Stewart, D. R., Lau, C. N., & Williams, R. S., “The mechanism of electroforming of metal oxide switches”, Nanotechnology 20, 215201 (2009)

[26] Noman, M., Sharma, A. A., Lu, Y. M., Skowronski, M., Salvador, P. A., & Bain, J. A., "Transient characterization of the electroforming process in TiO_2 based resistive switching devices." Applied Physics Letters 102.2 (2013): 023507.

[27] Bernard, Y., P. Gonon, and V. Jousseaume. "Resistance switching of Cu/SiO_2 memory cells studied under voltage and current-driven modes." Applied Physics Letters 96.19 (2010): 193502-193502.

[28] Simon, M., Nardone, M., Karpov, V. G., & Karpov, I. V. "Conductive path formation in glasses of phase change memory." Journal of Applied Physics 108 (2010): 064514.

[29] Ovshinsky, Stanford R. "Reversible electrical switching phenomena in disordered structures." Physical Review Letters 21 (1968): 1450.

[30] Karpov, I. V., Mitra, M., Kau, D., Spadini, G., Kryukov, Y. A., & Karpov, V. G., "Evidence of field induced nucleation in phase change memory." Applied Physics Letters 92 (2008): 173501.

[31] Zeldovich, J.B., "On the theory of new phase formation: cavitation", Acta Physicochimica USSR 18, 1 (1943)

[32] Bernard, Y., Gonon, P., and Jousseaume, V., "Resistance switching of Cu/SiO_2 memory cells studied under voltage and current-driven modes", Applied Physics Letters 96, 193502 (2010).

[33] Chakraverty, B.K., "Metal-insulator transition; nucleation of a conducting phase in amorphous semiconductors", Journal of Non-crystalline Solids 3 (1970) 317-326.

[34] Pevtsov, A. B., Medvedev, A. V., Kurdyukov, D. A., Il'Inskaya, N. D., Golubev, V. G., & Karpov, V. G, "Evidence of field-induced nucleation switching in opal: VO_2 composites and VO_2 films", Physical Review B 85, 024110 (2012).

[35] Strickland, James A., and Gordon Lang. "Time-domain reflectometry measurements" Tektronix, 1970.

[36] Waser, R. & Aono, M. Nanoionics-based resistive switching memories. *Nature. Mater*. **6,** 833–840 (2007)

[37] Strachan, J. P. et al., Direct identification of the conducting channels in a functioning memristive device. *Adv. Mat.,* **22(32),** 3573-3577, (2010).

[38] Tsendin, K. D. "Physics of Switching and Memory Effects in Chalcogenides." (2014).

[39] Hickmott, T. W., Impurity Conduction and Negative Resistance in Thin Oxide Films, *J. Appl. Phys.* **35,** 2118 (1964)

[40] Chopra, K. L., Avalanche-Induced Negative Resistance in Thin Oxide Films, *J. Appl. Phys.* **36,** 184 (1965)

[41] Argall, F., Switching phenomena in titanium oxide thin films, *Solid State Electronics,* **11,** 535 (1968)

[42] A. S. Alexandrov et al., Current-controlled negative differential resistance due to Joule heating in TiO2, *Appl. Phys. Lett.* **99,** 202104 (2011)

[43] Blonkowski, S., Regache, M., and Halimaoui, A., Investigation and modeling of the electrical properties of metal–oxide–metal structures formed from chemical vapor deposited Ta2O5 films, *J. Appl. Phys.* **90,** 1501 (2001)

[44] Zeng, W. et al., CVD of Tantalum Oxide Dielectric Thin Films for Nanoscale Device Applications, *J. Electrochem. Soc.* **151,** F172 (2004)

[45] Chaneliere, C., Autran, J. L., and Devine, R. A. B., Conduction mechanisms in Ta2O5/SiO2 and Ta2O5/Si3N4 stacked structures on Si, *J. Appl. Phys.* **86,** 480 (1999)

[46] Choi, W. K. and Ling, C. H., Analysis of the variation in the field-dependent behavior of thermally oxidized tantalum oxide films, *J. Appl. Phys.* **75,** 3987 (1994)

[47] Devine, R. A. B., Vallier, L., Autran, J. L., Paillet, P., and Leray, J. L., Electrical properties of Ta2O5 films obtained by plasma enhanced chemical vapor deposition using a TaF5 source, *Appl. Phys. Lett.* **68,** 1775 (1996)

[48] Fleming, R. M. et al., Defect dominated charge transport in amorphous Ta2O5 thin films, *J. Appl. Phys*. **88,** 850 (2000)

[49] Ielmini, Daniele. "Modeling the universal set/reset characteristics of bipolar RRAM by field- and temperature-driven filament growth." *Electron Devices, IEEE Transactions on* 58.12 (2011): 4309-4317.

[50] Russo, Ugo, et al. "Filament conduction and reset mechanism in NiO-based resistive-switching memory (RRAM) devices." *Electron Devices, IEEE Transactions on* 56.2 (2009): 186-192.

[51] Larentis, S., et al. "Filament diffusion model for simulating reset and retention processes in RRAM." *Microelectronic Engineering* 88.7 (2011): 1119-1123.

[52] Yalon, E., et al. "Evaluation of the local temperature of conductive filaments in resistive switching materials." *Nanotechnology* 23.46 (2012): 465201.

[53] Lu, Yi Meng, et al. "Thermographic analysis of localized conductive channels in bipolar resistive switching devices." *Journal of Physics D: Applied Physics*44.18 (2011): 185103.

[54] Yalon, Eilam, et al. "Thermometry of filamentary RRAM devices." (2015).

[55] D. Strukov et al., Appl. Phys. A 107.3 (2012): 509-518.

[56] B. Govoreanu et al., Electron Devices, IEEE Transactions on , vol.60, no.8, pp.2471,2478, Aug. 2013.

[57] R. Degraeve et al., 14-5, Symp. on VLSI-T 2015

[58] C.-S. Poon et al., Front. Neurosci, vol. 5, no. 108, pp. 1-3, 2011

[59] Nikonov, Dmitri E., et al. "Coupled-oscillator associative memory array operation." *arXiv preprint arXiv:1304.6125* (2013).

[60] Hoppensteadt, Frank C., and Eugene M. Izhikevich. "Oscillatory neurocomputers with dynamic connectivity." *Physical Review Letters* 82.14 (1999): 2983.

[61] Kaka, Shehzaad, et al. "Mutual phase-locking of microwave spin torque nano-oscillators." *Nature* 437.7057 (2005): 389-392.

[62] Parihar, Abhinav, et al. "Synchronization of pairwise-coupled, identical, relaxation oscillators based on metal-insulator phase transition devices: A model study." *Journal of Applied Physics* 117.5 (2015): 054902.

[63] Driscoll, T. et al, "Current oscillations in vanadium dioxide: Evidence for electrically triggered percolation", Phys. Rev. B, vol. 86-9, pp. 094203, 2012

[64] Nardone, M. et al., "Relaxation oscillations in chalcogenide phase change memory", Jour. App. Phys., vol.107, no.5, Mar 2010

[65] M. D. Pickett and R. S. Williams, "Sub-100fJ and sub-nanosecond thermally driven threshold switching in niobium oxide crosspoint nanodevices," Nanotechnology, vol. `, number 21, 215202

[66] Sakai, Joe, "High-efficiency voltage oscillation in VO2 planer-type junctions with infinite negative differential resistance", Journal of Applied Physics, 103, 103708 (2008)

[67] Schmidt, Pierre E., and Roberto C. Callarotti. "The operation of thin film chalcogenide glass threshold switches in the relaxation oscillation mode." Thin Solid Films 42.3 (1977): 277-282.

[68] Karpov, V. G., M. Nardone, and M. Simon. "Thermodynamics of second phase conductive filaments." Journal of Applied Physics 109.11 (2011): 114507.

[69] Lee, M.-J. et al., "A fast, high-endurance and scalable non-volatile memory device made from asymmetric Ta2O5−x/TaO2−x bilayer structures," Nature Materials 10, 625–630 (2011)

[70] Adler, David, Heinz K. Henisch, and Nevill Mott. "The mechanism of threshold switching in amorphous alloys." Reviews of Modern Physics 50.2 (1978): 209.

[71] Pryor, R. W., and H. K. Henisch. "Nature of the on-state in chalcogenide glass threshold switches." Journal of Non-Crystalline Solids 7.2 (1972): 181-191.

[72] Adler, D., et al. "Threshold switching in chalcogenide-glass thin films." Journal of Applied Physics 51.6 (1980): 3289-3309.

[73] Jackson, T.C., Sharma, A. A., Bain, J.A., Weldon, J.A., Pileggi, L., "An RRAM-Based Oscillatory Neural Network", 2015 IEEE 6th Latin American Symposium on Circuits and Systems (LASCAS), 2015

[74] M. P. Shaw and I. J. Gastman "Circuit controlled current instabilities in s-shaped negative differential conductivity elements", Appl. Phys. Lett., vol. 19, no. 7, pp.243 -245 1971

[74] Sharma, Abhishek A., Mohammad Noman, Marek Skowronski, and James A. Bain. "Comparison of electric field dependent activation energy for electroformation in TaO x and TiO x based RRAMs." In Integrated Reliability Workshop Final Report (IRW), 2013 IEEE International, pp. 146-149. IEEE, 2013.

[75] Kostylev, S. A., and V. A. Shkut. "Electronic switching in amorphous semiconductors." Kiev Izdatel Naukova Dumka 1 (1978).

[76] Ielmini, Daniele, and Yuegang Zhang. "Analytical model for subthreshold conduction and threshold switching in chalcogenide-based memory devices." Journal of Applied Physics 102.5 (2007): 054517.

[77] Sharma A. A., Karpov, I.V., Kotlyar, R., Skowronski, M., Bain, J.A., "Temporal Dynamics of Electroforming in Binary Metal Oxide-based Resistive Switching Memory", Jour. Appl. Phys. 2015

[78] Hoppensteadt, Frank C., and Eugene M. Izhikevich “Synchronization of laser oscillators, associative memory, and optical neurocomputing,” *Physical Review E*, vol. 62, no. 3, pp. 4010–4013, 2000.

[79] Kwon, Jonghan, Abhishek A. Sharma, James A. Bain, Yoosuf N. Picard, and Marek Skowronski. "Oxygen Vacancy Creation, Drift, and Aggregation in TiO2-Based Resistive Switches at Low Temperature and Voltage." Advanced Functional Materials 24.35 (2015): 1616-3028.

[80] Hoppensteadt, Frank C., and Eugene M. Izhikevich , “Synchronization of MEMS resonators and mechanical neurocomputing,” *Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on*, vol. 48, no. 2, pp. 133–138, 2001

[81] R. Hölzel and K. Krischer, “Pattern recognition with simple oscillating circuits,” *New Journal of Physics*, vol. 13, no. 7, p. 073031, 2011.

[82] Van der Pol, Balth. "LXXXVIII. On “relaxation-oscillations”." The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2, no. 11 (1926): 978-992.

[83] Shaw, M. P., H. L. Grubin, and I. J. Gastman. "Analysis of an inhomogeneous bulk" S-shaped" negative differential conductivity element in a circuit containing reactive elements." Electron Devices, IEEE Transactions on 20, no. 2 (1973): 169-178.

[84] Jackson, Thomas C., et al. "Oscillatory Neural Networks Based on TMO Nano-Oscillators and Multi-Level RRAM Cells." (2015).

[85] Hammarlund, Per, and Örjan Ekeberg. "Large neural network simulations on multiple hardware platforms." Journal of computational neuroscience 5.4 (1998): 443-459.

[86] Widrow, Bernard, David E. Rumelhart, and Michael A. Lehr. "Neural networks: Applications in industry, business and science." Communications of the ACM 37.3 (1994): 93-105.

[87] Kuzum, D., Jeyasingh, R. G., Lee, B., & Wong, H. S. P. (2011). Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing. Nano letters, 12(5), 2179-2186.

[88] Yu, S., Gao, B., Fang, Z., Yu, H., Kang, J., & Wong, H. S. (2012, December). A neuromorphic visual system using RRAM synaptic devices with sub-pJ energy and tolerance to variability: experimental characterization and large-scale modeling. In Electron Devices Meeting (IEDM), 2012 IEEE International (pp. 10-4). IEEE.

[89] Shibata, Tadashi, Renyuan Zhang, Steven P. Levitan, Dmitri E. Nikonov, and George I. Bourianoff. "CMOS supporting circuitries for nano-oscillator-based associative memories." In Cellular Nanoscale Networks and Their Applications (CNNA), 2012 13th International Workshop on, pp. 1-5. IEEE, 2012.

[90] Neogy, Arkosnato, and Jaijeet Roychowdhury. "Analysis and design of sub-harmonically injection locked oscillators." In Proceedings of the Conference on Design, Automation and Test in Europe, pp. 1209-1214. EDA Consortium, 2012.

[91] Strogatz, Steven H. "From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators." Physica D: Nonlinear Phenomena 143, no. 1 (2000): 1-20.

[92] Dal Toso, Stefano, Andrea Bevilacqua, Marc Tiebout, Stefano Marsili, Christoph Sandner, Andrea Gerosa, and Andrea Neviani. "UWB fast-hopping frequency generation based on sub-harmonic injection locking." Solid-State Circuits, IEEE Journal of 43, no. 12 (2008): 2844-2852

[93] Acar, Mustafa, Domine Leenaerts, and Bram Nauta. "A wide-band CMOS injection-locked frequency divider." In Radio Frequency Integrated Circuits (RFIC) Symposium, 2004. Digest of Papers. 2004 IEEE, pp. 211-214. IEEE, 2004.

[94] Yang, Tao, Richard A. Kiehl, and Leon O. Chua. "Tunneling phase logic cellular nonlinear networks." International Journal of Bifurcation and chaos 11, no. 12 (2001): 2895-2911.

[95] Lai, Xiaolue, and Jaijeet Roychowdhury. "Fast simulation of large networks of nanotechnological and biochemical oscillators for investigating self-organization phenomena." In Design Automation, 2006. Asia and South Pacific Conference on, pp. 6-pp. IEEE, 2006.

[96] Bersuker, G., et al. "Metal oxide resistive memory switching mechanism based on conductive filament properties." Journal of Applied Physics 110.12 (2011): 124518.

[97] Chang, S. H., et al. "Occurrence of both unipolar memory and threshold resistance switching in a NiO film." Physical review letters 102.2 (2009): 026801.

[98] Peng, Hai Yang, et al. "Deterministic conversion between memory and threshold resistive switching via tuning the strong electron correlation." Scientific reports 2 (2012).

[99] Bae, Jieun, et al. "Coexistence of bi-stable memory and mono-stable threshold resistance switching phenomena in amorphous NbOx films." Applied Physics Letters 100.6 (2012): 062902.

[100] Stefanovich, G., A. Pergament, and D. Stefanovich. "Electrical switching and Mott transition in VO2." Journal of Physics: Condensed Matter 12.41 (2000): 8837.

[101] Sharma, Abhishek A., et al. "Electronic Instabilities Leading to Electroformation of Binary Metal Oxide-based Resistive Switches." Advanced Functional Materials (2014).

[102] Guan, Ximeng, Shimeng Yu, and H-SP Wong. "On the switching parameter variation of metal-oxide RRAM—Part I: Physical modeling and simulation methodology." Electron Devices, IEEE Transactions on 59.4 (2012): 1172-1182.

[103] K. Kardell, C. Radehaus, R. Dohmen, and H.-G. Purwins, “Stable multifilament structures in semiconductor materials based on a kinetic model,” *Journal of Applied Physics*, vol. 64, no. 11, pp. 6336–6338, 1988.

[104] V. Kratyuk, P. K. Hanumolu, U.-K. Moon, and K. Mayaram, “A design procedure for all-digital phase-locked loops based on a charge-pump phase-locked-loop analogy,” *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS PART 2 EXPRESS BRIEFS*, vol. 54, no. 3, p. 247, 2007.

[105] J. J. Yang, M.-X. Zhang, M. D. Pickett, F. Miao, J. P. Strachan, W.-D. Li, W. Yi, D. A. Ohlberg, B. J. Choi, W. Wu *et al.*, “Engineering nonlinearity into memristors for passive crossbar applications,” *Applied Physics Letters*, vol. 100, no. 11, p. 113501, 2012.

[106] F. Chudnovskii, L. Odynets, A. Pergament, and G. Stefanovich, “Electroforming and switching in oxides of transition metals: The role of metal–insulator transition in the switching mechanism,” *Journal of Solid State Chemistry*, vol. 122, no. 1, pp. 95–99, 1996.

[107] L. Goux, A. Fantini, G. Kar, Y. Chen, N. Jossart, R. Degraeve, S. Clima, B. Govoreanu, G. Lorenzo, G. Pourtois *et al.*, “Ultralow sub-500na operating current high-performance tinal 2 o 3 hfo 2 hftin bipolar rram achieved through understanding-based stack-engineering,” in *VLSI Technology (VLSIT), 2012 Symposium on*. plus 0.5em minus 0.4emIEEE, 2012, pp. 159–160.

[108] Google Glass, URL: https://plus.google.com/+projectglass

[109] R. Szeliski, et al., "A comparative study of energy minimization methods for markov random fields with smoothness-based priors," IEEE Trans. on Pattern Analysis and Machine Intelligence, 2008.

[110] J. Choi and R. Rutenbar, “Hardware implementation of MRF map inference on an FPGA platform,” Field Programmable Logic and Applications, 2012.

[111] D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” International Journal of Computer Vision, 2002.

[112] D. Scharstein and R. Szeliski, “High-accuracy stereo depth maps using structured light,” IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003.

[113] V. Kolmogorov, “Convergent tree-reweighted message passing for energy minimization,” IEEE Transaction on Pattern Analysis and Machine Intelligence, 2006.

[114] M. J. Wainwright, et al., “MAP estimation via agreement on trees: message-passing and linear-programming approaches,” IEEE Transactions on Information Theory, 2005.

[115] Sharma, Abhishek A., et al. "Hardware-efficient stereo estimation using a residual-based approach." Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013.

Appendix A:

Review of thermometry & modeling parameters, and comparison with existing methodologies

This supplementary material seeks to serve two main purposes – first, we try to extend the arguments made for Ta_2O_{5-x} to TiO_2, in order to prove that the phenomena and proposed mechanisms are universal for all switching oxides; secondly, we explain the experimental details involved in sample preparation, characterization and, modeling and analysis needed for Chapters 2 and 3.

A.1. Presence of negative differential resistance in TiO_{2-x}-based memristors

We choose TiO_{2-x} in order to explore the generality of observations of NDR and associated filamentation. Titania was long regarded as a prototypical memristive material with early demonstrations of relevant phenomena. As a control experiment, we test a canonical TiO_{2-x} cross-bar device with a DC sweep from 0 to 17 V (at source), without electroforming it. Figure A.1 shows the DC *I-V* characteristics of a TiO_{2-x} device with a source resistance of 13.5 kΩ. Similar to Ta_2O_{5-x} samples, the device does show a clear CC-NDR behavior with thermal and electronic contributions. The SEM micrograph in the inset shows the state of the device after it has been subjected to the DC sweep. The vertical and horizontal rectangular traces represent the bottom and the top Pt electrodes respectively. The bright contrast on the left represents the grain

growth in the Pt top electrode, which occurs because of local temperature increase. This is consistent with the sharp temperature rise occurring due to electronic branch of filamentation. This is completely consistent with the observations in Ta_2O_{5-x}. This indicates that the voltage and temperature non-linearities observed in Ta_2O_{5-x} are not unique to a single material system but are general to most resistive switching glasses.

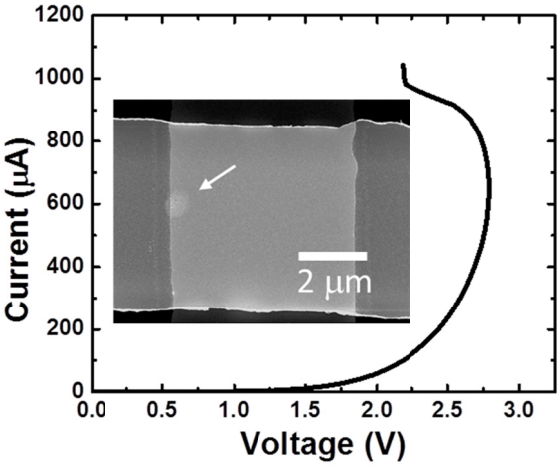


Figure A.1: Negative differential resistance observed in 20 nm thick TiO_{2-x} cross-bars (black curve), before electroforming. SEM image showing evidence of local heating in the pre-forming regime (shown as lighter contrast)

A.2. Pulsed I-V Characterization

In order to analyze the dynamics of the electroformation process, we employ time domain transmissometry (TDT). An optical micrograph of the device and the TDT setup schematic is shown in Fig. A.2. The method involves launching pulses from the pulse generator (Agilent 81110/81111A) through a transmission line with characteristic impedance (Z_0) of 50 Ω. The signal is delivered to the device using GSG RF probes. The transmission coefficient changes as a

function of device impedance. The transmitted pulse is then observed on an oscilloscope (Agilent MSO 6104A and DSO 80804A). Both, the pulse generator output impedance and the oscilloscope input impedance are set to 50 Ω to prevent reflections. We use the input pulse voltage and the transmission coefficient to calculate resistance, voltage across the device and current flowing in the circuit.

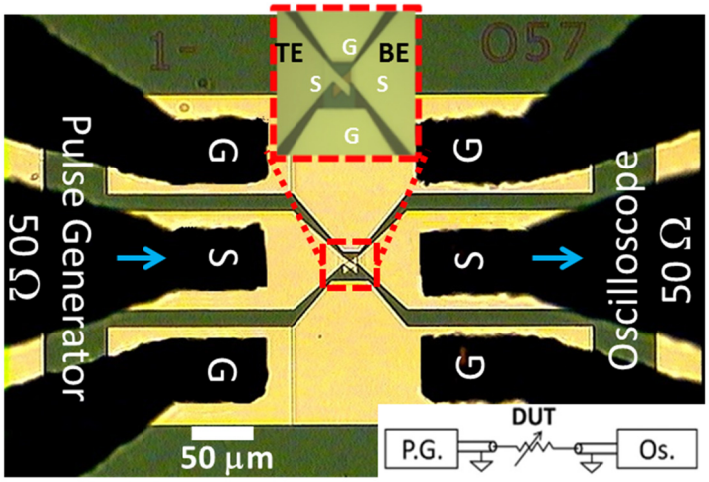


Figure A.2: 5 µm cross-bar Ta_2O_{5-x} and TiO_{2-x} devices. Ground-Signal-Ground (GSG) RF probes used for time domain transmissometry (TDT) measurements with device connected in series with the pulse generator and oscilloscope (both at 50 Ω). Circuit schematic shown in the inset.

In order to obtain the *I-V* characteristics at the stage temperature, narrow 5 ns pulses of increasing amplitude were delivered with this setup and the *I-V* characteristics were generated by sampling the current and voltage from the first 1 ns of the pulse. Such short pulses prevent the device from undergoing self-heating. Each *I-V* point during the dynamic excitation (long pulses) or DC sweeps can be mapped onto this *I-V-T* calibration. Thus, if the voltage and current are known, one can extract the device temperature.

I-V data from long 5 μs pulses was obtained in a similar way. Voltage and current through the device changes with temperature for the first 2.5 μs (thermal time constant). We can discard the temporal data and map the voltage and current on an *I-V* plane to give the evolution from pulsed *I-V* to DC *I-V* (shown in Chapter 2).

A.3. Temperature dynamics

Chapter 2 presents the evidence for current constriction before the onset of permanent resistance change. Specifically, we used the evidence of grain growth caused by local current flow during short pulses (Fig. 2.7). Similar conclusions can be reached using the thermometry technique described above. Figures A.3 (a) and A.3 (b) show the experimental voltage and current transients for a 1 μs long pulse.

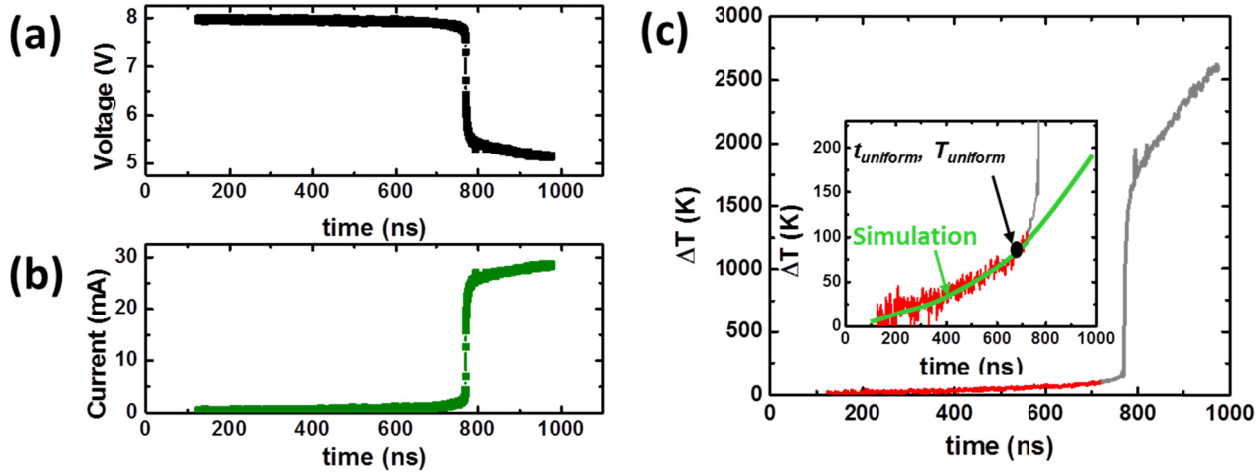


Figure A.3: Voltage **(a)** and current **(b)** response to a 1 μs long pulse that resulted in electroforming. The electroforming event is represented as the sudden decrease in the voltage across the device and increase in current (at ~ 780 ns). **(c)** Temperature response of the device undergoing electroforming shows correctly predicted (red) and simulated (green) temperature. The deviation of prediction from simulation represents the change in the conducting area i.e. filamentation.

Knowing the voltage and the current transient as a function of time we can determine the temperature transient. It is important to note that the procedure will give correct temperature values only if the current flow during the pulse is spatially uniform. We now use finite element thermal modeling to estimate the expected temperature excursion from the measured pulse power and known material parameters. Figure A.3(c) shows the extracted and simulated temperature

transients. The two transients are very consistent with each other till 780 ns. After 780 ns, for the very same measured input power, the extraction and the simulated temperatures deviate from each other indicating that the original assumption about uniform conduction through the film is no longer valid. Also, it must be noted that this deviation occurs before permanent change (conducting filament) has taken place. This behaviour is also seen in other oxides like TiO_{2-x}.

A.4. Extraction of filament size and estimate of temperature during filamentation

After filamentation onset, the temperature reached is a strong function of filament diameter, with greater current localization leading to higher temperature. Rather than simply postulating a filament size and then estimating the temperature based on that assumption, we have attempted to extract a filament size self-consistently from our data by reconciling temperature rise as estimated from thermal modeling and temperature rise estimated from conductivity change.

The electronic filaments are usually thought of as a continuous high current domain extending from the top electrode to the bottom electrode. Since the current flow is primarily through this current filament, the power dissipation also occurs inside this temporary filament. The R_{th} discussed in the main text represents the thermal resistance that is connected between the filament as the heat source and the thermal ground, which can be easily calculated from material properties and finite element simulation. We use Comsol Multiphysics finite element method (FEM) solver for the calculation of the R_{th} as the ratio of the rise in temperature experienced with unit increase in the power dissipated in the filament, at steady state. Figure S4(a) shows the simulation setup used for the FEM solver. The thermal properties assumed for the simulation are summarized in Table 1. The results of the simulation are summarized in Fig. A.4(b). From this

figure, then, it is possible to estimate the temperature given a filament radius and the measured power dissipated in the device after filamentation onset.

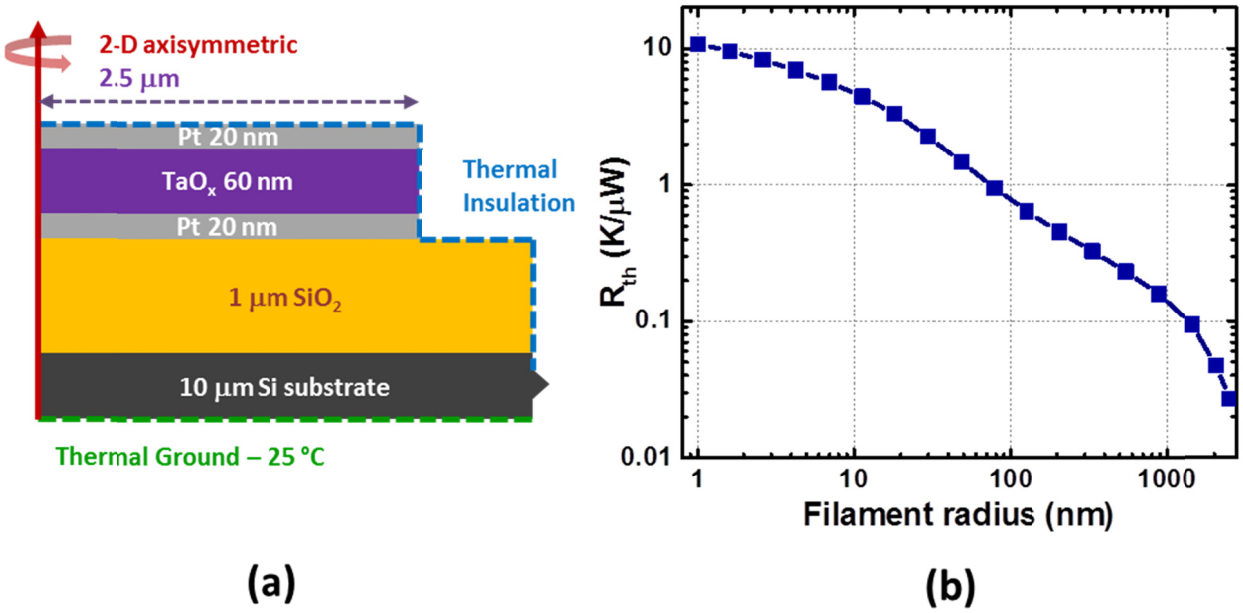


Figure A.4: (a) Schematic of the Ta_2O_{5-x} device used in Comsol Multiphysics electrothermal simulations. **(b)** Simulated thermal resistance (R_{th}) as a function of filament radius.

We also know that when we estimate ΔT from the pair of *I-V* coordinates, we assume that the current flows through a filament of radius 2.5 μm, i.e. uniform conduction. Thus, for a device undergoing filamentation, we will always underestimate the local filament temperature. Scaling the current axis in the *I-V-T* thermometer (Fig. 2.12(b)) by the ratio of the uniform device area (radius r = 2.5 μm) and the filament radius, r we get a new range of temperatures for an effectively higher current density. Thus, the corrected curve so constructed is a SECOND, independent figure we can consult to extract a temperature from an assumed filament radius and known *I* and *V* measurements.

Self-consistency is achieved by finding the single value of radius at each I, V point which simultaneously satisfies both of the above constructions. In the latter *I-V-T* curve it has been necessary to extrapolate from measured conductivities (room temperature to 200 °C) up to much higher temperatures (600 °C), but we believe the error introduced by this extrapolation is modest due to the very well-behaved dependence of conductivity on temperature in the measured range.

In the uniform conduction region, we the device starts off with a value of R_{th} of ~0.025 K/μW which corresponds to a filament radius of 2.5 μm. By enforcing the above constraints, we find that in the thermal NDR regime, the R_{th} is rises by an order of magnigtude (~0.2 K/μW) and the filament radius shrinks to 750 nm. As the device enters into the electronic NDR regime, the R_{th} increases by a factor of 10 again to a value > 2 K/μW which corresponds to a sub-10 nm diameter filament.

It must be noted that these calibrations (both experiment and modeling) have to be re-done for samples with different electrical and thermal properties.

A.5. Comparison with MIS-BT –based thermometry

Recently, two experimental temperature evaluation methods were reported that do not require assumptions on the filament properties. One method is based upon short pulse measurements (discussed in the thesis Chapters 2 and 3), and the other is based upon measurement of minority carrier thermionic emission current in a 3-terminal MIS structure [52]. It is the purpose of this section is to compare these two techniques and discuss their underlying physics.

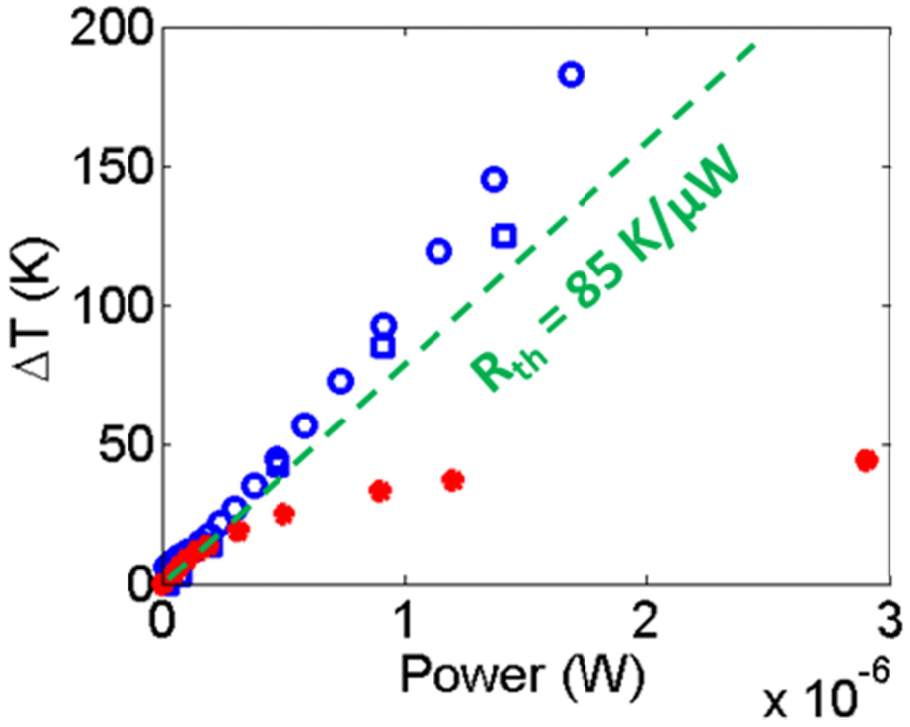


Fig. A.5. Comparison between extracted temperatures as a function of applied power for the two techniques.

Fig. A.5 shows the increase in temperature in the LRS plotted as a function of applied power, for both the techniques (MIS-BT in blue and pulsed thermometry in red). It is clear that at low-biases, the two techniques show somewhat similar slope (R_{th}), implying that the heated zone is the same and that the definitions of temperature for the two techniques are similar, namely most of the heating takes place at the filament, and local filament temperature is measured. As the bias increases, the temperature increase sensed by the pulsed thermometry reduces; perhaps due to the current spreading and associated averaging. On the other hand, the MIS-BT continues to detect the peak temperature at the tip of the filament. It is important to note that we attribute the changes in R_{th} to the expansion of heated zone, rather than changes in the dimensions of the filament since we consider volatile changes before any non-volatile switching takes place.

Nonetheless, the filaments in the two devices may have different diameters due to the difference in structure (MIM vs MIS, pulsed measurement sample twice thicker) and different forming conditions. It is expected that the filament in the MIS sample would have a gap, and therefore its R_{th} is expected to be higher. As the region experiencing peak temperature participates in switching events, the temperatures detected by the MIS-BT structure serve as important data points. Similarly, the non-linear increase in temperature (vs. power) detected by the pulsed thermometry, sheds light on the self-limiting mechanisms responsible device functioning at biases close to switching voltages.

Comparing between the two techniques, the temperatures extracted using the pulsed method are expected to be slightly lower. The reason is that the MIS-BT measures the local peak temperature at the filament tip, whereas the pulsed measurement averages the temperature across the device. Inspection of Fig. A.5 shows that for the same functional RRAM layer, measured under similar forming conditions the evaluated temperatures using the two techniques are quite comparable at low-biases, with the pulsed measurement showing a slightly lower temperature. It is evident that the two methods complement each other. The pulsed thermometry is particularly useful when the RRAM resistance (either LRS or HRS) is $<10/G_0$ while the MIS-BT works well when RRAM resistance $>10/G_0$. Moreover, the MIS-BT offers high-precision at low power regime, below ~10μW whereas the pulsed measurements can be carried out in a wider range of applied power.

RRAM filament thermometry has become a key for understanding the physics of RRAM devices due to the crucial role of the temperature in these devices. We have reviewed here the two main experimental methods to measure the filament temperature in RRAM devices. The two methods were compared on the basis of measurements carried out on the same functional RRAM layer.

Comparison between our experimental results and thermal simulations indicate that at low current compliance (~10 μA) and under low power conditions (<10 μW) the filament dimensions are below 5 nm. For higher values of the power the thermal resistance is reduced by a factor of ~100, perhaps, due to the expansion of heat generated zone around the filament (with an additional possibility of an increase in the thermal conductivity with temperature). We concluded that the two thermometry techniques are complementary; the MIS-BT method is useful when the resistance of the device is in the range ~100 kΩ - 100 MΩ, under operating power conditions relatively low (<10μW), whereas the pulsed thermometry is more suitable for devices having resistance <500 kΩ in the range of >1 μW applied power. These values of resistance can be either LRS or HRS of the RRAM device.

In order to extend the extraction range of the MIS-BT technique in future work, small area devices may be used to apply high speed pulsed measurements. Bipolar transistors having similar structures were demonstrated that exhibit cut-off frequency in the range of ~100 GHz, though the high frequency operation of a transistor having a filament emitter is yet to be evaluated.

The thermometry techniques offer insights into: (1) Peak temperature in the filament, (2) Geometry of the filament (3) Filament growth. (4) Dynamics of the heated zone. We point out that these thermometry tools can be applied, in principle, to any thin film having filamentary conduction. These experimental techniques are the tools to design materials and devices with optimized thermal and electrical characteristics. Furthermore, these methods can be used as characterization techniques to understand temperature-mediated physical processes during switching.

References

[1] M. Asheghi, M.N. Touzelbaev, K.E. Goodson, Y.K. Leung, and S.S. Wong, Journal of Heat Transfer 120, 30 (1998).

[2] D R M Crooks et al, Experimental measurements of mechanical dissipation associated with dielectric coatings formed using SiO2, Ta2O5 and Al2O3, Class. Quantum Grav. 23 4953, 2006

Tables:

Table I: Material properties for thermal simulations

	Thermal conductivity [W/(m K)]	Density [kg/m^3]	Heat Capacity [J/(kg K)]
Si	148[1]	2320	705
SiO_2	1.3	2648	733
Pt	71.6	21450	130
Ti	22	4500	520
Ta_2O_{5-x}	1.2*	6800[2]	693

All values are obtained from Comsol Multiphysics materials library, unless otherwise noted by explicit reference

*Measured independent using frequency domain thermoreflectance (FDTR). Also obtained and used thermal boundary resistance of 25 m^{2}K/GW between Ta_2O_{5-x} and Pt

Table 1: Table with assumed thermal parameters used in the simulation. The thermal conductivity of the functional layer was measured independently using frequency domain thermoreflectance (FDTR).

Appendix B:

Low-temperature forming process

In this appendix, we attempt to determine the lowest device temperature at which the electroformation can still occur. In order to accomplish this, we optimized the device structure and used the stage temperature of 10 K. The true device temperature, including self-heating, was determined using a novel thermometry approach.[12,13] The results , allow us to comment on the nature of processes involved. This section has been created in collaboration with Darshil K. Gala.

The devices used in this study were 5×5 μm^2 crossbars of 10 nm TiN/ 10 nm Ta/ TaO_x/10 nm TiN with 88 nm thick amorphous TaO_x functional layers. All layers have been deposited by sputtering, on a Si substrate with 1 μm thick thermal oxide. Details of the fabrication process have been discussed in our earlier publications.[12,13] Figure 1 shows the quasi-DC electroforming characteristics for TaO_x devices at stage temperatures from 10 K to 300 K. The circuit consisted of the voltage source, device under test and a load resistor of 25.5 kΩ connected in series. The direction of the voltage sweep is indicated by the arrows in the figure. Each test has been performed only once on different but nominally identical devices (< 1% variability in the threshold voltages across die, as fabricated) as the testing permanently changes the device characteristics. For example, the green curve in Fig. 1 represents the forming curve for a stage temperature of 200 K. *I-V* characteristics of the device shows four distinct regions. The first

region of the characteristics is the high resistance of the as-fabricated device with the current increasing super-linearly up to 6 V. In the second region at 7 V, the tangent to the *I-V* curve becomes vertical and for higher currents, the *I-V* slope turns negative corresponding to a negative differential resistance (NDR) exhibited by the device. Such characteristics are typically a consequence of the constriction of current flow to a narrow filament.[14] Also, our recent analysis of TaO_x devices confirmed the filament formation in the NDR region.[12] The filament, thus formed, is volatile[13] as the *I-V* curves are fully reversible up to this point. While the temperature may increase during the initial thermal NDR [Sharma Adv Func], the localization to nm sized filament does not occur until the snap has reached completion, making temperature rise very similar to the uniform conduction case [Sharma, Adv Func].

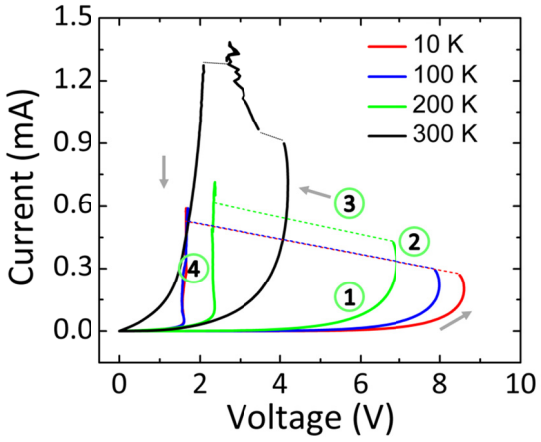


Fig. B.1. Quasi-DC electroforming curves for TaO_x devices as a function of stage temperature.

Typically, the threshold voltage is defined as the voltage at which the device undergoes an abrupt transition from the unformed high resistance state to the conducting state which could either be volatile or not. Frequently, this voltage corresponds to the highest voltage on the *I-V* curve at which the device differential resistance becomes zero. In a circuit with low series resistance, the

'snap' to high conductivity state would occur at this point. However, in our circuit, as shown in Fig. B.1, the voltage at which the abrupt transition takes place is lower (6.5 V) than zero differential resistance point and 'snap' does not occur until the value of negative differential resistance exceeds the series resistance in the circuit.[12] The 'snap' which is the third region in Fig. B.1 corresponds to the inclined load line of the series resistor. The fourth and last region of the characteristics is the nonlinear OFF state of the formed device.[12,15]

I-V characteristics observed for 10 K, 100 K and 300 K were similar to that obtained at 200 K. Threshold voltages monotonically decrease from about 8.2 V to 4 V between 10 K and 300 K, while the current at threshold increases from 0.3 mA to 0.9 mA. *I-V* characteristics for 300 K (black curve), after entering the NDR region, show two snaps (the first one occurs at 4 V) marked with dotted lines. The filament formed at this stage is volatile and the device returns back to the pristine state after the bias is reduced.[12] As the current is increased further, the device snaps again to the non-volatile OFF-state with permanently lowered resistance. After electroforming all devices exhibited stable switching.

The important conclusion of the above observations is that TaO_x devices undergo both threshold (volatile) and memory (permanent) switching at stage temperatures as low as 10 K. However, the true temperature of the filament where all important processes take place is considerably higher due to Joule heating. We have estimated this temperature using a thermometry approach discussed in detail in recent publications.[12,13] The true *I(V, T)* characteristics of the device have been obtained by measuring the voltage and current within the first 5 ns of a rectangular voltage pulse. In order to avoid reflections that typically obscure fast transients, the devices have been designed to be a matched 50 Ω waveguide in a Ground-Signal-Ground configuration with the measurements performed by time domain transmissometry (TDT) method. The measured and

simulated thermal time constant of our devices is about 2 μs[13] and is about three orders of magnitude longer than the pulse length. The estimated change of temperature during the pulse for all investigated voltages was below 5 K. Thus, one can consider the plots presented in Fig. 2 as true *I(V, T)* of the device, where the temperature is that of the stage without any distortion caused by the Joule heating. The results can be used as a look-up table to determine the true device temperature during the quasi-DC sweeps.

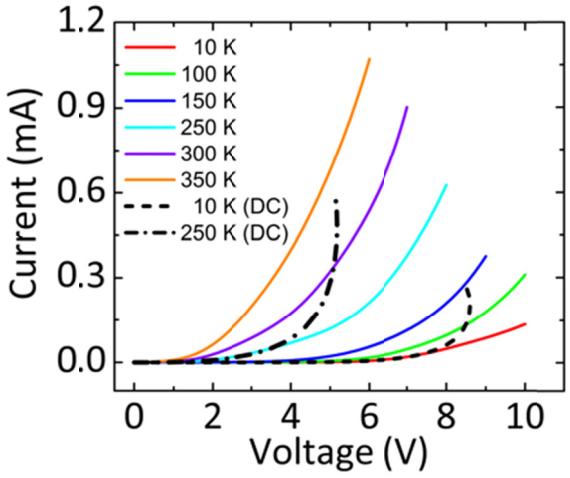


Fig. B.2. Pulsed $I(V,T_{stage})$ characteristics obtained using 5 ns pulses (continuous lines) obtained on unformed devices. The dashed and dash-dotted lines correspond to quasi-DC $I(V,T_{stage})$ characteristics including Joule heating.

The temperature at the point of switching was estimated superimposing pulsed *and* quasi-DC characteristics. In Fig. B.2, quasi-DC curve at 10 K stage temperature (dashed line) overlaps with the corresponding pulsed *I(V, T)*) curve up to 7.5 V indicating that the temperature of the device till this point is equal to the stage temperature. This is expected, as the effects of Joule heating are significant only at high dissipated power. Beyond 8 V, the quasi-DC line curves upward with

conductivity increasing with device temperature. At some point close to the knee on the *I-V* curve, the current constricts, forming a filament. For the sake of this discussion, let us assume that the constriction occurs at the snap. Since the current at this point is uniform and the DC curve intersects the pulsed *I-V* at the stage temperature of 150 K, we can assign this temperature to the device at the moment of filament formation. If the constriction happens earlier at lower voltage and current, the corresponding temperature would clearly be lower. The 150 K then represents the upper bound of temperature at which the filament forms. Similar arguments can be used at other stage temperatures giving the temperature at the point of switching between 150 K and 350 K for stage temperatures between 10 K and 250 K.

The temperature of the device right before threshold switching event was also simulated using the COMSOL Multiphysics finite element method solver for the range of stage temperatures. Simulation setup and the thermal parameters used have been discussed in our previous work[12,13] with the dissipated power taken from the experiment. The results were very similar to the temperatures extracted using experimental procedure described above.

The widely accepted mechanism for the formation of the conductive filament in TiO_x, TaO_x, and HfO_x-based devices is the creation, drift, and accumulation of oxygen vacancies. These are the elementary processes leading to formation of a secondary oxygen-deficient phase in the functional layer.[6-9] The drift velocity scales with the coefficient of diffusion, which has been extensively studied in oxides. Nakamura *et al.*[16] investigated the oxygen diffusion in amorphous oxides using pair distribution analysis by transmission electron microscopy and reported an activation energy (E_A) of 1.2 ± 0.1 eV. Lowest E_A calculated for orthorhombic Ta_2O_5 was approximately 0.7 eV, obtained by finding the minimum energy barrier from one lattice site to an adjacent one using the nudged elastic band method.[17] Also, most device modeling efforts

reproduced the device characteristics using values of $E_A \sim 1$ eV.[16, 18] The large value of the activation energy allows for the stability of the filament while still making it possible to switch in high electric fields and high temperatures. The equation describing the time between diffusion jumps of oxygen vacancy is:

$$\frac{1}{\tau} = \nu_0 \exp(-\frac{E_A}{k_B T}) \qquad (1)$$

where, ν_0 is the attempt frequency which is taken as 10^{13} s^{-1}, E_A is the activation energy of diffusion and k_B is the Boltzmann constant. The calculated value of τ at the extracted temperature of 150 K for an E_A of 1 eV is 4×10^{20} s. This time is many orders of magnitude longer than the age of the universe. This result in itself is a conclusive argument that the threshold switching in TaO_x-based devices cannot rely on diffusion of oxygen ions.

Additional information about the nature of the threshold switching during electroformation was obtained from the analysis of the switching dynamics. The process of threshold switching is not instantaneous and requires a certain incubation time before the device switches to the volatile ON-state (this corresponds to the non-volatile OFF state).[13,19] Incubation time in our devices was measured using TDT where the device under test was subjected to a series of rectangular pulses and the device voltage was monitored as a function of time (Fig. B.3). At a certain pulse amplitude, the resistance of the device decreases abruptly during the pulse and the device voltage drops. The delay between the leading edge of the pulse and the voltage drop was recorded as the incubation time. This was repeated for different pulse widths and stage temperatures. What is interesting is that the incubation time corresponding to the same voltage is identical for 10 K and 100 K. Inset shows the dependence of the incubation time on the inverse of applied voltage for stage temperatures from 10, 100 and 300 K. It is apparent that the incubation times strongly

depend on voltage at all temperatures but do not depend on temperature between 10 and 100K (corresponding to true device temperature of 150 K and 250 K). The activation energies determined for different voltages in this temperature range were below 30 meV. Thus, it can be inferred that if the mechanism leading to creation of the conductive filament is thermally activated, it has a very low activation energy. If we assume that the volatile ON state involves formation, drift, and coalescence of oxygen vacancies occurring in series, then the activation energy should be larger than the largest of the activation energies of individual processes. In other words it should exceed the activation energy for diffusion i.e. 1 eV. One should note that the activation energy for diffusion of ions decreases in high electric fields.[20] However, this effect is significant only for the values of field much larger than used in our experiments.

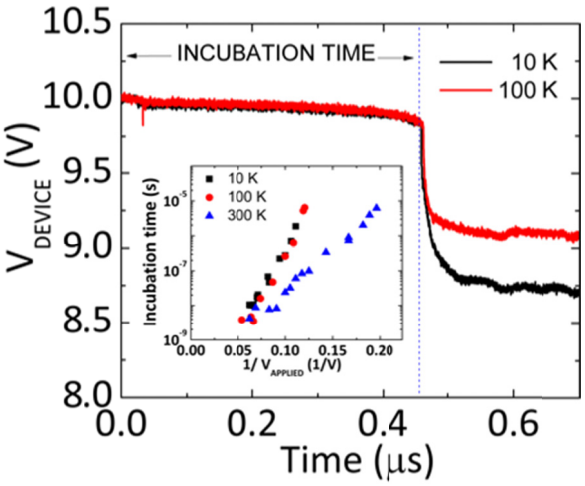


Fig. B.3. Transient Voltage at 10 K and 100 K with incubation time indicated as time to initiate threshold switching. (Inset) Incubation time versus inverse of applied voltage (1/ $V_{applied}$) for 10 K, 100 K and 300 K stage temperatures.

Most results on threshold switching are available for devices based on amorphous chalcogenides such as $Ge_2Sb_2Te_5$ with several different models proposed to explain this phenomenon. Adler *et al.*[21] suggested the double injection model with narrow barriers forming at both electrodes. Ielmini *et al*.[22] proposed a trap-hopping model in which the high electric fields lead to non-equilibrium distribution of trapped electrons that causes a sudden increase in the conductivity. The model features a critical power-density at which threshold switching can initiate. It is to be applicable not just to chalcogenide glasses but to most amorphous semiconductors featuring NDR. Noman *et al.*[13] presented the transient thermometry data, consistent with a model of filament formation based on charge trapping. They argued that the charge trapping can give rise to a local electric-field enhancement eventually causing a breakdown. Karpov *et al.*[23] suggested the field-induced nucleation model in which the critical size of conducting phase nucleates in presence of field and shunts the electrodes leading to threshold switching. Pergament *et al.*[24] proposed a threshold switching mechanism for VO_2 based on electronically induced Mott-Hubbard metal-insulator transition, occurring in the conditions of non-equilibrium carrier density in the presence of an applied electric-field. Several of these processes could have low activation energies and provide an acceptable origin of the threshold switching in binary oxides. While more experiments are needed to select a comprehensive predictive model, our data-driven analysis attempts to prove that the *onset* of forming does not lie in vacancy migration, and that if structural changes are to occur, they must be preceded by a reversible threshold switching process that cannot be mediated by creation or motion of vacancies.

In summary, quasi-DC measurements showed that TaO_x-based devices exhibit the typical electroforming characteristics at stage temperatures as low as 10 K. The temperatures at the point of transition to volatile low resistance state during the quasi-DC electroforming were extracted

using transient thermometry technique. The lowest temperature at which threshold switching was initiated was 150 K these device. The jump frequency of the oxygen vacancies at this temperature is 2.5×10^{-21} s^{-1}. This observation *excludes* the oxygen vacancies as being involved in threshold switching in oxide devices. Moreover, the incubation time needed to initiate threshold switching showed no change from 10 K to 100 K with extracted activation energies much lower than those needed for oxygen vacancy diffusion. Possible alternative models of the threshold switching have been discussed to explain the initiation of threshold switching in absence of oxygen vacancy migration.

References

[1] B. Govoreanu, G. S. Kar, Y. Chen, V. Paraschiv, S. Kubicek, A. Fantini, I. P. Radu, L. Goux, S. Clima, R. Degraeve *et al.*, IEEE IEDM 2011, (IEEE, Washington DC, 2011), p.31.6.1-31.6.4.

[2] R. Waser and M. Aono, Nature Mater. **6**, 833 (2007).

[3] H-S. P. Wong, H. Y. Lee, S. Yu, Y. S. Chen, Y. Wu, P. S. Chen, B. Lee, F. T. Chen, and M. J. Tsai, IEEE Proc. **100**, 1951 (2012).

[4] R. Waser, R. Dittmann, G. Staikov, and K. Szot, Adv. Mater. **21**, 2632 (2009).

[5] D-H. Kwon, K. M. Kim, J. H. Jang, J. M. Jeon, M. H. Lee, G. H. Kim, X-S Li, G-S Park, B. Lee, S. Han, *et al.*, Nat. Nanotechnol. **5**, 148 (2010).

[6] K. M. Kim, D. S. Jeong, and C. S. Hwang, Nanotechnology **22**, 254002 (2011).

[7] D. S. Jeong, H. Schroeder, U. Breuer, and R. Waser, J. Appl. Phys. **104**, 3716 (2008).

[8] J. J. Yang, F. Miao, M. D Pickett, D. A. A. Ohlberg, D. R. Stewart, C. N. Lau and R. S. Williams, Nanotechnology **20**, 215201 (2009).

[9] N. Ghenzi, D. Rubi, E. Mangano, G. Gimenez, J. Lell, A. Zelcer, P. Stoliar, and P. Levy, Thin Solid Films **550**, 683 (2014).

[10] A. S. Alexandrov, A. M. Bratkovsky, B. Bridle, S. E. Savel'ev, D. B. Strukov, and R. S. Williams, Appl. Phys. Lett. **99**, 202104 (2011).

[11] D. Ielmini, C. Cagli, and F. Nardi, Appl. Phys. Lett. **94**, 063511 (2009).

[12] A. A. Sharma, M. Noman, M. Abdelmoula, M. Skowronski, and J. A. Bain, Adv. Funct. Mater. **24**, 5522 (2014).

[13] M. Noman, A. A. Sharma, Y. M. Lu, R. Kamaladasa, M. Skowronski, P. A. Salvador, and J. A. Bain, Appl. Phys. Lett. **104**, 113510 (2014).

[14] B. K. Ridley, Proc. Phys. Soc. **82**, 954 (1963).

[15] M. Noman, A. A. Sharma, Y. Meng Lu, M. Skowronski, P. A. Salvador, and J. A. Bain, Appl. Phys. Lett. **102**, 023507 (2013).

[16] R. Nakamura, T. Toda, S. Tsukui, M. Tane, M. Ishimaru, T. Suzuk,i and H. Nakajima, J. Appl. Phys. **116**, 033504 (2014).

[17] R. Ramprasad, J. Appl. Phys. **95**, 954 (2004).

[18] S. Larentis, F. Nardi, S. Balatti, D. C. Gilmer, and D. Ielmini, IEEE Trans. Electron Devices **59**, 2468 (2012): 2468-2475.

[19] A. A. Sharma, I. V. Karpov, R. Kotlyar, J. Kwon, M. Skowronski, and J. A. Bain, J. Appl. Phys. **118**, 114903 (2015).

[20] M. Noman, W. Jiang, P. A. Salvador, M. Skowronski, and J. A. Bain, Appl. Phys. A **102**, 877 (2011).

[21] D. Adler, H. K. Henisch, and Sir N. Mott, Rev. Mod. Phys. **50**, 209 (1978).

[22] D. Ielmini, Phys. Rev. B **78**, 035308 (2008).

[23] V. G. Karpov, Y. A. Kryukov, S. D. Savransky and I. V. Karpov, Appl. Phys. Lett. 90, 123504 (2007).

[24] A.L. Pergament, P.P. Boriskov, A.A. Velichko, and N.A. Kuldin, J. Phys. Chem. Solids **71**, 874 (2010).

Appendix C:

Verilog-A model of S-NDR devices

In this appendix, we report a piece-wise linear fit based model that was developed in collaboration with Yunus Kesim.

```
`include "constants.vams"
`include "disciplines.vams"

module cubic_NDR(Vtop,Vbot);

inout Vtop,Vbot;

parameter Ith = 1e-6;
parameter Ih = 20e-6;
parameter R_ON = 500;
parameter R_OFF = 1000000;
parameter NDR = -3.157894736842105e+04;
parameter V1 = 1.031578947368421;
parameter V2 = 0.390000000000000;

electrical Vtop, Vbot;
real Rtemp, en_state, iout;

analog
     begin

          if (I(Vtop,Vbot) <= Ith)
               V(Vtop,Vbot) <+ R_OFF*I(Vtop,Vbot);
          if (Ith < I(Vtop,Vbot) && I(Vtop,Vbot) < Ih)
               V(Vtop,Vbot) <+ NDR*I(Vtop,Vbot) + V1;
          if (I(Vtop,Vbot) >= Ih)
               V(Vtop,Vbot) <+ R_ON*I(Vtop,Vbot) + V2;
     end

endmodule
```