Modeling the Spatio-Temporal Evolution of Oxygen Vacancies in Valence Change Memory Cells

---

Von der Fakultät für Elektrotechnik und Informationstechnik
der Rheinisch-Westfälischen Technischen Hochschule Aachen

zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigte Dissertation

vorgelegt von

Ching-Jung Chen, M.Sc.

aus Taiwan

Berichter

Univ.-Prof. Dr.-Ing. Christoph Jungemann
Univ.-Prof. Dr.-Ing. Rainer Waser

Tag der mündlichen Prüfung: 06.11.2024

Curriculum Vitae

06.11.2024	Doctoral Examination for Dr. rer. nat.
2020 - 2024	Studies of Electrical Engineering, Information Technology and Computer Engineering RWTH Aachen University, Germany
August 2018	Master of Science Department of Physics, National Tsing Hua University, Taiwan
2017 - 2018	Exchange student at Linköping University Linköping, Sweden
June 2016	Bachelor of Science Department of Physics, National Tsing Hua University, Taiwan
2013 - 2016	Ming-Dao High School Taichung, Taiwan
1994	Born in Taichung, Taiwan

Contents

Abstract

Motivation, Goal and Task of the Dissertation

Valence change memory is a promising type of non-volatile memory for next-generation applications. Compared to contemporary NAND Flash, valence change memory cells exhibit advantages such as lower power consumption and faster operating speeds. In addition, devices can be fabricated by existing semiconductor technologies. However, the underlying physical mechanisms intrinsically impose difficulties in manipulating the cell resistance precisely, leading to endurance and data retention issues. It has been observed that the variability of the electrical behavior can be reduced by adopting a large current compliance, which limits the maximum current flowing through the device, but theoretical interpretations are still incomplete. Specifically, most numerical models focus on devices with a large current compliance, while the impact of a small current compliance remains unclear.

From a statistical perspective, different tendencies in a wide range of current compliances have been observed in measurements. Different theoretical models have been proposed based on a simple scheme, where one conductive path exists in the oxide layer. However, none of these can explain the observed tendency in a small current compliance regime. In addition, devices with a small current compliance consume less power, thus offering significant advantages for practical applications.

The goal of this work is the theoretical investigation of the spatio-temporal evolution of oxygen vacancies resulting in a resistive change of the valence change memory cell. By treating oxygen vacancies as point defects, the same viewpoint as in the density functional theory, findings from *ab initio* calculations can be applied. This enriches the understanding of local structures and physical quantities during the oxygen migration. To this end, the measurements at a macroscopic level can be explained by the spatio-temporal evolution of oxygen vacancies at a microscopic level. The discussion sheds light

on engineering devices for a specialized functionality.

Major Scientific Contributions

It is well-accepted that vacancy migration is a stochastic process, and the existence of preferred paths due to local structures has been shown by *ab initio* calculations, but it is not clear what kind of migration patterns might exist. In this regard, the interplay between the vacancy distribution, local structures, and physical quantities involved in a dynamical process is explored. Since *ab initio* calculations cannot cover all possible vacancy distributions during a dynamical process, approximations are made to consider all these configurations in semi-classical kinetic Monte-Carlo simulations.

Possible migration patterns are proposed to explain the observed statistical tendencies. Three-dimensional device simulations are performed with relevant conditions aligned with measurements. The SET and RESET operations of the cell are simulated. Sometimes these operations fail and the corresponding local structures of the conductive filaments are identified. Furthermore, a simple yet physical interpretation is proposed to explain the statistical behavior in a wide current compliance regime. It is based on a scheme consisting of multiple filaments, which is different from existing interpretations. The scheme is supported by measurements. In addition, a new model of the charge state of the oxygen vacancies, which effectively leads to an extra charge factor, is proposed. The plausibility is discussed in the framework of detailed balance under equilibrium conditions.

List of publications

[1] C.-J. Chen, K. Z. Rushchanskii, and C. Jungemann, "Investigation of the Large Variability of HfO_2-Based Resistive Random Access Memory Devices with a Small Current Compliance by a Kinetic Monte Carlo Model," *physica status solidi (a)*, p. 2300403

Notation

x Scalar quantity

$\boldsymbol{x}$ Vector quantity

$\mathbf{X}$ Matrix quantity

$\langle x \rangle$ Average value

μ_x Median value

Symbols

α Symmetry factor

a Nearest hopping distance

a_0 Attenuation radius

β Sweep rate

b Bond polarization

c Speed of light

c_0 Vacuum speed of light

d Distance between two oxygen vacancy sites

$\boldsymbol{D}$ Electric flux density

ε Electrostatic permittivity

ε_0 Vacuum permittivity

ε_r Relative permittivity

e Magnitude of the elementary charge

$\boldsymbol{E}$ Electric field

E_{loc} Local electric field

E_{c} Energy level of the conduction band minimum

E_{D} Zero-field activation energy for an oxygen vacancy diffusion

ΔE_{D} Filed-modulated activation barrier for an oxygen vacancy diffusion

E_{G} Zero-field activation energy for an oxygen vacancy generation

E_{R} Zero-field activation energy for an oxygen vacancy recombination

g Density of the heat generation rate

h Miller-Abrahams hopping rate

$\hbar$ Reduced Planck constant

I Current

I^e Electron current

k_B Boltzmann factor

k_{th} Thermal conductivity

μ Elecetron Fermi level

μ_0 Vacuum permeability

μ_r Relative permeability

m^* Effective electron mass

m_0 Electron rest mass

ν_0 Attempt frequency for the vacancy generation, recombination and diffusion

ν_e Attempt frequency for an electron hopping

Ω Size of a finite volume

ρ Charge density

p Probability of an oxygen vacancy in the filled state

φ Electrostatic potential

R_0 Electrode coulping

$\boldsymbol{r}$ Position vector

t Time

T Temperature

V_{app} Applied voltage

V_{cell} Voltage across a cell

$\mathcal{Z}$ Grand partition function

Acronums

1T1R one-transistor–one-resistor

AE active electrode

ALD atomic layer deposition

BC boundary condition

BRS bipolar resistive switching

BS bipolar switching

BE bottom electrode

C2C cycle-to-cycle

CBM conduction band minimum

CBRAM conductive bridge random access memory

CDF cumulative distribution function

CF conductive filament

CIM computing-in-memory

CMOS complementary metal-oxide semiconductor

D2D device-to-device

DD drift diffusion

DFT density functional theory

DOS density of state

DRAM dynamic random-access memory

DUT device under test

ECC Error Correcting Code

ECM electrochemical metallization memory

FN Fowler-Nordheim

FP Frenkel pair

FTJ ferroelectric tunneling junction

FVM finite volume method

GB grain boundary

HRS high resistance state

KMC kinetic Monte Carlo

LRS low resistance state

MAC Multiple-Accumulate

MA Miller-Abrahams

MD molecular dynamics

MIM metal-insulator-metal

MRAM magnetoresistive random access memory

MTJ magnetic tunnel junction

MTTF mean time to failure

NEM Nano-electromechanical

NVM non-volatile memory

OE ohmic electrode

OEL oxygen exchange layer

OxRAM oxide-based random access memory

PCM phase change memory

PDE partial differential equation

PDF probability density function

P/E program/erase

PF Poole-Frenkel

PSD power spectral density

QPC quantum point contact

redox oxidation-reduction

RRAM resistive switching random access memory

ReRAM redox-based resistive switching random access memory

SA sense amplifier

SLC single-level cell

SSD solid-state drive

STO $SrTiO_3$

STT spin transfer torque

TAT trap-assisted-tunneling

TCM thermochemical memories

TDDB time dependent dielectric breakdown

TDTR time-domain thermoreflectance

TE top electrode

URS unipolar resistive switching

US unipolar switching

VBM valence band maximum

VCM valence change memory

V_O oxygen vacancy

WKB Wentzel–Kramers–Brillouin

1 Introduction

As information technology advances, the ability to process more data within a shorter time is an important demand. Over the past few decades, this demand has been addressed through scaling down fabrication technology. However, the down-scaling of complementary metal-oxide semiconductor (CMOS) architecture is approaching its physical limit. This presents challenges for contemporary memory devices such as NAND Flash and dynamic random-access memory (DRAM), where transistors often become bottlenecks in integrated circuits. On the one hand, the search for new materials and the vertical stacking of memory units are possibilities for improvements. On the other hand, solutions based on a different mechanism for both volatile and non-volatile memory devices are proceeding. Nowadays, the resistive switching random access memory (RRAM) is one of the promising candidates.

While studies of the resistance change date back to the 1960s [2, 3], research interest has gradually transitioned to Si-based integrated circuit technology since the late 1970s. Until the 1990s, this field started to regain attention triggered by Asamitsu *et al.* [4] and Beck *et al.* [5]. Later in the 2000s, RRAM integrated into the contemporary CMOS technologies have been reported [6, 7]. The reader is referred to Ref. [8, 9] for a detailed review up to the late 2000s. In this stage, it was originally designed for data storage applications. To be competitive with contemporary non-volatile memory devices, information must be stored for several years without loss. In addition, features such as low power consumption and fast read/write operations are desired for the next-generation electronics. These requirements have been met by a variety of RRAM devices. Nowadays, commercial products based on the RRAM can be found in the market.

Recently, the RRAM has found new applications, i. e., computing-in-memory (CIM) and neuromorphic computing, beyond pure data storage. Within the traditional von Neumann architecture, the processing and memory units are separated, leading to increased latency and power consumption due to data transfers. In contrast, CIM integrates these units, alleviating the aforementioned problems at the hardware level. This

is similar to the working principles of the biological brain, where both processing and memory functions are performed by neurons and synapses [10]. Drawing inspiration from the nervous system, a new architecture that emulates the neurons and synapses has been proposed with the aim of harnessing its energy efficiency. Remarkably, the human brain is estimated to conduct 10^{18} Multiple-Accumulate (MAC) using only 20 W. This energy efficiency outperforms those of contemporary supercomputers by about eight orders of magnitudes [11]. This biological efficiency highlights the potential of neuromorphic computing, which is applicable to artificial intelligence applications. The reader is referred to Ref. [12–15] for the architecture and benchmarks, and Ref. [10, 16, 17] for a review.

However, neither the solid-state drive (SSD) nor the contemporary von Neumann architecture based computing devices have been replaced by RRAM based devices. One of the challenging problems arises from the working principle, where a stochastic process of oxygen vacancy migration is involved. Although using a larger current compliance could reduce the variability, this would inevitably result in larger power consumption. This, in turn, raises the risk of damaging memory cells due to greater heat dissipation. Therefore, the trade-off between the current compliance and the reliability of a memory cell is critical. Since the oxygen vacancy is attributed to the electrical behavior of a valence change memory (VCM) cell, the evolution of oxygen vacancies is investigated. The focus is on devices in a small current compliance regime, i. e., $I_{\mathrm{cc}} < 15\,\mu\mathrm{A}$.

The second chapter provides measurement findings, a well-accepted theory for the resistive switching phenomenon of the studied device, and the theoretical background for the kinetic Monte Carlo (KMC) model. The model incorporates fingings from *ab initio* calculations at $T = 0\,\mathrm{K}$. This implies that entropy effects are absent, despite the influence of local high temperatures during resistive switching. The reader is referred to Ref. [18–22] for other numerical models discussing a resistive switching process.

In the third chapter, three-dimensional device simulations are performed and discussed. The chapter begins with simple schemes, where the temperature distribution and electrostatic potential distribution are studied independently. Physical quantity distributions under different modeling frameworks are compared. Based on the established results, a simple scenario for the observed variability is proposed and simulation results are provided. It starts with a specific current compliance value and then extends to a large window. In the meanwhile, a simple formulation based on the parallel connection is proposed for the explanation of the observed variability.

Lastly, the fourth chapter summarizes this simulation work and the outlook.

2 Background

This chapter provides the essential background for exploring the dynamics at the microscopic level involved in the operation of filamentary type VCM cells. It starts with a classification of different types of RRAM for an overview in Sec. 2.1. Later on, the focus shifts back to the filamentary type VCM cells. The electrical characteristics involved in a resistive switching process are provided in Sec. 2.2. However, a practical memory cell typically undergoes millions of read-write operations, and thus the statistical aspect is an important metric for examining a cell. Sec. 2.3 covers the fundamentals of the reliability aspect. The theoretical model, i. e., KMC model, is employed to explore the microscopic details during the dynamical process in this thesis. The KMC model incorporates physical processes such as charge transport and heat dissipation, covered in Sec. 2.4 and Sec. 2.5, respectively. The chemical reactions of oxygen vacancies, including the generation and recombination reactions, and the migration are detailed in Sec. 2.6 and Sec. 2.7, respectively. The statistical foundation for both numerical modeling and experimental data analysis is introduced in Sec. 2.8. Finally, the methodology for solving the three-dimensional numerical model is provided in Sec. 2.9.

2.1 Classification of the resistive switching random access memory

The classification aims to provide an overview of the technologies developed to date. A large variety of physical mechanisms are involved in the resistance change across different materials. Moreover, the interplay between physical mechanisms imposes difficulties in building a universal classification. Therefore, the provided classification is based on the principal physical effect widely accepted by the community.

The physical mechanisms behind the resistance change are generally divided into three groups: magnetic effects, electrostatic effects, and atomic configuration effects (see Fig. 2.1).

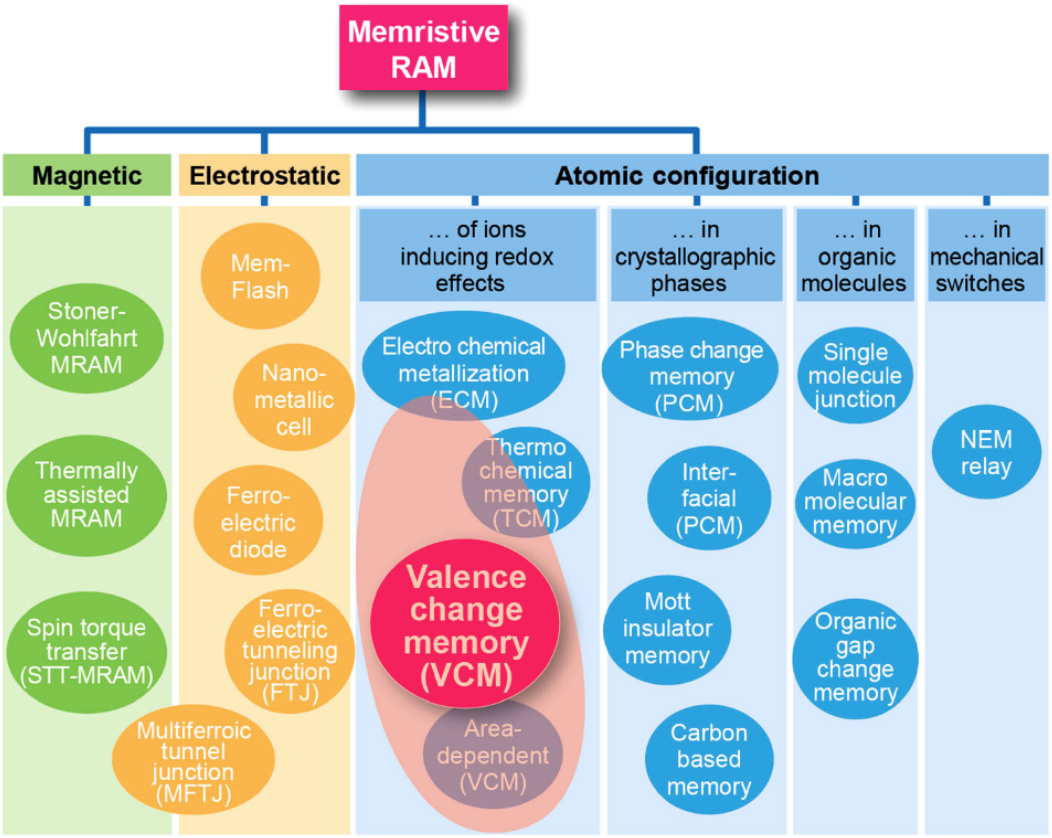

Figure 2.1: Survey of resistance-based non-volatile memories, also called memristive elements or memristive RAM. Reproduced with permission from [23]. Copyright © 2022 Taylor & Francis Group.

Magnetic effects responsible for resistance changes occur in magnetic tunnel junctions (MTJs), which consist of two ferromagnetic layers separated by a thin insulating layer [24]. The relative orientation of the magnetic fields changes the electron tunneling probability through the insulating layer, and thus data can be stored in terms of the relative orientation of magnetic fields leading to magnetoresistive random access memory (MRAM). More recently, the spin transfer torque (STT) technology utilizing the electron spin has been applied to develop the STT-MRAM [25, 26]. It has shown benefits such as reduced power consumption and increased storage density compared to the conventional MRAM. More details about the development of MRAMs can be found in Ref. [27].

In parallel to magnetic effects, resistances could be modulated by electric effects. A metal-insulator-metal (MIM) structure with a thin ferroelectric layer as the insulating layer first proposed in 1971 [28] was later known as the ferroelectric tunneling junction (FTJ). Conceptually, the manipulation of resistances is achieved through an applied voltage, which changes the electric polarization of the ferroelectric layer. The reader is referred to Ref. [29] for more details.

Resistance changes due to atomic effects are found in both organic [30] and inorganic materials with various mechanisms identified. The mechanical distortion due to the electrostatic force has been proposed for developing the Nano-electromechanical (NEM) RAM [31]. For the phase change memory (PCM) [32], different electrical resistances of amorphous and crystalline phases are used.

The redox-based resistive switching random access memory (ReRAM) based on the oxidation-reduction (redox) reactions can be further divided into three major types, i.e., thermochemical memories (TCM), electrochemical metallization memory (ECM), and VCM[1]. The dominant redox process of a TCM device is thermochemical rather than electrochemical. That is, the redox process is mainly triggered by the thermal effect. A detailed review of TCM can be found in Ref. [33]. On the other hand, in both ECM and VCM electrochemical processes play a dominant role. The distinction between ECM and VCM cells lies in the source of mobile ions. In an ECM cell, the electrochemically active electrode metal is involved in the redox process and thus the dynamics of metal ions is crucial for the resistance change. Under the pressure of an applied voltage, metal ions drift to the counter electrode, which is usually an inert metal. The reduction

[1]In some literature, conductive bridge random access memory (CBRAM) and oxide-based random access memory (OxRAM) are terminologies for ECM and VCM, respectively.

reaction that occurs upon contact with the inert electrode gives rise to the growth of a metallic conductive path towards the active electrode, thereby reducing the resistance. To increase the resistance, the opposite polarity is applied to reverse the above process. The reader is referred to Ref. [34] for more details about the ECM.

In VCM, a metal oxide layer sandwiched between two electrodes participates in the chemical reaction. The mobile ions are typically oxygen anions, referred to as oxygen vacancies. Furthermore, the dependence of the electrical behavior on the cross-section of a device can vary dramatically. That is, devices for which conductivity is proportional to their cross-section are said to be of an area-dependent type. In contrast, the filamentary type devices refer to those where the resistance has a weak dependence on the cross-section over almost three orders of magnitude [7, 35, 36]. The filamentary type VCM is discussed further in subsequent sections.

2.2 Operation of valence change memory devices

The simplest device structure for a VCM device that facilitates the resistance change can be realized by a MIM structure, i. e., an insulating layer sandwiched by metallic electrodes with the constituted materials expressed in the format, M/I/M. The insulating layer is a thin film with a typical thickness from several to tens of nanometers, see Fig. 2.2a. Up to date, there are a variety of binary transition metal oxides, such as HfO_2, ZrO_2, TiO_2, Ta_2O_5, and NiO, and ternary oxides, e.g. $SrTiO_3$ (STO), that are shown to be suitable for the insulating layer [9, 21][2]. Meanwhile, the materials of the two metallic electrodes are not necessarily identical. Moreover, the metals are found to be crucial for the creation of oxygen vacancies in the preliminary electroforming (FORMING) process (see Sec. 2.2.1) and the subsequent SET and RESET processes (see Sec. 2.2.2).

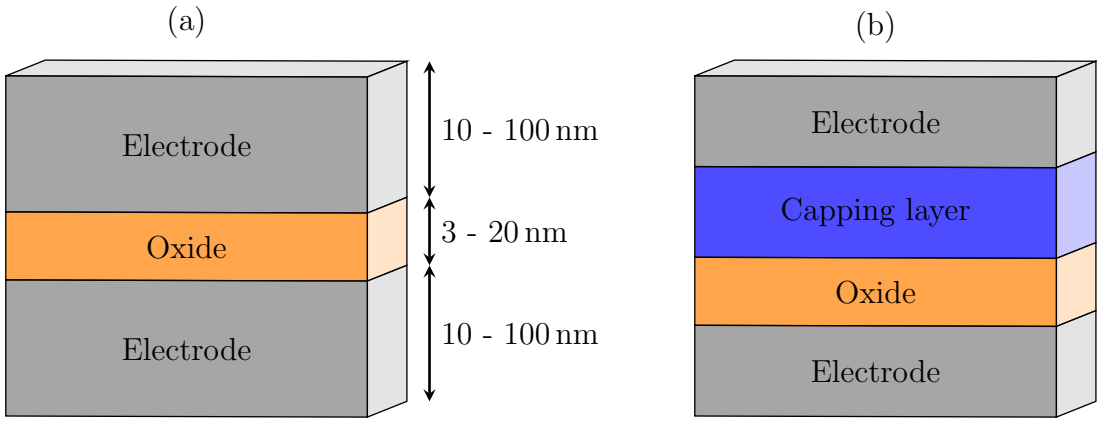

Figure 2.2: Schematic MIM structures for (a) a HfO_2-based cell with typical thickness and (b) a heterogeneous bilayer structure.

2.2.1 Electroforming

Preliminary to manipulating the resistance state of a VCM cell, the so-called electroforming or simply FORMING is required for most VCM devices. From a macroscopic point of view, a soft dielectric breakdown process is triggered by the large voltage in this step. The original resistance can be up to orders of magnitude higher than the

[2]In some literature, the subscript "x" is used for oxides, such as HfO_x, to emphasis the oxygen composition is nonstoichiometric.

resistance after a successful FORMING process; thus, the high resistance state (HRS) and the low resistance state (LRS) are used to refer to the state of a device possessing two distinct resistance levels. On the other hand, the introduction of oxygen vacancies (V_Os) into the oxide layer is observed after the FORMING process, leading to the resistance change from a microscopic point of view. Specifically, the V_Os inside the oxide layer provide intermediate sites for the electron tunneling between electrodes. With the introduction of a sufficient number of V_Os along the direction between electrodes, the effective tunneling length is greatly decreased. As a result, the resistance reduces in the presence of the establishment of a conductive path, which is coined as the conductive filament (CF) [37–46].

The generation of V_Os inside the oxide layer is closely related to the material of electrodes. Typically, a high-work-function metal, such as TiN or Pt, is required for one electrode. The other electrode may be either a high-work-function metal or a low-work-function metal such as Ti, Hf, Zr, Al, or Ta. In the former case, where both electrodes are high-work-function metals, a significantly higher voltage is required for the onset of the abrupt increase of current [47–49]. In the latter case, the oxygen affinity of a low-work-function metal is generally high. The oxidization of the metal is observed [49–51], and oxygen vacancies are created in the oxide layer close to the interface. The layer is referred to as the oxygen exchange layer (OEL). Furthermore, using a metal different from that in the oxide might create an extra sub-oxide layer [52–54][3], leading to the so-called heterogeneous bilayer structure. During the fabrication, the bilayer structure could be achieved by intentionally growing an easily oxidizable capping layer above the oxide layer as illustrated in Fig. 2.2b. The reader is referred to Ref. [55, 56] for the electrical performance aspects and the potential applications of bilayer RRAM cells.

Depending on the purpose, there are two approaches for applying an external voltage: the pulse and sweep mode, differing in the duration of the applied voltage. In principle, the sweep mode is adopted to investigate the I-V characteristics as a starting point. In contrast, the pulse mode is closer to practical applications since devices are switched under a short-duration voltage signal. In practice, both approaches are used in reliability tests, see Sec. 2.3.2. Regardless of the mode, applying a large voltage can potentially damage the device during the FORMING step which can be avoided by attaching a

[3]It is noted that the experimental data and physical processes strongly depend on the materials. In this introductory section, the materials for the MIM structure is presented.
$Ti/TiO_x/HfO_2/TiN$ [52, 54]. $Pt/Hf/ZrO_2/Pt$ and $Pt/Ti/ZrO_2/Pt$ [53]

current compliance to the circuit. By limiting the maximal current through the device, the energy dissipation and thus the positive feedback between the maximal temperature and the current can be controlled. Therefore, the risk of damaging a device due to the current overshoot is reduced. Herein we focus on the device where a FORMING step is required for subsequent resistive switching processes. In addition, the device structure with a single oxide layer subjected to the sweep mode is simulated.

2.2.2 Switching cycles

After the FORMING step, the RESET operation is performed, aiming to increase the resistance and to reach the expected HRS. From a microscopic point of view, the connection of both electrodes by the CF is ruptured, leading to the observed high resistance at the end of the process. Conversely, the re-connection of the ruptured CF is expected in the SET operation, leading to the expected LRS. Since the resistance is expected to be low, a current compliance is attached in the SET process as well. On the contrary, a thermal runaway is not expected when the value of an applied voltage is either small or large during the RESET process. When the voltage amplitude is still small, the Joule heating is not significant enough to damage the cell despite the low resistance of the cell. In the intermediate voltage regime, the migration of oxygen vacancies starts, leading to the rupture of a CF and thus the HRS. Ideally, the reduced current decreases the energy dissipation when the applied voltage is large. Therefore, the current compliance is not applied throughout the whole RESET process. However, this approach is accompanied by the risk of damaging a device if the decrease in current is slow or absent in the intermediate voltage regime. Since the maximum current during a RESET process is closely related to the current compliance during a SET process, adopting a small current compliance reduces the risk.

Depending on the polarity of the applied voltage to operate these two processes, two distinct approaches to manipulate the resistance states are available, i. e., the bipolar resistive switching (BRS) and unipolar resistive switching (URS). Specifically, BRS and URS refer to the switch of resistance states by opposite and identical voltage polarities, respectively. URS or simply unipolar switching (US) is commonly observed in TCM cells where the thermochemical reaction rather than the electrochemical reaction triggers the resistive change [9, 33]. In this case, the transition to the HRS arises from the high temperature in the absence of the current compliance in the RESET process while the

presence of the current compliance in the SET limits the temperature. In this scenario, the resistive switching depends on the asymmetrical usage of the current compliance but not the external polarities in the two processes. Interestingly, it has been reported that the VCM cells can be operated in US. In principle, US is observed for cells with high work-function metals for electrodes [47, 57][4]. The reader is referred to Ref. [58, 59] for theoretical models.

On the other hand, BRS or simply bipolar switching (BS) is more common for VCM cells. Typically, a BS operation relies on the asymmetrical arrangement of the work-function for two electrodes, namely, one electrode with a low-work-function metal and the other with a high-work-function metal [60]. The electrode with a low-work-function metal is called ohmic electrode (OE) due to the formation of the ohmic contact with the oxide in the ideal case. The electrode composed of a high-work function metal forms a Schottky contact. It is called active electrode (AE) since the switching phenomena are assumed to occur near the interface. The well-accepted scenario for BS-type cells involves the redistribution of oxygen vacancies, while the total number of vacancies remains unchanged after the FORMING process. Depending on the relative location and the vacancy density, the structure of a CF can be further divided into two parts, i. e., the plug and disc (see Fig. 2.3). The disc corresponds to the region close to the AE where the vacancy density changes drastically, and thus plays a crucial role in the resistive switching phenomena. The plug refers to the region where the vacancy density is high and remains (almost) constant throughout the entire switching process. It serves as a reservoir for providing and accepting vacancies to the disc in the SET and RESET processes, respectively. The reader is referred to Ref. [9, 61] for more details.

[4] $Pt/HfO_2/TiN$ and $TiN/HfO_2/TiN$ [47], $Pt/HfO_2/Pt$ [57]

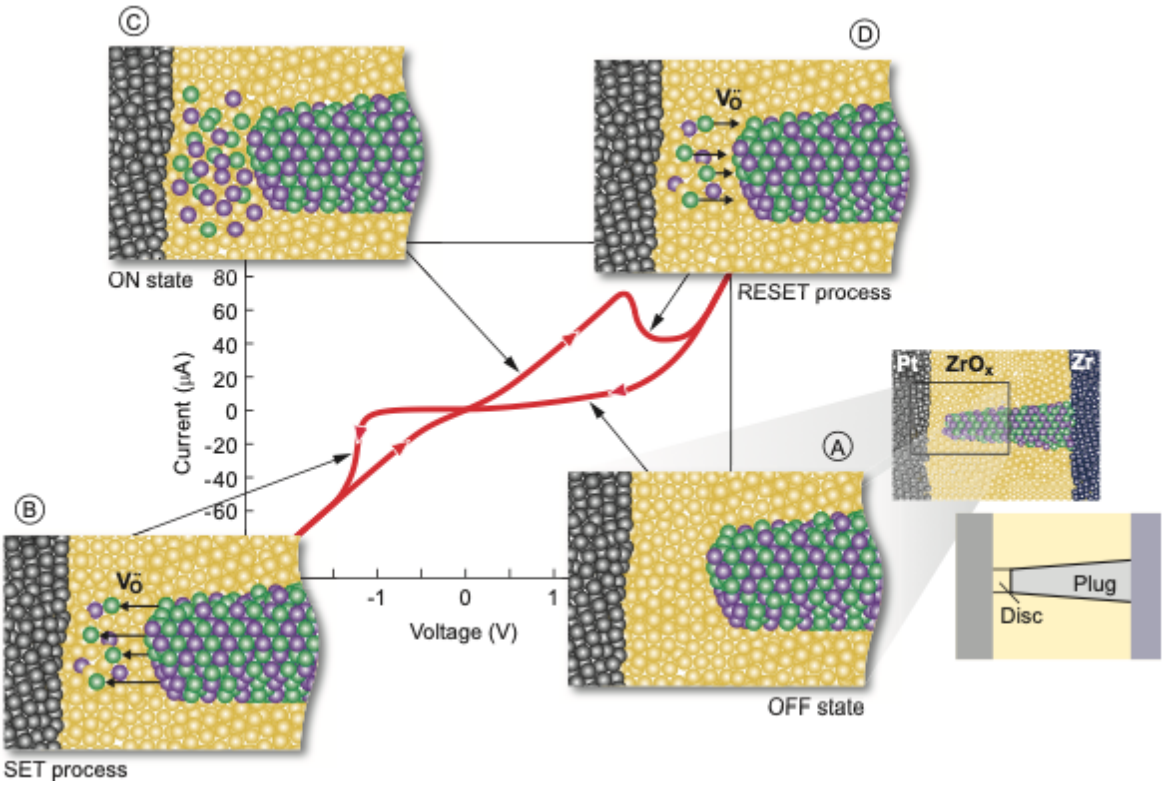

Figure 2.3: VCM switching mechanism. Schematic I–V characteristic in combination with sketches of the internal states and processes during switching. The bipolar switching is based on the redistribution of oxygen vacancies in the filament near the active electrode. The SET occurs when a negative potential is applied to the active electrode. Oxygen vacancies and immobile metal cations in a lower valence state, here Zr ions, are indicated by green and violet spheres, respectively. Yellow spheres represent the stoichiometric oxide. The active electrode, here Pt, is shown in gray on the left side. Reproduced with permission from [61], 2012 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim.

2.3 Requirements and reliability

To move on to commercialization for non-volatile storage applications, several criteria must be met. The following operations at the single-cell level are required for a single memory cell:

- Read operations: The resistance of each resistance state is obtained by measuring the read current at a read voltage, where a small voltage is aimed to avoid disturbing the V_O distribution and thus the resistance. However, the circuit design of a sense amplifier (SA) imposes a minimal read voltage of approximately one-tenth of the write voltage [23]. New circuit designs for a SA have been proposed to improve the sensing margin [62, 63].

- Write operations: The write voltage and the write time are two important values to compare with other memory devices. In general, non-volatile memory cells are operated under a higher write voltage and a slower write time, compared to volatile devices. For example, typical write voltages of NAND Flash cells and DRAM cells are about 3 V and about 1 V, respectively. The write time, which is defined as the duration of the write voltage pulse, is typically about 1 µs and 1 ns for Flash cells and DRAM cells, respectively. Up to date, the resistive switching of VCM cells can be operated at around 1 V with the write times at nanosecond scale [64] and below [65–68][5].

In addition, the scaling down of memory cells increases the density of logic elements and thus enhances the functionality of a device. Modern technology has been developed to maximize the space usage, which relates to the stackability. Lastly, since the established CMOS fabrication processes for small feature sizes are extraordinarily expensive, it is desirable to have fewer additional fabrication processes for a RRAM cell, which corresponds to a better CMOS compatibility.

- Scalability: The intrinsic limit in physical size is due to the underlying physical mechanisms. Contemporary technology has already encountered the geometrical limit of a single Flash or DRAM cell, which is roughly 10 nm [21, 69]. In comparison, HfO_2-based VCM cells were fabricated with widths of 40 nm×40 nm in 2012 [70] and 10 nm×10 nm [71] in 2014.

[5] Ta_2O_5-based [66], HfO_2-based and Ta_2O_5-based [67], HfO_2-based [64, 65, 68]

- Stackability: This refers to a solution to enhance scalability by stacking individual cells on top of one another, making a 2D crossbar array into a 3D circuit. This idea was first introduced to NAND Flash technology in the 2000s [72] and has dramatically boosted the usage of SSDs in data storage applications. The 3D NAND Flash cells were reported to be over 200 layers in 2022 [73] and over 300 layers in 2023 [74]. Since the early 2010s, the stackability of RRAM memory cells has been gaining much attention [75]. In 2020, devices composed of 8-layer HfO_2-based memory elements were proposed and fabricated [76].

- CMOS Compatibility: This refers to the additional fabrication processes for the RRAM with the established CMOS equipment. It involves the fabrication environments, materials, and the fabrication temperature. In 2004, Samsung showed the integration of NiO-based RRAM cells with the contemporary 0.18 µm CMOS technology into a one-transistor–one-resistor (1T1R) circuit [7]. Moreover, HfO_2, ZrO_2, TiO_2 were proposed as a high-κ dielectric material in replacement of SiO_2 for the gate dielectric material in 2001 [77]. Ta_2O_5-based memory cells have been successfully fabricated on the standard 0.18 µm CMOS in 2008 [78]. These materials show great compatibility with existing technologies.

The reader is referred to the Table 1 of Ref. [79] for a review of the progress in technologies. Aside from the aforementioned requirements, the number of program/erase (P/E) cycles of a NAND Flash cell is expected to be 10^5. For a volatile DRAM cell, there is no theoretical limit. However, an estimation of the operation lifetime could be conducted based on the refresh rate. Given a typical refresh time of 60 ms, a DRAM cell in the time span of 10 years would be operated about $5 \cdot 10^9$ times. Considering a large number of operations, additional requirements arise, i. e., endurance and variability.

2.3.1 Retention

The retention time is a measure of the long-term stability of a non-volatile memory (NVM) cell. It refers to the period during which the stored information can be kept without distortion. Take contemporary NAND Flash for example, both the P/E cycles for which the device has been used as well as the ambient temperature affect the retention. A common retention period for fresh NAND Flash cells is several years within the range of 75-125 °C. Given that the retention of NVM cells extends to years,

retention tests are typically not conducted at the above-mentioned temperatures. Instead, the devices under test (DUTs) are baked at elevated temperatures. The mean time to failure (MTTF), which represents the time to reach the defined failure criteria, is provided in the specifications by manufacturers. In a very simplified scenario, the determination of the (effective) retention time at a given temperature is derived from the Arrhenius relation.

The Arrhenius equation relates the reaction rate k to the temperature T by

$$k = A \cdot \exp\left(-\frac{E_\mathrm{a}}{k_\mathrm{B}T}\right), \tag{2.1}$$

where A is the Arrhenius factor, k_B is the Boltzmann constant and E_a is the activation energy for a reaction. Typically, an Arrhenius plot refers to the logarithm of the reaction rate versus the reciprocal temperature, showing the reaction rate at an extrapolated temperature. Herein, the reaction rate is related to time by the inverse relationship, $t_\mathrm{fail} = k^{-1}$, and thus the MTTF can be evaluated at a lower temperature, see Fig. 2.4. This is a common approach to evaluate the retention time for NAND Flash devices and VCM cells, even though underlying failure mechanisms differ.

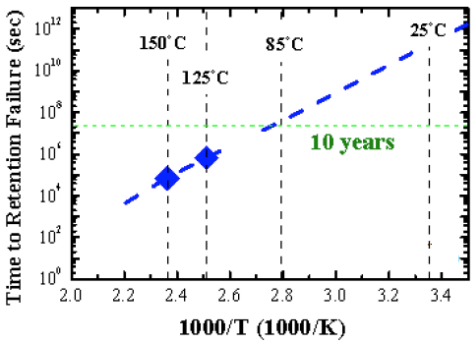

Figure 2.4: Arrhenius plot for tail bits of a retention test of a 256 kbit array. Reproduced with permission from [43], © 2011 IEEE.

Generally, the shift of resistances under a higher temperature is explained by the migration of oxygen vacancies, which is activated by the thermal fluctuation instead of an external electric field. Either a threshold of the read current [80] or the ratio of the final to initial resistances [81, 82] is adopted for the definition of data loss. A harsher criterion for failure based on a smaller part of the devices is employed to detect the fast retention failure [83, 84]. Specifically, the tail bits refer to devices with higher LRS

resistances and lower HRS resistances. The data retention of tail bits raises a crucial reliability concern, especially for a large memory array. In the early 2010s, common retention failure patterns were identified. That is, the shift of LRS and HRS resistances are often observed in the low- and high-current operations, respectively [43, 85]. Since the low-current operation is more desirable, more studies have been dedicated to the increase in LRS resistances referring to the LRS retention failure.

One explanation of the LRS retention failure is the highly mobile oxygen ions. In Ref. [83][6], extracted data indicated a high mobility of oxygen ions, aligning with measurements of the HfO_2 thin film [86]. Specifically, oxygen ions close to the constriction of the CF are responsible for the dissolution of the CF, leading to the early stage LRS retention failure in a small percentage of devices. Interestingly, different fitted mobilities are obtained due to the different criteria even for memory cells on the same stack [83, 87]. A similar argument regarding the mobility is proposed while it refers to the mobility of oxygen vacancies [88]. S. Clima *et al.* discussed a scenario where the different mobilities of vacancies create a heterogeneous distribution of diffusion barriers in space. In this scheme, resistance shifts could be attributed to the mobile vacancies migrating into or out of the constriction of the CF. The attribution of increased LRS resistances to the out-diffusion of vacancies is further supported by both measurements [43, 81, 83, 87, 89] and models [43, 81, 82, 90]. These works demonstrate that the LRS resistance stabilizes with increased current compliance due to a stronger CF mitigating the impact of vacancy outflow. In addition to the current compliance, the alloying of HfO_2 with Al has been shown to increase the stability of LRS resistances [82, 91–93]. Interestingly, using Ti as a dopant is shown to reduce the LRS retention [82, 92].

The requirement of a long-term stability of the stored information, a fast write operation, and a comparable read/write operation voltage is often called the *voltage-time dilemma*. On the one hand, the ratio is approximately 10^{14} with the retention being 10 years and the write operation within 10 ns. On the other hand, seeking low-power electronics by asking for the write voltage at around 1 V and the read voltage at around 0.1 V yields a ratio of 10. To fulfill the requirement, the resistance change must be highly non-linear in response to the applied voltage.

[6] HfO_2-based

2.3.2 Endurance

The endurance refers to the number of P/E operations until a NVM cell fails. For a VCM cell with only two resistance states, it is examined by repeating the cycle composed of the SET and RESET operations. Depending on the purpose, there are three common approaches for conducting an endurance test, as illustrated in Fig. 2.5 [94]. With the I-V sweep approach illustrated in Fig. 2.5a, one can obtain the detailed I-V relation in each cycle at the costs of a time-consuming process, a limited number of cycles, and a potential deviation from the realistic operation condition [95]. In the other two test schemes, the SET and RESET operations are conducted in the pulse mode. The difference is that the resistance states can be measured in each cycle only by advanced equipment, as illustrated in Fig. 2.5b. In contrast, a limited number of data can be obtained by standard hardware because the resistance is measured once every specific number of switching cycles, as illustrated in Fig. 2.5c.

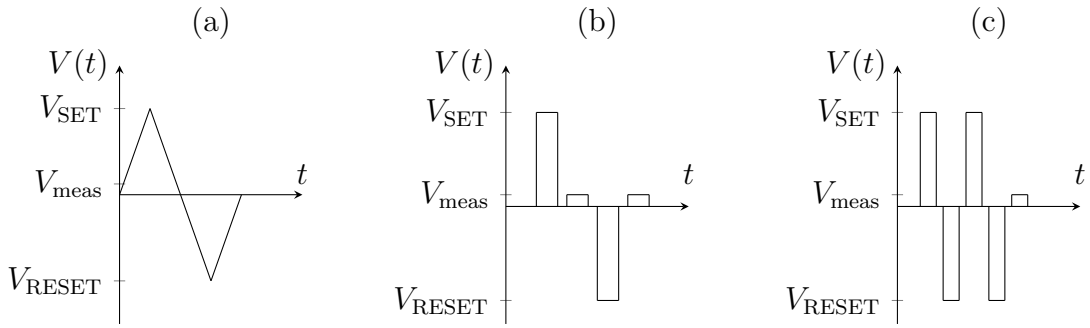

Figure 2.5: Schematic voltage signals for conducting endurance tests.

During the endurance test, electrical properties such as the LRS and HRS resistances are recorded. The test ends when the resistance ratio fails to reach a reliable window, or the resistance gets stuck at one of the resistance states. To develop next-generation applications, the endurance of a VCM cell must be comparable with contemporary technologies. For example, values of 10^5 and 10^{10} can be expected for a single-level cell (SLC) and a DRAM cell, respectively. Up to date, endurance tests of a VCM cell above 10^6 have been reported, see Table 2 of Ref. [96].

The increase of LRS resistances [44, 70, 85], decrease of HRS resistances [44, 70, 97, 98]

and both tendencies together [99, 100][7] are observed during endurance tests. The shift of resistances is explained by changes near the CF region, which can arise from either reversible or irreversible processes. The reversible processes are typically associated with the redistribution of CF constituents. For example, the decrease of HRS resistances is explained by the redistribution of oxygen vacancies [70]. Among switching cycles, the vacancy migration is more dominant in the SET process compared to that in the RESET process, leading to the gradual accumulation of vacancies at the switching interface. Consequently, the asymmetrical migration scenario leads to the cell being stuck in the LRS after the RESET process.

In the scheme of the vacancy redistribution, the failure of both a high resistance after the SET process and a low resistance after the RESET process can be recovered. It has been shown that both failures in the SET and RESET processes can be restored with a slow sweep rate in the SET and RESET operations, respectively [70]. Since the vacancy migration is primarily guided by the applied voltage, the optimization of the SET and RESET voltages is a promising approach to fine-tuning the endurance. More specifically, the amplitude and width of the voltages in both processes are investigated to improve the endurance [100, 101]. Under the optimized conditions, the endurance can achieve up to 10^{10} cycles [70].

However, failures can also be associated with irreversible possesses, e. g., an excessive amount of oxygen vacancies or atomic relaxation. In this case, a resistance state cannot be recovered by the above-mentioned approach. For example, it is found that devices with Hf or Ti as their OE demonstrate a RESET failure within the first 20 cycles [102]. In contrast, the uses of Ta or W demonstrate a much more stable resistive switching. By investigating the defect formation energy at the electrode/oxide interface, the RESET failure is attributed to excessive oxygen vacancies due to the negative formation energy. It is noteworthy that the coupling of reversible and irreversible mechanisms gives rise to intricate failure phenomena. In the analysis of 41 TiN/$TaO_{2\pm0.2}$/TiN VCM cells, it was found that 20 cells are stuck in the LRS while the other 21 cells are stuck in the HRS [44]. During the endurance test, the vacancy redistribution gradually leads to a thermodynamically favored configuration, where the vacancy mobility is decreased. Furthermore, the SET and RESET failure can take place due to the segregation of Ta- and O-rich regions along the CF, and crystallization at the gap of the CF, respectively.

[7] HfO_2-based [70, 85, 97, 99], Ta_2O_5-based [44, 100]

2.3.3 Variability

Failures of resistive switching can appear consecutively [70, 85, 97–99] or only occasionally [99, 103][8] during cycling. M. Lanza *et al.* pointed out that the LRS and HRS resistances, and other electrical properties should be measured in each cycle [96]. Only by doing this, the failures only in certain cycles can be detected and the results of endurance tests are reliable. That is, the reliability of a VCM cell should be assessed by the conventional endurance criteria of extensive operations, and by the cycle-to-cycle (C2C) variability.

The C2C variability refers to the variation of LRS or HRS resistances across switching cycles. It corresponds to the change of a CF arising from the stochastic migration of oxygen vacancies in a microscopic viewpoint. The C2C variability is typically accessed using the normalized deviation, calculated by dividing a deviation by either the average or the median value. Note that a deviation is calculated by the difference between the 30% and 70% values in Ref. [95]. In addition, the normalized deviation against either the resistance or the current compliance exists. It is evident that the normalized deviation is reduced at a higher current compliance, as shown in Fig. 2.6. Fig. 2.7a reveals the same trend after applying the inverse relationship between the resistance and current compliance. Constant products of LRS resistance and current compliance are fitted to be in the range of 0.4 V to 1 V [95, 104, 105]. Moreover, power-law relationships with exponents ranging from 0.5 to 1.0 are found for LRS resistances [95, 105–108] for relatively large current compliance, e. g., $I_{cc} > 10\,\mu A$. When the current compliance is lower than a certain value, a leveling-off of the normalized deviation emerges as seen in both Fig. 2.6 and 2.7a. On the other hand, exponents are closer to 0.5 for HRS resistances [95, 108].

On top of the C2C variability, the different electrical performance among different memory cells is mostly referred to as the device-to-device (D2D) (or cell-to-cell) variability [94, 95]. Due to the increasing number of memory cells in commercial applications, the uniformity of memory cells should also be under control. In contrast to the C2C variability, the D2D variability is considered to be associated with process variations. In this regard, a reduction in D2D variability should be plausible by better controls of fabrication processes. For instance, the deposition of the oxide layer is thought to be the dominant factor for the D2D variation [94], and various indicators are proposed to

[8] [103]: Ag/spin-on-glass/poly(3,4-ethylenedioxythiophene) polystyrene sulfonate

evaluate the uniformity of fabricated cells [109]. The reader is referred to Ref. [94] for the comparison of two common processes for VCM cells, i.e., atomic layer deposition (ALD) and sputtering.

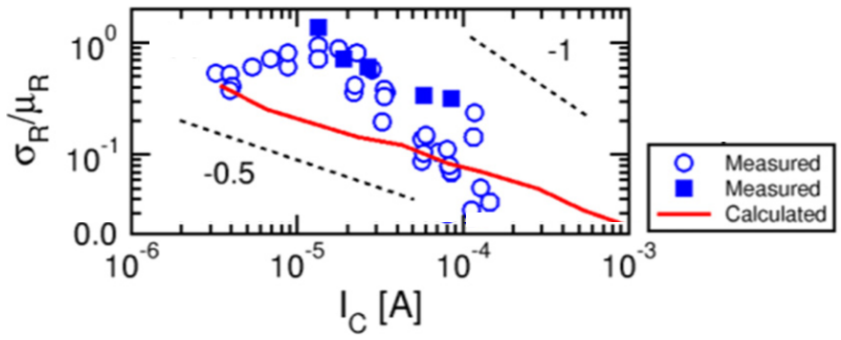

Figure 2.6: Normalized deviation against current compliance for LRS. Reproduced with permission from [108] with data in solid squares from [105], © 2014 IEEE.

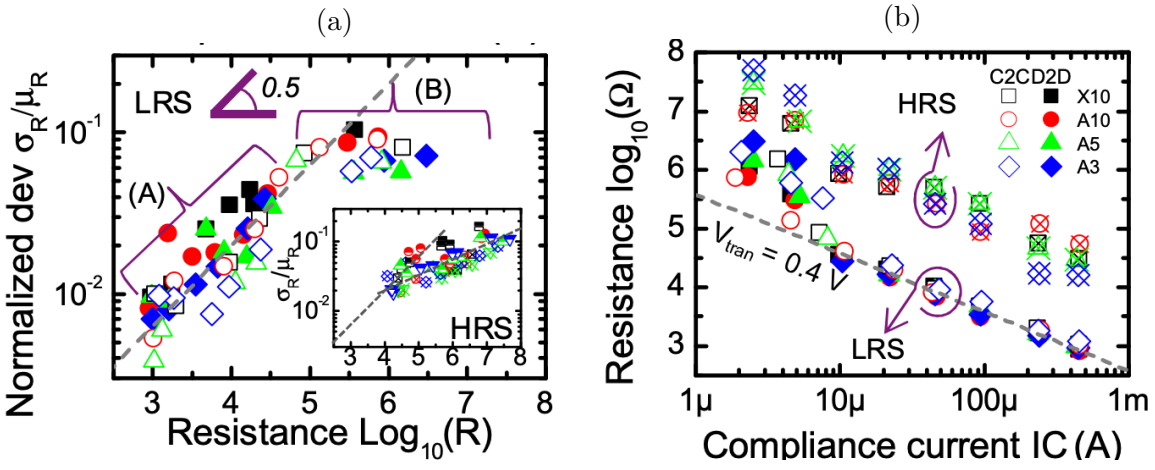

Figure 2.7: (a) Normalized deviation against LRS and HRS resistances. (b) LRS and HRS resistances against current compliance. Reproduced with permission from [95], © 2013 IEEE.

2.3.4 Short conclusion

Despite all the excellent properties of a single VCM cell, commercial applications are not yet mainstream in the market. To migrate towards the large-scale manufacturing phase, the reliability of commercial electronics is of central importance. Unfortunately,

the working principle of VCM cells, which involves stochastic vacancy migrations, has made the precise control of resistance states impossible. For the NAND Flash and DRAM cells, an Error Correcting Code (ECC) design utilizes a small number of logic elements to ensure the correctness of data [110, 111]. However, conventional ECC might not be sufficient for RRAM cells since their variability is expected to be larger [111]. The choice of an ECC design is left for future investigations to ensure its functionality without offsetting the benefits.

So far, measurements and theoretical models have provided a clear working principle for VCM cells. Specifically, density functional theory (DFT) is applied to associate materials with the corresponding formation energies of oxygen vacancies, which is useful for understanding the US and the BS phenomena as well as FORMING conditions. It is noteworthy that DFT is suitable for a structure containing few atoms over a short timescale due to complicated interactions. This would be sufficient to explore the physical properties around the CF, and a homogeneous material. However, the migration of vacancies during the resistive switching process could give rise to intricate atomic distributions, which might not be considered by DFT calculations. Moreover, physical properties have been found to vary dramatically in the presence of a cluster of vacancies which will be further discussed in Sec. 2.6 and 2.7. Therefore, DFT only provides a guideline and there are still missing pieces to explain the dynamics during the resistive switching process.

To this end, modeling the spatio-temporal distribution of vacancies in a larger dimensionality and a longer timescale is crucial for practical applications. Models have been developed for different regimes of space- and time- scales, see Fig. 2.8. For example, 1D compact models can simulate the electrical performance at a single cell level with the limitation of not accounting for discrete vacancy distributions [112, 113]. The spatial resolution of vacancies is improved by 3D continuous models [108, 114–116], which are useful for devices with a large current compliance. In contrast, KMC models treat oxygen vacancies as point defects, which lays a more intuitive picture of the dynamical process.

Within the framework of a KMC model, the time evolution is simulated by splitting into two steps. Firstly, the charge transport associated with physical quantities, such as temperature and electrostatic potential, are solved self-consistently under the given vacancy distribution. Once the solution is obtained, the change in vacancy distribution is regarded as a Poisson process, and the subsequent time step is determined accord-

ingly. The detailed description of the charge transport equation for current, and physical quantities are laid down in Sec. 2.4, and 2.5, respectively. In addition, the change in vacancy distribution is reserved for Sec. 2.6 and 2.7.

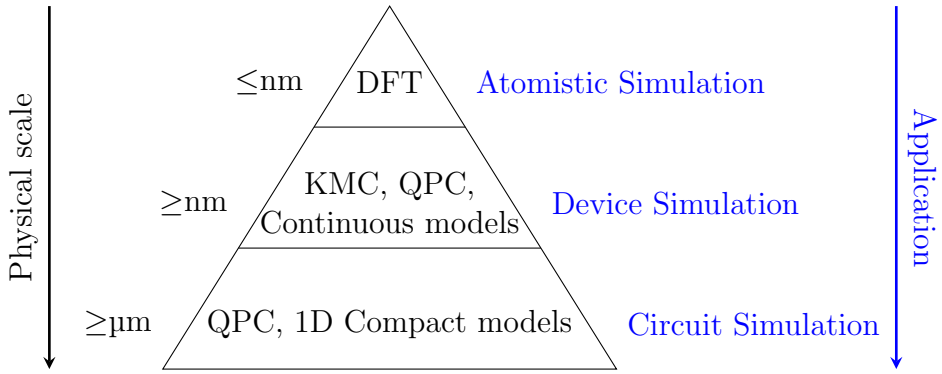

Figure 2.8: A hierarchy of physical scales and applications for different models.

2.4 Charge transport model

Due to the diversity of materials in VCM cells, several competing charge transport theories have been proposed. In general, these theories can be categorized into two groups: those that use a conventional drift diffusion (DD) mechanism and those that use a quantum mechanical tunneling mechanism. The choice of a transport model is determined by the energy levels of oxygen vacancies [117]. For an oxide layer with a shallow defect level, drift and diffusion are more appropriate for the conduction mechanism whereas quantum tunneling is more suitable for an oxide layer with a deep defect level. In this thesis, the simulator is developed primarily for HfO_2-based VCM cells, where quantum mechanical effects are assumed to dominate the charge transport process. Given the similar physical properties of HfO_2 and ZrO_2, it might apply to ZrO_2-based VCM cells as well. Unfortunately, a unified charge transport scheme is still missing even if the focus is narrowed to the specific oxide materials.

For example, Puglisi *et al.* modeled the resistance by a series connection of resistances corresponding to the unbroken and broken part of a CF in a 1D compact model [112]. The LRS and HRS resistances are explained by the length change of the broken segment. More generally, the contact potential is taken into account and the resistance of each segment is modeled to depend on the vacancy concentrations [113]. To resolve the spatio-temporal evolution of the vacancy density, a 3D DD model is proposed [108, 114–116]. It is worth noting that the drift and diffusion refer to the migration of oxygen vacancies instead of the charge carriers in conventional device simulations. The charge transport is treated as a continuous flow through the oxide layer. In contrast to models that regard vacancies as a continuous density, some treat vacancies as point defects. In the quantum point contact (QPC) model, the CF has an hour-glass shape, and the charge transport is formulated by the quantum tunneling through a constriction region [118–120]. However, the time evolution of the spatial distribution is not included in the QPC model, raising questions about the plausibility of the assumed geometry. On the other hand, the trap-assisted-tunneling (TAT) mechanism is typically employed by KMC models where the spatio-temporal evolution of vacancies is simulated. The simulation of a stochastic migration process is reserved for Sec. 2.9.3, and this section focuses on the TAT mechanism. In this framework, oxygen vacancies in the oxide act as traps since charge carriers from one electrode to the other are achieved by being captured and emitted from vacancy sites. This charge transport scheme is based on the quantum

mechanism, namely, the quantum tunneling effect.

2.4.1 Electron energy level

Similar to doped semiconductors, discrete energy levels emerge in the presence of oxygen vacancies. For cubic phase Hafnium oxide (c-HfO_2), the defect energy levels of neutrally and positively charged vacancies are approximately 2.1 eV, and approximately 1.0 eV below the conduction band minimum (CBM), respectively [121]. Meanwhile, the band gap of c-HfO_2 is approximately 6.0 eV [121]. For the monoclinic phase of HfO_2(m-HfO_2), the band gap is approximately 5.7 eV, and energy levels of neutrally charged and positively charged vacancies are approximately 2.3 eV and 0.6 eV below the CBM, respectively [122, 123]. This implies that holes are trapped much more deeply compared to electrons at vacancy sites. In this regard, neither the thermal excitation of holes to the valence band nor the tunneling of holes from one vacancy to another is significant. The latter could be understood as the high potential barrier seen by trapped holes. As a consequence, the current contribution from the hole transport is expected to be significantly less than that from the electron transport.

2.4.2 Trap-assisted-tunneling mechanism

Within the TAT mechanism, electron transport from one electrode to another occurs in multiple steps, with tunneling between traps enabled due to a small separation. Notably, the Poole-Frenkel (PF) emission [124, 125], Schottky emission, Fowler-Nordheim (FN) tunneling, and direct tunneling are not taken into account. The PF emission labeled as process 5 in Fig. 2.9 refers to electrons being excited to the conduction band. The Schottky emission labeled as process 1 refers to thermally activated electrons overcoming the energy barrier at the interface. Within the FN tunneling, electrons tunnel from a trapped site to the conduction band where it is lowered by the electric field, as shown by processes 2 and 6. These conduction mechanisms are not considered based on the following.

Yu *et al.* extracted measurement data and obtained a fitted value for a trap level of less than 0.1 eV [126]. Based on this unrealistic result, it is concluded that the PF emission is implausible for TiN/HfO_x/Pt cells with a thickness of 10 nm. However, this conclusion conflicts with the previous finding where the trap level of (1.5 ± 0.1) eV was determined

from the PF mechanism [127]. The discrepancy might arise from the much thinner HfO_2 layer with the thickness of 1.3 nm under investigation, where a much stronger electric field makes the PF mechanism prominent.

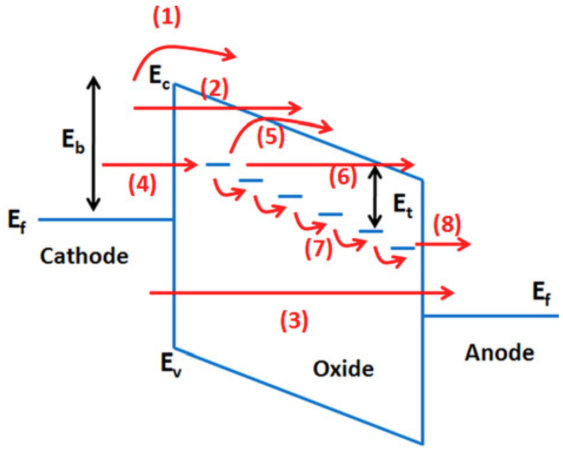

Figure 2.9: A summary of different charge transport schemes. (1) Schottky emission; (2) FN tunneling; (3) direct tunneling; (4) tunneling from cathode to traps; (5) emission from trap to conduction band, which is essentially the Poole-Frenkel emission; (6) FN like tunneling from trap to conduction band; (7) trap to trap tunneling; and (8) tunneling from traps to anode. Reproduced with permission from [126], © 2011 AIP Publishing.

On the other hand, one could argue that the FN tunneling is a minor effect due to the deep trap level, i. e., about 2 eV below the CBM according to DFT and measurements [122, 127, 128]. The distance for electrons tunneling to the sites with a sufficiently low conduction band is too far, hindering the scheme from being plausible. Similarly, the direct tunneling labeled as process 3 is excluded for the oxide thickness of 10 nm [126]. It is pointed out that the direct tunneling is negligible for a thickness roughly above 3 nm [128]. Lastly, the measured current is insensitive to temperatures, ruling out the Schottky emission [126].

Therefore, the TAT scheme is adopted where electrons enter and leave the oxide layer by processes 4 and 8, and tunnel among vacancy sites by process 7. The TAT mechanism is formulated in the master equation, which governs the time evolution of the electron occupancy at each vacancy site with the notation p. Under quasi-stationary conditions,

the master equation for the ith vacancy takes the form,

$$\begin{aligned}\frac{\partial p_i}{\partial t} = &-p_i \sum_{M=\mathrm{A,C}} R_{iM} + (1-p_i) \sum_{M=\mathrm{A,C}} R_{Mi} \\ &- p_i \sum_{j,j\neq i} (1-p_j) h_{ij} + (1-p_i) \sum_{j,j\neq i} p_j h_{ji} = 0,\end{aligned} \tag{2.2}$$

where R_{iM} and R_{Mi} are electron hopping rates from the trap to an electrode and vice versa. Both the anode ($M = \mathrm{A}$) and the cathode ($M = \mathrm{C}$) are taken into account. h_{ij} and h_{ji} are the electron hopping rates from the ith to the jth vacancy and vice versa. The current through the device is given by

$$I = e \sum_i p_i R_{i\mathrm{C}} - (1-p_i) R_{\mathrm{C}i}. \tag{2.3}$$

To further discuss Eq. (2.2), different terminologies are assigned for the charge transport between an electrode and a trap, and between one trap and another trap. The term "tunneling" refers to processes 4 and 8 in Fig. 2.9 where electrodes are involved, while "hopping" refers to process 7 where electrons transport between traps.

The R_{iM} and R_{Mi} in Eq. (2.2) are further split into three terms,

$$\begin{aligned}R_{iM} &= R_0 T_{iM} \mathcal{F}_{iM} \\ R_{Mi} &= R_0 T_{Mi} \mathcal{F}_{Mi}\end{aligned}, \tag{2.4}$$

where R_0 is an electrode coupling term, T_{Mi} and T_{iM} are tunneling probabilities, and $\mathcal{F}_{iM}$ and $\mathcal{F}_{Mi}$ are incomplete Fermi integrals.

Generally, the tunneling between electrodes and oxygen vacancies involves interactions between phonons and electrons [129–131]. However, the inclusion of phonon interactions increases the computation efforts. The reader is referred to Ref. [129–134] for multiphonon TAT model, where phonon interactions are taken into account. Since the objective of this work is the investigation of the C2C variability, where the capability of simulating more cycles is desired, phonon interactions are not considered. Rather, it is assumed that injected electrons tunnel with the same probability [135, 136]. That is, tunneling probabilities are independent of electron energies [135]. This decouples tunneling probabilities and incomplete Fermi integrals, leading to Eq. (2.4). Tunneling

probabilities given by the Wentzel-Kramers-Brillouin (WKB) approximation read

$$\begin{aligned} T_{iM} &= \exp\left(-2\int_{x_M}^{x_i} \frac{\sqrt{2m^*(-E^q_{i,\text{filled}} - e\varphi(x))}}{\hbar}\,\mathrm{d}x\right) \\ T_{Mi} &= \exp\left(-2\int_{x_M}^{x_i} \frac{\sqrt{2m^*(-E^q_{i,\text{empty}} - e\varphi(x))}}{\hbar}\,\mathrm{d}x\right) \end{aligned}, \tag{2.5}$$

where m^* is the effective electron mass, e is the magnitude of an elementary charge, φ is the electrostatic potential, and x_i and x_M are the 1D positions of the ith trap and an electrode, respectively. $E^q_{i,\text{filled}}$ and $E^q_{i,\text{empty}}$ are electron energy levels relative to the CBM for a vacancy in charge state q when filled and empty, respectively. It is noted that the definition of energy levels is not unique. In the case of energy levels are not referred to the CBM E_{CBM}, $E_{\text{CBM}} - e\varphi(x) - E^q_{i,\text{filled}}$ replaces the process involving a filled state. An analogous substitution applies to the process involving an empty state.

The incomplete Fermi integral for electrons injected into the ith vacancy reads

$$\mathcal{F}_{Mi} = \int_{E^q_{i,\text{empty}}}^{\infty} F(E)\rho(E)\,\mathrm{d}E, \tag{2.6}$$

and for electrons injected into an electrode,

$$\mathcal{F}_{iM} = \int_{-\infty}^{E^q_{i,\text{filled}}} (1 - F(E))\rho(E)\,\mathrm{d}E. \tag{2.7}$$

$F(E)$ is the Fermi-Dirac distribution, and $\rho(E)$ is the density of state (DOS) of the electrode. Eqs. (2.6) and (2.7) are evaluated above the empty energy level $E^q_{i,\text{empty}}$ and below the filled energy level $E^q_{i,\text{filled}}$ of the ith vacancy, respectively. Herein, the empty and filled energy levels refer to the electron energy levels of the vacancy site being unoccupied and occupied by electrons, respectively. Given that the energetically favored charge state is either doubly positive ($q = +2$) or neutral ($q = 0$), different charge states of vacancies are available under the stress of an external voltage. Specifically, the unoccupied state of a vacancy in the favored neutral charge state refers to the vacancy losing one electron, resulting in the vacancy being +1 charged. The occupied state of such a vacancy refers to the vacancy in the neutral charge state. A similar argument applies to a vacancy in the favored +2 charge state, where the unoccupied and occupied

states refer to a vacancy being +2 and +1 charged, respectively.

In theory, the free-electron model yields the DOS depending on the square root of the energy in 3D. However, the validity of the parabolic band structure within a thin-film electrode is questioned. To see this, the device dimensionality is compared with the de Broglie wavelength, which can be interpreted as the characteristic length below which quantum effects become important, e. g., the quantum confinement. A rough estimation of the de Broglie wavelength of electrons at room temperature yields $\lambda = 2\pi\hbar/\sqrt{3m^* k_\mathrm{B} T} \approx 6.2\,\mathrm{nm}$, where the electron rest mass is used. In this work, the constant DOS is used as in the 2D case since a typical thickness is comparable with the de Broglie wavelength. Guan *et al.* used a constant DOS for the incomplete Fermi integrals [135]. Lastly, the coupling constant R_0 serves as a fitted parameter [135].

The rate of electrons tunneling between two oxygen vacancies is formulated by the Miller-Abrahams (MA) hopping rate [137–139],

$$h_{ij} = \begin{cases} \nu_e \exp\left(-\frac{d_{ij}}{a_0}\right) & \varphi_i < \varphi_j \\ \nu_e \exp\left(-\frac{d_{ij}}{a_0} + \frac{e(\varphi_j - \varphi_i)}{k_\mathrm{B} T_i}\right) & \text{otherwise} \end{cases}, \tag{2.8}$$

where ν_e is the electron attempt frequency, d_{ij} is the distance between two traps. T_i is the temperature at the ith vacancy site. In contrast to the works where the potential obtained from solving the Laplace equation enters Eq. (2.8) [20, 135], we assume it is the potential due to space charges, i. e., charged vacancies, that modifies the electron energy levels. a_0 is the localization length of the trapped electron. More specifically, Kundstroem *et al.* showed the wave function of the electron trapped at a delta-function-like potential well has an asymptotic solution [140]. In spherical coordinates, it reads

$$\psi(r) = \left(\frac{K}{2\pi}\right)^{1/2} \frac{\exp(-Kr)}{r}, \tag{2.9}$$

where $K = \sqrt{2m^* E_t}/\hbar$, and E_t is the barrier height, i. e., the energy difference from the trapped level to the CBM. Care has to be taken when relating the K to a_0 in Eq. (2.8). From the unit analysis, it is naive to see $a_0 = 1/K$. However, a dimensionless factor could still be missing. To resolve it from an analytical perspective, let us consider the tunneling probability of an electron seeing an energy barrier height of E_t. To tunnel a

length of d_{ij} under a constant barrier, the probability is given by

$$\exp\left(-2\int_{x_0}^{x_0+d_{ij}} \frac{\sqrt{2m^*E_t}}{\hbar}\,\mathrm{d}x\right) = \exp(-2Kd_{ij}). \tag{2.10}$$

By comparing with Eq. (2.8), it is then clear that

$$a_0 = \frac{1}{2K} = \frac{\hbar}{2\sqrt{2m^*E_t}} \tag{2.11}$$

is a more physically sound assumption than $a_0 = 1/K$. a_0 is interpreted as the attenuation radius since it is the characteristic radius of the localization length. Since the diameter D instead of the radius is used in some work, the common exponential factor reads $\exp(-2d_{ij}/D)$ for the corresponding MA hopping rate. Since the derivations of Eq. (2.9) and (2.11) are not limited to electrons, the same argument can be applied to holes. With a barrier height of about 5 eV for a trapped hole in HfO_2 [121], it is clear that the wave function of a hole is much more localized than that of an electron. Thus hole transport via the TAT mechanism is less dominant than electron transport.

Lastly, it is noted the attenuation radius given by Eq. (2.11) is much smaller than that used in the work [135], where HfO_2-based devices are investigated. Meanwhile, parameters involved in the electron tunneling process are adopted from this work. To fill this gap, an extra factor is multiplied to Eq. (2.11) leading to

$$a_0 = \frac{3}{4K}. \tag{2.12}$$

The attenuation radius of the filled state is approximately 3.2 Å which is comparable to the 3.3 Å used in the cited work.

The deviation of a_0 in Eq. (2.12) from the theoretical one in Eq. (2.11) is interpreted as the impacts of an ideal 1D chain of vacancies. DFT calculations have shown electrons at these sites are more delocalized, see Fig. 2.10. In Chapter 3, simulation results show that a 1D vacancy chain exists during the dynamical process. Therefore, an extended attenuation radius is assumed to account for the 1D vacancy chain effect.

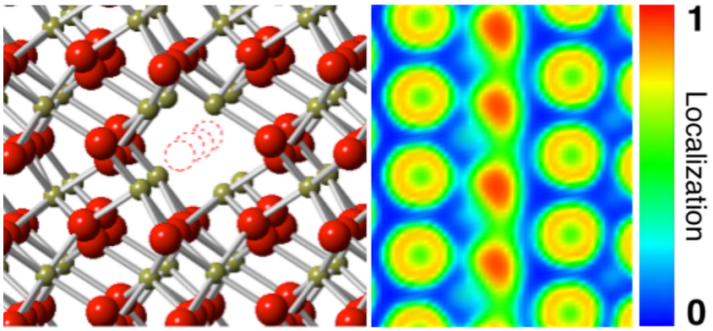

Figure 2.10: Structure and electron localization function of HfO_2 oxide with 1D vacancy chain. Reproduced with permission from [141], © 2016 IEEE.

2.5 Poisson-type equations

In the framework of the TAT mechanism, the charge transport depends on the occupation of charge carriers at vacancy sites, the electrostatic potential, and the temperature. With the fact that oxygen vacancies are inhomogeneously distributed in space and could be electrically charged, the corresponding modification to the electric potential should be considered. On the other hand, the confined cross-section, where most current flows through, makes the Joule heating around the CF region non-negligible. For both physical quantities, the corresponding governing equations are both in the general form,

$$\nabla \cdot (-m \nabla X) = f, \tag{2.13}$$

where X refers to any scalar field, e. g., the temperature or the electrostatic potential, m is the corresponding material property such as the permittivity or thermal conductivity, and f is the source for the corresponding quantity, e. g., the space charge density or heat generation rate density.

2.5.1 Poisson equation

In a medium, the speed of light c depends on the material properties and is given by

$$c = \frac{c_0}{\sqrt{\mu_r \varepsilon_r}}, \tag{2.14}$$

where $c_0 \approx 3 \cdot 10^8\,\mathrm{m/s}$ is the speed of light in vacuum, and μ_r and ε_r are the relative permeability and relative permittivity, respectively. For HfO_2 and ZrO_2 oxides, the typical values of material properties are $\varepsilon_r \approx 20$ and $\mu_r \approx 1$, leading to an estimation of speed in oxides,

$$c \approx 7 \cdot 10^7\,\mathrm{m/s}. \tag{2.15}$$

For VCM cells, the critical dimension is the thickness L, which is typically far below $1\,\mu\mathrm{m}$. Therefore, the propagation time of the electromagnetic wave through the critical dimension can be estimated as

$$t_{\mathrm{est}} = \frac{L}{c} \approx 1.4 \cdot 10^{-14}\,\mathrm{s} \tag{2.16}$$

for $L = 5\,\mathrm{nm}$. With the adoption of a time step larger than $t_{\mathrm{sim}} = 1\,\mathrm{ps} \approx 100 \cdot t_{\mathrm{est}}$ in simulations, $c \cdot t_{\mathrm{sim}} \gg L$ is fulfilled. That is, it implies that an electromagnetic signal travels almost instantaneously through the thickness dimension, which simplifies the Maxwell equation $-\frac{\partial \boldsymbol{B}}{\partial t} = \nabla \times \boldsymbol{E} = \boldsymbol{0}$. Specifically, the electrostatic potential φ can be defined and the electric field is given by

$$\boldsymbol{E} = -\nabla\varphi. \tag{2.17}$$

The substitution of $\boldsymbol{D} = \varepsilon \boldsymbol{E}$ into the Maxwell equation then yields

$$\nabla \cdot \boldsymbol{D} = \nabla \cdot (-\varepsilon \nabla\varphi) = \rho, \tag{2.18}$$

where $\boldsymbol{D}$ is the electric flux density, ε is the permittivity, and ρ is the charge density.

2.5.2 Fourier heat equation

The electron current confined in a small cross-section inevitably makes the Joule heating effect non-negligible. With the typical thermal response time being less than 1 ns [114, 142], the quasi-stationary approximation can also be applied to the Fourier heat equation provided the time step longer than approximately 0.1 µs in simulations. The governing equation reads

$$\nabla \cdot (-k_{\mathrm{th}} \nabla T) = g, \tag{2.19}$$

where T is the temperature, k_{th} is the thermal conductivity, and g is the heat generation rate density. In most applications, thermal properties including thermal conductivity are obtained from measurements adopting the so-called time-domain thermoreflectance (TDTR) technique [143]. For 5.6 nm-thick HfO_2 films, the thermal conductivity is close to $0.5\,\mathrm{Wm^{-1}K^{-1}}$ in the range of 300 K to 500 K [144].

2.6 Generation and recombination of vacancies

In physical chemistry, the occurrence of chemical reactions is explained in a scenario where reactants overcome a minimum energy barrier, known as the activation energy. The activation energy is a characteristic of a chemical reaction, which refers to the energy difference between the transition and the reactant states.

In the framework of KMC models, the common approach to explore the dynamics of oxygen vacancies involves three reactions, namely, the generation, recombination, and diffusion of vacancies. However, activation energies could vary dramatically for different configurations, such as crystal structures, local defect structures, or charge states of vacancies. It is then a question of how many configurations are covered by *ab initio* calculations. In addition, the impact of the external electric field is not discussed in many first principle calculations. Instead, an empirical thermochemical model is widely adopted to explain the shift of activation energies in the presence of the electric field.

In contrast to the consensus on the role of V_Os involved in processes of the resistance change, the physical effects leading to the creation of V_Os are still under debate. Two competing models exist for the source of V_Os, namely, the creation of Frenkel pairs (FPs) inside the bulk oxide and the reduction reaction at the oxide/electrode interface.

A FP refers to the displacement of an oxygen atom from its lattice site to an interstitial site. This leads to the creation of a vacancy-interstitial pair. In the Kröger-Vink notation, the reaction for a neutrally charged FP can be written as

$$\mathrm{O}_\mathrm{O}^\times \longrightarrow \mathrm{V}_\mathrm{O}^{\bullet\bullet} + \mathrm{O}_\mathrm{i}'', \tag{2.20}$$

where $\mathrm{O}_\mathrm{O}^\times$ is the neutrally charged oxygen atom at the lattice site, $\mathrm{V}_\mathrm{O}^{\bullet\bullet}$ is the doubly positively charged oxygen vacancy and the O_i'' is the doubly negatively charged interstitial oxygen ion. In literature, the FP is expressed with components enclosed by a set of square brackets, e. g., $[\mathrm{V}_\mathrm{O}^{\bullet\bullet} - \mathrm{O}_\mathrm{i}'']$ for the FP in Eq. (2.20).

On the other hand, the reduction reaction takes the form

$$\mathrm{O}_\mathrm{O}^\times \longrightarrow \mathrm{V}_\mathrm{O}^{\bullet\bullet} + \mathrm{O} + 2\,\mathrm{e}', \tag{2.21}$$

where O can either oxidize with the electrode or be released as oxygen gas after combining with a second oxygen atom. The oxidation of the electrode is observed in ZrO_x-based [41], and the release of oxygen gas is observed in TiO_2-based devices [38]

and STO [145]. It has been shown that the oxygen content is not homogeneous but lower at the oxide-metal interface where the oxygen affinity of the metal is relatively higher [48, 49]. This stresses the impact of the metallic electrode involved in the vacancy generation process.

2.6.1 Formation energy

From a theoretical perspective, the formation energy is determined by *ab initio* calculations to estimate the generation of vacancies. The general equation for the formation energy of a point defect in a charge state reads [146]

$$E_{\mathrm{form}}(q) = E_{\mathrm{tot}}(q) - E_{\mathrm{tot}}^{\mathrm{bulk}} + q(E_{\mathrm{F}} + E_{\mathrm{VBM}} - e\Delta V_{0/b}) + E_{\mathrm{corr}}(q) - \sum_i n_i \mu_i, \qquad (2.22)$$

where $E_{\mathrm{tot}}(q)$ and $E_{\mathrm{tot}}^{\mathrm{bulk}}$ are the total energies of the defect supercell and the pristine supercell, respectively. With the reference set to the valence band maximum (VBM) of the bulk material E_{VBM}, the $E_{\mathrm{F}} + E_{\mathrm{VBM}}$ is the Fermi level. The term $\Delta V_{0/b}$ is added to align the electrostatic potential of the defect supercell with that of the bulk supercell [147]. $E_{\mathrm{corr}}(q)$ is a correction term due to the finite size of a supercell [147, 148]. In some models, $-e\Delta V_{0/b}$ is included in a correction term [148]. The number change of the i-species is accounted for by the corresponding chemical potential μ_i with $n_i < 0$ for the removal from the bulk host. In addition, the chemical potential depends on the content of the species. For oxygen atoms, O-rich and O-poor conditions are two limits. That is, the O-rich condition refers to the O in the O_2 gas, setting the reference $\mu_O = 0$, while the O-poor condition refers to the chemical potential at the metal/oxide equilibrium. Typically, formation energies with both O-poor and O-rich conditions are plotted in a single defect energy diagram [149]. Close to the Hf-HfO_2 interface, Guo *et al.* have shown a decrease of μ_{O} by 5.9 eV compared to that of the bulk oxide, which is in agreement with Ref. [150]. Therefore, the creation of oxygen vacancies is dramatically promoted.

On the other hand, the consideration of FPs in different regions of the HfO_2, i. e. at the HfO_2-Hf interface, and in the bulk oxide, gives a similar conclusion [151]. In short, the formation of FPs in the bulk oxide is energetically unfavored. Even if a FP is formed, it will undergo recombination within a picosecond time scale, leading to a defect-free configuration [152, 153]. This excludes the impacts of FPs for the resistive

switching phenomena due to the short lifetime. Moreover, comparing a single V_O within the bulk oxide with a FP close to the Hf interface, the FP is shown to be favored from the energy perspective. Once a FP is formed, the interstitial oxygen gets more stable when migrating into the Hf-electrode [151]. According to Ref. [152], this scenario is not considered as a FP but is equivalent to Eq. (2.21), where the electrode acts as an oxygen reservoir.

In this work, the generation and recombination of V_Os is modeled by an interfacial model [19]. Specifically, V_Os can access and leave the simulated oxide region only via sites on the Hf-HfO_2 interface, which is modeled as a single layer. At these sites, generation and recombination rates are related to the activation energies as well as field acceleration corrections given by

$$k_{\mathrm{G}} = \nu_0 \cdot \exp\left(-\frac{E_{\mathrm{G}} - \alpha b E_{\mathrm{loc}}}{k_{\mathrm{B}} T}\right), \tag{2.23}$$

$$k_{\mathrm{R}} = \nu_0 \cdot \exp\left(-\frac{E_{\mathrm{R}} + (1-\alpha) b E_{\mathrm{loc}}}{k_{\mathrm{B}} T}\right), \tag{2.24}$$

where ν_0 is the attempt frequency. E_{G} and E_{R} are zero-field activation energies for the generation and recombination of oxygen vacancies, respectively. The zero-field activation energies, symmetry factor α, bond polarization b, and local electric field E_{loc} are detailed in the following sections.

2.6.2 Local structures

It has been pointed out that the interface conditions [154, 155], the aggregation of oxygen vacancies [156–161], and grain boundaries (GBs) [162] can lower the formation energy of an oxygen vacancy. With a rough oxide-metal interface fabricated by the co-sputtering technique, a much worse endurance is observed in Ta_2O_5-based devices [155]. This is attributed to the spots, introduced by rough interface, for creating oxygen vacancies. Consequently, a significant CF grows in a confined region, leading to a significant Joule heating and device degradation. In contrast, a smooth interface promotes homogeneous distribution of CFs with smaller cross-sections, leading to improved endurance.

On the other hand, DFT calculations have shown lower zero-field activation energies for vacancy generation and higher mobility along GBs. Based on these findings, the GBs have been widely assumed for simulating filamentary type VCM cells [132, 163, 164].

2.6.3 Bond polarization

The abrupt increase in current of a RRAM device is analogous to a dielectric breakdown under a high electric field in oxides [165, 166]. The principal model is a thermochemical model or the E-model in some literature [167]. Within the E-model, the required energy for breaking bonds is associated with the electric field and the atoms. Specifically, the differences in electronegativity of oxide constituents lead to atomic dipoles, which modulate the bond-breaking energy in the presence of an electric field. Furthermore, the electric field at the microscopic level is modulated by neighboring dipoles which gives rise to a difference in the electric field at the macroscopic level. The macroscopic electric field is related to the microscopic electric field[9] via the Lorentz relation [168–171]. Since E_{loc} is modulated by neighboring atomic dipoles, it depends on the arrangement of the local crystal structure. For a cubic structure, the analytical relation reads

$$\boldsymbol{E}_{\mathrm{loc}} = \frac{\varepsilon_r + 2}{3} \boldsymbol{E}_{\mathrm{macro}}, \tag{2.25}$$

where $\boldsymbol{E}_{\mathrm{macro}}$ is the macroscopic electric field, and the directions of these two fields are aligned. The bond-breaking energy shifts are attributed to the potential energy change arising from the molecular dipole $\boldsymbol{p_0}$ together with the local electric field [167]. More specifically, the inner product $\boldsymbol{p_0} \cdot \boldsymbol{E}_{\mathrm{loc}}$ is replaced by the scalar product bE_{loc} where b is the bond polarization with all involved atomic dipoles projected on to the direction of $\boldsymbol{E}_{\mathrm{loc}}$ [172]. Typically, the magnitude of the macroscopic field is not obtained by solving the Poisson equation. Instead, it is related to the applied voltage V_{app} by

$$E_{\mathrm{macro}} = \frac{V_{\mathrm{app}}}{L}, \tag{2.26}$$

where L is the thickness of the oxide layer.

For VCM cells, the E-model is extensively employed in time dependent dielectric breakdown (TDDB) measurements to extract activation energies under the stress of applied voltages. Similar to the approach to estimate the retention time, the Arrhenius equation is applied to extract the activation energy (see Fig. 2.11).

[9] In some literature, the terminology of the local electric field E_{loc} is used for the microscopic electric field.

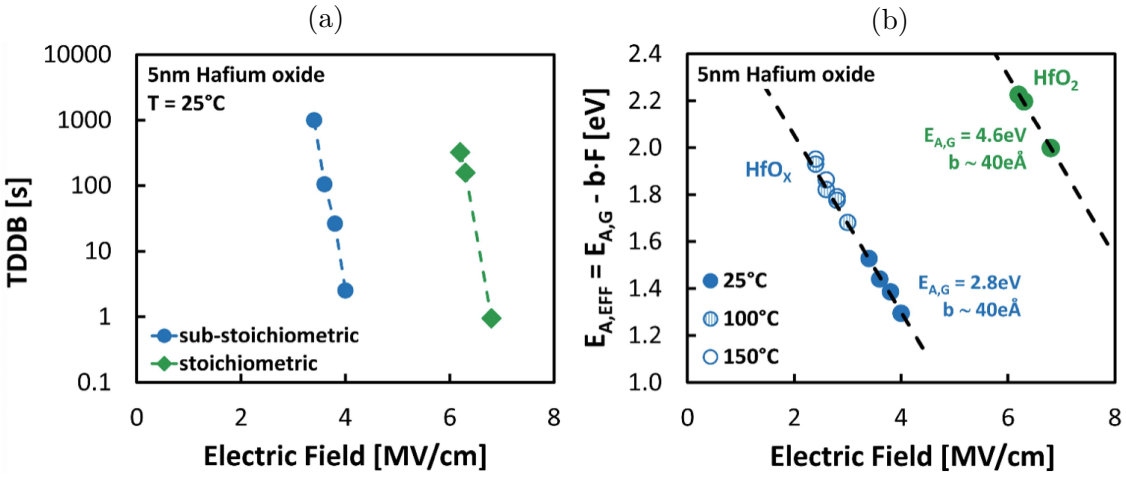

Figure 2.11: TDDB data to extract the zero-field activation energy and the corresponding bond polarization. (a) Arrhenius plot for devices under the stress of an electric field. (b) Effective activation energy versus the electric field. Reproduced with permission from [173], © 2013 IEEE.

It is noteworthy that Eq. (2.23) is an empirical model for the modulation of activation energies in the presence of an electric field. Schie *et al.* [152] mentioned the local field, i. e. Eq. (2.25), is not plausible for the filamentary type VCM cell from two perspectives. First, short-range interactions should be weaker than electrostatic interactions. Second, dielectric properties should be homogeneous within a range where the neighboring atomic dipoles contribute to the local field [174, 175]. Both requirements are not fulfilled near the CF in a HfO_2-based cell whereas a simple Arrhenius equation seems to be a good approximation in TDDB measurements. Since the requirements do not hold near the CF, the interface seems to be a reasonable region where it is assumed to be sufficiently away from the CF. This interpretation supports the interfacial model and hinders the generation of a FP within the oxide layer.

In contrast to the generation of oxygen vacancies, the recombination is more difficult to measure. From a theoretical perspective, the recombination of oxygen vacancies can also be explained as the recombination of a pair of a vacancy and an interstitial oxygen, or a vacancy and an oxygen atom in the reservoir. Again, one could argue the FPs might be a minor mechanism due to a fast recombination rate. The molecular dynamics (MD) calculations have shown annihilation of FPs at a picosecond time scale after being generated in HfO_2 [152, 153]. The lifetime of a $[V_O^{\bullet\bullet} - O_i'']$ pair and of a

$[V_O^{\times} - O_i'']$ pair are comparable. That is, the recombination is still fast even in the absence of Coulomb interaction. The short lifetime of FPs suggests negligible impacts during the operation. As a consequence, the recombination with oxygen atoms in the reservoir leaves the interfacial model the only plausible process.

The recombination and generation of oxygen vacancies are reverse reactions; thus the electric field has opposite effects on the change in activation energy (see Eqs. (2.23) and (2.24)). In theory, the symmetry factor α is a free parameter that describes the position of a barrier peak in a reaction coordinate. Typically, a value close to 0.5 is used for α, indicating an almost equal shift in the energy barrier for reversible reactions due to the electric field. However, the $\alpha = 1$ is assumed in this work. This is identical to Ref. [130, 176] and is adopted based on two findings. First, the linear dependence of the effective generation barrier to the electric field is confirmed in TDDB experiments [173]. Second, the total number of vacancies is believed to be almost unchanged after the FORMING process [120, 177]. Any $\alpha \neq 1$ is effectively treated as a choice of a new zero-field generation barrier together with a high zero-field recombination barrier. Therefore, the $\alpha = 1$ is chosen to simplify further discussions.

With the interfacial model, the abrupt current increase in the FORMING stage can not be solely explained since it predicts only the accumulation of vacancies at the interface but no CF. To have the CFs, migration of vacancies from the interface into the bulk oxide must be taken into account.

2.7 Vacancy diffusion

Similar to the generation and recombination processes, the diffusion is modeled by the Arrhenius equation in the form

$$k_{\mathrm{D}} = \nu_0 \exp\left(-\frac{E_{\mathrm{D}} + \Delta E_{\mathrm{D}}}{k_{\mathrm{B}} T}\right), \tag{2.27}$$

where the E_{D} is the zero-field activation energy of the diffusion process. From theoretical calculations, a much lower activation energy of a positively charged $\mathrm{V_O}$ than of a neutrally charged one has been found. The values for a doubly positively and neutrally charged vacancy in m-HfO_2 are approximately 0.7 eV [178, 179] and 2.0 eV [178, 180, 181], respectively. Moreover, *ab initio* calculations have investigated the impact of GBs [179], 1D vacancy chains [141, 182] and the crystal orientation [181], suggesting shifts of the zero-field activation energy. Modulations due to these local structures can be categorized by migration directions. Therefore, the diffusion barriers are not only inhomogeneous in space but also anisotropic near the above-mentioned structures. These effects are detailed in the next section.

2.7.1 Anisotropic diffusion

While the CF refers to the description from a macroscopic viewpoint, it is not strictly an idealized 1D structure from a microscopic level. In the following discussion, the vacancy chain refers to an idealized 1D structure. From a theoretical perspective, the impact of a CF is then investigated by a vacancy chain. Rushchanskii *et al.* considered the c-HfO_2 and showed the diffusion of a vacancy out of the chain, i. e., the normal direction to the chain axis, is suppressed [182]. On the other hand, the diffusion in the parallel direction is enhanced. Specifically, the first vacancy to detach the rest of the chain is 1.1 ev. However, the barrier for the original second vacancy to detach the remaining chain decreases to the value of 0.67 eV. The barrier saturates to the value of 0.51 eV for the third vacancy of the original vacancy chain. Similar trends are found in the m-HfO_2. Duncan *et al.* considered both neutrally and positively charged states and showed that outwards migration is suppressed and the inwards migration is enhanced [141]. In addition, the migration in parallel to the chain is enhanced as shown in Fig. 2.12.

Interestingly, such anisotropic modulations are also found in a different local structure,

i. e., GBs. McKenna *et al.* considered a twin-boundary inside m-HfO_2 and showed the migration of positively charged vacancies inside the GB being promoted with a barrier of 0.57 eV [179]. Similarly, migration of vacancies towards and outwards the GB in the normal direction is enhanced and suppressed, respectively. Furthermore, the modulation of barriers is significant only within 3 Å of the twin-boundary, which is roughly the nearest distance of two oxygen atoms in the HfO_2. Admittedly, not only the anisotropic but also isotropic zero-field diffusion barriers are not thoroughly studied for either c-HfO_2 or m-HfO_2, and results from both crystal structures are adopted in this thesis.

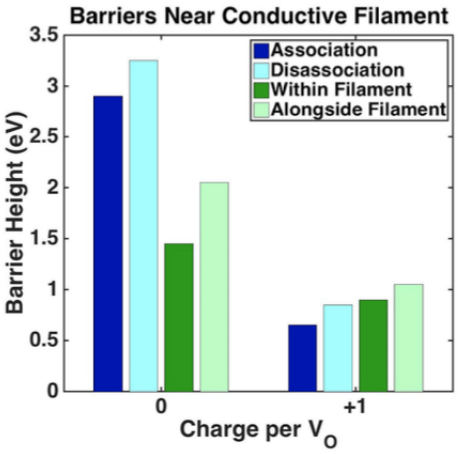

Figure 2.12: Migration barriers for V_Os near and inside V_O chain. A larger barrier height corresponds to a lower vacancy mobility. Reproduced with permission from [141], © 2016 IEEE.

2.7.2 Field acceleration

Herein, a simplified ion hopping scheme is applied, where the energy barrier peak is assumed at the middle of the migration path [183]. Gradients of temperature [184] and the electrostatic potential are considered as driving forces. The modulation of activation energy in Eq. (2.27) takes the form

$$\Delta E_\mathrm{D} = \frac{Q_\mathrm{eff}\Delta V - k_\mathrm{B}\Delta T}{2}, \tag{2.28}$$

where Q_eff is the effective charge of the charged vacancy, accounting for its charge state as well as the electron trapped at the site. The factor $1/2$ comes from the assumption that fields shift only the value of a barrier peak but not the position of the peak, though

the electric field could in general shift the position [185]. The potential energy change at the barrier peak is proportional to the electrostatic potential difference

$$\Delta V(\boldsymbol{r}, \boldsymbol{r'}) = \Big[\varphi(\boldsymbol{r'}) - \varphi^{\boldsymbol{r}}_{\text{self}}(\boldsymbol{r'}) + \varphi^{\boldsymbol{r}}_{\text{im}}(\boldsymbol{r'})\Big] - \Big[\varphi(\boldsymbol{r}) - \varphi^{\boldsymbol{r}}_{\text{self}}(\boldsymbol{r}) + \varphi^{\boldsymbol{r}}_{\text{im}}(\boldsymbol{r})\Big], \tag{2.29}$$

where $\varphi(\boldsymbol{r})$ and $\varphi(\boldsymbol{r'})$ are the potential at the initial and final position, respectively. The final position $\boldsymbol{r'}$ refers to the destination site where the vacancy will reside after the diffusion event. The potential change should not include the electric field due to the vacancy itself, and thus the self-potentials $\varphi^{\boldsymbol{r}}_{\text{self}}(\boldsymbol{r})$ and $\varphi^{\boldsymbol{r}}_{\text{self}}(\boldsymbol{r'})$ are excluded. Specifically, $\varphi^{\boldsymbol{r}}_{\text{self}}(\boldsymbol{r})$ and $\varphi^{\boldsymbol{r}}_{\text{self}}(\boldsymbol{r'})$ are solutions of the Poisson equation, where only a single vacancy is placed at the initial site, subjected to zero applied voltages. However, the exclusion of the self-potential also removes the impact of boundaries, or equivalently, image potential in this case. Therefore, the image potential $\varphi^{\boldsymbol{r}}_{\text{im}}(\boldsymbol{r'})$ and $\varphi^{\boldsymbol{r}}_{\text{im}}(\boldsymbol{r})$ should be added back. Since the charged vacancy is placed between two metallic electrodes, the projection into the top electrode simultaneously creates a paired image charge in the bottom electrode, and vice versa. However, the potential of an image charge decays as the distance increases, leaving a finite number of image charges n_{im} important. Denote the distance to the top and bottom electrode by x, and $L - x$, respectively. The image potential at the initial position takes the form

$$\varphi^{\boldsymbol{r}}_{\text{im}}(\boldsymbol{r}) = \frac{1}{16\pi\varepsilon}\left(\sum_{i=-n_{\text{im}}}^{n_{\text{im}}} \frac{-Q_{\text{eff}}}{|2x - 2iL|} + \sum_{i=-n_{\text{im}}, i\neq 0}^{n_{\text{im}}} \frac{Q_{\text{eff}}}{|2iL|}\right). \tag{2.30}$$

Meanwhile, the image potential at $\boldsymbol{r'}$ depends on the position change, where the following cases need to be considered.

1. The distance to both electrodes is unchanged.

$$\varphi^{\boldsymbol{r}}_{\text{im}}(\boldsymbol{r'}) = \frac{1}{16\pi\varepsilon}\left(\sum_{i=-n_{\text{im}}}^{n_{\text{im}}} \frac{-Q_{\text{eff}}}{\sqrt{a^2 + (2x - 2iL)^2}} + \sum_{i=-n_{im}, i\neq 0}^{n_{\text{im}}} \frac{Q_{\text{eff}}}{\sqrt{a^2 + (2iL)^2}}\right). \tag{2.31}$$

2. The distance to the top electrode is decreased by a.

$$\varphi^{\boldsymbol{r}}_{\text{im}}(\boldsymbol{r'}) = \frac{1}{16\pi\varepsilon}\left(\sum_{i=-n_{\text{im}}}^{n_{\text{im}}} \frac{-Q_{\text{eff}}}{|2x - a - 2iL|} + \sum_{i=-n_{\text{im}}, i\neq 0}^{n_{\text{im}}} \frac{Q_{\text{eff}}}{|a - 2iL|}\right). \tag{2.32}$$

3. The distance to the top electrode is increased by a.

$$\varphi^{\boldsymbol{r}}_{\mathrm{im}}(\boldsymbol{r'}) = \frac{1}{16\pi\varepsilon}\left(\sum_{i=-n_{\mathrm{im}}}^{n_{\mathrm{im}}} \frac{-Q_{\mathrm{eff}}}{|2x+a-2iL|} + \sum_{i=-n_{\mathrm{im}},i\neq 0}^{n_{\mathrm{im}}} \frac{Q_{\mathrm{eff}}}{|a-2iL|}\right). \tag{2.33}$$

In this work, $n_{\mathrm{im}} = 50$ is used for Eqs. (2.30) - (2.33), see Fig. 2.13 for the convergence of Eq. (2.30).

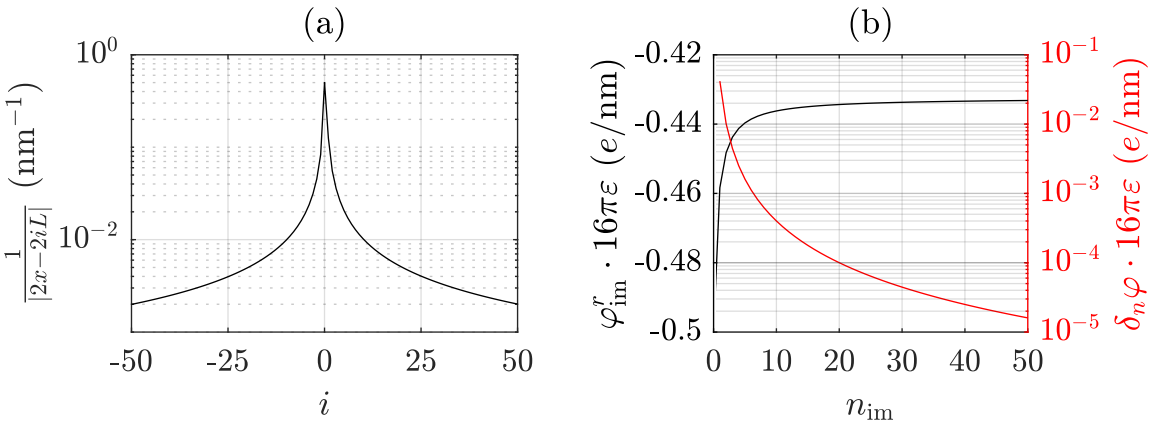

Figure 2.13: Convergence of Eq. (2.30) with $x = 1\,\mathrm{nm}$ and $L = 5\,\mathrm{nm}$, $Q_{\mathrm{eff}} = e$. (a) The absolute value of the first term in the series. (b) The sum of the series (left axis), and the absolute change of the series (right axis).

Lastly, the thermal effect is considered. Known as the Soret effect, thermodiffusion, or thermophoresis effect, the temperature gradient is a driving force for the ion migration [186]. Note that the direction of preferred migration under a temperature gradient depends on the ion species in general. Strukov *et al.* pointed out that vacancies will move to a hotter region [184], and it takes the form

$$\Delta T(\boldsymbol{r}, \boldsymbol{r'}) = T(\boldsymbol{r'}) - T(\boldsymbol{r}). \tag{2.34}$$

2.8 Statistics

2.8.1 Normal distribution and probit function

C2C variability and D2D variability are commonly visualized by the plot of cumulative distribution function (CDF). Specifically, the function describes the proportion of values no larger than its argument, see Fig. 2.14a.

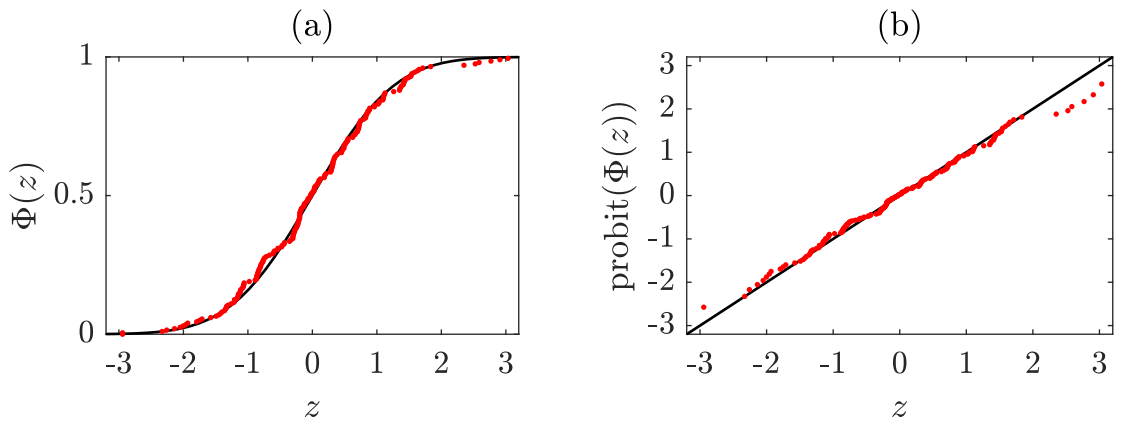

Figure 2.14: (a) CDF and (b) probit plot of an ideal and empirical standard normal distribution. The red dots are 100 points randomly drawn from the normal distribution, and the black lines are the ideal curves.

With a suitable transformation of the linear proportion, a random variable that follows the normal distribution can be easily identified from the CDF plot. This is achieved by the probit function, which is defined as the inverse function of the CDF of a normal distribution. With it, a linear relation will be seen in the probit plot if the argument follows a normal distribution, see Fig. 2.14b. In measurements, the logarithm of both LRS and HRS resistances are found to follow the normal distribution [95, 187–189]; and thus they are said to follow the log-normal distribution[10]. The reader is referred to Sec. 5.1 for the mathematical expression of the CDF of a normalized distribution.

[10]In addition to the log-normal distribution, the LRS resistance follows the Gaussian distribution are also observed [190].

2.8.2 Generation of a random process

The Poisson process among random processes is chosen to simulate the stochastic migration process of oxygen vacancies. Within the framework of the Poisson process, points are randomly distributed in the 1D time interval (or the n-dimensional space). The points are typically referred to as events, where two requirements need to be fulfilled, i. e., homogeneity and independence[11] [191]. To satisfy the two requirements, the interval of two subsequent events x needs to follow the exponential distribution

$$\mathrm{Exp}(x,\lambda) = \begin{cases} \lambda \exp(-\lambda x) & x \geq 0 \\ 0 & x < 0 \end{cases}, \tag{2.35}$$

where λ is the rate, or equivalently, the inverse of the expected interval of two subsequent events. The 1D Poisson process is a series of events, where the interval of two consecutive events is determined based on a uniformly sampled random variable $0 \leq r' \leq 1$.

Since the probability $\Pr(r' \leq R) = R$ holds provided that $0 \leq R \leq 1$, the CDF $F(x)$ can be expressed in the form

$$F(x) = \Pr(r' \leq F(x)). \tag{2.36}$$

In addition, a CDF is a monotonously increasing function and its inverse function exists, Eq. (2.36) is equivalent to

$$F(x) = \Pr(F^{-1}(r') \leq x). \tag{2.37}$$

This recovers to the definition of a CDF, suggesting that $F^{-1}(r')$ acts as the argument of the distribution F. Once the algebraic equation $F(x) = r'$ is solved for x, $F^{-1}(r')$ is then evaluated at the given value $r' = r$. In this way, the value following a known distribution is selected based on r. For the exponential distribution, $x = F^{-1}(r')$ is analytic. To see this, we need to integrate the exponential distribution for its CDF. The $F(x)$ is simply $\int_0^x \lambda \exp(-\lambda x')\,\mathrm{d}x' = 1 - \exp(-\lambda x)$. The inverse function is simply given by

$$x = -\frac{\ln(1-r')}{\lambda}. \tag{2.38}$$

It is noteworthy that aside from Eq. (2.38), the numerator $\ln(1-r')$ could be replaced

[11] In some literature, independence is called memoryless.

by $\ln(r')$,

$$x = -\frac{\ln(r')}{\lambda} \tag{2.39}$$

for the exponential distribution [192, 193]. Eq. (2.39) lays a foundation for simulating a stochastic process in the KMC simulator, where specific details are reserved for Sec. 2.9.3.

2.9 Simulation method

2.9.1 Spatial discretization

To solve a continuous partial differential equation (PDE) numerically for the spatial distribution of a physical quantity, the system needs to be discretized. In this regard, a finite number of points are introduced, which allows to approximate the solution to the PDE by solving a set of algebraic equations. The form of an algebraic equation set is dependent on the discretization approach. For semiconductor device simulations and thus this work, the flux conservation is critical in relevant governing equations. To this end, the finite volume method (FVM) method is adopted since the flux conservation can be fulfilled with fewer efforts. The implementation of the FVM involves the construction of finite volumes based on grids. In contrast to the general cases, where unstructured grids are taken into account, we only consider structured grids. More specifically, the tensor product of three 1D grids yields

$$\boldsymbol{r}_{i,j,k} = \begin{pmatrix} x_i \\ y_j \\ z_k \end{pmatrix}, \tag{2.40}$$

where positive integers i, j and k are bounded by their maximum numbers, i. e., $1 \leq i \leq N_x$, $1 \leq j \leq N_y$ and $1 \leq k \leq N_z$. That is, the boundary in the x-direction is represented by x_1 and x_{N_x}, and so as for the y-direction and the z-direction. Moreover, denote three half integers for the intersection point of bisections,

$$\boldsymbol{r}_{i\pm1/2,j\pm1/2,k\pm1/2} = \begin{pmatrix} x_{i\pm1/2} \\ y_{j\pm1/2} \\ z_{k\pm1/2} \end{pmatrix} = \frac{1}{2}\begin{pmatrix} x_i \\ y_j \\ z_k \end{pmatrix} + \frac{1}{2}\begin{pmatrix} x_{i\pm1} \\ y_{j\pm1} \\ z_{k\pm1} \end{pmatrix}. \tag{2.41}$$

The finite volume $\Omega_{i,j,k}$ is then the volume enclosed by $x \in (x_{i-1/2}, x_{i+1/2})$, $y \in (y_{j-1/2}, y_{j+1/2})$ and $z \in (z_{k-1/2}, z_{k+1/2})$. Denote the length in a vector notation,

$$\Delta\boldsymbol{r}_{i+1/2,j+1/2,k+1/2} = \begin{pmatrix} \Delta x_{i+1/2} \\ \Delta y_{j+1/2} \\ \Delta z_{k+1/2} \end{pmatrix} = \begin{pmatrix} x_{i+1} - x_i \\ y_{j+1} - y_j \\ z_{k+1} - z_k \end{pmatrix}, \tag{2.42}$$

the grid primitive of (i, j, k) is the cuboid centered at $\boldsymbol{r}_{i+1/2,j+1/2,k+1/2}$ with the length $\Delta\boldsymbol{r}_{i+1/2,j+1/2,k+1/2}$ in each dimension. Conversely, the extent of a finite volume can be defined in terms of half integers,

$$\Delta\boldsymbol{r}_{i,j,k} = \frac{1}{2}(\Delta\boldsymbol{r}_{i+1/2,j+1/2,k+1/2} + \Delta\boldsymbol{r}_{i-1/2,j-1/2,k-1/2}). \tag{2.43}$$

Fig. 2.15 summarizes the spatial discretization with notations in 2D.

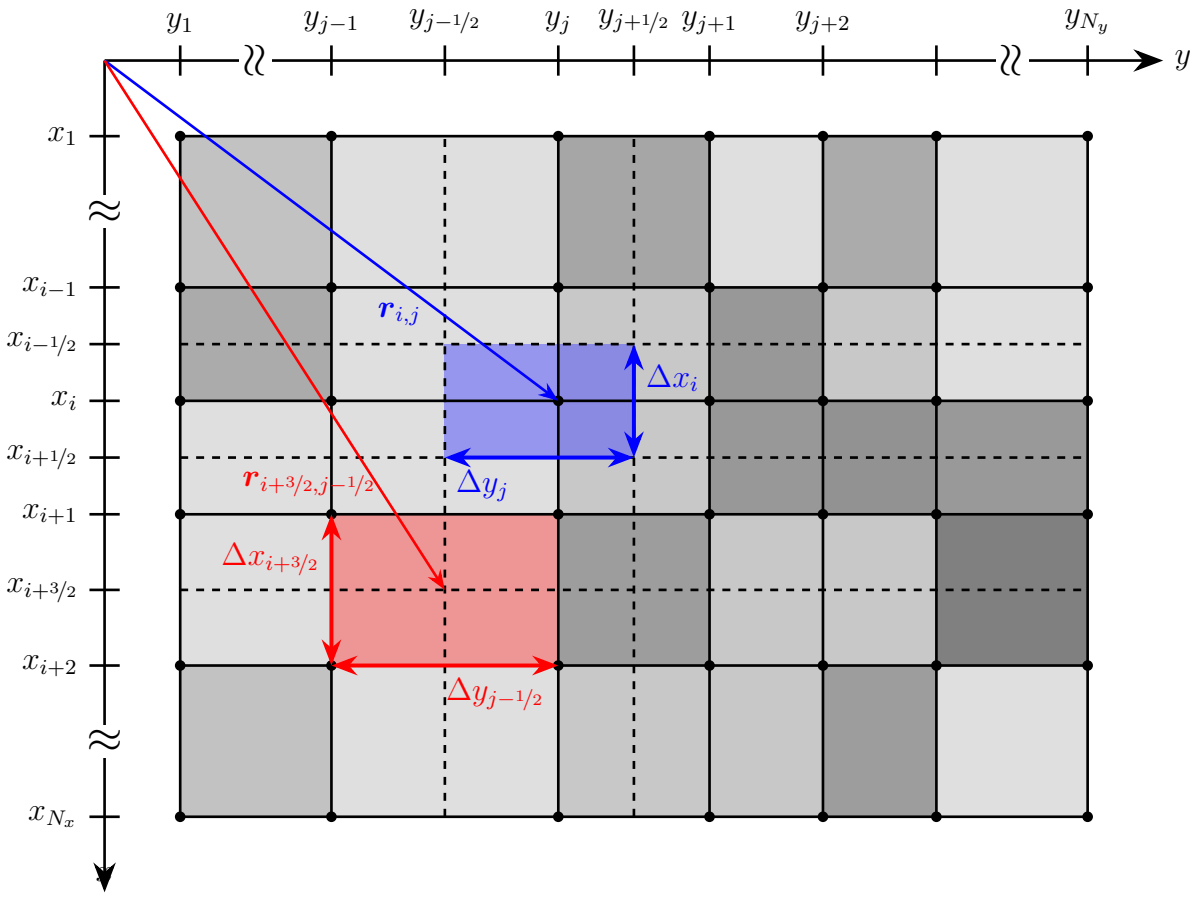

Figure 2.15: An exemplified spatial discretization scheme in 2D. An inhomogeneous material property is visualized by the shading of grid cells.

2.9.2 Discretization of Poisson-type equations

With the introduction of points, each PDE is broken into one algebraic equation at each point. In the framework of the FVM, physical quantities are assumed to be constant

within each finite volume. Using the simple structure of tensor grids, three indices can be mapped into a single index

$$d = i + (j-1)N_x + (k-1)N_xN_y. \tag{2.44}$$

Therefore, any quantity defined by three indices can also be expressed in a 1D vector form. The discretization of Eq. 2.13 with a standard FVM yields a set of equations in the form

$$\mathbf{M}\boldsymbol{X} = \boldsymbol{F}, \tag{2.45}$$

where $\mathbf{M}$ is associated with the corresponding material property, $\boldsymbol{F}$ is the source enclosed in each finite volume, and $\boldsymbol{X}$ is the physical quantity at each point to be solved.

Since Eq. (2.45) is discretized from a partial differential equation, boundary conditions are required. Specifically, Dirichlet boundary conditions (BCs) are used at contact interfaces of each metallic electrode, i. e., $i = 1$ and $i = N_x$. The electrostatic potential shift due to the metal-oxide work function difference, i. e., the Schottky barrier, is neglected. This leads to the potential of the points contacting metallic electrodes

$$\varphi_{ct} = V_M, \tag{2.46}$$

where V_M is the external voltage of the electrode. However, a BS-type VCM cell can be composed of a high-work-function electrode (e. g., Pt), or a lower work-function electrode (e. g., TiN) as the oxygen-inert electrode[12]. The contact of such metal to the oxide layer is expected to give rise to the Schottky barrier. However, the Schottky barrier is not taken into account herein as typically adopted in simulation works [135, 164][13]. For the Fourier heat equation, Dirichlet BCs at the contact points yield

$$T_{ct} = T_{\text{room}}, \tag{2.47}$$

where $T_{\text{room}} = 300\,\text{K}$ is the room temperature. On the remaining surfaces, homogeneous Neumann BCs with zero flux are used for both equations.

Though each governing equation can be solved separately, they are coupled to each other. Therefore, the electron occupation probability obtained by solving the master

[12]Work functions of Hf, TiN, and Pt are about 3.8, 4.6, and 5.6 in the unit of eV, respectively. [177, 194]
[13]The contact potential is not provided [20, 131, 136].

equation must also satisfy the other governing equations. This requirement applies to the potential and the temperature as well. That is to say, the self-consistent solution to an equation set composed of Eqs. (2.18), (2.19), and (2.2) is sought. In principle, the equation set is solved by the Newton-Raphson method. However, convergence is not guaranteed especially when the initial guess is far away from the solution. This problem could be generally alleviated by solving each governing equation iteratively in the first few steps. That is, the solution obtained from one governing equation is immediately adopted to solve another equation. After certain criteria are met, the complete equation set composed of all governing equations will be solved by the full Newton-Raphson method.

2.9.3 Time steps

For the evolution towards the next time step, a forward Euler scheme is applied. Specifically, physical quantities remain invariant in a short period of time [195]

$$\Delta t = -\frac{\ln(r_1)}{k_{\text{tot}}}, \tag{2.48}$$

where $0 \leq r_1 \leq 1$ is a uniformly sampled random number, and $k_{\text{tot}} = \sum_i k_i$ is the total rate accounting for the rate of each possible event k_i, including the vacancy generation, recombination, and diffusion. Note that in comparison to Eq. (2.39), λ is replaced by the total rate instead of being the rate of a single event [193, 196, 197]. Also note that vacancy generation rates, recombination rates, and diffusion rates are unchanged due to the stationary potential and temperature within this time interval. This assumption naturally requires Δt to be small. A large Δt arises typically when the external voltage is weak. In this case, the temperature of the oxide layer is low and field acceleration effects are negligible. Consequently, reactions are unlikely to occur and thus Eq. (2.48) gives an unfeasibly large time step. A maximum time step t_{max} is defined to limit Δt. If $\Delta t > t_{\text{max}}$, the $\Delta t = t_{\text{max}}$ is chosen.

With the simulation time step determined, the applied voltage is updated at the given sweep rate. Care has to be taken when the current is limited by the current compliance during a FORMING or a SET process. During either process, the voltage across the VCM cell V_{cell} could be smaller than the applied voltage V_{app} to limit the current. Therefore, extra trial loops are required to find V_{cell} when the resistance is low.

Analogous to the procedure explained in Sec. 2.8.2, another uniformly sampled random number $0 \leq r_2 \leq 1$ is generated for the temporal evolution of vacancies. Mathematically, the inequality

$$\frac{1}{k_{\text{tot}}} \sum_{i=1}^{n-1} k_i < r_2 \leq \frac{1}{k_{\text{tot}}} \sum_{i=1}^{n} k_i \tag{2.49}$$

is solved for n [192, 198]. The vacancy configuration is updated based on the chosen event provided that $\Delta t \leq t_{\text{max}}$. Otherwise, the temporal evolution of vacancies is discarded. The procedure of choosing the event is schematically illustrated in Fig. 2.16.

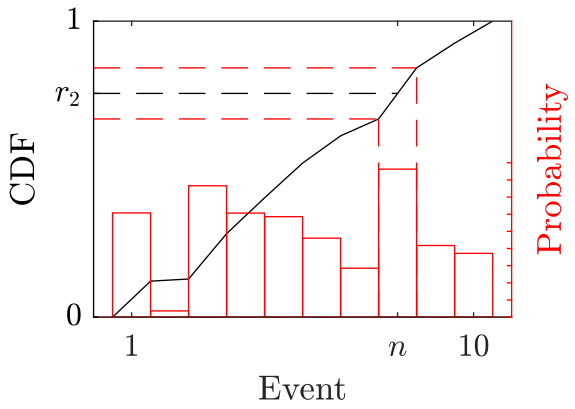

Figure 2.16: Scheme for choosing the nth event to occur.

Important steps are summarized in Fig. 2.17.

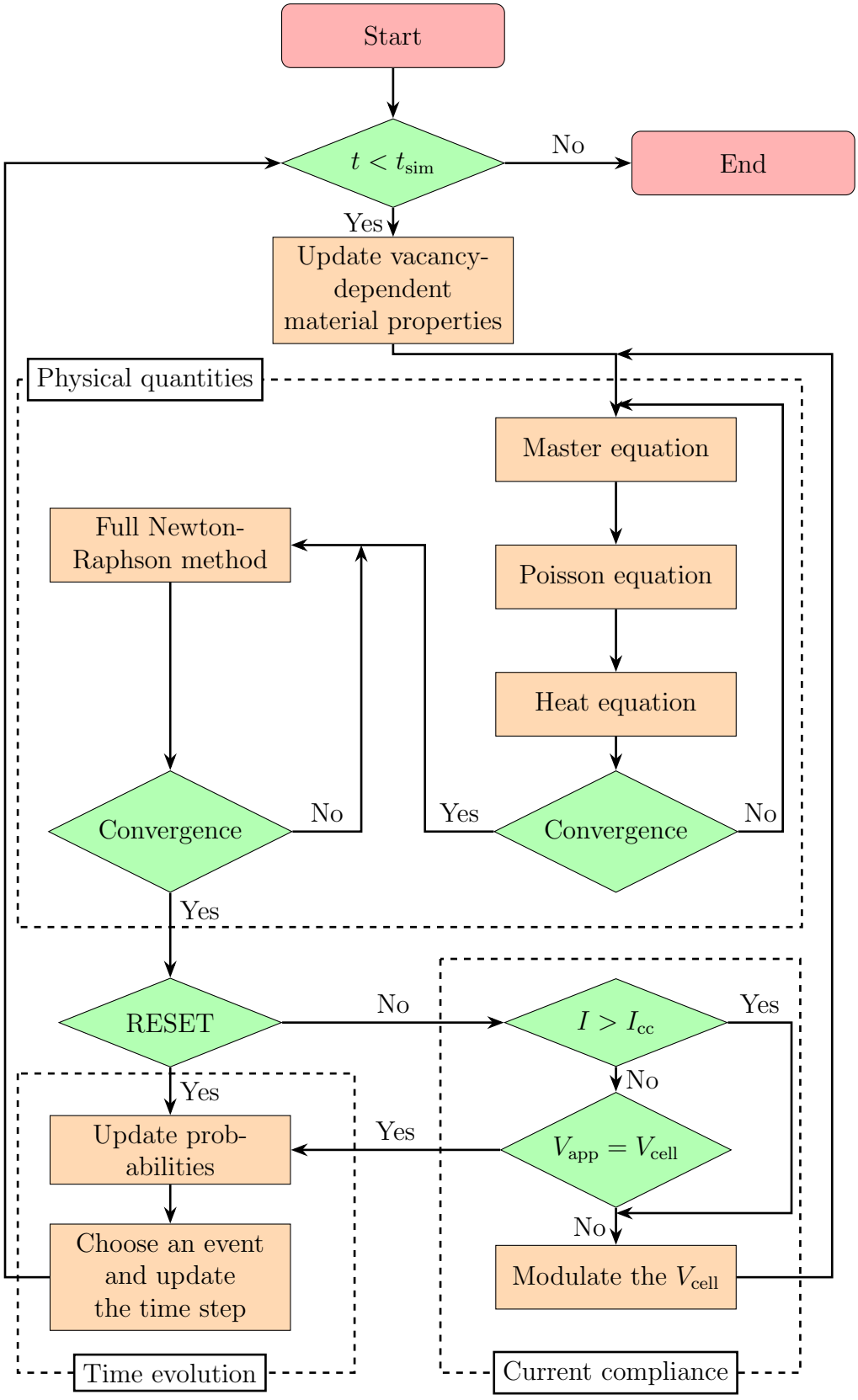

Figure 2.17: Flowchart for the KMC simulator.

3 Three-dimensional device simulation

This chapter presents simulation results of a three-dimensional HfO_2-based filamentary type VCM cell, focusing on the spatio-temporal evolution of oxygen vacancies in the oxide layer. Since the spatial distribution of oxygen vacancies is difficult to capture in real-time using contemporary microscopic imaging technologies, the objective is to explore the relationship between the evolution of vacancies and the observed phenomena from a modeling perspective. Unlike existing works, which primarily discuss the evolution of vacancies in terms of vacancy density, this work focuses on the point-defect nature. Within the small current compliance regime, the spatial variation of vacancies could lead to relatively large changes in electrical characteristics, making the stochastic nature of the migration process important. Since variations are typically smoothed in a continuous model, the point-defect point of view is more suitable for simulating devices with small current compliance to capture the impact of a limited number of vacancies.

To begin with, the impacts of different factors are discussed from Sec. 3.2 to 3.4. Specifically, the impact of vacancies in different charge states during the charge transport process is investigated in Sec. 3.2. As a CF emerges, thermal properties around it start deviating from those of the host oxide. The modeling of an inhomogeneous thermal conductivity and the temperature distribution are discussed in Sec. 3.3. Furthermore, anisotropic diffusion barriers have been shown by DFT, though the discussion on the dynamical process was still missing. The impact of anisotropic diffusion barriers is detailed in Sec. 3.4. Based on the aforementioned findings, a possible scenario involving a single CF is proposed to explain the observed large C2C variability, and simulations from the FORMING process are presented in Sec. 3.5. However, the scheme only works for the specific current compliance of 2 µA. The reason for this limitation is analyzed at the end of the section. The extension to larger current compliance is discussed in Sec. 3.7. It is investigated under a proposed electron hopping scheme within the framework of MA hopping, detailed in Sec. 3.6. The C2C variability is extended to a larger current compliance regime in Sec. 3.8.

3.1 Simulation setup

The 3D simulation focuses on the dynamics of oxygen vacancies and other physical quantities within the oxide layer. The electrodes are treated as contact surfaces, neglecting microscopic changes within the electrode materials. This simplification may be relevant for phenomena such as oxygen exchange at the oxide-electrode interface, particularly at the interface with the bottom electrode (BE). It is worth noting that the oxygen exchange at the bottom interface involves using easily oxidizable metals, such as Ti, Hf, or Zr, for the BE. Although using a metal other than Hf might introduce a heterogeneous bilayer structure, this is not considered by this KMC simulator. Conversely, the top electrode (TE) is typically composed of Pt or TiN[1], which are less oxidizable. With these materials, the VCM cell is expected to demonstrate a BS-type I-V characteristics. The simulation setup of the simulated device is illustrated in Fig. 3.1. Note that the arrangement of electrodes in the simulation might be opposite to those in measurements.

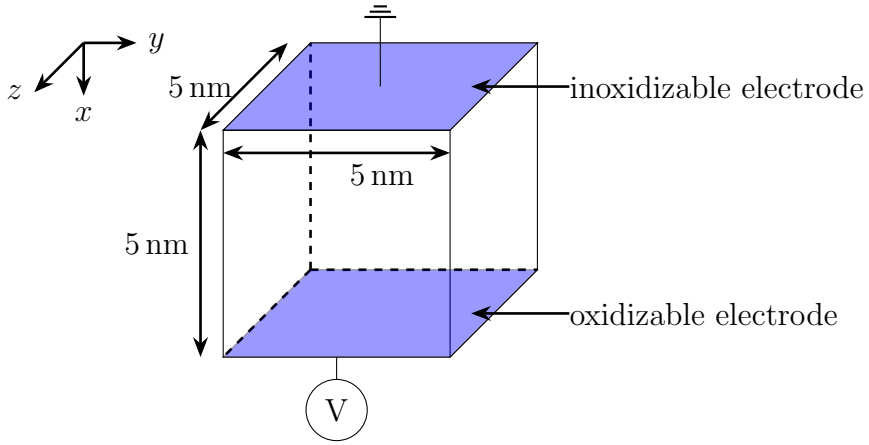

Figure 3.1: Simulated 3D HfO_2-based cell with sizes labeled in each dimensionality.

The external voltage is always applied to the bottom electrode, while the top electrode is always grounded during all operations, namely, FORMING, SET, and RESET processes. This arrangement might differ from some measurements, where the voltage is applied to either the top or bottom electrode depending on the operation. In addition, the current compliance is not depicted as it is not part of the oxide layer. However, the current compliance sets an upper limit for current through the VCM cell during the SET

[1]The TiN could also be the oxidizable electrode when the other electrode is composed of Pt. However, FORMING voltages of this set of electrodes are typically much higher.

and FORMING processes achieved by reducing the voltage across the cell $V_{\text{cell}} < V_{\text{app}}$.

In this work, an ideal current compliance is assumed corresponding to the lack of current overshoot due to an instantaneous modulation of V_{cell}. Practically, this is achieved by attaching a VCM cell to a transistor. It is noted that another approach exists where an external resistor is connected in series. However, this approach is not considered in simulations.

3.2 Impact of multiple charge states

3.2.1 Electron transport process

Although the macroscopic Poisson equation has been employed in device simulations for decades, the downscaling of modern electronic devices raises questions regarding its validity. One example is the attempt to resolve the potential at nanometer-scale resolution. At this scale, even an elementary charge inside a 1-nm cube results in a large charge density, leading to a peak in the potential. This issue arises from a classical perspective on charge density. When quantum mechanical effects are considered, the spatial variation of the electric potential becomes gradual [199]. However, the consideration of quantum mechanical effects increases complexity and thus extends the simulation time beyond practical usages. Therefore, several methods have been proposed to leverage the high efficiency of conventional formulations. For example, the charge density is spread out [199] or averaged [200] for the Poisson equation.

In this work, the introduction of the neutral charge state effectively smooths out the peaks, leading to simulation results comparable to measurements. Fig. 3.2a compares the current with and without the neutral charge state. Vacancies are fixed in space to exclude complicated dynamical processes and the temperature is fixed at 300 K for simplicity. In the presence of an evident CF structure, the current is expected to be high during the voltage sweep. This is seen only from the simulation that includes the neutral charge state, while the current is much smaller without the neutral charge state. To further discuss, potential distributions along the CF are plotted. In Fig. 3.2b, the gradient of the potential is seen to be monotonic, and thus the direction of the electric field is unchanged along the electron transport path. In contrast, a peak in the potential at the middle of the transport path is seen in Fig. 3.2c. The electric field produced by charged vacancies opposes the external voltage at the anode interface. Conceptually, the (electron) current is driven by the electric field. Therefore, the electric field at the anode interface effectively blocks the electron flow out of the oxide layer. As a result, the current is much smaller and deviates from well-established understandings.

The inclusion of the neutral charge state gives a more physically intuitive interaction between charged vacancies, the electrostatic potential, and electron transport. This approach differs from the scheme of some existing works, where the energy level within the oxide layer is modulated from the electrostatic potential obtained by solving the

Laplace equation. Herein, the Poisson equation is solved for the potential, which shifts the energy level. The analysis, where all vacancies are fixed in space and the intrinsic charge state is either neural or doubly positive, provides a starting point. In the following section, a model of the charge states is developed to account for intricate distributions during operations.

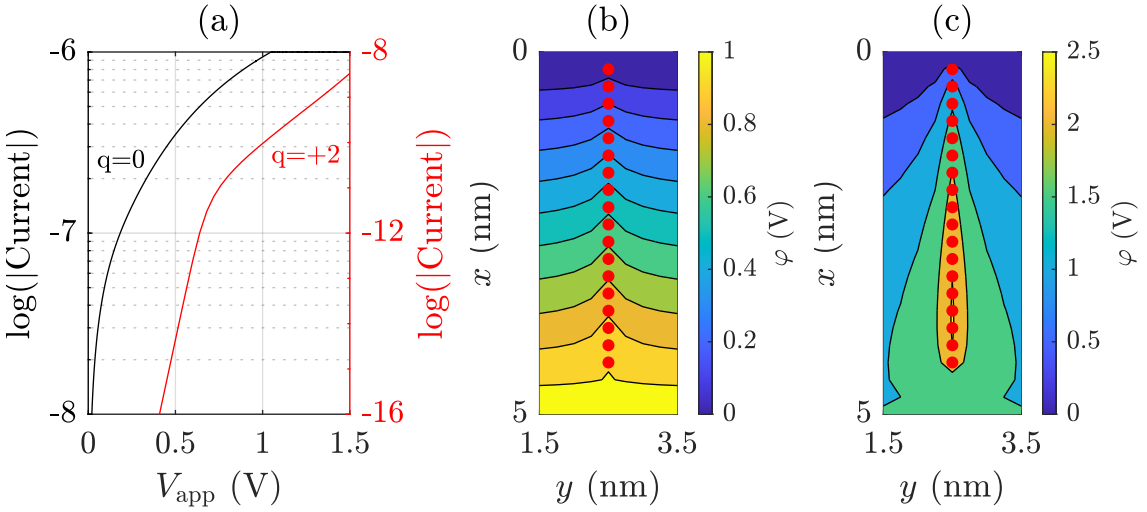

Figure 3.2: (a) Simulated I-V characteristic curves for all vacancies in the neutral ($q = 0$) or doubly positive ($q = +2$) charge states. Electrostatic potential distributions at $V_{app} = 1.5\,\mathrm{V}$ for all vacancies in (b) the neutral charge state and (c) the doubly positive charge state. Vacancies are shown in red solid circles.

3.2.2 Modeling the charge states

In Ref. [160], the intrinsically neutral charge state is stable with a Fermi energy above approximately 3.5 eV of its VBM. Within a Fermi energy range of approximately 2.5 eV to 3.5 eV, an energetically preferred charge state depends on the local structure (see Fig. 3.3). For Fermi energies within this range, a perfect 1D chain of vacancies referred to as the V_O-chain favors the neutral charge state over positive charge states from the formation energy perspective (see Fig. 3.3). The analysis of its band structure shows a significant dispersion along the direction of the vacancy chain (see Fig. 3 of Ref. [160]), giving rise to a high electrical conductivity.

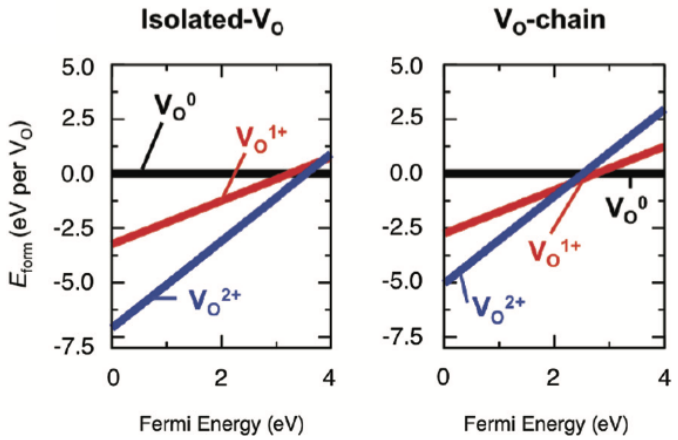

Figure 3.3: Formation energy diagram per one V_O for the isolated-V_O and V_O-chain models as a function of Fermi energy. Note that the Fermi energy refers to the VBM. Reprinted Fig. 2 with permission from [160]. Copyright (2013) by the American Physical Society.

The Fermi energy of a HfO_2-based device can be estimated given that the band gap from experiments[2] and electron affinity are 5.7 eV [201] and (2.00 ± 0.25) eV [202], respectively. An evaluation yields a Fermi level of (3.70 ± 0.25) eV referring to the VBM. Considering a Fermi level might fall below 3.5 eV due to the variation, the energetically preferred charge state is modeled to depend on the existence of an V_O-chain. In this work, a vacancy chain is modeled by the connection of adjacent vacancies only in the x-direction, i. e., the direction connecting the two electrodes. This is assumed to investigate the resistive switching phenomena due to CFs growing in this direction.

In Ref. [160], the V_O-chain possesses three consecutive vacancies. However, it is not clear whether the two connected vacancies would modulate the formation energy. Herein, the minimum length of two vacancies is assumed to be sufficient for the modulation. This aims to avoid a local electrostatic potential peak due to the doubly positive charge state. In addition, it is assumed that an V_O-chain has an impact on its nearest neighboring sites. Therefore, an V_O adjacent to an existing V_O-chain in the y- or z- direction is assumed to prefer the neutral charge state as well. Note that an existing V_O-chain is a prerequisite condition. The modeling of intrinsic charge states is illustrated in Fig. 3.4. Within this scheme, the issue of the charge transport process in the presence of an electrostatic potential peak is alleviated.

[2]DFT for different phases of HfO_2 yields values within the range of 5.6 eV to 6.4 eV [121, 123]

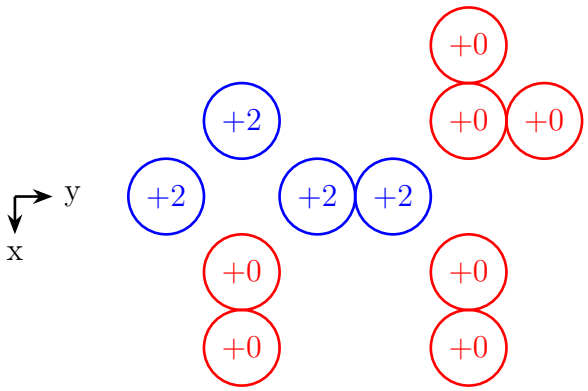

Figure 3.4: A summary of modeling intrinsic charge states.

3.3 Impact of the thermal effects

3.3.1 Modeling the thermal conductivity

In Ref. [144], the thermal conductivity is measured on amorphous HfO_2 (a-HfO_2) for different thicknesses at various ambient temperatures. For a 5.6-nm-thick HfO_2 sample, a thermal conductivity ranges from $0.49\,\mathrm{Wm^{-1}K^{-1}}$ to approximately $0.6\,\mathrm{Wm^{-1}K^{-1}}$ in the range of 300 K to 500 K. In comparison, values of $0.27\,\mathrm{Wm^{-1}K^{-1}}$ to $0.49\,\mathrm{Wm^{-1}K^{-1}}$ are obtained for a 3-nm-thick sample by the scanning thermal microscopy measurement [203]. It is noteworthy that the devices under measurement are not subjected to an applied voltage. Under this condition, the creation of oxygen vacancies is not expected. Therefore, the thermal conductivity is close to that of an ideally perfect oxide, which differs from the thermal conductivity of its metallic form, namely, the oxide with low oxygen content. Specifically, the thermal conductivity of Hf is in a range of $21\,\mathrm{Wm^{-1}K^{-1}}$ to $23\,\mathrm{Wm^{-1}K^{-1}}$ in the range of 300 K to 500 K [204].

Larentis *et al.* linearly interpolated the thermal conductivity as a function of oxygen vacancy density [114]. The values of $k_{\mathrm{th,HfO_2}} = 0.5\,\mathrm{Wm^{-1}K^{-1}}$ and $k_{\mathrm{th,Hf}} = 23\,\mathrm{Wm^{-1}K^{-1}}$ are adopted for the densities of 0 and $1.2 \cdot 10^{21}\,\mathrm{cm^{-3}}$, respectively. In this work, the thermal conductivity is formulated as a parabolic function of vacancy density n given by

$$k_{\mathrm{th}} = \begin{cases} k_0 & n \leq n_0 \\ k_0 + k_1(n - n_0) + k_2(n - n_0)^2 & n > n_0 \end{cases}, \tag{3.1}$$

where vacancies within a 1-nm cube are counted for density and $n_0 = 10^{-21}\,\mathrm{cm^{-3}}$ is chosen. Note that the chosen n_0 corresponds to one vacancy within a cube and the thermal conductivity is assumed to be that of an oxide. Moreover, using the k_1 from Larentis's work together with $k_2 = 0$ gives $k_{\mathrm{th}} \approx 0.83 k_{\mathrm{th,Hf}}$ for two vacancies within a cube. This increment is considered too rapid with respect to n. Therefore, a smaller k_1 together with a quadratic modification term is assumed in replacement of a linear interpolation. Lastly, the thermal conductivity is bounded below the $k_{\mathrm{th,Hf}}$. The coefficient of k_0 is fixed at $k_{\mathrm{th,HfO_2}} = 0.5\,\mathrm{Wm^{-1}K^{-1}}$ and the k_1 and k_2 are positive parameters in the following sections.

3.3.2 Modeling the energy dissipation

Similar to the formulation of the TAT mechanism, phonons are absent during the energy transport process. This implies that changes in the electron energy are assumed to occur only at the initial or the final vacancy sites (see Fig. 3.5). In addition, the energy change of electrons is not quantized but arbitrary.

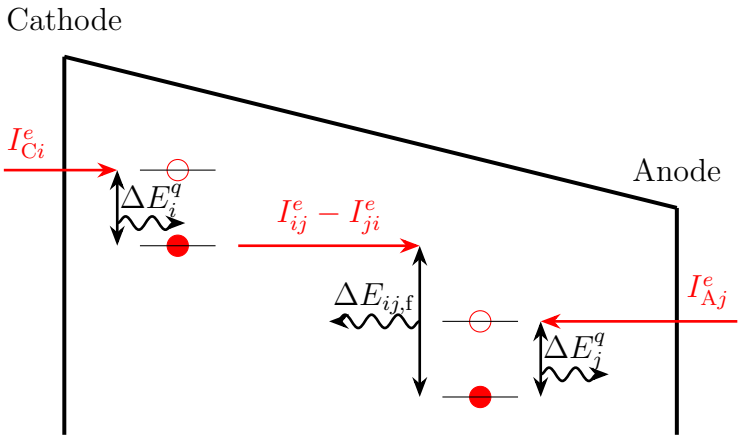

Figure 3.5: Heat dissipation scheme during charge transport process.

Based on these assumptions, the heat generation rate density associated with electron energy changes is divided into two parts: g_i^{tun} and g_{ij}^{hop}. The heat generation rate density at the ith vacancy site due to the jth vacancy in the hopping process reads

$$g_{ij}^{\mathrm{hop}} = \begin{cases} \frac{1}{e\Omega_i}(I^e_{ji} - I^e_{ij})(E^q_{j,\mathrm{filled}} - E^q_{i,\mathrm{filled}}) & E^q_{j,\mathrm{filled}} > E^q_{i,\mathrm{filled}} \\ 0 & \text{otherwise} \end{cases}, \tag{3.2}$$

where I^e_{ij} and I^e_{ji} are the electron current from the ith vacancy to the jth vacancy and vice versa. Ω_i is the finite volume size of the ith grid, within which the energy is assumed to spread homogeneously. Note that Eq. (3.2) includes the energy dissipation due to electrons from the higher-energy site $I^e_{ji}(E^q_{j,\mathrm{filled}} - E^q_{i,\mathrm{filled}})/e$ and the energy absorption due to electrons from the lower-energy site $-I^e_{ij}(E^q_{j,\mathrm{filled}} - E^q_{i,\mathrm{filled}})/e$. The total heat generation rate density at the ith vacancy site from a hopping process is then given by

$$g_i^{\mathrm{hop}} = \sum_{j \neq i} g_{ij}^{\mathrm{hop}}, \tag{3.3}$$

where all other vacancies are taken into account.

On the other hand, an injected electron always dissipates its energy at the ith vacancy site in a tunneling process. This is based on the assumption that the electron with an energy higher than the empty state energy level of the vacancy site can tunnel. Additionally, an injected electron eventually occupied the filled state, where the energy level is lower than that of the empty state as illustrated in Fig. 3.5. Since electrons are assumed to be injected with an identical tunneling probability, they are interpreted as being equal energy. This simplifies the heat generation rate at the ith vacancy site, which reads

$$\Omega_i g_i^{\mathrm{tun}} = \frac{1}{e}(I_{\mathrm{C}i}^e + I_{\mathrm{A}i}^e)\Delta E_i^q, \tag{3.4}$$

where $\Delta E_i^q = E_{i,\mathrm{empty}}^q - E_{i,\mathrm{filled}}^q > 0$, and $I_{\mathrm{C}i}^e$ and $I_{\mathrm{A}i}^e$ refer to the electron current from the cathode and anode, respectively. Unlike the hopping process where energy conservation is preserved, there is no conservation law in Eq. (3.4). This is due to the truncation of energy changes at electrodes in correspondence to the assumption of a constant temperature, i. e., the Dirichlet boundary condition.

3.3.3 Temperature distribution

To explore the role of temperature during vacancy migration, let us first consider two representative vacancy distributions during the RESET operation. That is, the rupture of a vacancy chain occurs at different positions as shown in Fig. 3.6. Here, a negative voltage is applied to the bottom electrode. The right panel depicts the outcome of a successful migration from the initial distribution shown in the left panel. Note that the migration does not necessarily take place in a dynamical process. While the spatio-temporal evolution of vacancies will be further discussed in the next section, the discussion is restricted to static vacancy distribution here.

In the left panel of Fig. 3.6a, a high temperature around the rupture region is observed, which aligns with findings in Ref. [114]. This is expected due to the significant electrostatic potential drop between two segments (see the left panel of Fig. 3.6b), leading to large power dissipation. Note that high temperature arises from a large potential change and a non-negligible current. This scenario is typical in an early RESET stage when the resistance remains low after a successful SET operation.

In the case of successful downward vacancy migrations, the rupture moves upward. This leads to the high-temperature region moving into the middle of a vacancy chain,

as shown in the right panel of Fig. 3.6a. However, vacancies in the chain can not move downwards further. Alternatively, vacancies out-flow the chain as shown in Fig. 3.7a even though it is not energetically preferred. An escaped vacancy undergoes a transition to a doubly positive charge state, and thus the mobility increases. In the bottom panel of Fig. 3.7a, the presence of the isolated vacancy distorts the local electrostatic potential. Specifically, the potential around it is raised due to the doubly positive charge state. Similar migration continues provided that temperature is still sufficiently high. This is seen in the upper panel of Fig. 3.7b, where the highest temperature is about 500 K. Eventually, the initially perfect 1D vacancy chain broadens, as shown in Fig. 3.7c.

To assess the temperature being sufficiently high, the temperature dependency of the Arrhenius equation is plotted in Fig. 3.8. For the rate of 10^3 Hz, estimated temperatures are 300 K, 400 K and 600 K for activation energies of 0.7 eV, 0.9 eV and 1.3 eV, respectively. It is noted that these are rough estimations, particularly for migrations in the x-direction since the electric field is excluded.

A shortening process of a 1D vacancy chain is provided from a point defect point of view. The discussion reveals an important role of the temperature in understanding the C2C variability. However, the scenario of a recovery in the length of a vacancy chain is still missing. Without recovery, resistive switching can not be established since resistance increases monotonically. This will be discussed in the following sections, where the anisotropic diffusion barrier is introduced.

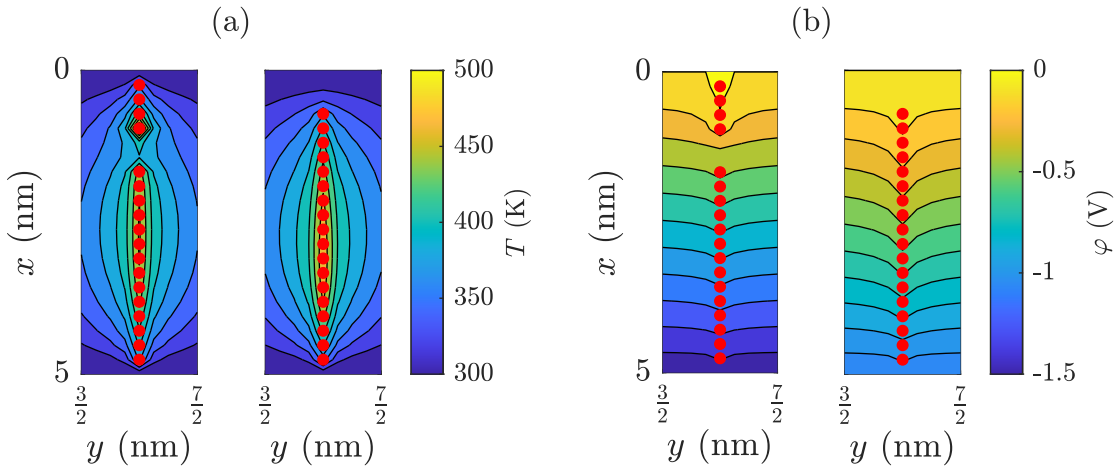

Figure 3.6: (a) Temperature distributions and (b) potential distributions in two exemplified vacancy chains.

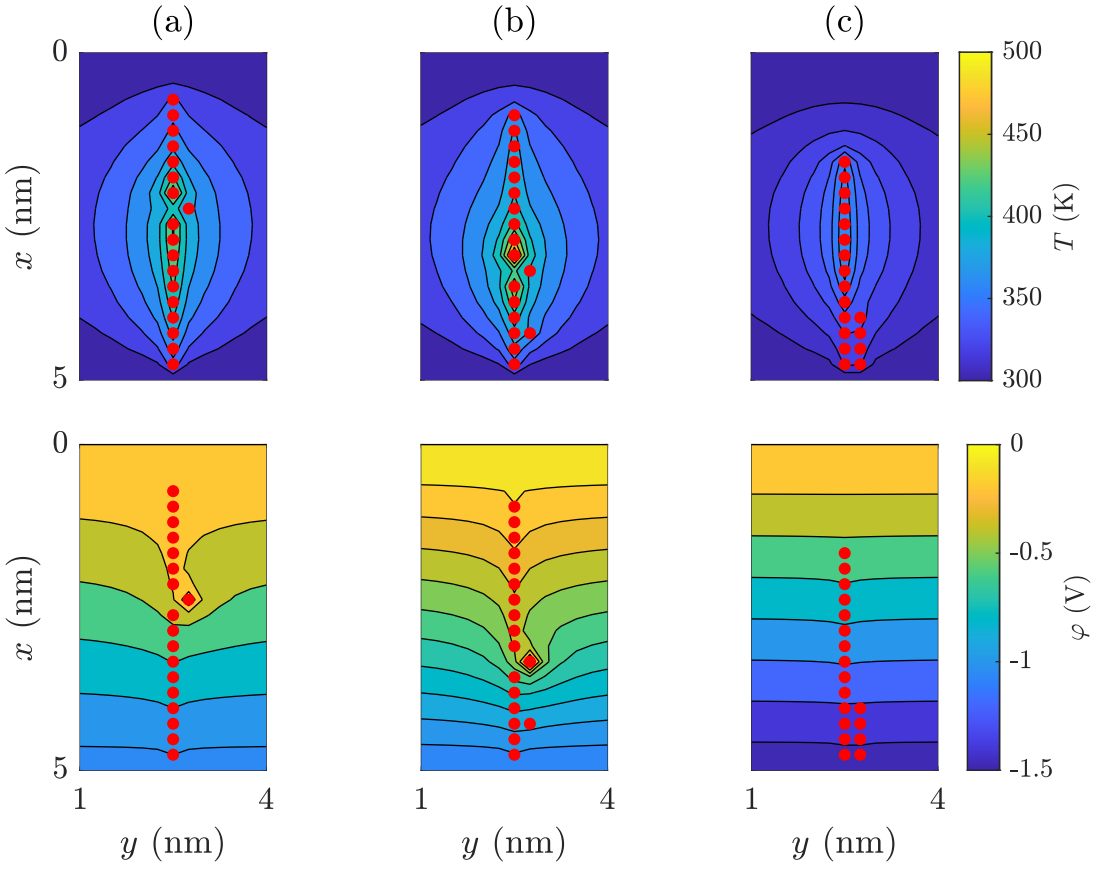

Figure 3.7: Temperature and potential distributions in an ordered time series.

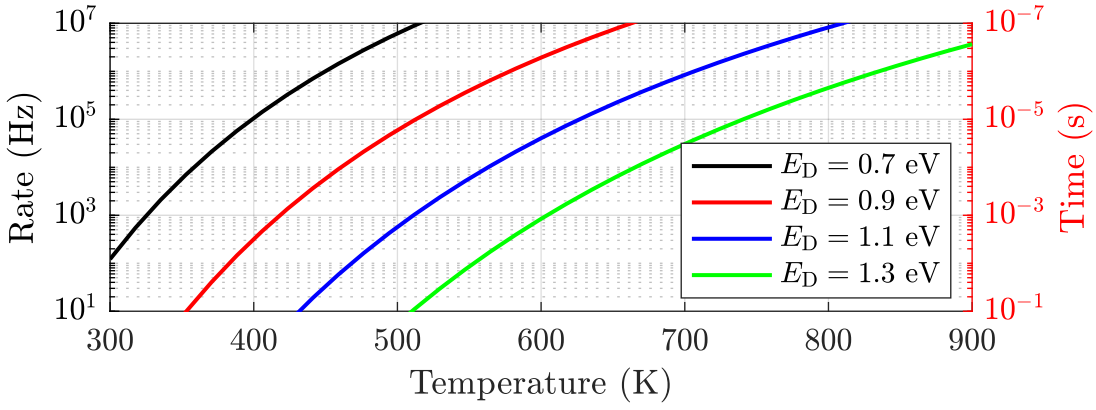

Figure 3.8: Temperature-dependent rates given by the Arrhenius equation with the hopping frequency of 70 THz for different zero-field activation energies.

3.4 Impact of the anisotropic diffusion

DFT calculations suggest that either GBs or vacancy chains lead to anisotropic vacancy migration. However, the combined effect of a vacancy chain within a grain boundary is not fully understood. This section addresses this gap by first modeling effects from each local structure depicted in Sec. 2.7.1. Within each type of local structure, the zero-field activation energy in Eq. (2.27) takes the form

$$E_{\mathrm{D}} = E^{q}_{\mathrm{D,iso}} + \Delta E^{dir}_{\mathrm{D},str}, \tag{3.5}$$

where $E^{q}_{\mathrm{D,iso}}$ is the isotropic zero-field activation energy. $E^{q}_{\mathrm{D,iso}}$ is modeled to depend on the intrinsic charge state q where only the doubly positive and neutral charge states are accounted for. Anisotropic modulations $\Delta E^{dir}_{\mathrm{D},str}$ depending on the local structure as well as the direction are indicated by the subscript and superscript, respectively. Moreover, it is assumed to be independent of its intrinsic charge state. This is based on a similar shift for different charge states observed in Ref. [141]. Lastly, E_{D} is bound to be larger than $E_{\mathrm{D,min}} = 0.51\,\mathrm{eV}$, which is the zero-field activation energy of a vacancy deeply inside a vacancy chain.

3.4.1 Vacancy chain effect

In Ref. [182], anisotropic migration barriers are found for vacancies in a vacancy chain. The detachment of the second vacancy along the chain direction is easier than that of the first vacancy. A minimal activation energy of 0.51 eV is found for the third vacancy of a chain. The modeling of modulations in the chain direction is sketched in Fig. 3.9a, where at least four vacancies are required for the CF effect. This minimal number of vacancies is assumed based on the symmetry. Specifically, the modulation for the second $\mathrm{V_O}$ moving downward is identical to that for the third $\mathrm{V_O}$ moving upward provided there are at least two vacancies within the lower segment. This implies that the upper segment plays an equal role to the lower segment, as illustrated by the rightmost vacancy chain in Fig. 3.9a.

On the other hand, outward migrations in y- and z- directions that break a perfect chain are suppressed. In this case, a length of at least three $\mathrm{V_O}$s in a row is required to increase the activation energy. Conversely, the inward migration is promoted as sketched in Fig. 3.9b. Relevant anisotropic modulations of the activation energy are

listed in Table 3.1.

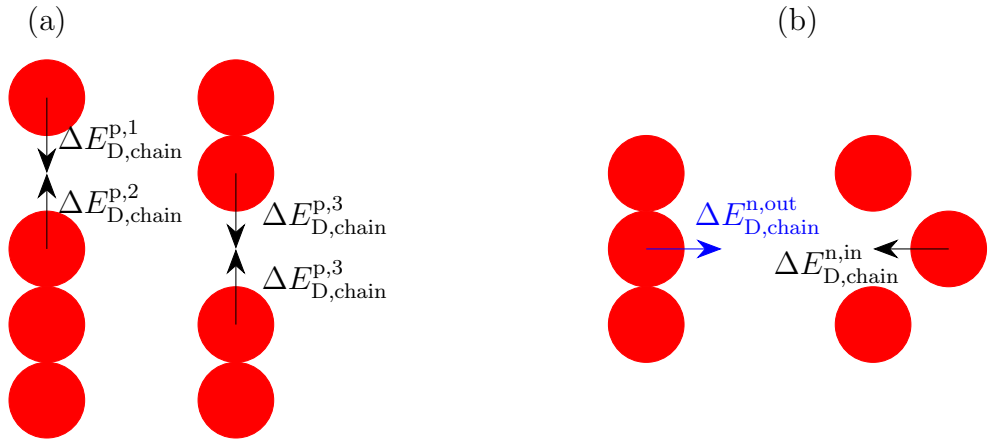

Figure 3.9: Modeling the vacancy chain effect for anisotropic migration barriers in (a) the parallel direction and (b) the normal direction. The modulations in black color refer to a promotion in migrating along the indicated directions while the blue color refers to a suppression in migrating along the indicated direction.

3.4.2 Grain boundary effect

Modulations due to GBs are similar to those due to vacancy chains. Namely, the migration within a GB is promoted while the out-flow is suppressed. However, DFT investigation does not suggest an asymptotic modulation in the activation energy in neither the x nor z- direction [179], where the GB extends on a xz-plane. The invariant energy shift along both directions might arise from the considered vacancy migrating deeply inside the GB. It is unclear whether the vacancy close to the boundary of a GB would still exhibit the same anisotropy.

In this work, a GB region is modeled as a collection of 1D regions. Within each 1D region, anisotropic properties are identical. However, these anisotropic properties may differ between individual 1D regions.

3.4.3 Combination of both effects

The GBs, including their positions and associated anisotropic properties, are assumed to be unchanged over time. Due to their attraction to vacancies, vacancies tend to accu-

Symbol	Value	Symbol	Value
$\Delta E^{\mathrm{n,in}}_{\mathrm{D,chain}}$	-0.2	$\Delta E^{\mathrm{in}}_{\mathrm{D,GB}}$	-0.2
$\Delta E^{\mathrm{n,out}}_{\mathrm{D,chain}}$	0.2	$\Delta E^{\mathrm{out}}_{\mathrm{D,GB}}$	0.2
$\Delta E^{p,1}_{\mathrm{D,chain}}$	-0.3	$E^{+2}_{\mathrm{D,iso}}$	0.7
$\Delta E^{p,2}_{\mathrm{D,chain}}$	-0.43	$E^{0}_{\mathrm{D,iso}}$	1.1
$\Delta E^{p,3}_{\mathrm{D,chain}}$	-0.59	$E_{\mathrm{D,min}}$	0.51

Table 3.1: Shifts of the zero-field activation energy in the unit of eV.

mulate within GBs. Moreover, the generation of vacancies is promoted at the interface between GBs and the BE. These factors lead to the growth of CFs within GBs, and thus the CF and GB effects are not exclusive.

To avoid double-counting effects of CFs and GBs, modulations due to each local structure are first calculated separately. The effective modulation is assumed to be the larger one of these two values. In addition, the modulation of an $\mathrm{V_O}$ from one chain to another in the y- or z- diction is also unclear. When a vacancy migrates from its original chain, the migration is suppressed from the original chain perspective. In contrast, migration is promoted from the neighboring chain perspective. Moreover, the above argument is not restricted to a vacancy chain but also applies to a GB. To address this complexity, we simplify the scenario by assuming a superposition of the modulations due to the corresponding local structures.

3.4.4 Homogeneous anisotropic modulations

To begin investigations of vacancy dynamics during operations, the simple geometry of a GB is considered. Specifically, there is only one single GB composed of five 1D regions, referred to as channels. Each channel extends from the BE to the TE and possesses an identical modulation to diffusion barriers. However, an $\mathrm{V_O}$ can leave or enter a GB in the y- or z- directions. In either direction, the modulation is obtained by superimposing the modulations of the initial and final regions discussed in Sec. 3.4.3. Fig. 3.10 shows the anisotropic modulations on the plane where the GB is located. Specifically, the finite volumes are plotted as rectangles, and vacancies are defined at the center of finite volumes. The figure illustrates the anisotropic modulation for a vacancy occupying the finite volume. Fig. 3.10a shows the modulation in the $\pm x$-direction, and Fig. 3.10b and 3.10c show the modulations for migrations in the $+y$ and $-y$ directions, respectively.

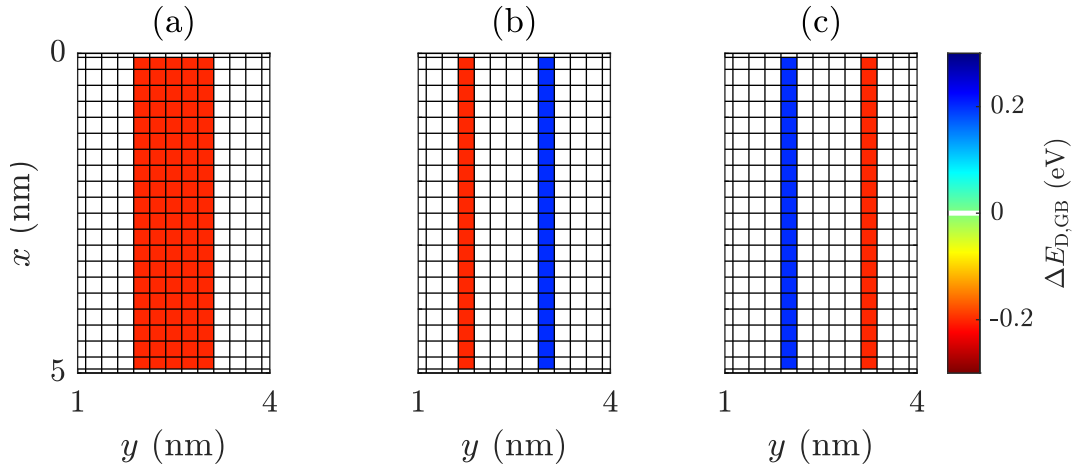

Figure 3.10: Anisotropic modulations for migrations in (a) the $\pm x$-directions, (b) the $+y$-direction and (c) the $-y$-direction.

To ensure the following discussion aligns with practical conditions, simulations start with a FORMING stage. Discussed vacancy distributions will be associated with FORMING conditions rather than being convenient ones for a model. Maximum applied voltages and sweep rates are summarized in Table 3.2.

Symbol	Value	Symbol	Value
V_{FORM}	2.5 V	V_{RESET}	-1.5 V
V_{SET}	1.5 V	$\lvert\beta\rvert$	$0.5\,\mathrm{V}\cdot\mathrm{s}^{-1}$

Table 3.2: The sweeping rates β and the maximal values of applied voltages in the FORMING, SET, and RESET operations.

Fig. 3.11a shows the vacancy distribution at the last simulation step of a FORMING process. It is observed that the middle vacancy chain has the maximum length, which can be explained by the following procedures. Firstly, the emergence of a perfect 1D chain leads to an abrupt increase in current. Accompanied by the increased current, the Joule heating becomes significant. The temperature around the vacancy chain, especially at the bottom of the chain, becomes sufficiently high to break the perfect 1D structure. This is similar to the situation discussed in Sec. 3.3.3, specifically Fig. 3.7. The migration pattern will eventually stop with the emergence of long chains, as the resistance decreases and V_{cell} is subsequently limited due to the current compliance.

Therefore, the geometry of a CF is similar to an upside-down pyramid where the tip is expected to be located at the chain with the lowest generation barrier, as shown in Fig. 3.11a. The statistical significance is demonstrated by simulating fifty realizations of a FORMING process. Specifically, the number of times that each finite volume is occupied by an V_O is plotted in Fig. 3.11b. The chain with a maximum length is seen to be one with a lower generation barrier.

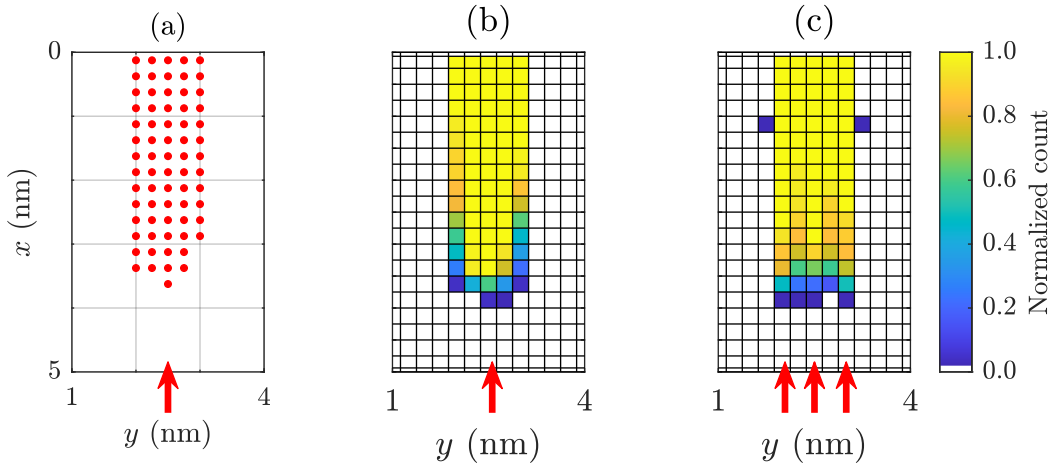

Figure 3.11: Vacancy distributions of (a) a single FORMING process. Statistical vacancy distributions out of fifty realizations of FORMING processes for (b) one fragile spot and (c) three fragile spots. The fragile spots with the lowest generation barrier are indicated by red arrows.

During subsequent operations, the variation in lengths gradually smooths out. This is illustrated by the location of high temperature during the RESET process together with the layer-by-layer migration pattern. To demonstrate this, twenty switching cycles after a FORMING process are simulated.

At the early stage of vacancy migration in a RESET process, vacancies closer to the BE migrate first. Vacancies that detach the CF are initially triggered by high temperatures. As these vacancies arrive at the BE interface, they remain in the doubly positive charge state since they are not connected to any vacancy chain. These vacancies are highly mobile, allowing them to move within the GB area on the interface plane. The longer vacancy chain then loses vacancies via the horizontal migration at the BE interface shown in Fig. 3.12a. Specifically, the migration is visualized by counting the number of times a vacancy goes through (i. e., entering and leaving) each site over a period of time.

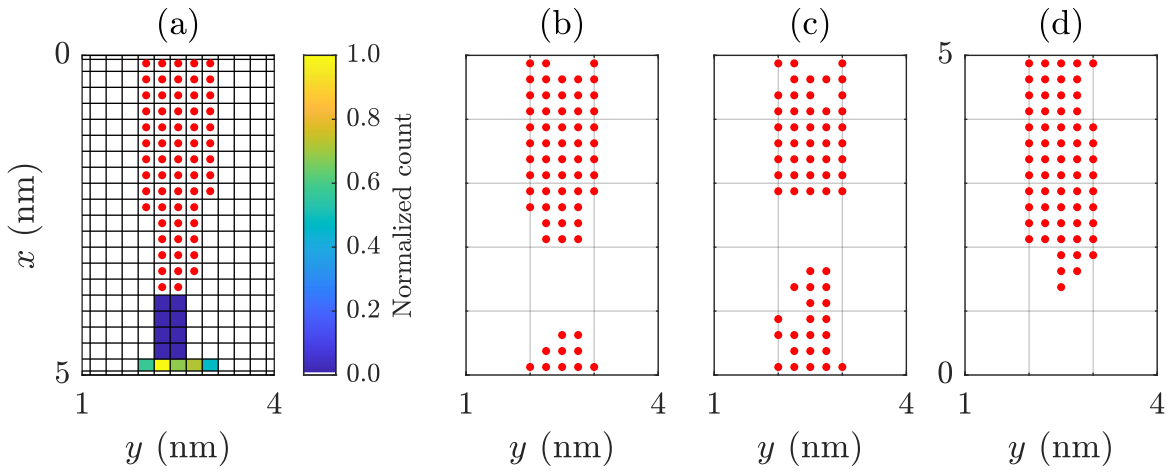

Figure 3.12: Vacancy distributions (a) at $V_{\mathrm{app}} = 0\,\mathrm{V}$, (b) at $V_{\mathrm{app}} = -1.13\,\mathrm{V}$, (c) at $V_{\mathrm{app}} = -1.22\,\mathrm{V}$, and (d) at $V_{\mathrm{app}} = 0\,\mathrm{V}$. The evolution in (a) collects migrations from $V_{\mathrm{app}} = 0\,\mathrm{V}$ to $V_{\mathrm{app}} = -1.02\,\mathrm{V}$ during the ramp-up stage of a RESET process. Note that the x-direction is turned upside down in (d).

As the magnitude of applied voltage increases, the downward migration gradually dominates. During this stage, the original CF still remains connected to the TE while the CF connecting to the BE starts growing, as shown in Fig. 3.12b and 3.12c. Interestingly, the length variation of the upper CF decreases or even disappears due to a layer-by-layer migration pattern. The length, top position, and bottom position of a gap over twenty cycles are shown in Fig. 3.13a, 3.13b, and 3.13c, respectively. A gap region persists since the gap length is always non-zero. The position of the gap is characterized by the top position and the shift due to the polarity of V_{app} is seen.

The length, top position, and bottom position of a gap at the first and the last step of RESET processes are plotted in Fig. 3.14. In the figure, D_1 and D_2 refer to devices with $I_{\mathrm{cc}} = 2\,\mu\mathrm{A}$ and $I_{\mathrm{cc}} = 4\,\mu\mathrm{A}$, respectively. Notably for the D_1 device, the top position from the TE and the bottom position from the BE are frequently exceeding 1 nm. This leads to an intermediate resistance state instead of two distinct states for two reasons. Firstly, the vacancy distributions in early RESET processes are similar to those in late RESET processes in an upside-down sense. Secondly, the gap lengths of these two stages are comparable. Thus, the LRS and HRS resistances are comparable, implying the failure of manipulating a resistance state by the applied voltage. This simple scheme reproduces the failure to switch between the desired LRS and HRS.

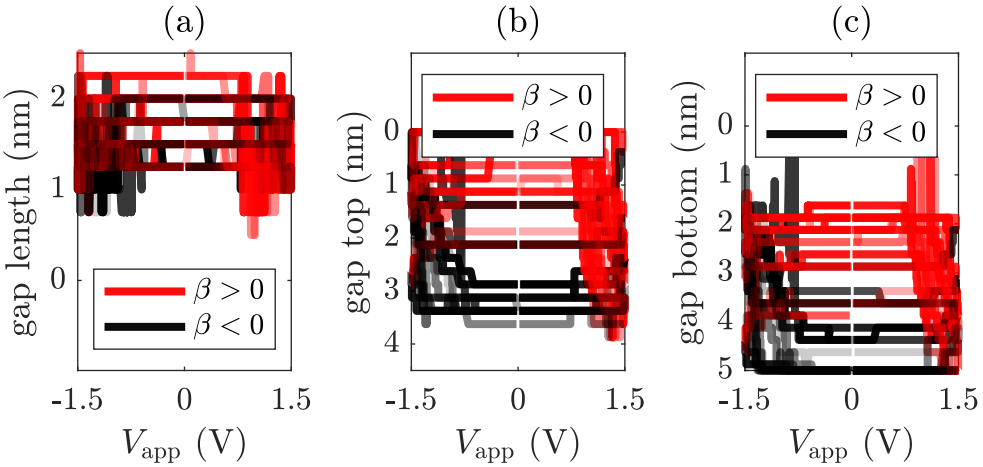

Figure 3.13: (a) Length, (b) top position, and (c) bottom position of a gap for increasing applied voltage ($\beta > 0$) and decreasing applied voltage ($\beta < 0$) stages within twenty switching cycles. A latter cycle is plotted in a lighter color.

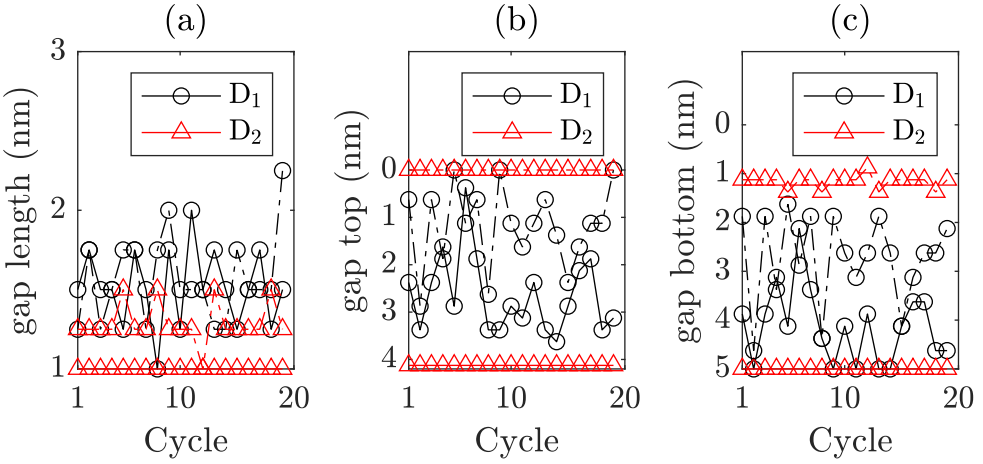

Figure 3.14: (a) Length, (b) top position, and (c) bottom position of a gap. Solid and dash-dotted lines refer to early and late RESET processes, respectively.

With this conclusion, the outcome of a homogeneous activation energy for the generation process is straightforward. Since vacancies are generated at multiple spots, growths of chains are comparable, and thus multiple long chains emerge within a short time. This effectively reduces the variation of vacancy chains after FORMING is done as shown in Fig. 3.11c. Consequently, the failure scheme is analogous to the above one.

It is noted that discussions of the width of a GB region and the magnitude of I_{cc} are still missing. Simulation results of changing only one parameter are shown in Fig. 3.15. Specifically, the D_3 and D_4 devices are denoted to devices with GB spanning 0.5 nm and 2 nm, respectively. In comparison, the width of D_1 and D_2 devices is 1.0 nm. For the D_2 device, an ordered trajectory in all I-V sweep cycles is seen in Fig. 3.15a, and the top and bottom positions demonstrate low variations in Fig 3.14. A similar trend is observed for D_3 device in Fig. 3.15b. On the other hand, a device with a wider GB in Fig. 3.15c demonstrates a more chaotic trajectory similar to that in Fig. 3.13.

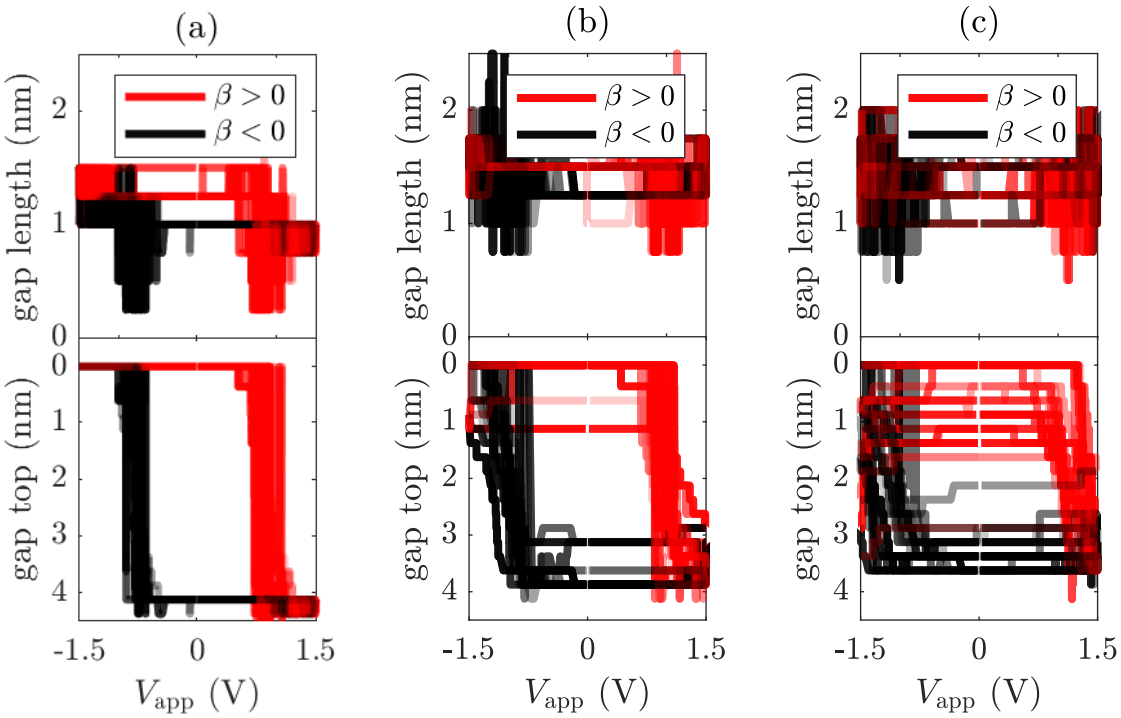

Figure 3.15: Lengths and top positions of a gap in the (a) D_2, (b) D_3, and (c) D_4 devices.

In summary, the assumption of a GB region with homogeneous properties could only serve as a starting point. For an oxide layer with a wide GB, this assumption leads to a degraded memory cell as implied from Fig. 3.15c. That is, the vacancies spread out in the lateral dimension within the predefined GB region resulting in a decreased length of a CF. This leads to an increase of LRS resistances, or equivalently, an intermediate state instead of two distinct states after applying applied voltages. In contrast, a larger I_{cc} or a narrow GB yields a small C2C variability since the evolution of a gap follows a similar trajectory. This might contradict observations. In this section, the C2C variability is

discussed on the evolution of a gap while the resistance aspect is reserved for Sec. 3.5.3.

3.4.5 Impact of variational anisotropic modulations

To break the layer-by-layer migration pattern, it is straightforward to introduce the differentiation of the channels' anisotropic modulations. Specifically, modulations of one channel are more significant than those of the other channel. Under this condition, the vacancy mobility in the x-direction of the strong channel is higher than that of the weak channel. The differentiation in the mobility in the x-direction disrupts the layer migration. In addition, vacancies in the weak channel experience a weak attraction to the strong channel, essentially turning the weak channel into a supplier of oxygen vacancies. This is illustrated in an exemplified SET process in Fig. 3.16.

In Fig. 3.16a, the top vacancy migrates upward within the strong 1D channel. Opposed to random migrations on the interface plane, it is confined within the strong channel. Note that a vacancy can still escape the strong channel but it tends to be drawn back as highlighted by the red arrow. Also, note that the vacancies of the strong channel do not always migrate upwards before that of the weak channel. In Fig. 3.16b, high temperature can trigger the migration of the vacancy in the weak channel as highlighted by the white arrow. Afterward, the vacancy merges into the strong channel in a normal direction as the next movement of the configuration of Fig. 3.16c.

Since migration in the normal direction is driven by the attraction to a strong channel instead of an electric field, such migration can also occur for a neutrally charged vacancy. However, a high temperature is required for such migration. This is captured in Fig. 3.16d, where the top vacancy of the weak channel is surrounded by approximately 400 K. An V_O merges into the strong channel at the next simulation step. Therefore, the vacancy leakage of the strong channel in the former RESET process is supplied during the subsequent SET process. In contrast, the weak channel experiences a recovery of vacancies during a RESET process discussed earlier. Therefore, the pair of a strong and a weak channel contributes to a recoverable length of the longest vacancy chain.

Interestingly, both the GB and CF effects are important during the evolution illustrated in Fig. 3.17. In Fig. 3.17a, the upward migration of the vacancy indicated by the white arrow is promoted due to the CF effect. Similarly, vacancies within the green box shown in Fig. 3.17b will migrate upwards. This type of migration is expected to drive the vacancies of a weak channel away from the BE. In this cycle, the upward migration

continues for 0.5 nm further. Consequently, it leads to an increased LRS resistance.

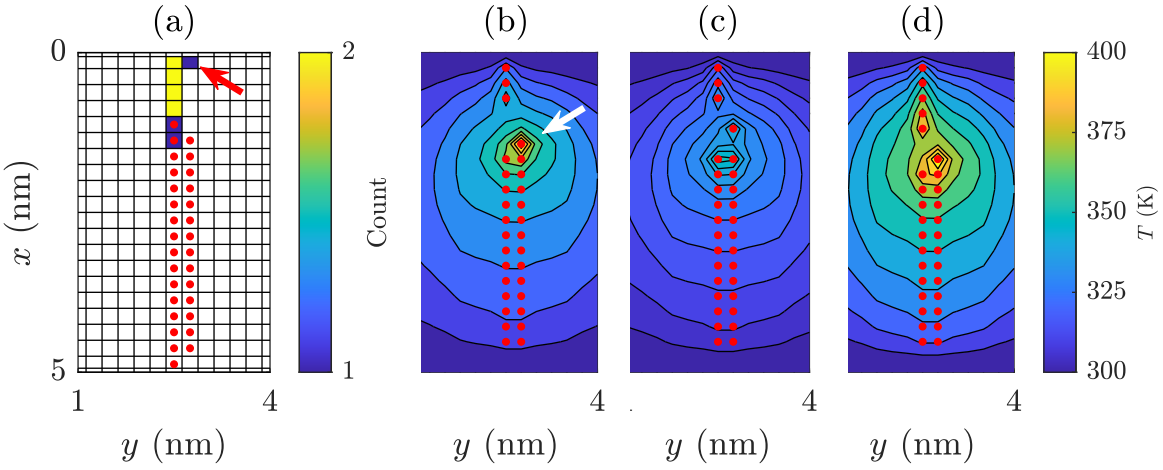

Figure 3.16: (a) Vacancy distribution at the first simulation step of a SET process and the evolution ends at $V_{\text{app}} = 0.84V$. Vacancy distribution superimposed to the temperature distribution (b) at $V_{\text{app}} = 0.86\,\text{V}$, (c) at the next simulation step afterward, (d) at the fourth simulation steps afterward.

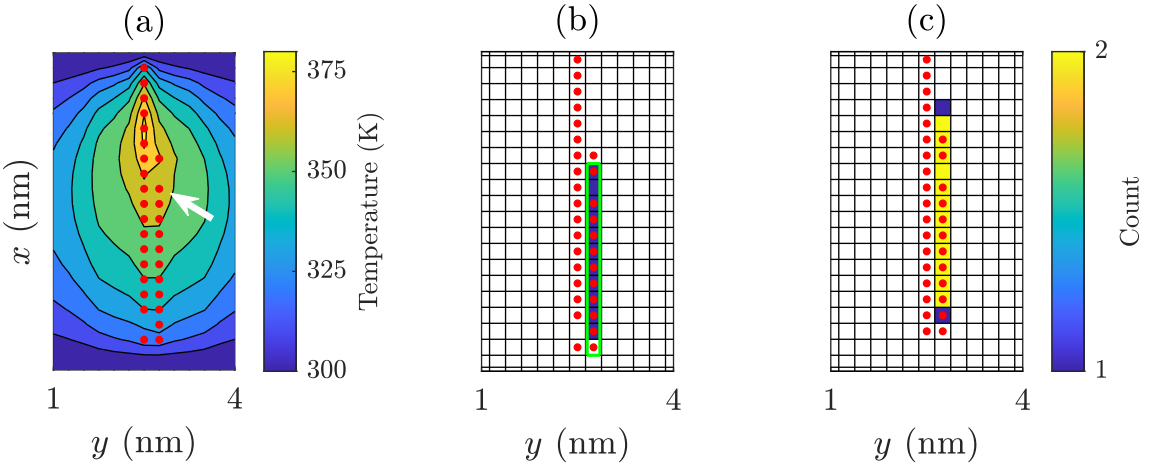

Figure 3.17: (a) Vacancy distribution superimposed to temperature distribution at $V_{\text{app}} = 0.87\,\text{V}$. (b) Vacancy distribution at the next simulation step and evolution within ten further simulation steps. (c) Vacancy distribution at $V_{\text{app}} = 1.32\,\text{V}$ and evolution ends at the last simulation step of a SET process.

3.4.6 Comparison to continuous models

A point-defect viewpoint allows for a fully depleted gap region where no oxygen vacancies exist. In contrast, a continuous model allows a microscopically vacancy-depleted region to possess a residual vacancy density as shown in Fig. 3.18. This treatment effectively reduces the gap length. In addition, a continuous treatment inevitably smooths the spatial variation, especially around the gap. Consequently, a more complete CF exists during the dynamical process, leading to a lower resistance and reduced variability. It is unclear whether the lack of investigation of small current compliance by continuous models stems from the assumption of a finite vacancy density within a gap region.

In addition, it is also unclear how to properly interpret discrete point defects based on vacancy density, where a choice of characteristic length scale is critical. For instance, the characteristic length may be around 1 nm to accurately resolve the gap in a typical 5-nm-thick HfO_2 layer. However, a single vacancy within a cube extending 1 nm already gives the vacancy density of $10^{21}\,\mathrm{cm}^{-3}$. This value is comparable to $1.2\cdot10^{21}\,\mathrm{cm}^{-3}$, which is the maximum density within a CF assumed in Ref. [114]. In comparison, the vacancy density of a Hf_4O_7 is $4\cdot10^{21}\,\mathrm{cm}^{-3}$ given the lattice constant of 5 Å. While DFT calculation suggests a high electrical resistance for a Hf_4O_7 [182], the resistance of a CF with the lower vacancy density of $1.2\cdot10^{21}\,\mathrm{cm}^{-3}$ is expected to be even higher. Therefore, regarding the vacancy density of $1\cdot10^{21}\,\mathrm{cm}^{-3}$ as a criterion for a CF is questionable. To alleviate this discrepancy, one can interpret the vacancy density in a larger characteristic size, as illustrated in Fig. 3.18b. In this scheme, more vacancies contained in a larger characteristic size reproduce the same density while a lower electrical resistance is expected. However, this larger size sacrifices the spatial resolution needed to capture a fully depleted gap.

In conclusion, models treating vacancies as point defects can capture a fully depleted gap region in a thin-film structure. This lays a feasible approach to studying a large variability at a small current compliance regime. Moreover, the point-defect viewpoint aligns with DFT calculations, where the atomic interactions are accounted for. Thus, it is believed to reveal more details during the dynamical process.

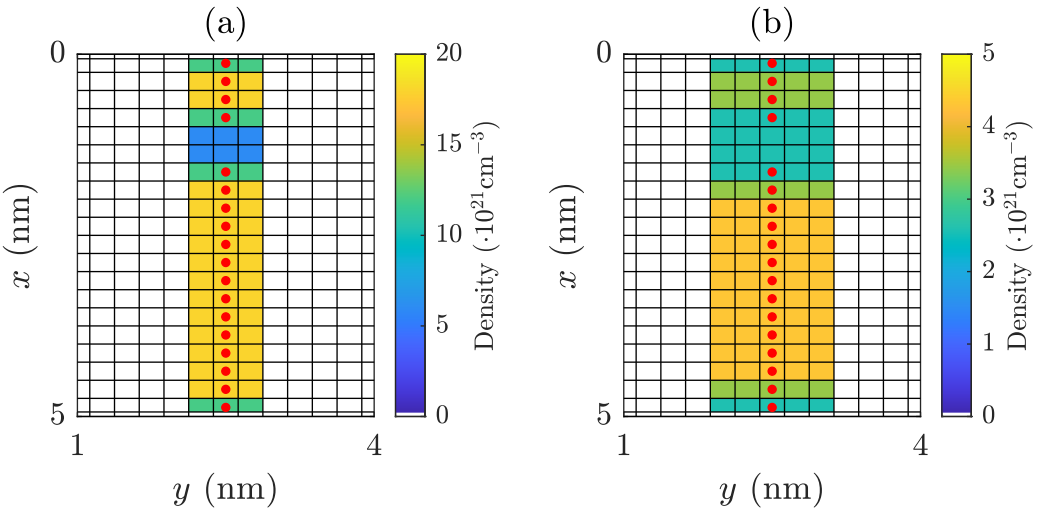

Figure 3.18: Oxygen vacancy densities calculated based on cubes with lengths of (a) 0.5 nm and (b) 1.0 nm.

3.5 Cycle-to-cycle variability at the small current compliance

It is noted that the proposed scenario that occurs within a few cycles does not necessarily keep occurring in subsequent cycles. To attribute the C2C variability in measurements to the anisotropic diffusion barriers, one hundred cycles that follow the cycle of a FORMING and a RESET are simulated. Relevant simulation parameters are listed in Table 5.1. The anisotropic modulation due to the GB effect is visualized in Fig. 3.19. To compare with measurements, the smallest value of $I_{\mathrm{cc}} = 2\,\mu\mathrm{A}$ in the measurement [95] is chosen to simulate. On the one hand, it reduces the computational effort. As a starting point, it is assumed that a similar migration pattern occurs for another current compliance, and thus the C2C variability follows the same statistical behavior. That is, the normalized deviation is approximately unchanged for a comparable current compliance [95, 107]. On the other hand, simulations at a larger current compliance do not provide a comparable C2C variability as that of the measurement. The reason for inconsistent simulation results is briefly discussed at the end of this section, while a generalization to larger current compliance is reserved for Sec. 3.7.

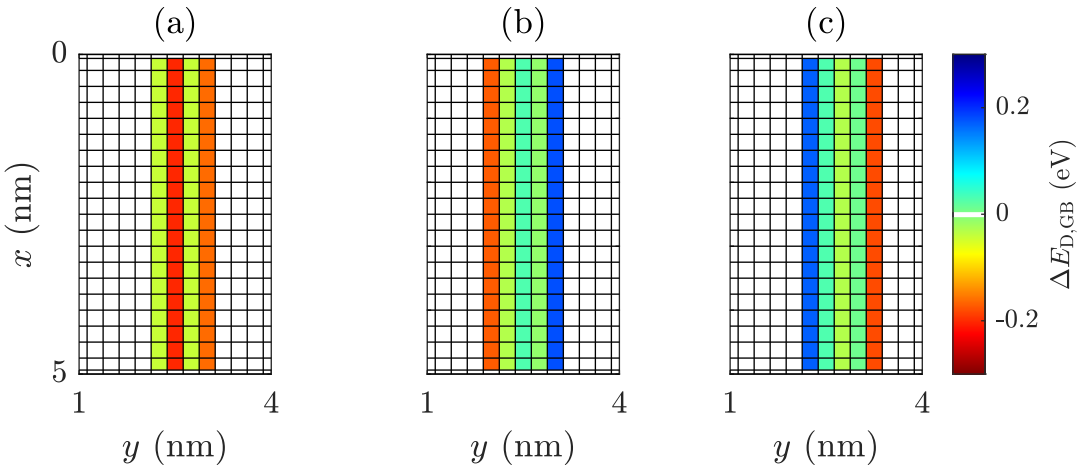

Figure 3.19: Anisotropic modulations for migrations in (a) $\pm x$-directions, (b) the positive y-direction, and (c) the negative y-direction.

3.5.1 Switching cycles

Fig. 3.20 shows simulation results of both LRS and HRS resistances over cycles, where the data from the FORMING cycle is excluded. The LRS and HRS resistances are calculated in the early and the late RESET process at $V_{\mathrm{meas}} = -0.1\,\mathrm{V}$, respectively. The deviation σ_R is defined as the difference between the 70% and 30% of the overall data. In addition, the median resistance μ_R rather than the average resistance is adopted to represent resistances from a statistical perspective. The V_{meas}, μ_R and σ_R are defined to align with the measurement [95].

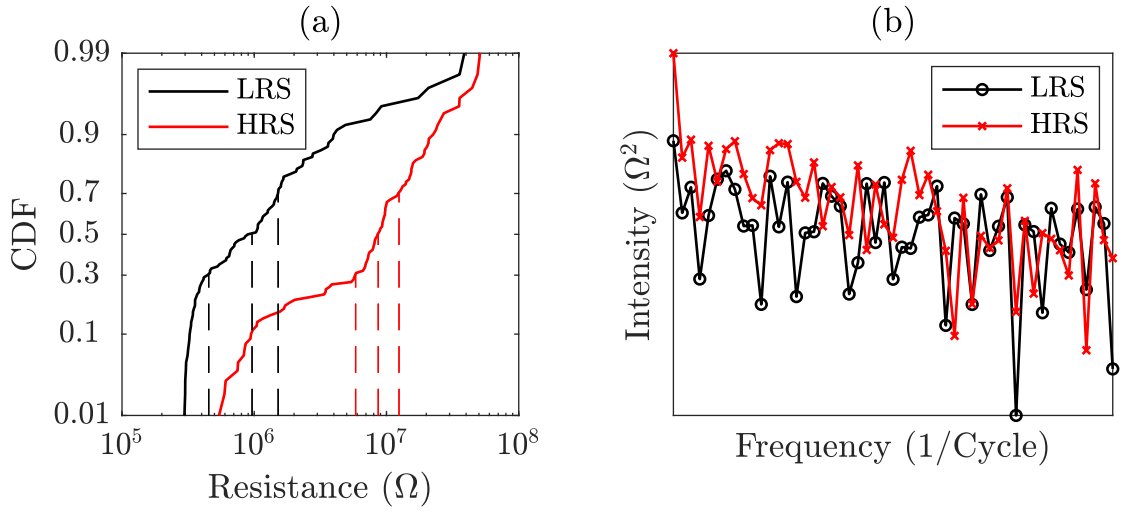

Figure 3.20: Simulation results of LRS and HRS resistances out of one hundred switching cycles after a FORMING process where the $I_{\mathrm{cc}} = 2\,\mu\mathrm{A}$ is used. (a) CDF in the probit-scale. The 30%, 50%, and 70% values are indicated by dashed lines. Reproduced from Ref. [1], licensed under CC BY 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0/). (b) The power spectral density (PSD) of resistances suggests the lack of a fixed frequency for both resistances over cycles.

From the simulation result, the median and the deviation of the LRS resistances are 0.96 MΩ and 1.1 MΩ, respectively. In comparison, the measurement gives approximately 0.5 MΩ and 0.7 MΩ, according to Fig. 3.21a. Meanwhile, the median and the deviation of the HRS resistances from the simulation are 8.6 MΩ and 6.6 MΩ, respectively. From the measurement, it reads approximately 4 MΩ and 3 MΩ for the median and deviation from Fig. 3.21b, respectively. The simulation of $I_{\mathrm{cc}} = 2\,\mu\mathrm{A}$ reproduces a comparable statistical result except for a consistent factor of approximately 2.

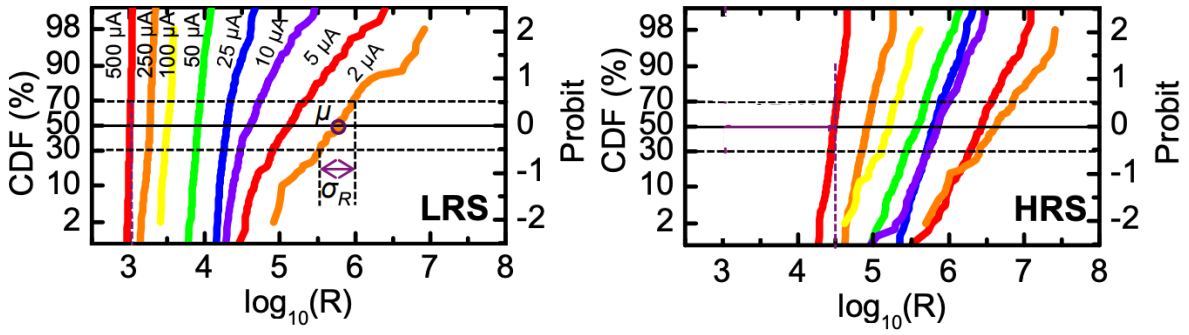

Figure 3.21: CDF plot of LRS and HRS resistances out of measurements. Reproduced with permission from [95], © 2013 IEEE.

It is noted that the information on changes in resistances over cycles is missing in a CDF plot. To ensure the change is stochastic over cycles, the LRS and HRS resistances in a sequential order are plotted in Fig. 3.22. Furthermore, the PSD in Fig. 3.20b provides the information of a correlation in cycles. The absence of a peak in the PSD implies a stochastic resistance distribution within one hundred cycles.

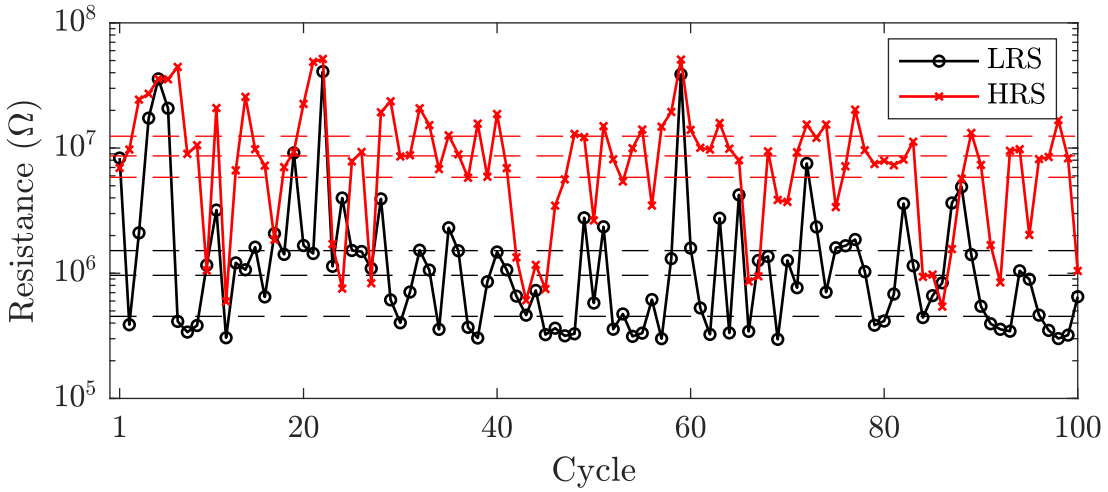

Figure 3.22: Simulation results of LRS and HRS resistances over cycles. The 30%, 50%, and 70% values are indicated by dashed lines. Reused from Ref. [1], licensed under CC BY 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0/).

3.5.2 Dynamical processes

Three cycles are chosen to represent low, high, and intermediate values of LRS resistances, and the corresponding I-V characteristic curves are shown in Fig. 3.23.

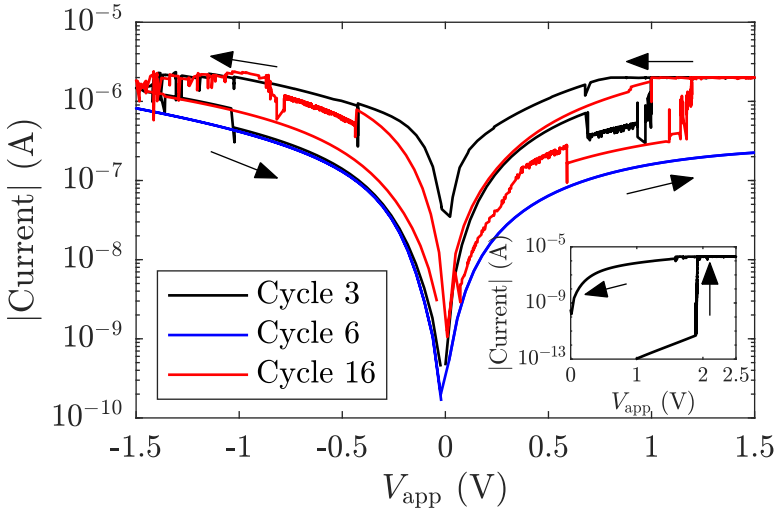

Figure 3.23: I-V characteristic curves of three cycles. The inset shows the I-V characteristic curve of the FORMING process. Reproduced from Ref. [1], licensed under CC BY 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0/).

It is seen that a long CF yields a low value of the LRS resistance shown in Fig. 3.24a while a disruption of the CF yields a high value of the LRS resistance shown in Fig. 3.24b. Between two extreme situations, an intermediate resistance arises from a moderate gap region shown in Fig. 3.24c. So far, it is consistent with the well-established understanding that the length of a gap is crucial to resistance.

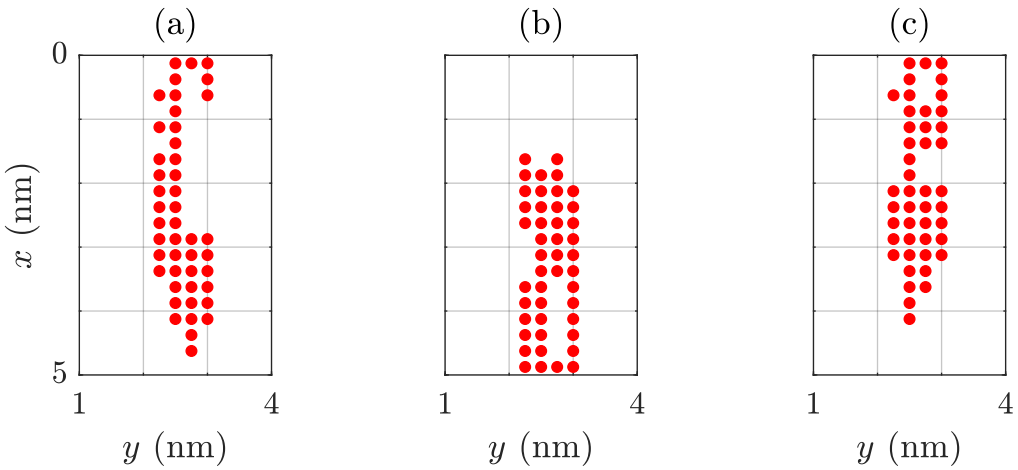

Figure 3.24: Vacancy distributions at $V_{app} = V_{meas} = -0.1\,V$ of the (a) third cycle, (b) sixth cycle, and (c) sixteenth cycle. Reproduced from Ref. [1], licensed under CC BY 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0/).

However, a deeper investigation of the evolution during the third SET operation reveals the anisotropic migration pattern. In Fig. 3.25a, the vacancy distribution at the early SET operation is a reference for later distributions. Fig. 3.25b shows the vacancy distribution before the current reaches the current compliance. To illustrate the migration process, the evolution of vacancies is plotted in Fig. 3.26. In the following discussion, dummy indexes from 1 to 4 are assigned for the vacancy chains from left to right. Fig. 3.26a and 3.26b show that only the vacancies of the first and second vacancy chains are involved in the migration. More importantly, the vacancy migration of the second chain is more frequent due to the assumption of anisotropic diffusion (see Fig. 3.19a). This leads to the break of a layer-by-layer migration, as discussed in Sec. 3.4.4. The gap is eventually enclosed with further vacancies migrating upwards, as shown in Fig. 3.25c as well as Fig. 3.26c. When the current reaches the current compliance, the V_{cell} might deviate from the V_{app}, depending on the resistance of the cell. The amount of reduction in the V_{cell} is related to the power dissipation along the CF, which is critical for the subsequent migration. In this exemplified cycle, a few vacancies close to the bottom interface keep moving, leading to the vacancy distribution shown in Fig. 3.24a. Interestingly, it is seen from Fig. 3.26d that the vacancy migration of the fourth chain is more active than that of the third. The break of layer-by-layer migration is again observed, which creates a minor gap to the BE interface and thus increases the LRS resistance.

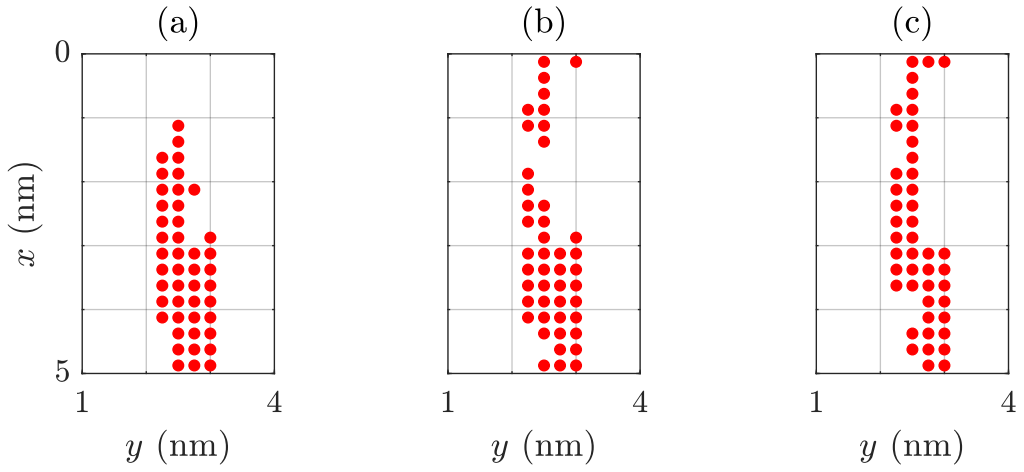

Figure 3.25: Vacancy distributions at (a) $V_{\mathrm{app}} = 0.64\,\mathrm{V}$, (b) $V_{\mathrm{app}} = 1.0\,\mathrm{V}$ and (c) $V_{\mathrm{app}} = 1.5\,\mathrm{V}$ during the ramp-up stage of the third SET process. Reproduced from Ref. [1], licensed under CC BY 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0/).

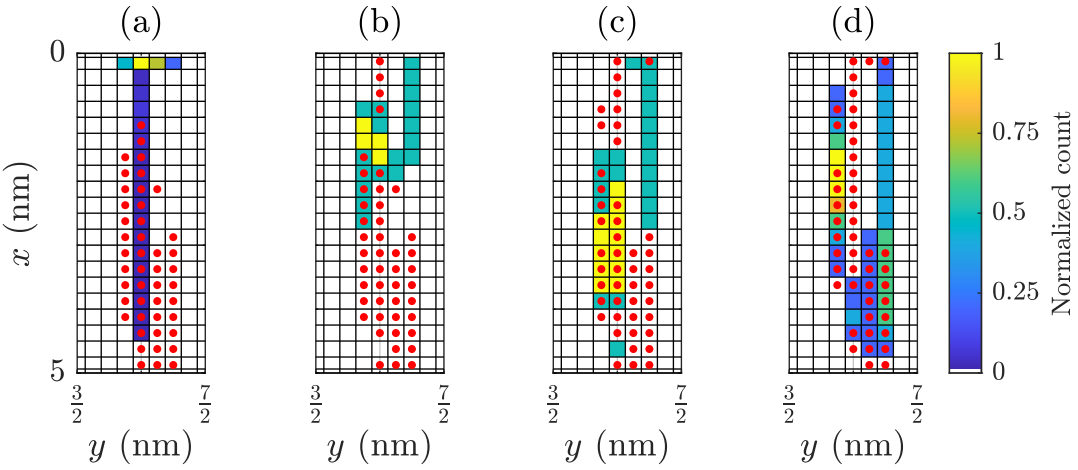

Figure 3.26: Vacancy distributions superimposed by the evolution during the SET process of the third cycle. The vacancy distribution at (a) $V_{\mathrm{app}} = 0.64\,\mathrm{V}$, (b) $V_{\mathrm{app}} = 0.98\,\mathrm{V}$, (c) $V_{\mathrm{app}} = 1.0\,\mathrm{V}$, and (d) $V_{\mathrm{app}} = 1.5\,\mathrm{V}$. Evolution from the contemporary stage towards the next stage is illustrated in colors (a), (b), and (c), while the evolution of (d) ends at $V_{\mathrm{app}} = V_{\mathrm{meas}}$.

Lastly, the process involving a decrease in the LRS resistance of one cycle compared to that of the previous cycle is shown in Fig. 3.27. Specifically, Fig. 3.27a, 3.27b and 3.27c show the vacancy and temperature distributions in sequential time order in the RESET process, while Fig. 3.27d and 3.27e show distributions in the next SET process. Since the LRS resistance in the early RESET process is relatively high, the gap close to the BE interface is expected as shown in Fig. 3.27a. That is to say, most vacancies are located in the middle-height region as indicated by the white rectangle. As proceeding in the RESET process, a high temperature coincides in the region where vacancy chains tend to migrate downwards. This leads to the CF effect, where the migration of the vacancies within the vacancy chain is promoted shown by the black arrow in Fig. 3.27b. This could trigger a series of migration events, leading to a CF contains more vacancies near the BE interface, as shown in Fig. 3.27c. In Fig. 3.27d, the positively charged vacancy indicated by the white circle has the degree of freedom to migrate in either the $-x$ or y- direction. It is noted that the migration in $-y$ and $\pm z$ directions are not considered, since they result in the vacancy escaping out of the assumed GB region. With a high temperature, the guided direction of an electric field is hindered. In this exemplified scenario, this vacancy migrates into the second channel. As a result, a conduction path composed of the second vacancy chain and the bottom segments of other vacancy chains

is built, see Fig. 3.27e. It is noteworthy that the positively charged vacancy in Fig. 3.27d does not necessarily migrate into the second vacancy chain. However, the decrease in LRS resistance does not qualitatively depend on its migration direction. The dominant factor is the large number of vacancies in the bottom region, leading to a significant reduction in the resistance.

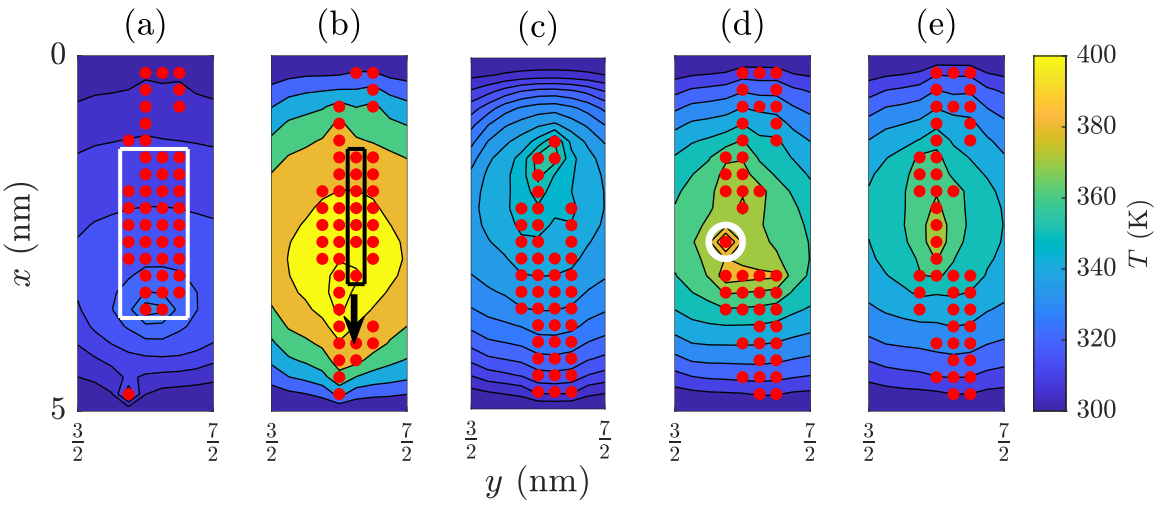

Figure 3.27: Vacancy distributions involved in a decreased LRS resistance process. Reused from Ref. [1], licensed under CC BY 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0/).

The opposite process involving an increased LRS resistance is shown in Fig. 3.28. Fig. 3.28a and 3.28b show the distributions at the beginning and the middle of the RESET process, while Fig. 3.28c, 3.28d and 3.28e sketch the distributions in the following SET process. Since the previous SET sets a lower LRS resistance, some vacancies residing in the lower region, which are indicated by the white rectangle in Fig. 3.28a. In contrast to the previous scenario, the migration of short chains is significant in the SET process, as indicated by the black arrow in Fig. 3.28c. That is, these short vacancy chains move away from the bottom interface, leading to an increase in the LRS resistance. This is seen in Fig. 3.28d, where the vacancy chain indicated by a black box is in the top region. By comparing to Fig. 3.28e, the vacancy chain indicated by the white box is about to migrate upwards. This is due to the CF effect. At the end of the SET process, a larger gap in the BE interface, which is compared to the early RESET process, is seen in Fig. 3.28e. Thus, the LRS resistance of the latter cycle increases.

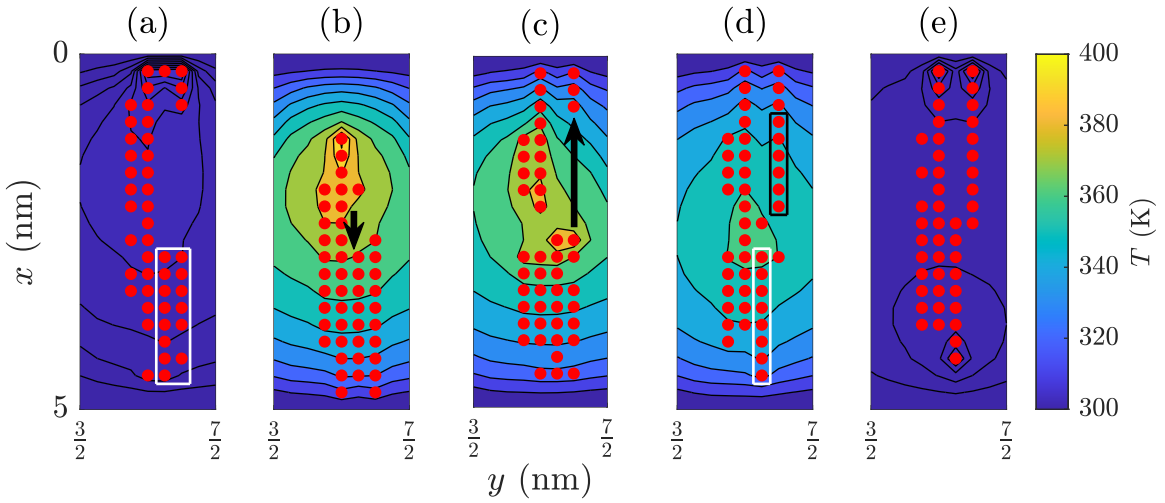

Figure 3.28: Vacancy distributions involved in an increased LRS resistance process. Reused from Ref. [1], licensed under CC BY 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0/).

3.5.3 Failure to larger values of current compliance

Although the simulation reproduces comparable results for $I_{cc} = 2\,\mu A$, the variability counters observations at an elevated current compliance. For example, the use of $I_{cc} = 3\,\mu A$ shown as D_1 in Fig. 3.29 leads to a narrow spread of LRS resistances yet an averagely higher value.

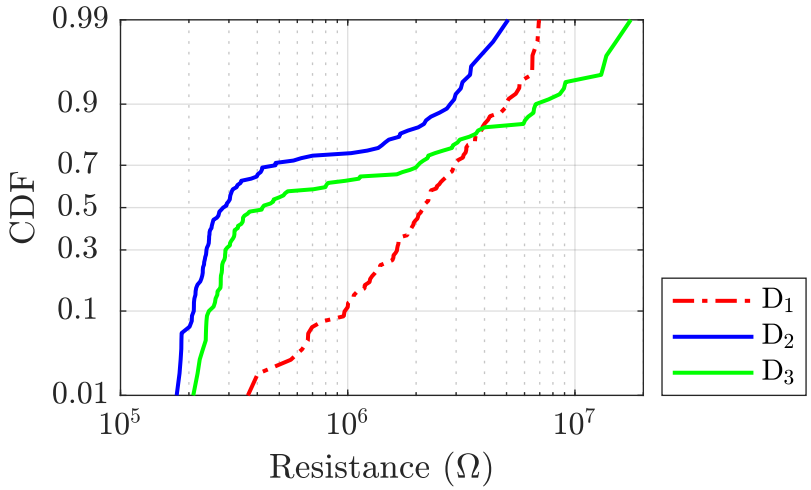

Figure 3.29: CDF plot of LRS resistances out of one hundred cycles. D_1, D_2, and D_3 refer to the devices with the original GB, a wide GB with one fragile spot, and a wide GB with three fragile spots, respectively. Additionally, D_1 is subjected to $I_{cc} = 3\,\mu A$ while D_2 and D_3 are subjected to $I_{cc} = 2\,\mu A$.

The statistical vacancy distribution out of fifty FORMING processes is shown in Fig. 3.30a. It is seen that not only strong channels but also weak channels are filled with vacancies. Interestingly, a large number of vacancies within the GB does not always result in the expected LRS. Fig. 3.30c shows the statistical vacancy distributions at the end of one hundred RESET processes. It is seen that vacancies can reach the BE interface during RESET processes. However, a clear gap at the BE interface is seen in all cycles from Fig. 3.30b. That is, the vacancies of weak chains do not stay in the bottom region but keep migrating upwards during SET processes. An intricate picture arises where the median (or average) resistance can not be solely explained by the number of vacancies. On the other hand, the variability in resistances is confined. Given that a gap always emerges at the end of a SET process, the large number of vacancies filled up the upper GB region and thus reduces the shape variation of a CF.

This discrepancy of the average (or median) LRS resistance might be alleviated by adaptively assuming a larger width of the GB area for the corresponding increased I_{cc}. Therefore, the accumulation of vacancies in the weak channels can be avoided. However, the C2C variability accounts for the same device at different I_{cc}. Such an assumption deviates from a practical condition, and thus it is not considered in this work.

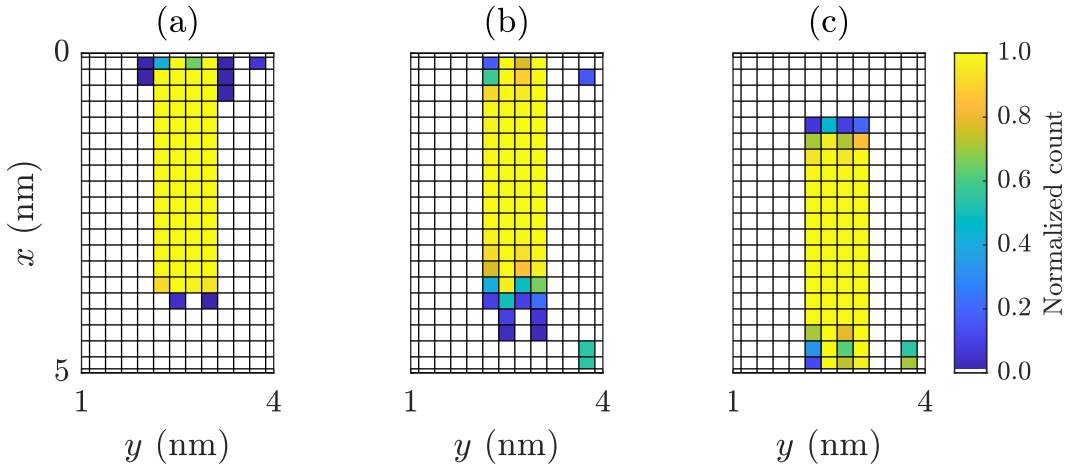

Figure 3.30: Statistical vacancy distributions of $I_{cc} = 3\,\mu A$ out of (a) fifty FORMING processes for the D_1 device. Statistics of vacancy distributions of one hundred (b) SET processes and (c) RESET processes.

On the other hand, the assumption of a fixed GB area with a large width, i. e., a region

composed of four sets of a strong channel being next to a weak channel, is investigated. Herein, the $I_{cc} = 2\,\mu A$ is applied to D_2 and D_3, and the impact of one and three fragile spots is investigated. In Fig. 3.29, devices with a larger GB seemingly lead to a reasonable C2C variability. However, it is noted an abrupt change after approximately $R = 4 \cdot 10^5\,\Omega$ in the CDF plot. To identify whether the devices are stuck in a HRS, the LRS resistance against switching cycles is plotted in Fig. 3.31. The LRS resistances of both devices are observed to be stuck at a high value in a window of approximately twenty cycles, as highlighted by black rectangles.

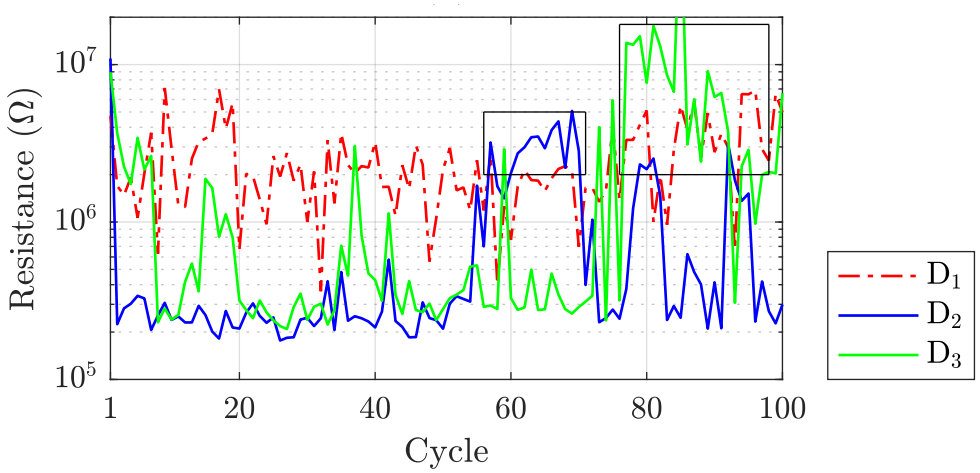

Figure 3.31: LRS resistances in sequential order for the three devices defined in Fig. 3.29.

To attribute the failure to the vacancy migration pattern, statistics of vacancy distribution at the end of SET processes are plotted. Specifically, Fig. 3.32b and 3.33b show the statistics of the failure windows for D_2 and D_3 devices, respectively. A gap at the middle height at the end of SET processes for both devices is seen. Together with the statistics at the end of RESET processes shown in Fig. 3.32c and 3.33c, it suggests that the failure of enclosing a CF leads to elevated LRS resistances. Since this phenomenon is seen in both devices with a wider GB, the width is suspected to be the reason. This is interpreted as the outcome of the re-distribution of vacancies. Given an increased space of the GB, vacancies tend to cluster at both TE and BE interfaces. Under this condition, the CF effectively shrinks in length, leading to decreases in current and heat dissipation. In addition, a local cluster of vacancies is modeled to increase the thermal conductivity, further suppressing the temperature rise. Consequently, vacancy migration is suppressed and the gap persists even after a SET operation is finished.

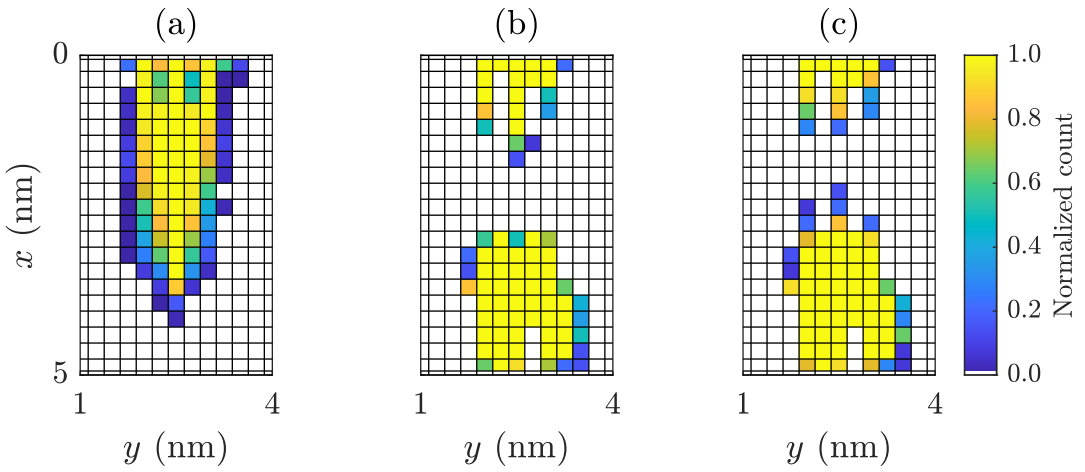

Figure 3.32: Statistical vacancy distributions of $I_{cc} = 2\,\mu A$ out of (a) fifty FORMING processes for the D_2 device. Statistics of (b) the first simulation step and (c) the last simulation step of RESET processes. The statistics of (b) and (c) account for cycles from the fifty-seventh to the seventieth cycle.

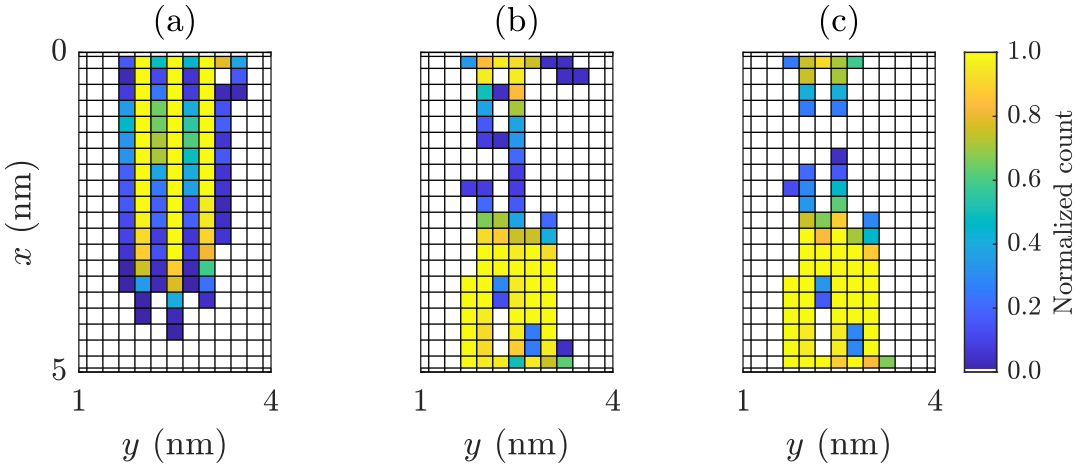

Figure 3.33: Statistical vacancy distributions of $I_{cc} = 2\,\mu A$ out of (a) fifty FORMING processes for the D_3 device. Statistics of (b) the first simulation step and (c) the last simulation step of RESET processes. The statistics of (b) and (c) account for cycles from the seventy-seventh to the one-hundredth cycle.

In comparison, fewer strong channels as extensively investigated in this section reduce the probability of vacancies merging into strong channels. In this scenario, the length

of an effective CF preserves above a certain value, and a large window of failure is not observed. This is supported by the statistical vacancy distributions in Fig. 3.34, where the length of channel 2 is long in all late SET processes.

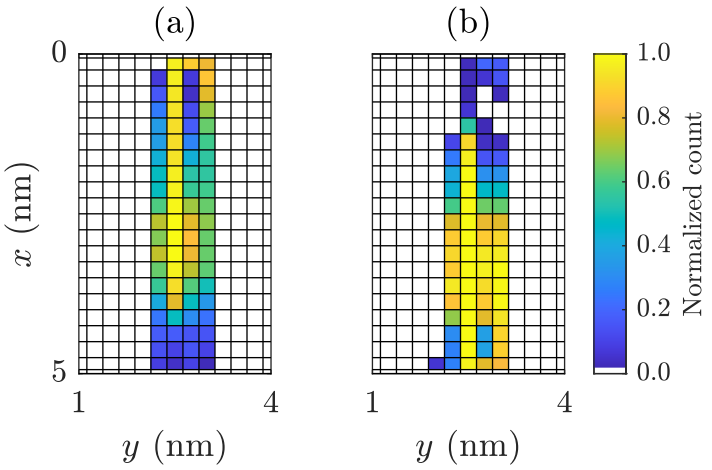

Figure 3.34: Statistical vacancy distributions of (a) the first simulation step and (b) the last simulation step of RESET processes. The statistics account for cycles from all switching cycles of the D_1 device subjected to $I_{cc} = 2\,\mu A$.

3.6 Generalized electron hopping scheme

In this section, an extra pre-factor in the master equation is proposed. The scheme is examined by requiring the detailed balance between each pair of vacancies to be held, and then see if the original MA hopping formalism is recovered.

The electron hopping from the ith to the jth vacancy is proportional to four independent factors: the probability of the original site being occupied, the probability of the final site being empty, the transition rate between these two sites, and the constant attempt frequency. The hopping event per second is determined by the product of these factors, as they are independent. However, this treatment might not account for the situation where the original site is occupied by two electrons. Consider two configurations where only the original site differs. In the first configuration, the original site can be occupied by two electrons, while in the second configuration, only one electron can occupy it. In addition, only one electron can hop out at a time. Whenever the final site is empty and the original site is fully occupied, either one of the two electrons in the first configuration can hop. This is interpreted as having two possible paths, as schematically illustrated in Fig. 3.35a. Hereby, a factor of 2 is proposed to account for this additional degree of freedom in choosing an electron. On the contrary, no extra degree of freedom is expected in the second configuration, as illustrated in Fig. 3.35b. Therefore, the hopping event per second of the first configuration is expected to be twice of the second configuration.

Similarly, the same argument applies to empty states, where at most two empty orbitals and only one empty orbital are assigned for the first and the second configurations, respectively. With one of the two empty orbitals occupied by an injected electron, the hopping events per unit time of the first configuration is twice of the second configuration. This is schematically shown in Fig. 3.35c and 3.35d for the first and second configurations, respectively.

Denote the notation f_t^q for the extra factor, and the discussed cases for $q = \{0, +2\}$, and $t = \{\mathrm{d}, \mathrm{a}\}$ are summarized in Table 3.3.

f_{d}^{0}	f_{d}^{+2}	f_{a}^{0}	f_{a}^{+2}
2	1	1	2

Table 3.3: Extra pre-factors accounting for the degree of freedom involved in donating and accepting an injected electron.

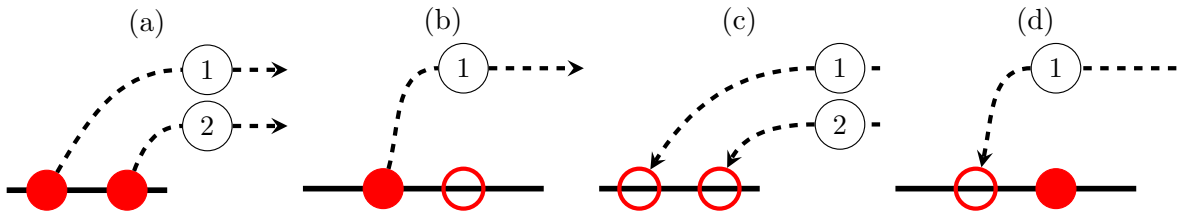

Figure 3.35: Schemes for (a) and (b) donating an electron, and (c) and (d) accepting an electron.

In the next section, the MA hopping process

$$-p_i f^q_{i,\mathrm{d}} \sum_j f^q_{j,\mathrm{a}}(1-p_j)h_{ij} + (1-p_i) f^q_{i,\mathrm{a}} \sum_j f^q_{j,\mathrm{d}} p_j h_{ji} = 0 \tag{3.6}$$

will be examined by asking for the detailed balance under equilibrium conditions.

3.6.1 Grand partition function and the probability

To examine the detailed balance under equilibrium conditions, the probability at each state must first be determined analytically. Within the theory of grand canonical ensemble, particle numbers and the energy of a sub-system can be exchanged with its reservoir. The grand partition function takes the form

$$\mathcal{Z} = \sum_{N=0}^{\infty} \sum_{E(N)} \exp\left(\frac{N\mu - E}{k_\mathrm{B}T}\right), \tag{3.7}$$

where N is the number of electrons, μ is the Fermi level and E is the number-dependent energy of the sub-system. The energy levels are schematically shown in Fig. 3.36. From the energy perspective, the formation energy of a doubly positive charge state is higher than that of the other two charge states given that the neutral charge state is energetically favored. Conversely, the neutral charge state possesses a higher formation energy if the doubly positive charge state is favored. For both cases, the formation energies of a singly positive charge state are between the other two charge states, see Fig. 3.3. In the following derivation, the fully ionized state is excluded for the intrinsically neutral charge state, and the fully occupied state is excluded for the intrinsically doubly positive state for simplicity.

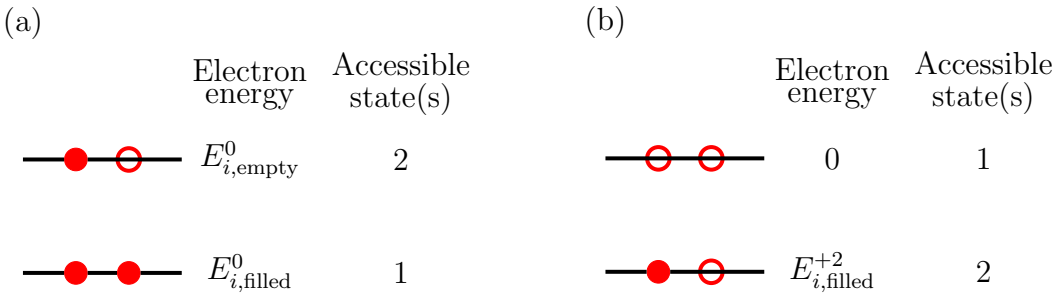

Figure 3.36: Electron energy level scheme for (a) intrinsically neutral and (b) intrinsically doubly positive charge states. The filled and unfilled circles refer to the occupied and unoccupied orbitals, respectively. Note that energies are referred to the CBM.

For the neutral charge state, the grand partition function of the ith vacancy site reads

$$\mathcal{Z}_i = 2\exp\left(\frac{\mu - E^0_{i,\text{empty}}}{k_\text{B}T}\right) + \exp\left(\frac{2\mu - 2E^0_{i,\text{filled}}}{k_\text{B}T}\right). \tag{3.8}$$

The pre-factor of the first exponential function comes from the two accessible states while the factor of μ in the exponent comes from the fact that only one electron is involved. Similarly, the pre-factor the second exponential function, and the factor of μ result from one accessible state and two electrons, respectively. The probability of the ith vacancy at the filled state then reads

$$\begin{aligned} p^0_{i,\text{filled}} &= \exp\left(\frac{2\mu - 2E^0_{i,\text{filled}}}{k_\text{B}T}\right) \Bigg/ \left[2\exp\left(\frac{\mu - E^0_{i,\text{empty}}}{k_\text{B}T}\right) + \exp\left(\frac{2\mu - 2E^0_{i,\text{filled}}}{k_\text{B}T}\right)\right] \\ &= \left[1 + 2\exp\left(\frac{2E^0_{i,\text{filled}} - E^0_{i,\text{empty}} - \mu}{k_\text{B}T}\right)\right]^{-1}. \end{aligned} \tag{3.9}$$

Note that $p^0_{i,\text{filled}}$ is the probability of the vacancy being fully occupied by two electrons, or equivalently, the probability of the vacancy being neutrally charged.

For the doubly positive charge state, the grand partition function reads

$$\mathcal{Z}_i = 2\exp\left(\frac{\mu - E^{+2}_{i,\text{filled}}}{k_\text{B}T}\right) + 1. \tag{3.10}$$

Note that energies are referred to CBM and no electron is removed from the reservoir, giving rise to the latter term, 1. This yields the probability

$$\begin{aligned} p^{+2}_{i,\text{filled}} &= 2\exp\left(\frac{2\mu - E^{+2}_{i,\text{filled}}}{k_\text{B}T}\right) \Bigg/ \left[2\exp\left(\frac{\mu - E^{+2}_{i,\text{filled}}}{k_\text{B}T}\right) + 1\right] \\ &= \left[1 + \frac{1}{2}\exp\left(\frac{E^{+2}_{i,\text{filled}} - \mu}{k_\text{B}T}\right)\right]^{-1}. \end{aligned} \tag{3.11}$$

It is noted that the shift due to the electrostatic potential is missing in Eqs. (3.9) and (3.11). Since the electron energy levels are shifted by $-e\varphi$ due to the potential, $E^q_{i,\text{filled}}$ and $E^q_{i,\text{empty}}$ are replaced by $E^q_{i,\text{filled}} - e\varphi_i$ and $E^q_{i,\text{empty}} - e\varphi_i$, respectively. This argument holds regardless of the intrinsic charge state. In turn, the electrostatic potential shifts the exponents and the probabilities of staying at the filled states reads

$$\begin{aligned} p^0_{i,\text{filled}} &= \left[1 + 2\exp\left(\frac{2E^0_{i,\text{filled}} - E^0_{i,\text{empty}} - e\varphi_i - \mu}{k_\text{B}T}\right)\right]^{-1} \\ p^{+2}_{i,\text{filled}} &= \left[1 + \frac{1}{2}\exp\left(\frac{E^{+2}_{i,\text{filled}} - e\varphi_i - \mu}{k_\text{B}T}\right)\right]^{-1} \end{aligned}. \tag{3.12}$$

3.6.2 Detailed balance among vacancies

Under equilibrium conditions, the requirement of the detailed balance between a pair of vacancies yields

$$f^q_{i,\text{d}} p_i f^q_{j,\text{a}} (1 - p_j) h_{ij} = f^q_{j,\text{d}} p_j f^q_{i,\text{a}} (1 - p_i) h_{ji}. \tag{3.13}$$

Eq. (3.13) is then examined in four cases, categorized by the set of intrinsic charge states of the ith and jth vacancies.

Hopping between vacancies in a neutral charge state

By substituting $p_i = p^0_{i,\text{filled}}$ and $p_j = p^0_{j,\text{filled}}$ into Eq. (3.13), one obtains

$$\begin{aligned} \frac{h_{ji}}{h_{ij}} &= \frac{f^0_{i,\text{d}} p^0_{i,\text{filled}}}{f^0_{j,\text{d}} p^0_{j,\text{filled}}} \cdot \frac{f^0_{j,\text{a}}(1 - p^0_{j,\text{filled}})}{f^0_{i,\text{a}}(1 - p^0_{i,\text{filled}})} = \frac{2 \cdot p^0_{i,\text{filled}}}{1 \cdot p^0_{j,\text{filled}}} \cdot \frac{1 \cdot (1 - p^0_{j,\text{filled}})}{2 \cdot (1 - p^0_{i,\text{filled}})} \\ &= \frac{\exp\left(\frac{2E^0_{j,\text{filled}} - E^0_{j,\text{empty}} - e\varphi_j - \mu}{k_\text{B}T}\right)}{\exp\left(\frac{2E^0_{i,\text{filled}} - E^0_{i,\text{empty}} - e\varphi_i - \mu}{k_\text{B}T}\right)} = \exp\left(-e\frac{\varphi_j - \varphi_i}{k_\text{B}T}\right). \end{aligned} \tag{3.14}$$

Since the intrinsic charge states of both vacancies are identical, the energy level terms of the ith vacancy cancel with those of the jth vacancy. In addition, the $\varphi_j - \varphi_i > 0$ is assigned without loss of generality. The exchange of dummy indexes i and j does not change the result that the larger MA hopping term arises from the flow to a lower-energy site. This leads to the function form recovering to the MA formalism as Eq. (2.8).

Hopping between vacancies in a doubly positive charge state

Similarly, the substitution of $p_i = p^{+2}_{i,\mathrm{filled}}$ and $p_j = p^{+2}_{j,\mathrm{filled}}$ yields

$$\frac{h_{ji}}{h_{ij}} = \frac{\exp\left(\frac{E^{+2}_{j,\mathrm{filled}} - e\varphi_j - \mu}{k_\mathrm{B}T}\right)}{\exp\left(\frac{E^{+2}_{i,\mathrm{filled}} - e\varphi_i - \mu}{k_\mathrm{B}T}\right)} = \exp\left(-e\frac{\varphi_j - \varphi_i}{k_\mathrm{B}T}\right). \tag{3.15}$$

Therefore, the original MA hopping term is again recovered.

Hopping between vacancies in different charge states

For the electron hopping between vacancies in different charge states, the attenuation radius are different. According to Eq. (2.11) with the absolute values of $(E^0_{i,\mathrm{filled}}, E^{+2}_{i,\mathrm{filled}}) = (2.11, 1.97)$ in the unit of eV, this results in the relative difference 3.5% referred to that of the neutral charge state. The difference is assumed to be sufficiently small and further discussion is based on the truncation of the difference. Under this condition, the ratio of the MA hopping rates is still the ratio of probabilities.

The ratio of MA rates is given by

$$\begin{aligned}
\frac{h_{ji}}{h_{ij}} &= \frac{f^0_{i,\mathrm{d}} p^0_{i,\mathrm{filled}}}{f^{+2}_{j,\mathrm{d}} p^{+2}_{j,\mathrm{filled}}} \cdot \frac{f^{+2}_{j,\mathrm{a}}(1 - p^{+2}_{j,\mathrm{filled}})}{f^0_{i,\mathrm{a}}(1 - p^0_{i,\mathrm{filled}})} = \frac{2 \cdot p^0_{i,\mathrm{filled}}}{1 \cdot p^{+2}_{j,\mathrm{filled}}} \cdot \frac{2 \cdot (1 - p^{+2}_{j,\mathrm{filled}})}{1 \cdot (1 - p^0_{i,\mathrm{filled}})} \\
&= \frac{4}{1} \cdot \frac{\frac{1}{2}\exp\left(\frac{E^{+2}_{j,\mathrm{filled}} - e\varphi_j - \mu}{k_\mathrm{B}T}\right)}{2\exp\left(\frac{2E^0_{i,\mathrm{filled}} - E^0_{i,\mathrm{empty}} - e\varphi_i - \mu}{k_\mathrm{B}T}\right)} \qquad (3.16) \\
&= \exp\left(\frac{\Delta E^0_i}{k_\mathrm{B}T}\right) \cdot \exp\left(\frac{(E^{+2}_{j,\mathrm{filled}} - e\varphi_j) - (E^0_{i,\mathrm{filled}} - e\varphi_i)}{k_\mathrm{B}T}\right), \qquad (3.17)
\end{aligned}$$

where $\Delta E^0_i = E^0_{i,\mathrm{empty}} - E^0_{i,\mathrm{filled}} > 0$. Two points are note-worthy. Firstly, pre-factors of exponential functions for the probabilities are canceled with the proposed pre-factors

in Eq. (3.16). Secondly, energy levels are shifted by the corresponding electrostatic potentials, and an extra exponent $\Delta E_i^0/k_\mathrm{B}T$ emerges in Eq. (3.17). The shift in energy levels is an extension of the comparison of electrostatic potentials between two identically charged vacancies. The MA hopping term that satisfies the detailed balance reads

$$h_{ij} = \begin{cases} \nu_e \exp\left(-\frac{d_{ij}}{a_0}\right) & \widetilde{E}^q_{i,\mathrm{filled}} > \widetilde{E}^q_{j,\mathrm{filled}} \\ \nu_e \exp\left(-\frac{d_{ij}}{a_0}\right) \exp\left(-\frac{\widetilde{E}^{+2}_{j,\mathrm{filled}} - \widetilde{E}^0_{i,\mathrm{filled}}}{k_\mathrm{B}T_i}\right) & \text{otherwise} \end{cases}, \tag{3.18}$$

where $\widetilde{E}^q_{i,\mathrm{filled}}$ evaluated at $q = 0$ gives $E^0_{i,\mathrm{filled}} - \Delta E^0_i - e\varphi_i$, and $\widetilde{E}^q_{i,\mathrm{filled}}$ evaluated at $q = +2$ gives $\widetilde{E}^{+2}_{j,\mathrm{filled}} = E^{+2}_{j,\mathrm{filled}} - e\varphi_j$.

To complete the discussion, the charge states of vacancies are now exchanged. This operation is equivalent to exchanging indexes i, j in Eq. (3.17). Note that the substitutions of j by $\tilde{i}$ and i by $\tilde{j}$ yield Eq. (3.18) in terms of new dummy indexes. In this condition, values of energy levels are $E^{+2}_{i,\mathrm{filled}} - e\varphi_i$ for the $\widetilde{i}$th vacancy and $E^0_{j,\mathrm{filled}} - \Delta E^0_j - e\varphi_j$ for the $\widetilde{j}$th one.

In short, the requirement of the detailed balance among each pair of vacancies is consistent with the MA formalism, provided vacancies are in an identical intrinsic charge state. For two vacancies in different charge states, the MA hopping term is in the form of Eq. (3.18) with an aligned a_0 term. Moreover, the tunneling processes are assumed to follow the same physical picture. Thus, the pre-factors apply to the tunneling process as well. In the following sections, the master equation is applied in the form

$$\begin{aligned} -p_i f^q_{i,\mathrm{d}} \sum_j f^q_{j,\mathrm{a}}(1-p_j)h_{ij} + (1-p_i) f^q_{i,\mathrm{a}} \sum_j f^q_{j,\mathrm{d}} p_j h_{ji} \\ - p_i f^q_{i,\mathrm{d}} \sum_M R_{iM} + (1-p_i) f^q_{i,\mathrm{a}} \sum_M R_{Mi} = 0. \end{aligned} \tag{3.19}$$

To investigate the impact of extra factors, a SET process is simulated for two assumed vacancy distributions shown in Fig. 3.37. In a dynamical process, the vacancy distribution in Fig. 3.37a is followed by that in Fig. 3.37b. For comparison of charge transport schemes, vacancies are intentionally kept fixed in space with only one vacancy in a doubly positive charge state in each distribution. Fig. 3.37a shows a larger current for TAT2 compared to TAT1, where TAT2 and TAT1 refer to the TAT mechanism with and without extra factors, respectively. The same tendency is observed in Fig. 3.37b.

However, it is noted that numerical instability could be induced due to the modeling of energy dissipation. Suppose that energy dissipates at jth vacancy site at an intermediate step while it dissipates at ith vacancy site at the next step during Newton's iteration. According to Eqs. (3.2) and (3.18), this is due to the exchange of a higher energy level, or equivalently, an oscillating sign of $\widetilde{E}^q_{i,\text{filled}} - \widetilde{E}^q_{j,\text{filled}}$. When both vacancies are in the neutral charge state, a smooth electrostatic potential is expected, thus suppressing oscillations. However, the potential peak due to a positively charged vacancy disrupts this scheme. To see this, assume $q = 0$ and $q = +2$ for the ith and jth vacancies, respectively. Then, $\varphi_j > \varphi_i$ is expected given that these vacancies are not far away. With the positive ΔE^0_i, the difference between $E^0_{i,\text{filled}} - \Delta E^0_i - e\varphi_i$ and $E^{+2}_{j,\text{filled}} - e\varphi_j$ reduces, making it susceptible to oscillation. Since the problem results from the positive ΔE^0_i, a truncation $\Delta E^0_i = 0$ is assumed. More generally, the truncation is assumed for the vacancies in the neutral charge state. The Impact is examined by the resultant I-V characteristic curves denoted as TAT2t in Fig. 3.37. Insets show a negligible deviation ($\approx 1 \cdot 10^{-11}$ A) with and without ΔE^0_i at a small applied voltage regime. Since the detailed balance is discussed at equilibrium, the truncation method is assumed to reproduce the same physical process.

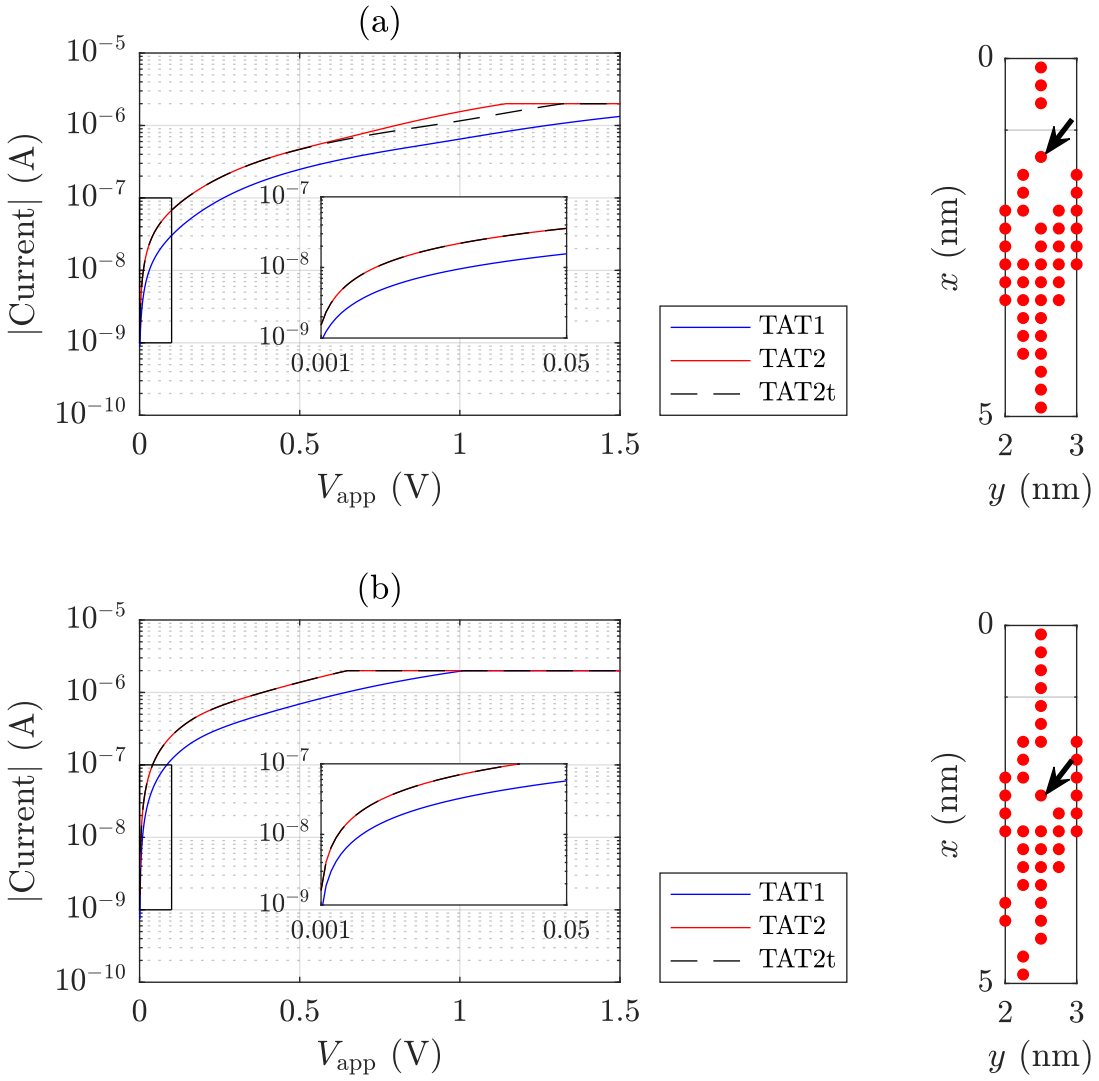

Figure 3.37: I-V characteristic curves and the corresponding vacancy distributions. The vacancy in the doubly positive charge state is indicated by an arrow.

3.7 Cycle-to-cycle variability for a larger current compliance

In Sec. 3.5.3, the reason for the failure of reproducing a large C2C variability at an elevated I_{cc} was discussed. Simply put, the variation of a single GB area can not lead to a reasonable C2C variability for a larger current compliance. However, the assumption of multiple GBs adopted by simulation works [130, 132] alleviates this problem.

Multiple GBs are assumed in the oxide layer. Properties of each GB are kept identical with generation barriers as the only exception for simplicity. More specifically, each GB possesses only one fragile spot for introducing oxygen vacancies and energy barriers differ by a small value of 0.02 eV. GB1 contains the spot with the lowest generation barrier while GB3 contains the one with the highest barrier. Fig. 3.38 shows the anisotropic modulations of channels that constitute a GB. According to the Arrhenius equation, the difference of 0.02 eV in activation energies yields a factor of 2.2 at room temperature, and thus CFs roughly grow in the order from GB1 to GB3. Impacts of the assumed differentiation in generation barriers are reserved for Sec. 3.8.

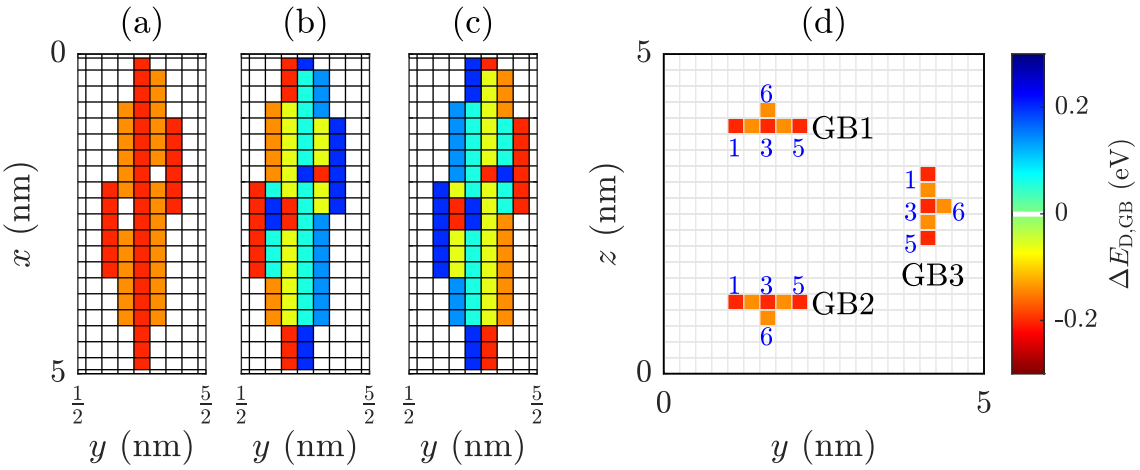

Figure 3.38: Anisotropic modulations for migrations in (a) $\pm x$-directions, (b) the positive y-direction, and (c) the negative y-direction. (d) Anisotropic modulations at $x = 2.3$ nm for visualizing GBs in the oxide layer.

For further discussion, indices 1 to 6 are introduced for 1D regions as labeled in Fig. 3.38d. Different from the assumption in the previous investigation, it is noteworthy

that only channel 3 extends from the TE to the BE. The other two strong channels now shrink in length and are assumed to reside in the middle height, aiming to avoid a persistent gap shown in Fig. 3.32 and Fig. 3.33. Since the vacancy migration along the x-direction is promoted even in a weaker channel, vacancies flowing into strong channels, e. g., channel 1 and channel 3, in a $\pm y$-direction might be hindered. Thus, weak channels, e. g., channel 2, are assumed to be broken into segments. Under this assumption, vacancy migration along the $\pm x$-direction is suppressed at the broken sites, and thus the horizontal migration into strong channels is possible. Lastly, the sixth 1D region extending from $x = 1.5\,\mathrm{nm}$ to $x = 3.0\,\mathrm{nm}$ is assumed to break a perfect 2D GB.

3.7.1 Switching cycles

To assess the statistical significance, multiple realizations of stochastic processes are simulated for each current compliance. The motivation is to demonstrate that simulation results are not a special case of the proposed vacancy migration scheme. Note that these results are not related to the D2D variability, which originates from the device variation.

In Fig. 3.39a, the product $V_{\mathrm{trans}} = I_{\mathrm{cc}} \cdot \mu_R$ is guided by the dashed line with $V_{\mathrm{trans}} = 2.0\,\mathrm{V}$. The simple relationship is seen to be a good approximation with the exception at $I_{\mathrm{cc}} = 1.5\,\mu\mathrm{A}$. It is noted that this value is larger than the reported ones, i. e., 0.4 V [95, 142] and 1.0 V [105, 106]. However, the LRS resistance of a small current compliance $I_{\mathrm{cc}} \leq 5\,\mu\mathrm{A}$ is mentioned to deviate from this empirical relationship (see Fig. 2.7b) [95]. Within a small interval of I_{cc}, an empirical value of a larger V_{trans} is observed from the measurement which qualitatively agrees with the simulation result. In addition, the median value and normalized deviation in multiple realizations for $I_{\mathrm{cc}} = 2\,\mu\mathrm{A}$ are approximately $1\,\mathrm{M}\Omega$ and 1.0, respectively. These values are comparable to the simulation result in Sec. 3.5. From a statistical perspective, the introduction of extra factors involved in electron transport and the assumed GB distributions do not undermine the established findings.

On the other hand, a bending tendency of the normalized deviation is reproduced in a small current compliance regime, i. e., $I_{\mathrm{cc}} \leq 3\,\mu\mathrm{A}$, as shown in Fig. 3.39b. Fig. 3.40 shows the CDF of LRS resistances over cycles for different current compliance values. It is seen that the C2C variability keeps large at an elevated value of current compliance. Therefore, the introduction of multiple GBs alleviates the problem discussed in Sec. 3.5.3. The details of multiple CFs are reserved for Sec. 3.7.4.

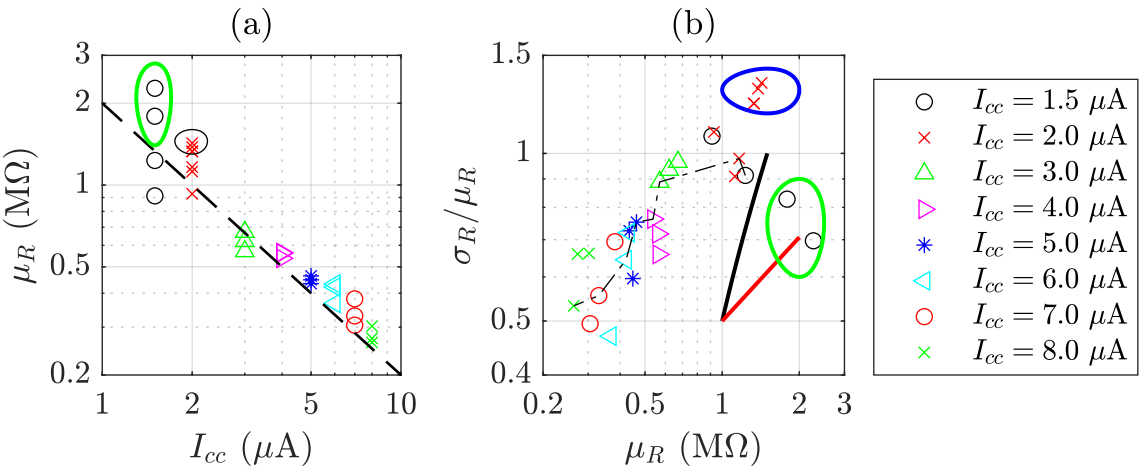

Figure 3.39: (a) Median resistance against current compliance and (b) normalized deviation against median resistance. The exponents of 0.5 and 1 are sketched in the red and black lines, respectively.

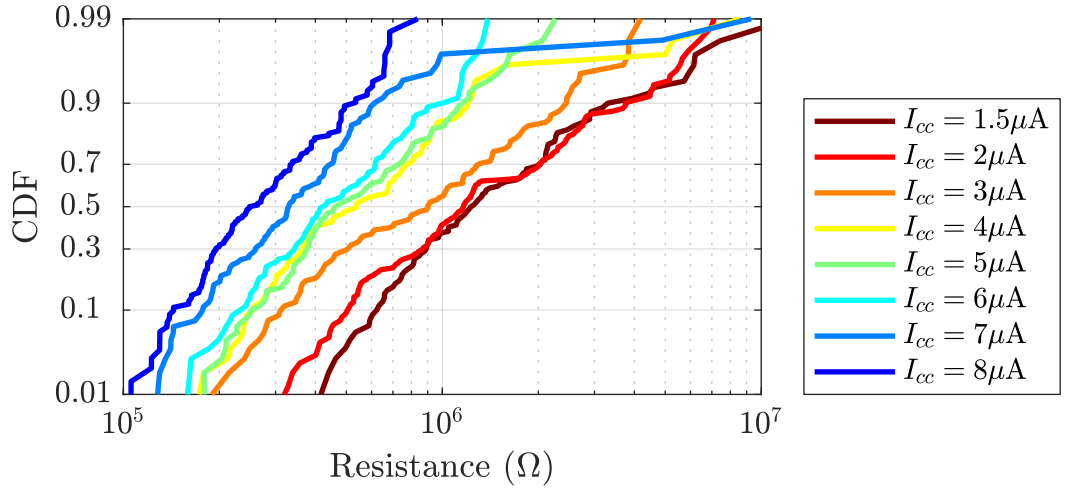

Figure 3.40: CDF plot for current compliance with the realizations indicated by the dashed-dotted line in Fig. 3.39b.

3.7.2 Single conductive path

Fig. 3.41 shows the current through each GB during the ramp-up stage of a FORMING process. It is seen that most current flows through one GB area for $I_{cc} \leq 2\,\mu A$, even though three GBs are available. In addition, an abrupt increase in the vacancy number is seen. Given that a 5-nm-thick oxide layer contains at most approximately twenty

vacancies in a perfect 1D chain, Fig. 3.41a implies the maximum temperature is not sufficiently high to break the 1D vacancy chain for $I_{cc} = 1.5\,\mu A$. This leads to the deceptive phenomenon reserved for Sec. 3.7.3.

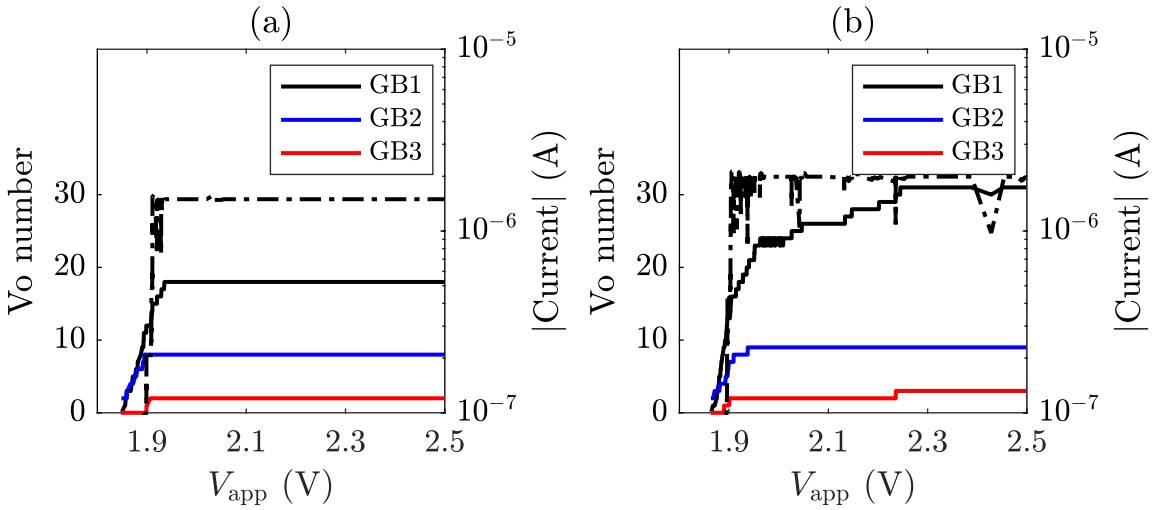

Figure 3.41: Vacancy numbers (left axis) and currents (right axis) through each GB for (a) $I_{cc} = 1.5\,\mu A$ and (b) $I_{cc} = 2\,\mu A$. The currents are shown as dashed-dotted lines, where those less than 10% of the total current are not plotted.

In contrast, Fig. 3.41b shows a significant increase in vacancy number for $I_{cc} = 2\,\mu A$, suggesting a different geometry of the CF. It is noted there are vacancies within GB2 and GB3. However, less than 10% of the total current flows through either GB2 or GB3 which makes the vacancies within these GBs negligible. Therefore, the previous scenario with only one GB still holds. To study the impact of GBs in a different geometry, the dynamics during a SET process is again investigated. Five stages are labeled in the inset of the upper panel of Fig. 3.42 and the evolution of vacancies is presented in Fig. 3.43. From Fig. 3.43a, it is seen the upward migration of vacancies which reduces the gap length to the TE. This corresponds to the increase in current observed in the top panel of Fig. 3.42. In Fig. 3.43b, the vacancies of channel 5 provide intermediate sites for electron transport, leading to a sufficiently high temperature to make vacancies migrate into the weak channel 4. At stage C, a gap emerges starting from approximately $x = 1.5\,nm$ to $x = 2.0\,nm$ as shown in Fig. 3.43c. The emergence prevents a further resistance drop which reflects on an approximately unchanged V_{cell} shown in the lower inset of Fig. 3.42. Meanwhile, the upward migration of channel 3 continues, eventually enclosing the gap

as shown in Fig. 3.43d. A significant V_{cell} drop followed by a temperature drop is seen in the bottom panel of Fig. 3.42. During this ramp-up process, the temperature is high enough to trigger the migration. Although this continuous migration creates a gap to the BE at stage E as shown in Fig. 3.43e, the V_{cell} is approximately unchanged.

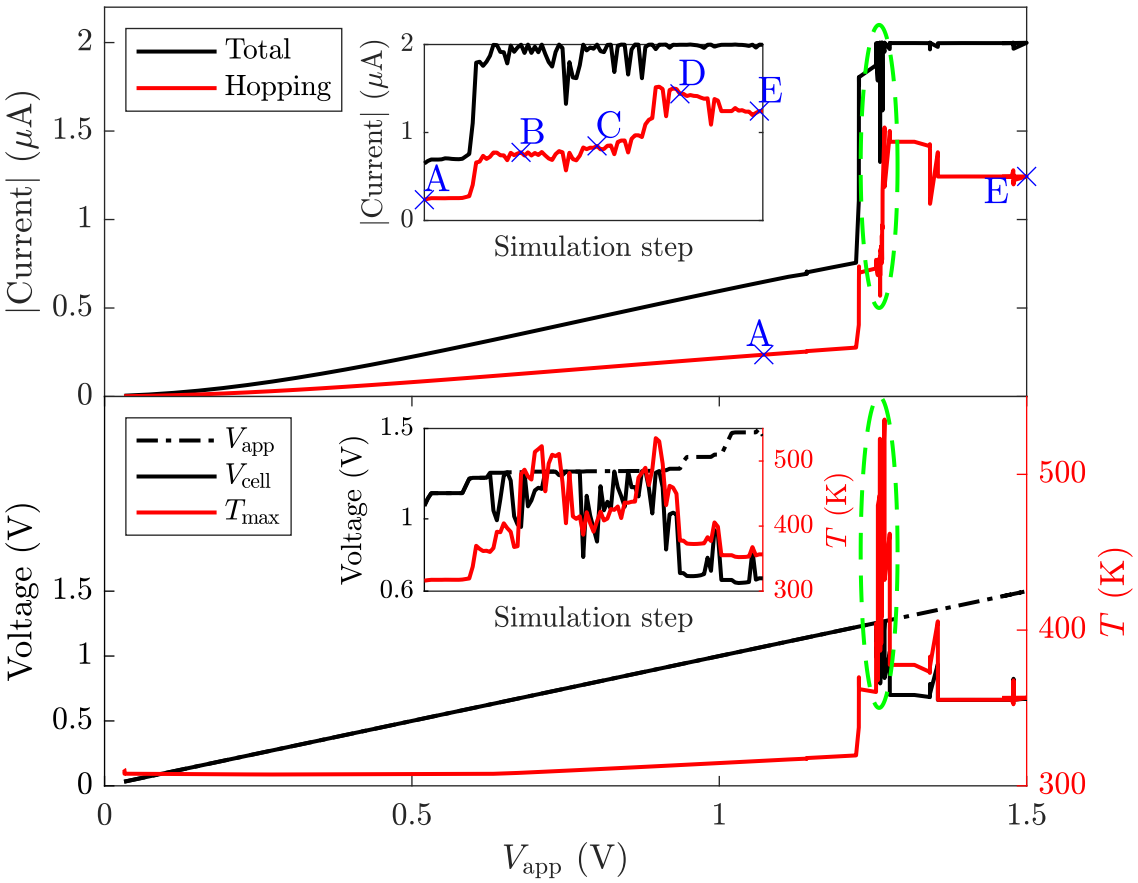

Figure 3.42: I-V characteristic curve in the top panel, and evolution of voltages and maximum temperature in the bottom panel. The regions where the hopping current is accounted for are indicated by green boxes in Fig. 3.43e.

This is discussed by investigating the evolution of a bottleneck of the charge transport path. Specifically, the large distance along the electron transport path occurs between the upper channel 3 and channel 1 at stage D. As this gap shrinks, a new gap near the BE develops. Consequently, the resistance is effectively compensated, leading to the unchanged V_{cell}. To verify this interpretation, hopping current by the MA hopping out of

channel 1 is plotted in the top panel of Fig. 3.42. From the figure, an increased hopping current is seen between stages C and D. This suggests that vacancies of channel 1 provide intermediate sites for electron transport. Therefore, channel 1 is a vacancy-rich region and the vacancies of it do not exchange with an existing CF. The underlying reason stems from the stronger anisotropic modulation in y-direction than that of channel 2. Vacancies in channel 1 prevent a gap in an intermediate period and improve the stability of the electrical performance. A similar argument applies to vacancies of upper channels 4 and 5 between stages B and C. The slight difference is that some vacancies are not fixed in space during the whole ramp-up process.

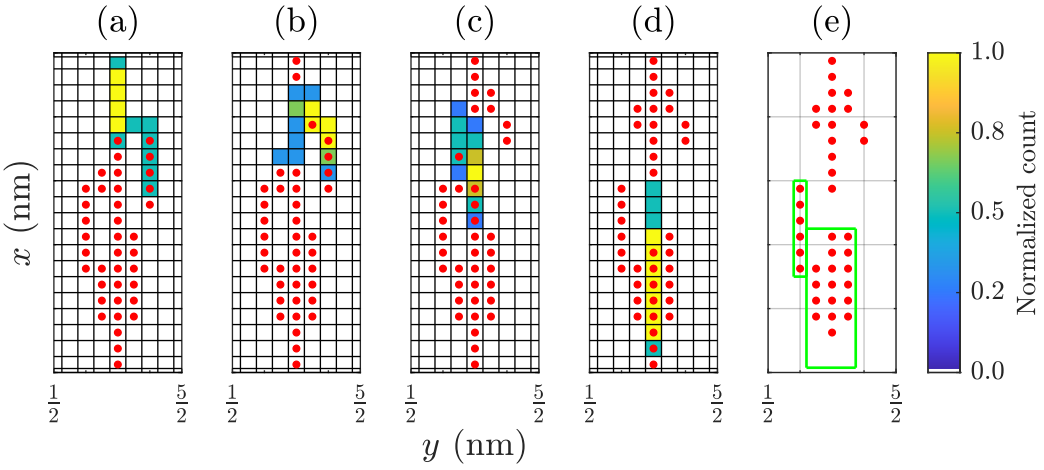

Figure 3.43: (a) - (d) Vacancy distributions superimposed to evolution towards the next stage, where stages are labeled in the upper panel of Fig. 3.42.

To this end, devices with two assumed GB distributions in the corresponding charge transport schemes are compared. From a statistical perspective, the analysis of C2C variability shows comparable median values and normalized deviations. Furthermore, one SET process is chosen to examine the migration pattern.

3.7.3 Deceptive success in a SET operation

In Fig. 3.39b, the bending tendency in normalized deviation is observed for $\mu_R > 0.5\,\mathrm{M\Omega}$, corresponding to $I_{\mathrm{cc}} \leq 3.0\,\mu\mathrm{A}$. However, it is noted that the LRS resistances could be stuck in a value without notice in a CDF plot, as the failure schemes discussed in Sec. 3.5.3. To see this, LRS resistance of a realization of $I_{\mathrm{cc}} = 1.5\,\mu\mathrm{A}$ in the green

circle in Fig. 3.39a is plotted. In Fig. 3.44, the resistance of 2.0 MΩ is highlighted by the green line. Resistances continuously over this value are seen from the twentieth to thirty-eighth cycles as well as after the seventieth cycle for $I_{cc} = 1.5\,\mu A$. In addition, Fig. 3.45a shows that a SET operation typically leads to a gap deep inside the oxide layer. By comparing to the statistical distribution from RESET processes, it suggests the upward vacancy migration is suppressed during SET processes. Thus, lower channel 2 and channel 4 can not supply their vacancies, leading to steadily high values of the LRS resistance. It is concluded that the device is stuck in an intermediate resistance state, where a resistance change is attributed to the gap around $x = 2.5\,nm$. Since the resistive switching does not occur during approximately half cycles, this data point should be excluded in the CDF plot and normalized deviation plot. This scenario is applicable to explain that a repetitive resistive switching under 2 µA on a HfO_2-based device is rarely, if any, reported. On the other hand, the device using $I_{cc} = 2.0\,\mu A$ in the CDF plot does not reproduce large resistances in two subsequent cycles. This is supported by the statistical vacancy distributions in Fig. 3.45b, where no distinct gap deep inside the oxide layer is observed.

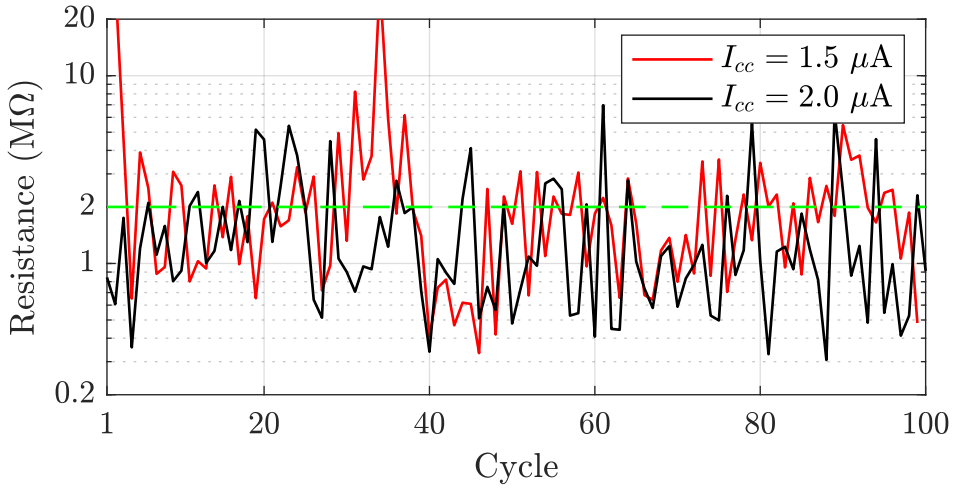

Figure 3.44: LRS resistances in sequential order for small current compliance.

Interestingly, the deviation arising from the persistent gap is approximately constant. From the data highlighted in the green circle in Fig. 3.39b, the deviation is estimated to be 1.5 MΩ. And, the deviation in the blue circle is estimated to be 1.7 MΩ. A constant deviation seems to be a good indicator of a failed device.

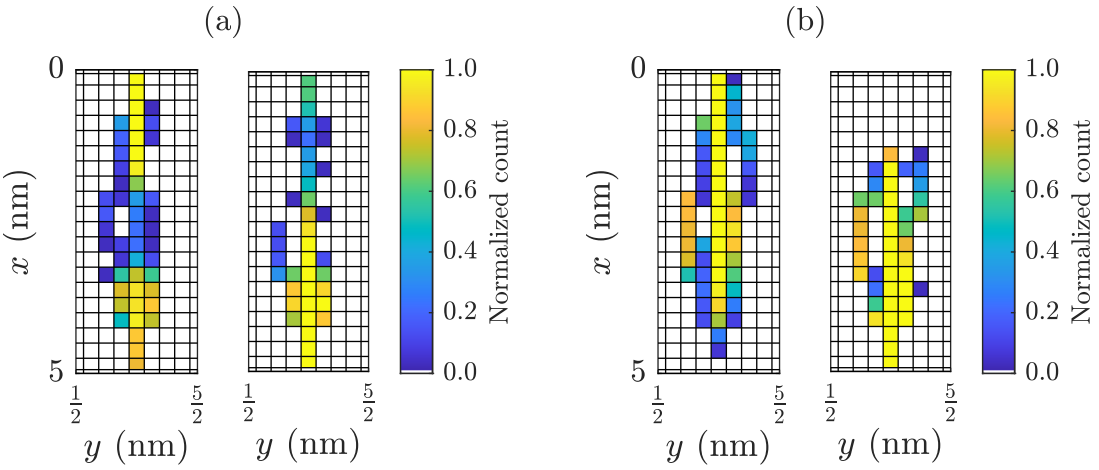

Figure 3.45: Statistical vacancy distributions of the last simulation steps of SET and RESET processes out of the last thirty cycles for (a) $I_{cc} = 1.5\,\mu A$ and (b) $I_{cc} = 2\,\mu A$.

3.7.4 Multiple conductive paths

It is seen that an adoption of $I_{cc} \geq 4\,\mu A$ leads to a distinct reduction in normalized deviation in Fig. 3.39b. By comparing Fig. 3.41b with Fig. 3.46b, it seems that the leveling-off in the normalized deviation exists when one GB contributes most current. As more than one GB becomes conductive, the normalized deviation decreases.

However, it is noted that the use of $I_{cc} = 3\,\mu A$, where two GBs are conductive as shown in Fig. 3.46a, also leads to the normalized deviation in a bending level. It is suspected whether the vacancy number within GB2 is limited, making this CF analogous to the one seen for $I_{cc} = 1.5\,\mu A$. In Fig. 3.47a, it is clearly seen that the vacancy number within the GB2 is less than that within the GB1. However, the number of approximately thirty is comparable to that within GB1 for $I_{cc} = 2\,\mu A$ (see Fig. 3.41b). Therefore, the scenario for $I_{cc} = 1.5\,\mu A$ does not apply to GB2 for $I_{cc} = 3\,\mu A$ in spite of fewer vacancies within GB2. A normalized deviation for $I_{cc} = 3\,\mu A$ in the bending level is interpreted as the outcome of one well-developed and one under-developed CFs. As a comparison, Fig. 3.47b shows the vacancy numbers for $I_{cc} = 4\,\mu A$, where a comparable vacancy number in two GBs is seen. Consequently, the normalized deviation is reduced. A similar finding is found in Fig. 3.47c for $I_{cc} = 8\,\mu A$ where three CFs are comparable.

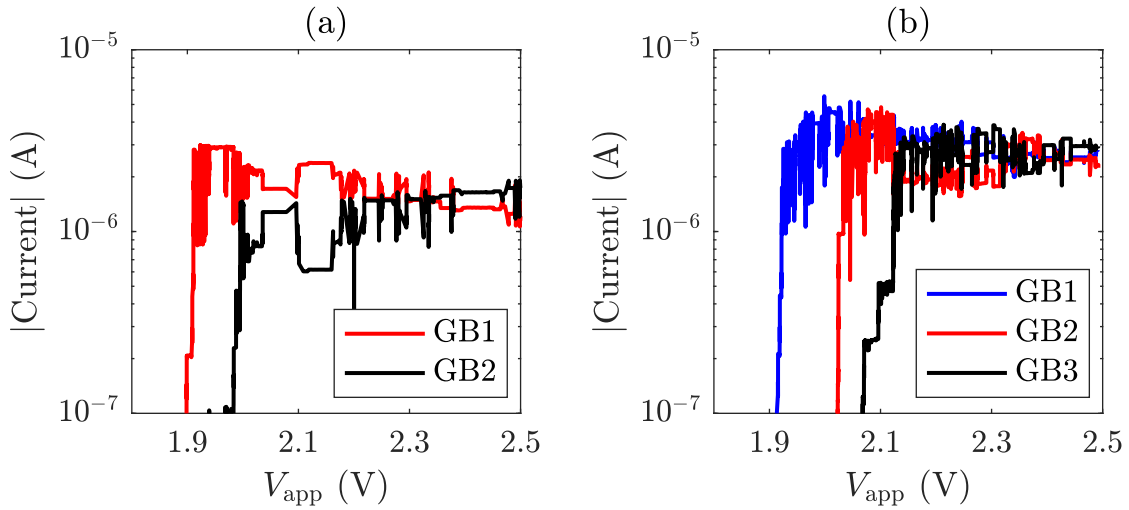

Figure 3.46: Current through each GB area for (a) $I_{cc} = 3\,\mu A$ and (b) $I_{cc} = 8\,\mu A$. The current through the third GB area is less than 10% of the other two GB and thus is not shown in (a).

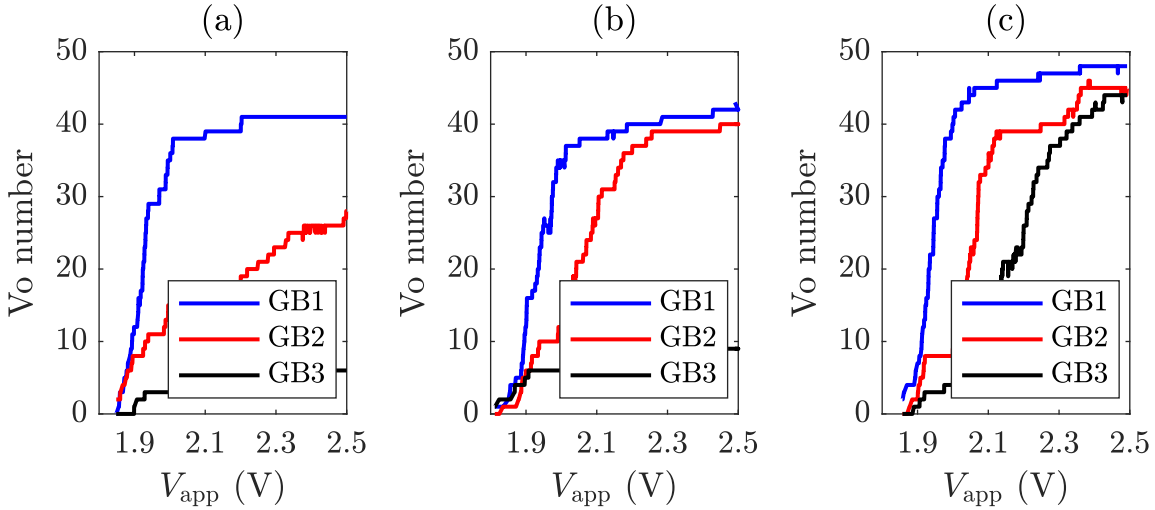

Figure 3.47: Vacancy numbers within each GB for (a) $I_{cc} = 3\,\mu A$, (b) $I_{cc} = 4\,\mu A$ and (c) $I_{cc} = 8\,\mu A$.

The observation of a reduced normalized deviation can be understood by the single CF picture at $I_{cc} = 2\,\mu A$ together with the error propagation analysis. Denote $R_{i,j}$ as the resistance of the ith CF at the jth cycle, and $\langle R_i \rangle$ as the average over cycles. Model

the effective resistance R by parallel connections of n CFs. With the error propagation and the assumption of non-correlated CFs, the average and the variance of an effective resistance take the form

$$\begin{aligned} \langle R \rangle &= \left(\sum_{i=1}^{n} \frac{1}{\langle R_i \rangle} \right)^{-1} \\ \sigma_R^2 &= \sum_{i=1}^{n} \left(\frac{\partial R}{\partial R_i} \right)^2 \sigma_i^2 \end{aligned}, \tag{3.20}$$

where the partial derivative terms are evaluated at the average resistance of each CF. With the parallel connection formulation, a partial derivative term yields

$$\frac{\partial R}{\partial R_i} = \frac{\partial}{\partial R_i} \left(\sum_{j=1}^{n} R_j^{-1} \right)^{-1} = \left(\sum_{j=1}^{n} R_j^{-1} \right)^{-2} \langle R_i \rangle^{-2} = \left(\frac{\langle R \rangle}{\langle R_i \rangle} \right)^2. \tag{3.21}$$

Herein, median values replace the roles of average values to represent the LRS resistances out of one hundred cycles. By applying the assumption that one CF develops after another when $I_{\mathrm{cc}} > 2\,\mu\mathrm{A}$, $\sigma_i = \sigma_0$ and $\mu_{R_i} = R_0$ are assumed for fully developed ones. Under this condition, Eq. 3.20 are parameterized based on n in the form

$$\begin{aligned} \mu_R &= R_0 \cdot n^{-1} \\ \sigma_R &= \sigma_0 \cdot n^{-3/2}. \end{aligned} \tag{3.22}$$

Therefore, the normalized deviation is proportional to the $n^{-1/2}$, or equivalently, $\mu_R^{1/2}$ as seen in experimental data. In the above argument, each CF is assumed to have sufficient vacancies. The situation of at least one CF having fewer vacancies is not discussed. However, a lower LRS resistance is attributed to more CFs, and thus the impact of one CF deviating from a fully developed one decreases in the presence of more CFs. The deviation leaves the power law in the large current compliance regime intact. However, this might change the normalized deviation for small current compliance where fewer CFs are available. The related discussion is reserved for Sec. 3.8.1.

The proposed analysis is different from previous works, where the normalized deviation is estimated by the shape variation [107] or by the vacancy number variation [106] of a single CF. The modeling of parallel connections provides another interpretation for the reduced variability in a large current compliance regime. It is reported from Ta_2O_5-based VCM devices, that uniformly spread-out CFs instead of a significantly localized CF can

lead to the resistive switching [155]. Multiple CFs are found in other filamentary type devices [205–207]. Recently, multiple filaments have been observed in HfO_2-based VCM devices with micrometer-level cross-sections using a novel photon emission microscopy technique [208, 209]. However, the spatial resolution of this technique is limited to above 300 x 300 nm, leaving it unclear whether devices with nanometer-scale cross-sections follow the same conclusion.

It is noted that the parallel connection should be based on CFs being non-interacting. Specifically, an electron flows through a single CF instead of hopping between different ones. As illustrated in Fig. 3.38, the separation between each pair of CFs is approximately 3 nm. This distance is expected to be sufficiently large to stop electron transport between two CFs. To justify this, the electron transport via the MA hopping mechanism of $I_{cc} = 8\,\mu A$ is plotted in Fig. 3.48. Specifically, the figure shows the hopping electron current from one CF to another. All of the three pairs are compared with values normalized by the total current. The hopping current is less than 0.2% of the current through the device, suggesting electron hopping between CFs is negligible.

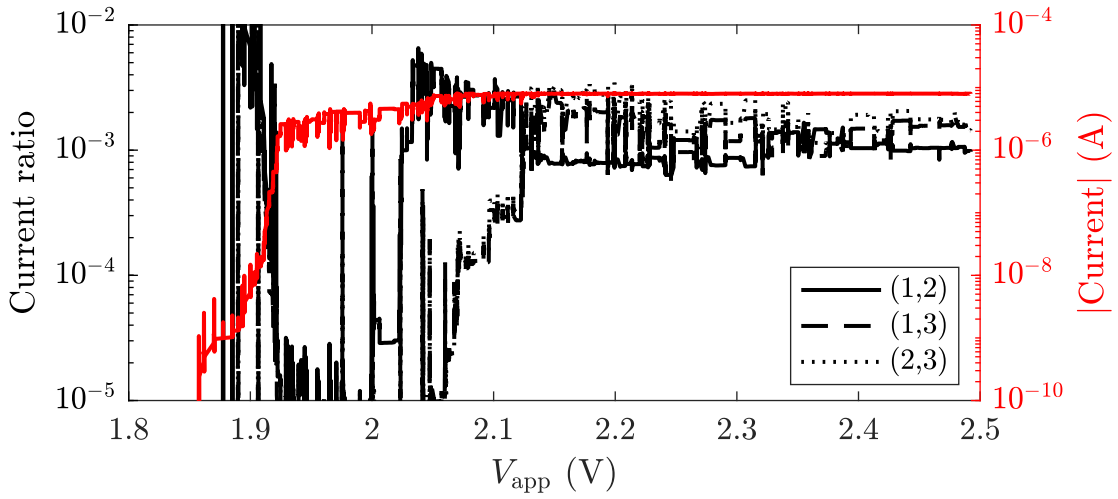

Figure 3.48: Electron hopping current between each pair of GBs normalized by the total current during the ramp-up stage of a FORMING process. The right axis shows the total current through the device.

3.8 Impact of grain boundary properties

The growth condition of CFs significantly influences the C2C variability. In turn, the pattern of one CF growing after another is related to the assumption of generation barriers of fragile spots. The impact of generation barriers on the transition from a leveling-off to a power-law regime is studied in Sec. 3.8.1.

On the other hand, the simulation reproduces a power law in $4\,\mu A \leq I_{cc} < 8\,\mu A$. Since the power-law regime is interpreted as the parallel connection of CFs, the limited number of assumed GBs is the reason for the cessation of the power-law relationship. With the adoption of $I_{cc} \geq 8\,\mu A$, simulations show an increase in either the normalized deviation or the median value, leading to discrepancies in experimental findings. This is analogous to the failure scheme earlier shown in Sec. 3.5.3 where the assumed GBs can not reproduce a sufficient number of CFs. Intuitively, the assumption of more GBs enlarges the current compliance window for the power law being held. Meanwhile, it does not change the result of a small I_{cc} given that one CF develops after another. However, both the generation of vacancies and the migration are important for forming CFs during a FORMING process. Consider a scenario in which a vacancy is generated but immediately stuck at the original site during the whole FORMING process, there would be no CF. The details are reserved for Sec. 3.8.2. In Sec. 3.8.3, an assumption is proposed to address the problem. Together with findings in Sec. 3.8.1, simulation results show the power law in a larger window with an intact bending regime. In the following discussion, a CF with an index refers to the CF within the GB with the same index.

3.8.1 Extended leveling-off window

According to simulation results, the existence of a fully-developed CF and an under-developed CF seems to yield its normalized deviation in the bending regime. In this section, the smaller variation in the generation barriers of 0.01 eV is adopted. According to the Arrhenius equation, this value yields a factor of approximately 1.5 at room temperature.

Fig. 3.49a shows the vacancy number for $I_{cc} = 4\,\mu A$, where the number within the GB2 is less than that of the GB1. The development of CFs of this R_1 device is different from the one with the same I_{cc} shown in Fig. 3.47b. Instead, this is analogous to that of $I_{cc} = 3\,\mu A$ in Fig. 3.47a. The roles of vacancies within the GB2 and GB3 are

justified by the contribution to current, as shown in Fig. 3.50a. It is seen that the GB2 but not both GBs is sufficient to contribute to the current. In return, this yields the $(\mu_R, \sigma_R/\mu_R) = (0.58\text{M}\Omega, 0.8)$, corresponding to a value in the bending value shown in Fig. 3.52b.

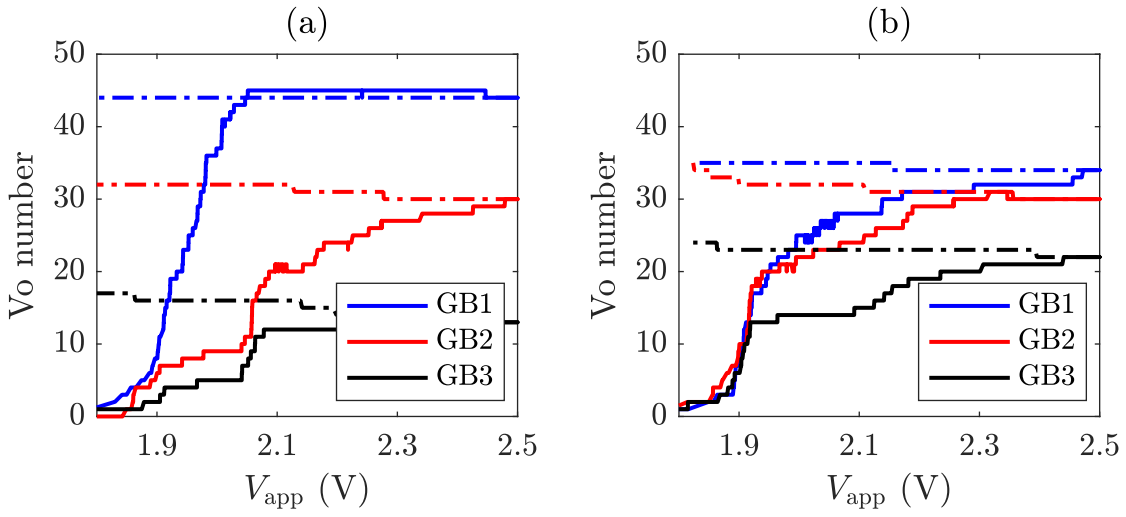

Figure 3.49: Vacancy numbers within each GB for $I_{\text{cc}} = 4\,\mu\text{A}$ during a FORMING process. Solid lines and dashed-dotted lines refer to the ramp-up and the ramp-down stages, respectively. The normalized deviation of (a) is higher than that of (b).

As a comparison, the vacancy number of the other realization denoted as the R_2 device is shown in Fig. 3.49b. Two comparable CFs emerges, which is qualitatively similar to that of Fig. 3.47b. The minor reduction in the vacancy number is compensated by the growth of CF3. Consequently, this leads to a reduced normalized deviation value, where the corresponding median resistance and the normalized deviation are $0.5\,\text{M}\Omega$ and 0.6, respectively. The current through each GB as identification of being a CF is plotted in Fig. 3.50b. The role of CFs during switching cycles is justified in Fig. 3.51, where statistical vacancy distributions out of one hundred SET processes are plotted. The CFs within the GB3 of both devices are not stable since the regions indicated by the purple arrows are not always closed. Moreover, a detailed comparison implies that the CF3 of R_1 device is less stable than that of R_2 device. From a statistical perspective, more current flowing through the CF3 of R_2 device compared to that of R_1 device.

Equivalently, less current flowing through the other two CFs of R_2 device. Therefore, maximum temperatures around CF1 and CF2 of R_2 device are expected to be lower compared to those of R_1 device. Consequently, vacancy migration is expected to be less frequent, leading to less C2C variability as observed.

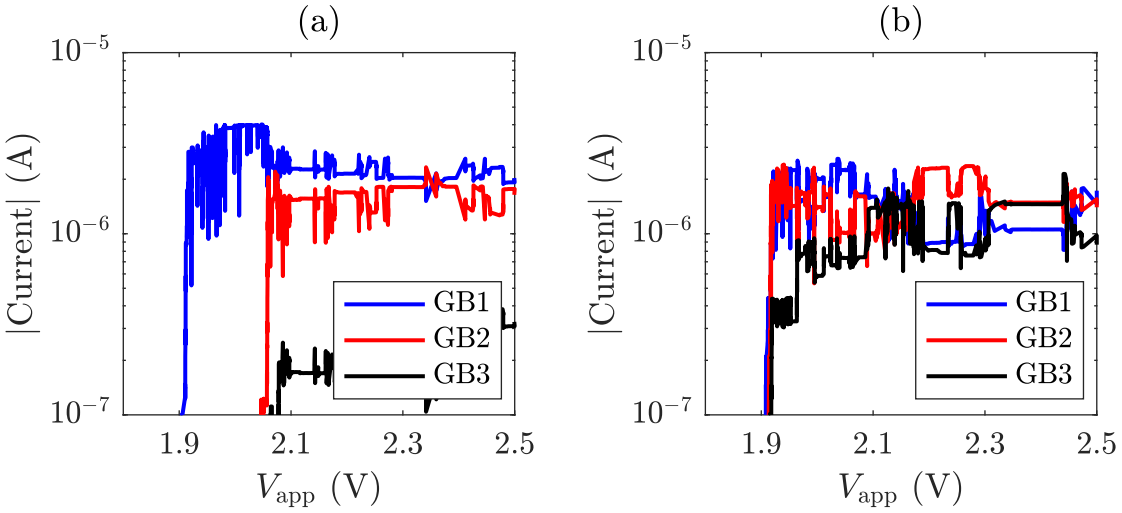

Figure 3.50: Current through each GB region for $I_{cc} = 4\,\mu A$ during the ramp-up stage of the FORMING process. The normalized deviation of (a) is higher than that of (b).

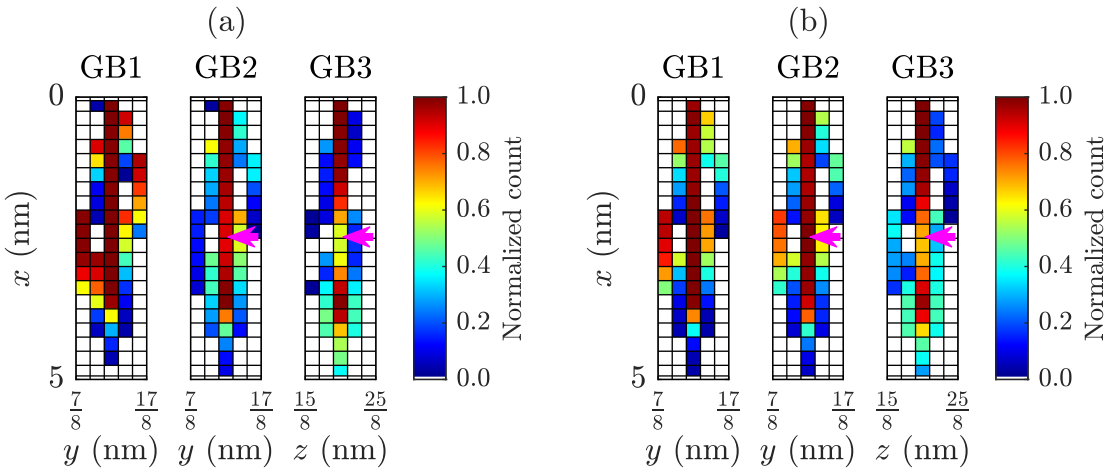

Figure 3.51: Statistical vacancy distributions out of one hundred SET processes for (a) R_1 and (b) R_2 devices.

Fig. 3.52b shows the normalized deviation against the median value of LRS resistances. The data highlighted by the blue and green circles are analogous to those in Fig. 3.39b. As discussed in Sec. 3.7.3, such data should not be accounted for. Moreover, simulations imply that the application of $I_{cc} = 1.5\,\mu A$ is not stable after multiple switching cycles. A persistent gap emerges in fifteen realizations while only one exception is found to possess a very close statistical behavior as the use of $I_{cc} = 2.0\,\mu A$. Therefore, the data of $I_{cc} = 1.5\,\mu A$ is not shown here. On the other hand, the leveling-off regime in the interval of $\mu_R > 0.4\,M\Omega$ is seen. Compared to the previous result, it starts from a lower value of median value. It is noted there is a group of data highlighted by the red circle that shows reduced normalized resistances. The reason is analogous to fragile spots with larger differentiation in the generation barriers, where three CFs emerge. Therefore, the assumption of a reduced variation in generation barriers extends the leveling-off regime while the transition to the power-law regime becomes abrupt.

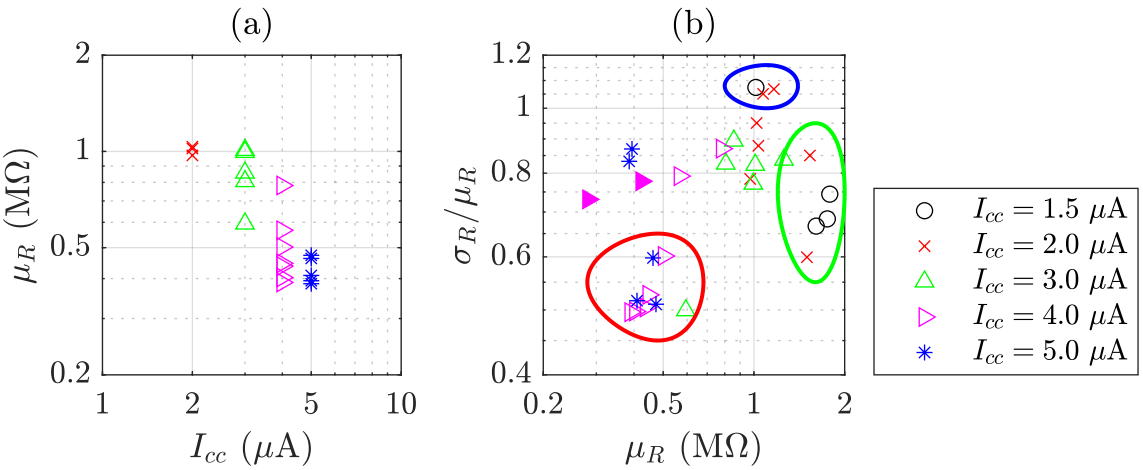

Figure 3.52: C2C statistics for multiple realizations. (a) Median LRS resistance against current compliance. (b) Normalized deviation against median resistance.

3.8.2 Failure in forming conductive filaments

During FORMING processes, a vacancy is generated at the BE interface and then migrates upwards. This is due to the migration being more probable than the generation. One critical factor for such a condition is a lower migration energy barrier compared to the generation energy barrier. This is expected for an V_O in the positive charge state in

most cases. However, the electric field plays a different role in Arrhenius equations that the migration and the generation processes follow. Specifically, the bond polarization term (see Eq. (2.23)) enhances the impact of an electric field in the generation process. This can lead to a generation process preferred over a migration process, even if an V_O is in the positive charge state. Consequently, once a vacancy exits the BE interface, a new V_O is generated. These two vacancies undergo a transition to the neutral charge state, making both less mobile and stuck at the bottom region. This is captured in Fig. 3.53a and 3.53b, where three short vacancy chains with merely two vacancies for each chain are see at the BE interface. These short chains are located within GB3, GB4, and GB5, corresponding to the spots with the three highest generation barriers of the device. Since the fragile spots are stuck, no more vacancies can be created, and thus the formation of CF at these GBs is suppressed. The assumed GBs are sketched in Fig. 3.54 and the generation barriers of the fragile spot within GB4 and GB5 are 0.14 eV and 0.16 eV higher than that of the lowest fragile spot, respectively.

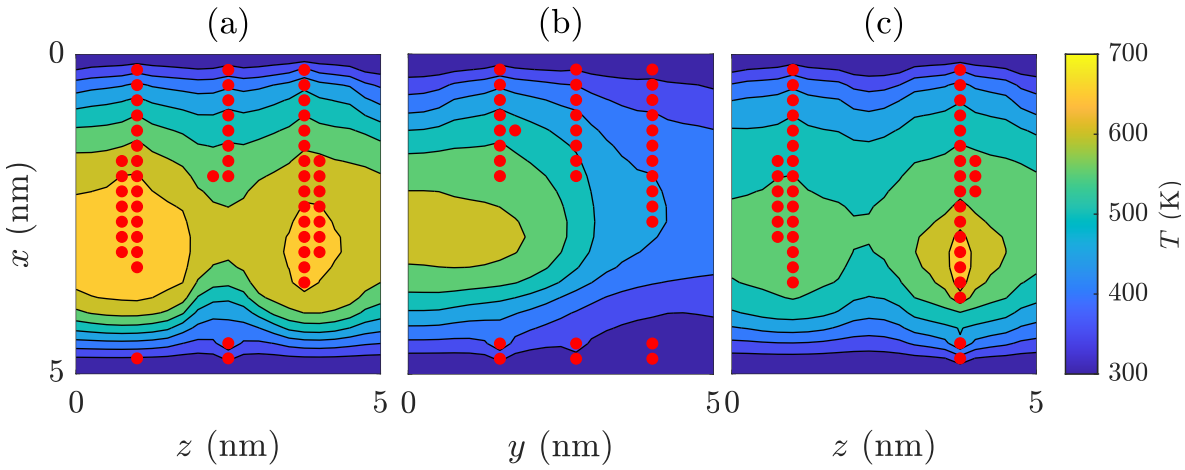

Figure 3.53: Vacancy and temperature distributions on (a) the $y = 1.5\,\text{nm}$ plane, (b) the $z = 2.5\,\text{nm}$ plane at $V_{\text{app}} = V_{\text{FORMING}}$ and (c) the $y = 1.5\,\text{nm}$ plane at $V_{\text{app}} = 2.1\,\text{V}$.

In contrast, the GB possessing the lowest generation barrier fragile spot is expected to develop a long CF. Hence, a sufficiently high temperature to activate the immobile vacancies is expected. Fig. 3.53c shows a short rupture of a vacancy chain at V_{app}=2.1 V of the early FORMING while the upward migration occurs in less than 1 ms.

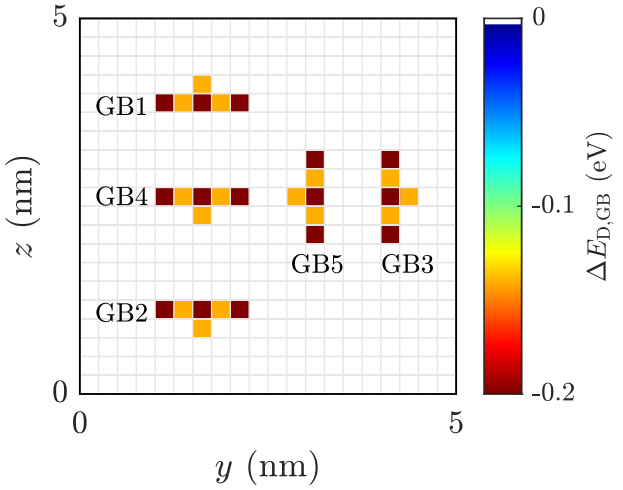

Figure 3.54: Anisotropic modulations at $x = 2.3\,\text{nm}$ for visualizing GBs.

It is noteworthy that the choice of the generation energy barrier, the migration energy barrier, and the dipole term together compose the scenario. In this work, the choice of parameters leads to an abrupt current increase at $V_{\text{app}} \approx 2.0\,\text{V}$, which is in good agreement with HfO_2-based VCM cells in the same thickness in measurements.

3.8.3 Interacting conductive filaments

Even the newly assumed GB possesses a higher generation barrier, the temperature around its fragile spot is raised due to the established CF as earlier shown in Fig. 3.53c. Depending on the raised temperature, CFs might grow within the newly assumed GB. To verify this idea, three GBs are assigned to the existing GBs by a short distance of 0.5 nm as sketched in Fig. 3.55d. The anisotropic modulations are assumed to be weaker since generation barriers of the newly assumed GBs are much lower than that of the first three GBs. Values are visualized in Fig. 3.55. The generation barriers of GB4, GB5 and GB6 are 0.14 eV, 0.16 eV and 0.18 eV higher than the lowest value, respectively.

Fig. 3.56 shows the statistical vacancy distributions within six GBs for $I_{\text{cc}} = 10\,\mu\text{A}$. Although there is no long CF grown in the newly assumed GBs, vacancies are seen to attach to the BE interface. A distinct offset in the data before and after $I_{\text{cc}} = 9\,\mu\text{A}$ is seen in Fig. 3.57a. Moreover, the reduction in normalized deviation extends to a wider I_{cc} window shown in Fig. 3.57b. An exponent of approximately 1 is found for the power-law regime. This tendency is comparable with measurements, where exponents in the

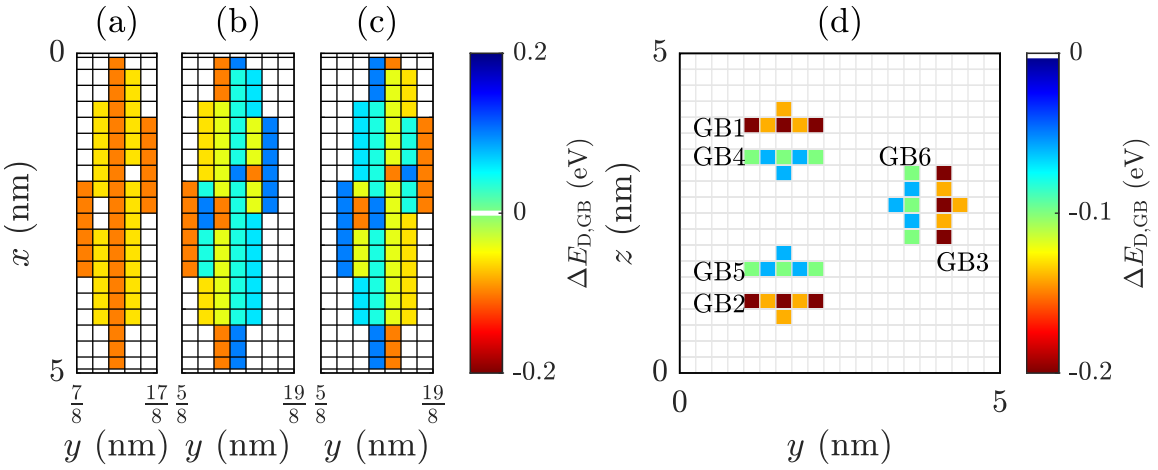

Figure 3.55: Anisotropic modulations of the newly assumed GBs for migrations in (a) $\pm x$-directions, (b) the positive y-direction, and (c) the negative y-direction. (d) The anisotropic modulations at $x = 2.3\,\mathrm{nm}$ for visualizing GBs in the oxide layer.

range of 0.5 to 1 have been reported. As a transition to the interacting CFs picture, two CFs being close to each other can be regarded as a single effective resistor. Note that the effective CF is composed of two components differentiating in vacancy mobilities. The CFs consisting of slow vacancies ensures that the bottom gap is enclosed when a SET process is done, as shown in Fig. 3.56. Therefore, the emergence of vacancies within GB4, GB5, and GB6 which couples to the long CFs is attributed to the transition in C2C variability. Notably, Fig. 3.57a successfully reproduces a transition in V_{trans} from 1.0 V to 0.5 V at around $I_{\mathrm{cc}} = 8\,\mu\mathrm{A}$, which is in a good agreement with measurements (see Fig. 2.7b).

On the other hand, the inclusion of new GBs has a minor impact on results of smaller current compliance, saying $I_{\mathrm{cc}} \leq 4\,\mu\mathrm{A}$. Interestingly, this is still due to the higher generation barrier. For example, the formation of CF2 is earlier than CF4. With CF2 completely grown, the current reaches its current compliance level, leading to a decrease in V_{cell}. Therefore, the growth of subsequent CFs is not expected. This ensures that the bending tendency is intact in the interacting CF picture.

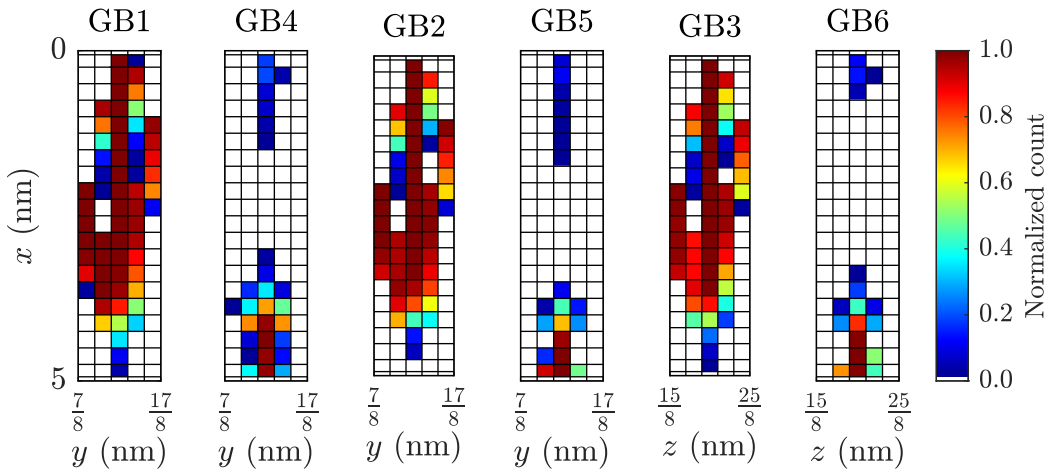

Figure 3.56: Statistical vacancy distributions out of one hundred SET processes for $I_{\mathrm{cc}} = 10\,\mu\mathrm{A}$.

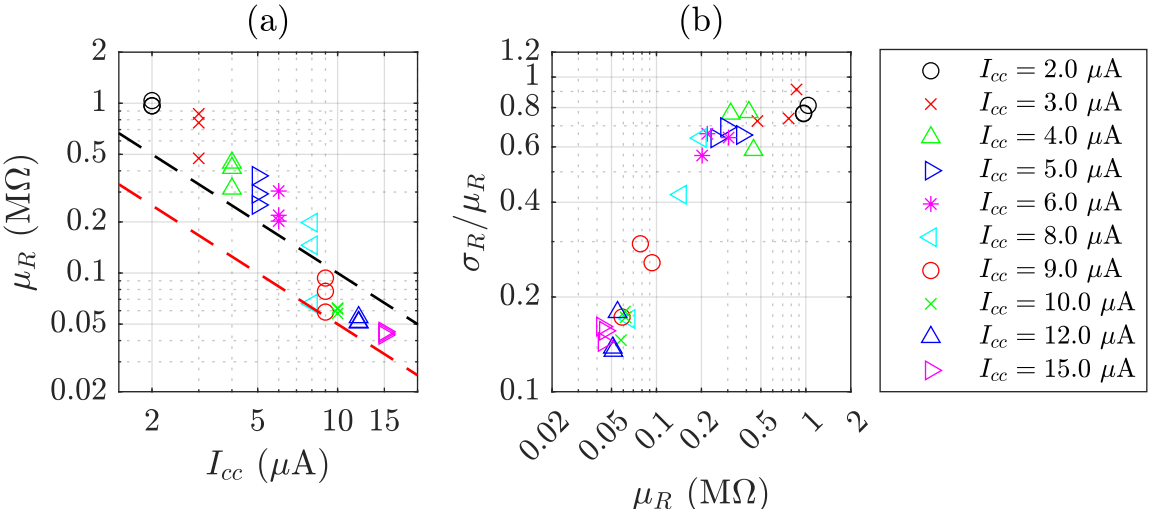

Figure 3.57: C2C statistics for multiple realizations. (a) Median LRS resistance versus current compliance. The black and red dashed-lines show $V_{\mathrm{trans}} = 1.0\,\mathrm{V}$ and $0.5\,\mathrm{V}$ for $\mu_R = V_{\mathrm{trans}}/I_{\mathrm{cc}}$, respectively. (b) Normalized deviation versus median resistance.

4 Conclusion and outlook

While existing knowledge of variability in a large current compliance regime has been built, a theoretical interpretation of the bending tendency in a small regime remains elusive. However, the application of small current compliance is crucial for the development of next-generation devices, particularly from a power consumption perspective.

To investigate variability, a device under the stress of a slow-varying applied voltage is simulated. 3D simulations start from a FORMING process instead of an assumed vacancy distribution. To this end, the spatio-temporal evolution of vacancies is discussed in a more general sense since an initial vacancy distribution convenient for the proposed interpretation is not employed. In addition, local structures are considered to introduce anisotropy in the migration path. The impact on assumptions of GBs consisting of 1D channels is investigated. Successful and unsuccessful SET operations are found to depend on the width of assumed GB. For repetitive resistive switching operations within one hundred cycles at $I_{cc} = 2\,\mu A$, the vacancy evolution involved in increasing and decreasing the LRS resistances is discussed in detail.

Furthermore, extra factors involved in electron transport are proposed and employed for the subsequent analysis. The discussion extends to current compliance within the wider regime, i. e., $1.5\,\mu A \leq I_{cc} \leq 8\,\mu A$. The data of $I_{cc} = 1.5\,\mu A$ seemingly shows a large C2C variability but it is identified to be deceptive data due to the device being stuck in an intermediate resistance state after tens of switching cycles. This is consistent with the fact that repetitive resistive switching is rarely observed for HfO_2-based VCM devices below the current compliance of $2\,\mu A$. For the use of larger current compliance, simulations successfully reproduce the power-law relationship and the leveling-off in the normalized deviation of LRS resistances. A new interpretation based on the parallel connection of CFs for the large current compliance regime is proposed. Lastly, both the bending and the power law are shown to depend on the assumed GBs.

C2C variability is discussed based on limited changes to the assumed GBs in this work. These are starting points for investigating D2D variability. A more general problem is

seeking an appropriate distribution for the assumed GBs. However, the assumption of GBs enables one to adopt DFT calculations for constructing a model to explore the dynamical process. It is noted that the HfO_2 layer may exhibit an amorphous crystal structure, and equilibrium conditions can be disrupted under the stress of a large voltage or elevated temperature. The entropy effects at a finite temperature is mainly unknown. These conditions are typically not captured by DFT calculations. The presented results are therefore obtained under specific assumptions. In this regard, the vacancy migration inside such local structures is a complicated process and a better understanding is critical for resolving the evolution of point defects. While a KMC model is a valuable tool for studying resistive switching, it is important to recognize its limitations. Other models may be more suitable for specific research questions.

5 Appendix

5.1 Cumulative distribution function of a normalized distribution

In literature, the notation $\varphi(x)$ is used for either the normal distribution or standard normal distribution. And, the same situation occurs for the corresponding CDF $\Phi(z)$. In this section, $\varphi(\tilde{x})$ and $\Phi(\tilde{z})$ refer to unnormalized forms while $\varphi(x)$ and $\Phi(z)$ refer to normalized forms. Mathematically speaking, the standard normal distribution is one of analytic probability density functions (PDFs) given by

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right). \tag{5.1}$$

However, the closed form of the corresponding CDF does not exist, and thus the probit function can not be expressed by elementary functions. In light of the wide usage of normal distributions, the corresponding CDF can be formulated by the error function

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp\bigl(-x^2\bigr)\,\mathrm{d}x. \tag{5.2}$$

With Gaussian integral $\int_{-\infty}^{\infty} \exp(-x^2)dx = \sqrt{\pi}$ and the upper incomplete gamma function

$$\Gamma(s, x) = \int_x^{\infty} t^{s-1} \exp(-t)\,\mathrm{d}t, \tag{5.3}$$

the error function in a variant form can be expressed by the upper incomplete gamma function

$$\operatorname{erf}(z) = 1 - \frac{1}{\sqrt{\pi}}\,\Gamma\left(\frac{1}{2}, z^2\right). \tag{5.4}$$

By applying the above expression to the standard normal distribution, the following form

$$\Phi(z) = \int_{-\infty}^{z} \varphi(x)\,\mathrm{d}x = \begin{cases} \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{z}{\sqrt{2}}\right)\right] & z \geq 0 \\ \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{-z}{\sqrt{2}}\right)\right] & z < 0 \end{cases}, \tag{5.5}$$

is often seen in the literature. The numerical values of the $\Phi(z)$ and corresponding probit function can be looked up and important values are labeled in Fig.5.1.

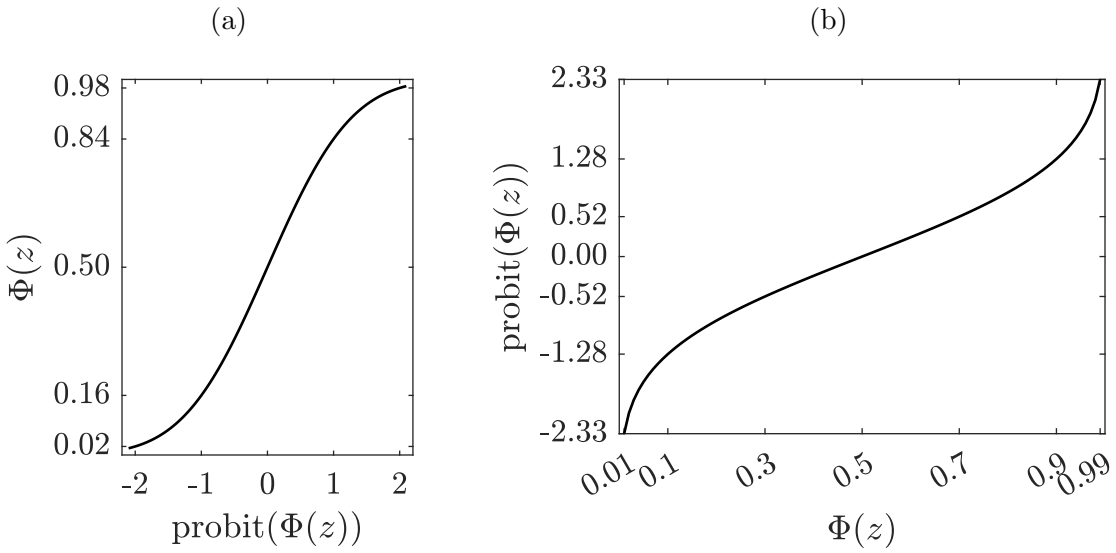

Figure 5.1: The conversion between the probit function and $\Phi(z)$.

5.2 Simulation parameters

Symbol	Value	Symbol	Value
$E_{\mathrm{G},0}$	$5.7\,\mathrm{eV}$	ν_0	$70\,\mathrm{THz}$
E_{R}	$1.1\,\mathrm{eV}$	ν_e	$70\,\mathrm{THz}$
$E^0_{\mathrm{D,iso}}$	$1.1\,\mathrm{eV}$	b	$8.7\,\mathrm{eÅ}$
$E^{+2}_{\mathrm{D,iso}}$	$0.7\,\mathrm{eV}$	a	$2.5\,\mathrm{Å}$
E^{+2}_{empty}	$-1.83\,\mathrm{eV}$	ε	$29\ \varepsilon_0$
E^{+2}_{filled}	$-1.97\,\mathrm{eV}$	m^*	$0.1m_0$
E^{0}_{filled}	$-2.11\,\mathrm{eV}$	R_0	$100\,\mathrm{THz}$
μ	$-1.9\,\mathrm{eV}$	$k_{\mathrm{th},0}$	$0.5\,\mathrm{WK^{-1}m^{-1}}$
$\lvert\beta\rvert$	$0.5\,\mathrm{Vs^{-1}}$	k_1	$8\cdot10^{-28}\,\mathrm{Wm^2K^{-1}}$
V_{FORMING}	$2.5\,\mathrm{V}$	k_2	$0\,\mathrm{Wm^5K^{-1}}$
V_{SET}	$1.5\,\mathrm{V}$	V_{RESET}	$-1.5\,\mathrm{V}$

Table 5.1: The parameters adopted in Sec. 3.5. Note that the energy levels are referred to the CBM.

Symbol	Value	Symbol	Value
$E_{\mathrm{G},0}$	$3.8\,\mathrm{eV}$	k_1	$6\cdot10^{-28}\,\mathrm{Wm^2K^{-1}}$
E_{R}	$2.1\,\mathrm{eV}$	k_2	$1.1\cdot10^{-55}\,\mathrm{Wm^5K^{-1}}$
b	$3.9\,\mathrm{eÅ}$		

Table 5.2: The parameters adopted in Sec. 3.7 and Sec. 3.8.

Bibliography

[1] C.-J. Chen, K. Z. Rushchanskii, and C. Jungemann, "Investigation of the Large Variability of HfO_2-Based Resistive Random Access Memory Devices with a Small Current Compliance by a Kinetic Monte Carlo Model," *physica status solidi (a)*, p. 2300403.

[2] T. W. Hickmott, "Low-Frequency Negative Resistance in Thin Anodic Oxide Films," *Journal of Applied Physics*, vol. 33, no. 9, pp. 2669–2682, 1962.

[3] J. Gibbons and W. Beadle, "Switching properties of thin Nio films," *Solid-State Electronics*, vol. 7, no. 11, pp. 785–790, 1964.

[4] A. Asamitsu, Y. Tomioka, H. Kuwahara, and Y. Tokura, "Current switching of resistive states in magnetoresistive manganites," *Nature*, vol. 388, no. 6637, pp. 50–52, 1997.

[5] A. Beck, J. G. Bednorz, C. Gerber, C. Rossel, and D. Widmer, "Reproducible switching effect in thin oxide films for memory applications," *Applied Physics Letters*, vol. 77, no. 1, pp. 139–141, 2000.

[6] W. Zhuang, W. Pan, B. Ulrich, J. Lee, L. Stecker, A. Burmaster, D. Evans, S. Hsu, M. Tajiri, A. Shimaoka, K. Inoue, T. Naka, N. Awaya, A. Sakiyama, Y. Wang, S. Liu, N. Wu, and A. Ignatiev, "Novel colossal magnetoresistive thin film nonvolatile resistance random access memory (RRAM)," in *Digest. International Electron Devices Meeting,*, 2002, pp. 193–196.

[7] I. Baek, M. Lee, S. Seo, M. Lee, D. Seo, D.-S. Suh, J. Park, S. Park, H. Kim, I. Yoo, U.-I. Chung, and J. Moon, "Highly scalable nonvolatile resistive memory using simple binary oxide driven by asymmetric unipolar voltage pulses," in *IEDM Technical Digest. IEEE International Electron Devices Meeting, 2004.*, 2004, pp. 587–590.

[8] A. Sawa, "Resistive switching in transition metal oxides," *Materials Today*, vol. 11, no. 6, pp. 28–36, 2008.

[9] R. Waser, R. Dittmann, G. Staikov, and K. Szot, "Redox-based resistive switching memories – nanoionic mechanisms, prospects, and challenges," *Advanced Materials*, vol. 21, pp. 2632–2663, 2009.

[10] C. D. Schuman, S. R. Kulkarni, M. Parsa, J. P. Mitchell, P. Date, and B. Kay, "Opportunities for neuromorphic computing algorithms and applications," *Nature Computational Science*, vol. 2, no. 1, pp. 10–19, Jan 2022.

[11] B. J. Shastri, A. N. Tait, T. Ferreira de Lima, M. A. Nahmias, H.-T. Peng, and P. R. Prucnal, *Principles of Neuromorphic Photonics*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2018, pp. 1–37.

[12] P. Chi, S. Li, C. Xu, T. Zhang, J. Zhao, Y. Liu, Y. Wang, and Y. Xie, "Prime: A novel processing-in-memory architecture for neural network computation in reram-based main memory," in *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*, 2016, pp. 27–39.

[13] W.-H. Chen, C. Dou, K.-X. Li, W.-Y. Lin, P.-Y. Li, J.-H. Huang, J.-H. Wang, W.-C. Wei, C.-X. Xue, Y.-C. Chiu, Y.-C. King, C.-J. Lin, R.-S. Liu, C.-C. Hsieh, K.-T. Tang, J. J. Yang, M.-S. Ho, and M.-F. Chang, "CMOS-integrated memristive non-volatile computing-in-memory for AI edge processors," *Nature Electronics*, vol. 2, no. 9, pp. 420–428, 2019.

[14] W. Wan, R. Kubendran, C. Schaefer, S. B. Eryilmaz, W. Zhang, D. Wu, S. Deiss, P. Raina, H. Qian, B. Gao, S. Joshi, H. Wu, H.-S. P. Wong, and G. Cauwenberghs, "A compute-in-memory chip based on resistive random-access memory," *Nature*, vol. 608, no. 7923, pp. 504–512, 2022.

[15] T. Gokmen and Y. Vlasov, "Acceleration of Deep Neural Network Training with Resistive Cross-Point Devices: Design Considerations," *Frontiers in Neuroscience*, vol. 10, 2016.

[16] D. V. Christensen, R. Dittmann, B. Linares-Barranco, A. Sebastian, M. L. Gallo, A. Redaelli, S. Slesazeck, T. Mikolajick, S. Spiga, S. Menzel, I. Valov, G. Milano, C. Ricciardi, S.-J. Liang, F. Miao, M. Lanza, T. J. Quill, S. T. Keene,

A. Salleo, J. Grollier, D. Marković, A. Mizrahi, P. Yao, J. J. Yang, G. Indiveri, J. P. Strachan, S. Datta, E. Vianello, A. Valentian, J. Feldmann, X. Li, W. H. P. Pernice, H. Bhaskaran, S. Furber, E. Neftci, F. Scherr, W. Maass, S. Ramaswamy, J. Tapson, P. Panda, Y. Kim, G. Tanaka, S. Thorpe, C. Bartolozzi, T. A. Cleland, C. Posch, S. Liu, G. Panuccio, M. Mahmud, A. N. Mazumder, M. Hosseini, T. Mohsenin, E. Donati, S. Tolu, R. Galeazzi, M. E. Christensen, S. Holm, D. Ielmini, and N. Pryds, "2022 roadmap on neuromorphic computing and engineering," *Neuromorphic Computing and Engineering*, vol. 2, no. 2, p. 022501, 2022.

[17] V. Milo, G. Malavena, C. Monzio Compagnoni, and D. Ielmini, "Memristive and cmos devices for neuromorphic computing," *Materials*, vol. 13, no. 1, 2020.

[18] S. Menzel, U. Böttger, M. Wimmer, and M. Salinga, "Physics of the switching kinetics in resistive memories," *Advanced Functional Materials*, vol. 25, no. 40, pp. 6306–6325, 2015.

[19] E. Abbaspour, S. Menzel, A. Hardtdegen, S. Hoffmann-Eifert, and C. Jungemann, "Kmc simulation of the electroforming, set and reset processes in redox-based resistive switching devices," *IEEE Transactions on Nanotechnology*, vol. 17, no. 6, pp. 1181–1188, 2018.

[20] S. Dirkmann, J. Kaiser, C. Wenger, and T. Mussenbrock, "Filament growth and resistive switching in hafnium oxide memristive devices," *ACS Applied Materials & Interfaces*, vol. 10, 2018.

[21] F. Zahoor, T. Z. A. Zulkifli, and F. A. Khanday, "Resistive random access memory (rram): an overview of materials, switching mechanism, performance, multilevel cell (mlc) storage, modeling, and applications," *Nanoscale Research Letters*, vol. 15, 2020.

[22] D. G. Pahinkar, P. Basnet, M. P. West, B. Zivasatienraj, A. Weidenbach, W. A. Doolittle, E. Vogel, and S. Graham, "Experimental and computational analysis of thermal environment in the operation of hfo2 memristors," *AIP Advances*, vol. 10, no. 3, p. 035127, 2020.

[23] R. Dittmann, S. Menzel, and R. Waser, "Nanoionic memristive phenomena in metal oxides: the valence change mechanism," *Advances in Physics*, vol. 70, no. 2, pp. 155–349, 2021.

[24] J. S. Moodera, L. R. Kinder, T. M. Wong, and R. Meservey, "Large Magnetoresistance at Room Temperature in Ferromagnetic Thin Film Tunnel Junctions," *Physical Review Letters*, vol. 74, pp. 3273–3276, 1995.

[25] J. Slonczewski, "Current-driven excitation of magnetic multilayers," *Journal of Magnetism and Magnetic Materials*, vol. 159, no. 1, pp. L1–L7, 1996.

[26] L. Berger, "Emission of spin waves by a magnetic multilayer traversed by a current," *Physical Review B*, vol. 54, pp. 9353–9358, 1996.

[27] D. Apalkov, B. Dieny, and J. M. Slaughter, "Magnetoresistive Random Access Memory," *Proceedings of the IEEE*, vol. 104, no. 10, pp. 1796–1830, 2016.

[28] L. Esaki, R. Laibowitz, and P. Stiles, "Polar switch," *IBM Tech. Discl. Bull*, vol. 13, no. 2161, p. 114, 1971.

[29] V. Garcia and M. Bibes, "Ferroelectric tunnel junctions for information storage and processing," *Nature Communications*, vol. 5, no. 1, p. 4289, 2014.

[30] Y. Li, Q. Qian, X. Zhu, Y. Li, M. Zhang, J. Li, C. Ma, H. Li, J. Lu, and Q. Zhang, "Recent advances in organic-based materials for resistive memory applications," *InfoMat*, vol. 2, no. 6, pp. 995–1033.

[31] W. Kwon, "Nano-electromechanical random access memory (RAM) devices," in *Advances in Non-volatile Memory and Storage Technology*, Y. Nishi, Ed. Woodhead Publishing, 2014, pp. 415–433.

[32] M. Wuttig and N. Yamada, "Phase-change materials for rewriteable data storage," *Nature Materials*, vol. 6, no. 11, pp. 824–832, 2007.

[33] D. Ielmini, R. Bruchhaus, and R. Waser, "Thermochemical resistive switching: materials, mechanisms, and scaling projections," *Phase Transitions*, vol. 84, no. 7, pp. 570–602, 2011.

[34] I. Valov, R. Waser, J. R. Jameson, and M. N. Kozicki, "Electrochemical metallization memories—fundamentals, applications, prospects," *Nanotechnology*, vol. 22, no. 25, p. 254003, 2011.

[35] F. Nardi, S. Larentis, S. Balatti, D. C. Gilmer, and D. Ielmini, "Resistive Switching by Voltage-Driven Ion Migration in Bipolar RRAM—Part I: Experimental Study," *IEEE Transactions on Electron Devices*, vol. 59, no. 9, pp. 2461–2467, 2012.

[36] B. Govoreanu, G. Kar, Y.-Y. Chen, V. Paraschiv, S. Kubicek, A. Fantini, I. Radu, L. Goux, S. Clima, R. Degraeve, N. Jossart, O. Richard, T. Vandeweyer, K. Seo, P. Hendrickx, G. Pourtois, H. Bender, L. Altimime, D. Wouters, J. Kittl, and M. Jurczak, "10×10nm2 Hf/HfOx crossbar resistive RAM with excellent performance, reliability and low-energy operation," in *2011 International Electron Devices Meeting*, 2011, pp. 31.6.1–31.6.4.

[37] Y. Zhang, G.-Q. Mao, X. Zhao, Y. Li, M. Zhang, Z. Wu, W. Wu, H.-J. Sun, Y. Guo, L. Wang, X. Zhang, Q. Liu, H. Lv, K.-H. Xue, G. Xu, X. S. Miao, S. Long, and M. Liu, "Evolution of the conductive filament system in HfO2-based memristors observed by direct atomic-scale imaging," *Nature Communications*, vol. 12, 2021.

[38] J. J. Yang, F. Miao, M. D. Pickett, D. A. A. Ohlberg, D. R. Stewart, C. N. Lau, and R. S. Williams, "The mechanism of electroforming of metal oxide memristive switches," *Nanotechnology*, vol. 20, no. 21, p. 215201, 2009.

[39] C. Lenser, A. Kuzmin, J. Purans, A. Kalinko, R. Waser, and R. Dittmann, "Probing the oxygen vacancy distribution in resistive switching Fe-SrTiO3 metal-insulator-metal-structures by micro-x ray absorption near-edge structure," *Journal of Applied Physics*, vol. 111, no. 7, p. 076101, 2012.

[40] F. Miao, J. P. Strachan, J. J. Yang, M.-X. Zhang, I. Goldfarb, A. C. Torrezan, P. Eschbach, R. D. Kelley, G. Medeiros-Ribeiro, and R. S. Williams, "Anatomy of a nanoscale conduction channel reveals the mechanism of a high-performance memristor," *Advanced Materials*, vol. 23, no. 47, pp. 5633–5640, 2011.

[41] A. Kindsmüller, C. Schmitz, C. Wiemann, K. Skaja, D. J. Wouters, R. Waser, C. M. Schneider, and R. Dittmann, "Valence change detection in memristive oxide

based heterostructure cells by hard X-ray photoelectron emission spectroscopy," *APL Materials*, vol. 6, no. 4, p. 046106, 2018.

[42] Y. Ma, D. Li, A. A. Herzing, D. A. Cullen, B. T. Sneed, K. L. More, N. T. Nuhfer, J. A. Bain, and M. Skowronski, "Formation of the conducting filament in taox-resistive switching devices by thermal-gradient-induced cation accumulation," *ACS Applied Materials & Interfaces*, vol. 10, no. 27, pp. 23 187–23 197, 2018.

[43] Z. Wei, T. Takagi, Y. Kanzawa, Y. Katoh, T. Ninomiya, K. Kawai, S. Muraoka, S. Mitani, K. Katayama, S. Fujii, R. Miyanaga, Y. Kawashima, T. Mikawa, K. Shimakawa, and K. Aono, "Demonstration of high-density ReRAM ensuring 10-year retention at 85°C based on a newly developed reliability model," in *2011 International Electron Devices Meeting*, 2011, pp. 31.4.1–31.4.4.

[44] Y. Ma, P. P. Yeoh, L. Shen, J. M. Goodwill, J. A. Bain, and M. Skowronski, "Evolution of the conductive filament with cycling in TaOx-based resistive switching devices," *Journal of Applied Physics*, vol. 128, no. 19, p. 194501, 2020.

[45] H. Du, C.-L. Jia, A. Koehl, J. Barthel, R. Dittmann, R. Waser, and J. Mayer, "Nanosized Conducting Filaments Formed by Atomic-Scale Defects in Redox-Based Resistive Switching Memories," *Chemistry of Materials*, vol. 29, no. 7, pp. 3164–3173, 2017.

[46] U. Celano, Y. Yin Chen, D. J. Wouters, G. Groeseneken, M. Jurczak, and W. Vandervorst, "Filament observation in metal-oxide resistive switching devices," *Applied Physics Letters*, vol. 102, no. 12, p. 121602, 03 2013.

[47] X. P. Wang, Y. Y. Chen, L. Pantisano, L. Goux, M. Jurczak, G. Groeseneken, and D. J. Wouters, "Effect of anodic interface layers on the unipolar switching of HfO2-based resistive RAM," in *Proceedings of 2010 International Symposium on VLSI Technology, System and Application*, 2010, pp. 140–141.

[48] C. Cagli, J. Buckley, V. Jousseaume, T. Cabout, A. Salaun, H. Grampeix, J. F. Nodin, H. Feldis, A. Persico, J. Cluzel, P. Lorenzi, L. Massari, R. Rao, F. Irrera, F. Aussenac, C. Carabasse, M. Coue, P. Calka, E. Martinez, L. Perniola, P. Blaise, Z. Fang, Y. H. Yu, G. Ghibaudo, D. Deleruyelle, M. Bocquet, C. Müller, A. Padovani, O. Pirrotta, L. Vandelli, L. Larcher, G. Reimbold, and B. de Salvo,

"Experimental and theoretical study of electrode effects in HfO2 based RRAM," in *2011 International Electron Devices Meeting*, 2011, pp. 28.7.1–28.7.4.

[49] A. Padovani, L. Larcher, P. Padovani, C. Cagli, and B. De Salvo, "Understanding the Role of the Ti Metal Electrode on the Forming of HfO2-Based RRAMs," in *2012 4th IEEE International Memory Workshop*, 2012, pp. 1–4.

[50] D.-Y. Cho, M. Luebben, S. Wiefels, K.-S. Lee, and I. Valov, "Interfacial metal–oxide interactions in resistive switching memories," *ACS Applied Materials & Interfaces*, vol. 9, no. 22, pp. 19 287–19 295, 2017.

[51] Y. Y. Chen, G. Pourtois, S. Clima, L. Goux, B. Govoreanu, A. Fantini, R. Degreave, G. S. Kar, G. Groeseneken, D. J. Wouters, and M. Jurczak, "Hf cap thickness dependence in bipolar-switching TiN\HfO2\Hf\TiN rram device," *ECS Transactions*, vol. 50, no. 34, p. 3, 2013.

[52] M. Sowinska, T. Bertaud, D. Walczyk, S. Thiess, M. A. Schubert, M. Lukosius, W. Drube, C. Walczyk, and T. Schroeder, "Hard x-ray photoelectron spectroscopy study of the electroforming in Ti/HfO2-based resistive switching structures," *Applied Physics Letters*, vol. 100, no. 23, p. 233509, 2012.

[53] A. Kindsmüller, A. Schönhals, S. Menzel, R. Dittmann, R. Waser, and D. J. Wouters, "The influence of interfacial (sub)oxide layers on the properties of pristine resistive switching devices," in *2018 Non-Volatile Memory Technology Symposium (NVMTS)*, 2018, pp. 1–4.

[54] T. Bertaud, M. Sowinska, D. Walczyk, S. Thiess, A. Gloskovskii, C. Walczyk, and T. Schroeder, "In-operando and non-destructive analysis of the resistive switching in the Ti/HfO2/TiN-based system by hard x-ray photoelectron spectroscopy," *Applied Physics Letters*, vol. 101, no. 14, p. 143501, 2012.

[55] C. Bengel, F. Cüppers, M. Payvand, R. Dittmann, R. Waser, S. Hoffmann-Eifert, and S. Menzel, "Utilizing the switching stochasticity of HfO_2/TiOx-based reram devices and the concept of multiple device synapses for the classification of overlapping and noisy patterns," *Frontiers in Neuroscience*, vol. 15, 2021.

[56] A. Kindsmüller, A. Meledin, J. Mayer, R. Waser, and D. J. Wouters, "On the role of the metal oxide/reactive electrode interface during the forming procedure of valence change reram devices," *Nanoscale*, vol. 11, pp. 18 201–18 208, 2019.

[57] S. Long, L. Perniola, C. Cagli, J. Buckley, X. Lian, E. Miranda, F. Pan, M. Liu, and J. Suñé, "Voltage and Power-Controlled Regimes in the Progressive Unipolar RESET Transition of HfO2-Based RRAM," *Scientific Reports*, vol. 3, no. 1, p. 2929, 2013.

[58] S. Yu and H.-S. P. Wong, "A Phenomenological Model for the Reset Mechanism of Metal Oxide RRAM," *IEEE Electron Device Letters*, vol. 31, no. 12, pp. 1455–1457, 2010.

[59] B. Gao, J. F. Kang, Y. S. Chen, F. F. Zhang, B. Chen, P. Huang, L. F. Liu, X. Y. Liu, Y. Y. Wang, X. A. Tran, Z. R. Wang, H. Y. Yu, and A. Chin, "Oxide-based rram: Unified microscopic principle for both unipolar and bipolar switching," in *2011 International Electron Devices Meeting*, 2011, pp. 17.4.1–17.4.4.

[60] F. Nardi, S. Balatti, S. Larentis, D. Gilmer, and D. Ielmini, "Complementary switching in oxide-based bipolar resistive-switching random memory," *IEEE Transactions on Electron Devices*, vol. 60, pp. 70–77, 2013.

[61] R. Waser, *Nanoelectronics and Information Technology*, 3rd ed. Wiley-VCH, 2012.

[62] H. Chee, Y. Kok, N. Thulasiraman, and H. Almurib, "Sense Amplifier for ReRAM-based Crossbar Memory Systems," *International Journal of Electronics Letters*, vol. 11, 2022.

[63] T. Na, B. Song, J. P. Kim, S. H. Kang, and S.-O. Jung, "Offset-Canceling Current-Sampling Sense Amplifier for Resistive Nonvolatile Memory in 65 nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 2, pp. 496–504, 2017.

[64] H. Y. Lee, Y. S. Chen, P. S. Chen, T. Y. Wu, F. Chen, C. C. Wang, P. J. Tzeng, M.-J. Tsai, and C. Lien, "Low-Power and Nanosecond Switching in Robust Hafnium Oxide Resistive Memory With a Thin Ti Cap," *IEEE Electron Device Letters*, vol. 31, no. 1, pp. 44–46, 2010.

[65] H. Y. Lee, Y. S. Chen, P. S. Chen, P. Y. Gu, Y. Y. Hsu, S. M. Wang, W. H. Liu, C. H. Tsai, S. S. Sheu, P. C. Chiang, W. P. Lin, C. H. Lin, W. S. Chen, F. T. Chen, C. H. Lien, and M.-J. Tsai, "Evidence and solution of over-reset problem for hfox based resistive memory with sub-ns switching speed and high endurance," in *2010 International Electron Devices Meeting*, 2010, pp. 19.7.1–19.7.4.

[66] A. C. Torrezan, J. P. Strachan, G. Medeiros-Ribeiro, and R. S. Williams, "Sub-nanosecond switching of a tantalum oxide memristor," *Nanotechnology*, vol. 22, no. 48, p. 485203, 2011.

[67] M. Abedin, N. Gong, K. Beckmann, M. Liehr, I. Saraf, O. Van der Straten, T. Ando, and N. Cady, "Material to system-level benchmarking of CMOS-integrated RRAM with ultra-fast switching for low power on-chip learning," *Scientific Reports*, vol. 13, no. 1, p. 14963, 2023.

[68] H. Y. Lee, Y. S. Chen, P. S. Chen, P. Y. Gu, Y. Y. Hsu, S. M. Wang, W. H. Liu, C. H. Tsai, S. S. Sheu, P. C. Chiang, W. P. Lin, C. H. Lin, W. S. Chen, F. T. Chen, C. H. Lien, and M.-J. Tsai, "Evidence and solution of over-RESET problem for HfOX based resistive memory with sub-ns switching speed and high endurance," in *2010 International Electron Devices Meeting*, 2010, pp. 19.7.1–19.7.4.

[69] Y. Park, J. Lee, S. S. Cho, G. Jin, and E. Jung, "Scaling and reliability of NAND flash devices," in *2014 IEEE International Reliability Physics Symposium*, 2014, pp. 2E.1.1–2E.1.4.

[70] Y. Y. Chen, B. Govoreanu, L. Goux, R. Degraeve, A. Fantini, G. S. Kar, D. J. Wouters, G. Groeseneken, J. A. Kittl, M. Jurczak, and L. Altimime, "Balancing SET/RESET Pulse for 10^{10} Endurance in HfO2 1T1R Bipolar RRAM," *IEEE Transactions on Electron Devices*, vol. 59, no. 12, pp. 3243–3249, 2012.

[71] A. Fantini, L. Goux, A. Redolfi, R. Degraeve, G. Kar, Y. Chen, and M. Jurczak, "Lateral and vertical scaling impact on statistical performances and reliability of 10nm TiN/Hf(Al)O/Hf/TiN RRAM devices," in *2014 Symposium on VLSI Technology (VLSI-Technology): Digest of Technical Papers*, 2014, pp. 1–2.

[72] S.-M. Jung, J. Jang, W. Cho, H. Cho, J. Jeong, Y. Chang, J. Kim, Y. Rah, Y. Son, J. Park, M.-S. Song, K.-H. Kim, J.-S. Lim, and K. Kim, "Three Dimensionally

Stacked NAND Flash Memory Technology Using Stacking Single Crystal Si Layers on ILD and TANOS Structure for Beyond 30nm Node," in *2006 International Electron Devices Meeting*, 2006, pp. 1–4.

[73] M. Kim, S. W. Yun, J. Park, H. K. Park, J. Lee, Y. S. Kim, D. Na, S. Choi, Y. Song, J. Lee, H. Yoon, K. Lee, B. Jeong, S. Kim, J. Park, C. A. Lee, J. Lee, J. Lee, J. Y. Chun, J. Jang, Y. Yang, S. H. Moon, M. Choi, W. Kim, J. Kim, S. Yoon, P. Kwak, M. Lee, R. Song, S. Kim, C. Yoon, D. Kang, J.-Y. Lee, and J. Song, "A 1Tb 3b/Cell 8th-Generation 3D-NAND Flash Memory with 164MB/s Write Throughput and a 2.4Gb/s Interface," in *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 65, 2022, pp. 136–137.

[74] B. Kim, S. Lee, B. Hah, K. Park, Y. Park, K. Jo, Y. Noh, H. Seol, H. Lee, J. Shin, S. Choi, Y. Jung, S. Ahn, Y. Park, S. Oh, M. Kim, S. Kim, H. Park, T. Lee, H. Won, M. Kim, C. Koo, Y. Choi, S. Choi, S. Park, D. Youn, J. Lim, W. Park, H. Hur, K. Kwean, H. Choi, W. Jeong, S. Chung, J. Choi, and S. Cha, "28.2 A High-Performance 1Tb 3b/Cell 3D-NAND Flash with a 194MB/s Write Throughput on over 300 Layers," in *2023 IEEE International Solid-State Circuits Conference (ISSCC)*, 2023, pp. 27–29.

[75] S. Yu, H.-Y. Chen, B. Gao, J. Kang, and H.-S. P. Wong, "HfOx-Based Vertical Resistive Switching Random Access Memory Suitable for Bit-Cost-Effective Three-Dimensional Cross-Point Architecture," *ACS Nano*, vol. 7, no. 3, pp. 2320–2325, 2013.

[76] P. Lin, C. Li, Z. Wang, Y. Li, H. Jiang, W. Song, M. Rao, Y. Zhuo, N. K. Upadhyay, M. Barnell, Q. Wu, J. J. Yang, and Q. Xia, "Three-dimensional memristor circuits as complex neural networks," *Nature Electronics*, vol. 3, no. 4, pp. 225–232, 2020.

[77] G. D. Wilk, R. M. Wallace, and J. M. Anthony, "High-κ gate dielectrics: Current status and materials properties considerations," *Journal of Applied Physics*, vol. 89, no. 10, pp. 5243–5275, 2001.

[78] Z. Wei, Y. Kanzawa, K. Arita, Y. Katoh, K. Kawai, S. Muraoka, S. Mitani, S. Fujii, K. Katayama, M. Iijima, T. Mikawa, T. Ninomiya, R. Miyanaga, Y. Kawashima, K. Tsuji, A. Himeno, T. Okada, R. Azuma, K. Shimakawa, H. Sugaya, T. Takagi, R. Yasuhara, K. Horiba, H. Kumigashira, and M. Oshima, "Highly reliable

TaOx ReRAM and direct evidence of redox reaction mechanism," in *2008 IEEE International Electron Devices Meeting*, 2008, pp. 1–4.

[79] D. Zhang, B. Peng, Y. Zhao, Z. Han, Q. Hu, X. Liu, Y. Han, H. Yang, J. Cheng, Q. Ding, H. Jiang, J. Yang, and H. Lv, "Sensing Circuit Design Techniques for RRAM in Advanced CMOS Technology Nodes," *Micromachines*, vol. 12, no. 8, 2021.

[80] E. Perez, M. K. Mahadevaiah, C. Zambelli, P. Olivo, and C. Wenger, "Data retention investigation in Al:HfO2-based resistive random access memory arrays by using high-temperature accelerated tests," *Journal of Vacuum Science and Technology B*, vol. 37, no. 1, p. 012202, 2019.

[81] S. Ambrogio, S. Balatti, Z. Q. Wang, Y.-S. Chen, H.-Y. Lee, F. T. Chen, and D. Ielmini, "Data retention statistics and modelling in HfO2 resistive switching memories," in *2015 IEEE International Reliability Physics Symposium*, 2015, pp. MY.7.1–MY.7.6.

[82] B. Traoré, P. Blaise, E. Vianello, H. Grampeix, S. Jeannot, L. Perniola, B. De Salvo, and Y. Nishi, "On the Origin of Low-Resistance State Retention Failure in HfO2-Based RRAM and Impact of Doping/Alloying," *IEEE Transactions on Electron Devices*, vol. 62, no. 12, pp. 4029–4036, 2015.

[83] C. Y. Chen, A. Fantini, L. Goux, R. Degraeve, S. Clima, A. Redolfi, G. Groeseneken, and M. Jurczak, "Programming-conditions solutions towards suppression of retention tails of scaled oxide-based RRAM," in *2015 IEEE International Electron Devices Meeting (IEDM)*, 2015, pp. 10.6.1–10.6.4.

[84] C. Wang, H. Wu, B. Gao, L. Dai, D. C. Sekar, Z. Lu, G. Bronner, D. Wu, and H. Qian, "The Statistical Evaluation of Correlations between LRS and HRS Relaxations in RRAM Array," in *2016 IEEE 8th International Memory Workshop (IMW)*, 2016, pp. 1–4.

[85] Y. Y. Chen, R. Degraeve, S. Clima, B. Govoreanu, L. Goux, A. Fantini, G. S. Kar, G. Pourtois, G. Groeseneken, D. J. Wouters, and M. Jurczak, "Understanding of the endurance failure in scaled HfO2-based 1T1R RRAM through vacancy mobility

degradation," in *2012 International Electron Devices Meeting*, 2012, pp. 20.3.1–20.3.4.

[86] S. Zafar, H. Jagannathan, L. F. Edge, and D. Gupta, "Measurement of oxygen diffusion in nanometer scale HfO2 gate dielectric films," *Applied Physics Letters*, vol. 98, no. 15, p. 152903, 2011.

[87] Y. Y. Chen, M. Komura, R. Degraeve, B. Govoreanu, L. Goux, A. Fantini, N. Raghavan, S. Clima, L. Zhang, A. Belmonte, A. Redolfi, G. S. Kar, G. Groeseneken, D. J. Wouters, and M. Jurczak, "Improvement of data retention in HfO2/Hf 1T1R RRAM cell under low operating current," in *2013 IEEE International Electron Devices Meeting*, 2013, pp. 10.1.1–10.1.4.

[88] S. Clima, Y. Y. Chen, A. Fantini, L. Goux, R. Degraeve, B. Govoreanu, G. Pourtois, and M. Jurczak, "Intrinsic Tailing of Resistive States Distributions in Amorphous HfOx and TaOx Based Resistive Random Access Memories," *IEEE Electron Device Letters*, vol. 36, no. 8, pp. 769–771, 2015.

[89] D. Ielmini, F. Nardi, C. Cagli, and A. L. Lacaita, "Size-Dependent Retention Time in NiO-Based Resistive-Switching Memories," *IEEE Electron Device Letters*, vol. 31, no. 4, pp. 353–355, 2010.

[90] S. Yu, Y. Yin Chen, X. Guan, H.-S. Philip Wong, and J. A. Kittl, "A Monte Carlo study of the low resistance state retention of HfOx based resistive switching memory," *Applied Physics Letters*, vol. 100, no. 4, p. 043507, 2012.

[91] A. Fantini, L. Goux, S. Clima, R. Degraeve, A. Redolfi, C. Adelmann, G. Polimeni, Y. Y. Chen, M. Komura, A. Belmonte, D. J. Wouters, and M. Jurczak, "Engineering of Hf1-xAlOy amorphous dielectrics for high-performance RRAM applications," in *2014 IEEE 6th International Memory Workshop (IMW)*, 2014, pp. 1–4.

[92] Y. Y. Chen, R. Roelofs, A. Redolfi, R. Degraeve, D. Crotti, A. Fantini, S. Clima, B. Govoreanu, M. Komura, L. Goux, L. Zhang, A. Belmonte, Q. Xie, J. Maes, G. Pourtois, and M. Jurczak, "Tailoring switching and endurance / retention reliability characteristics of HfO2 / Hf RRAM with Ti, Al, Si dopants," in *2014*

Symposium on VLSI Technology (VLSI-Technology): Digest of Technical Papers, 2014, pp. 1–2.

[93] B. Traoré, P. Blaise, E. Vianello, H. Grampeix, A. Bonnevialle, E. Jalaguier, G. Molas, S. Jeannot, L. Perniola, B. DeSalvo, and Y. Nishi, "Microscopic understanding of the low resistance state retention in HfO2 and HfAlO based RRAM," in *2014 IEEE International Electron Devices Meeting*, 2014, pp. 21.5.1–21.5.4.

[94] M. Lanza, H.-S. P. Wong, E. Pop, D. Ielmini, D. Strukov, B. C. Regan, L. Larcher, M. A. Villena, J. J. Yang, L. Goux, A. Belmonte, Y. Yang, F. M. Puglisi, J. Kang, B. Magyari-Köpe, E. Yalon, A. Kenyon, M. Buckwell, A. Mehonic, A. Shluger, H. Li, T.-H. Hou, B. Hudec, D. Akinwande, R. Ge, S. Ambrogio, J. B. Roldan, E. Miranda, J. Suñe, K. L. Pey, X. Wu, N. Raghavan, E. Wu, W. D. Lu, G. Navarro, W. Zhang, H. Wu, R. Li, A. Holleitner, U. Wurstbauer, M. C. Lemme, M. Liu, S. Long, Q. Liu, H. Lv, A. Padovani, P. Pavan, I. Valov, X. Jing, T. Han, K. Zhu, S. Chen, F. Hui, and Y. Shi, "Recommended methods to study resistive switching devices," *Advanced Electronic Materials*, vol. 5, no. 1, p. 1800143, 2019.

[95] A. Fantini, L. Goux, R. Degraeve, D. Wouters, N. Raghavan, G. Kar, A. Belmonte, Y.-Y. Chen, B. Govoreanu, and M. Jurczak, "Intrinsic switching variability in HfO2 RRAM," in *2013 5th IEEE International Memory Workshop*, 2013, pp. 30–33.

[96] M. Lanza, R. Waser, D. Ielmini, J. J. Yang, L. Goux, J. Suñe, A. J. Kenyon, A. Mehonic, S. Spiga, V. Rana, S. Wiefels, S. Menzel, I. Valov, M. A. Villena, E. Miranda, X. Jing, F. Campabadal, M. B. Gonzalez, F. Aguirre, F. Palumbo, K. Zhu, J. B. Roldan, F. M. Puglisi, L. Larcher, T.-H. Hou, T. Prodromakis, Y. Yang, P. Huang, T. Wan, Y. Chai, K. L. Pey, N. Raghavan, S. Dueñas, T. Wang, Q. Xia, and S. Pazos, "Standards for the characterization of endurance in resistive switching devices," *ACS Nano*, vol. 15, no. 11, pp. 17 214–17 231, 2021.

[97] W.-M. Chung, Y.-F. Chang, Y.-L. Hsu, Y. C. D. Chen, C.-C. Lin, C.-H. Lin, and J. Leu, "A study of the relationship between endurance and retention reliability for a hfox-based resistive switching memory," *IEEE Transactions on Device and Materials Reliability*, vol. 20, no. 3, pp. 541–547, 2020.

[98] D. Alfaro Robayo, G. Sassine, Q. Rafhay, G. Ghibaudo, G. Molas, and E. Nowak, "Endurance statistical behavior of resistive memories based on experimental and theoretical investigation," *IEEE Transactions on Electron Devices*, vol. 66, no. 8, pp. 3318–3325, 2019.

[99] S. Balatti, S. Ambrogio, Z. Wang, S. Sills, A. Calderoni, N. Ramaswamy, and D. Ielmini, "Voltage-controlled cycling endurance of hfox-based resistive-switching memory," *IEEE Transactions on Electron Devices*, vol. 62, no. 10, pp. 3365–3372, 2015.

[100] C. Y. Chen, L. Goux, A. Fantini, S. Clima, R. Degraeve, A. Redolfi, Y. Y. Chen, G. Groeseneken, and M. Jurczak, "Endurance degradation mechanisms in TiN Ta2O5
Ta resistive random-access memory cells," *Applied Physics Letters*, vol. 106, no. 5, p. 053501, 2015.

[101] S. Balatti, S. Ambrogio, Z.-Q. Wang, S. Sills, A. Calderoni, N. Ramaswamy, and D. Ielmini, "Pulsed cycling operation and endurance failure of metal-oxide resistive (rram)," in *2014 IEEE International Electron Devices Meeting*, 2014, pp. 14.3.1–14.3.4.

[102] W. Kim, S. Menzel, D. J. Wouters, Y. Guo, J. Robertson, B. Roesgen, R. Waser, and V. Rana, "Impact of oxygen exchange reaction at the ohmic interface in Ta2O5-based ReRAM devices," *Nanoscale*, vol. 8, pp. 17 774–17 781, 2016.

[103] B. Huber, P. B. Popp, M. Kaiser, A. Ruediger, and C. Schindler, "Fully inkjet printed flexible resistive memory," *Applied Physics Letters*, vol. 110, no. 14, p. 143503, 2017.

[104] A. Fantini, D. J. Wouters, R. Degraeve, L. Goux, L. Pantisano, G. Kar, Y.-Y. Chen, B. Govoreanu, J. A. Kittl, L. Altimime, and M. Jurczak, "Intrinsic Switching Behavior in HfO2 RRAM by Fast Electrical Measurements on Novel 2R Test Structures," in *2012 4th IEEE International Memory Workshop*, 2012, pp. 1–4.

[105] S. Balatti, S. Ambrogio, D. Ielmini, and D. C. Gilmer, "Variability and failure of set process in HfO_2 RRAM," in *2013 5th IEEE International Memory Workshop*, 2013, pp. 38–41.

[106] S. Balatti, S. Ambrogio, D. C. Gilmer, and D. Ielmini, "Set Variability and Failure Induced by Complementary Switching in Bipolar RRAM," *IEEE Electron Device Letters*, vol. 34, no. 7, pp. 861–863, 2013.

[107] D. Ielmini, "Resistive switching memories based on metal oxides: mechanisms, reliability and scaling," *Semiconductor Science and Technology*, vol. 31, no. 6, p. 063002, 2016.

[108] S. Ambrogio, S. Balatti, A. Cubeta, A. Calderoni, N. Ramaswamy, and D. Ielmini, "Statistical Fluctuations in HfOx Resistive-Switching Memory: Part I - Set/Reset Variability," *IEEE Transactions on Electron Devices*, vol. 61, no. 8, pp. 2912–2919, 2014.

[109] D. Niu, Y. Chen, C. Xu, and Y. Xie, "Impact of process variations on emerging memristor," in *Design Automation Conference*, 2010, pp. 877–882.

[110] D. Niu, Y. Xiao, and Y. Xie, "Low power memristor-based ReRAM design with Error Correcting Code," in *17th Asia and South Pacific Design Automation Conference*, 2012, pp. 79–84.

[111] N. Mielke, T. Marquart, N. Wu, J. Kessenich, H. Belgal, E. Schares, F. Trivedi, E. Goodness, and L. R. Nevill, "Bit error rate in NAND Flash memories," in *2008 IEEE International Reliability Physics Symposium*, 2008, pp. 9–19.

[112] F. Puglisi, L. Larcher, G. Bersuker, A. Padovani, and P. Pavan, "An empirical model for rram resistance in low- and high-resistance states," *IEEE Electron Device Letters*, vol. 34, p. 387, 2013.

[113] C. La Torre, "Physics-based compact modeling of valence-change-based resistive switching devices," Dissertation, Rheinisch-Westfälische Technische Hochschule Aachen, Aachen, 2019, veröffentlicht auf dem Publikationsserver der RWTH Aachen University; Dissertation, Rheinisch-Westfälische Technische Hochschule Aachen, 2019.

[114] S. Larentis, F. Nardi, S. Balatti, D. C. Gilmer, and D. Ielmini, "Resistive Switching by Voltage-Driven Ion Migration in Bipolar RRAM—Part II: Modeling," *IEEE Transactions on Electron Devices*, vol. 59, no. 9, pp. 2468–2475, 2012.

[115] S. Kim, S. Choi, and W. Lu, "Comprehensive physical model of dynamic resistive switching in an oxide memristor," *ACS Nano*, vol. 8, no. 3, pp. 2369–2376, 2014, pMID: 24571386.

[116] S. Kim, S.-J. Kim, K. M. Kim, S. R. Lee, M. Chang, E. Cho, Y.-B. Kim, C. J. Kim, U. In Chung, and I.-K. Yoo, "Physical electro-thermal model of resistive switching in bi-layered resistance-change memory," *Scientific Reports*, vol. 3, no. 1, p. 1680, 2013.

[117] K. Schnieders, C. Funck, F. Cüppers, S. Aussen, T. Kempen, A. Sarantopoulos, R. Dittmann, S. Menzel, V. Rana, S. Hoffmann-Eifert, and S. Wiefels, "Effect of electron conduction on the read noise characteristics in ReRAM devices," *APL Materials*, vol. 10, p. 101114, 2022.

[118] R. Degraeve, L. Goux, S. Clima, B. Govoreanu, Y. Chen, G. Kar, P. Rousse, G. Pourtois, D. Wouters, L. Altimime, M. Jurczak, G. Groeseneken, and J. Kittl, "Modeling and tuning the filament properties in rram metal oxide stacks for optimized stable cycling," in *Proceedings of Technical Program of 2012 VLSI Technology, System and Application*, 2012, pp. 1–2.

[119] R. Degraeve, A. Fantini, S. Clima, B. Govoreanu, L. Goux, Y. Chen, D. Wouters, P. Roussel, G. Kar, G. Pourtois, S. Cosemans, J. Kittl, G. Groeseneken, M. Jurczak, and L. Altimime, "Dynamic 'hour glass' model for SET and RESET in HfO2 RRAM," in *2012 Symposium on VLSI Technology (VLSIT)*, 2012, pp. 75–76.

[120] R. Degraeve, A. Fantini, G. Gorine, P. Roussel, S. Clima, C. Y. Chen, B. Govoreanu, L. Goux, D. Linten, M. Jurczak, and A. Thean, "Quantitative model for post-program instabilities in filamentary rram," in *2016 IEEE International Reliability Physics Symposium (IRPS)*, 2016, pp. 6C–1–1–6C–1–7.

[121] K. Xiong, J. Robertson, M. C. Gibson, and S. J. Clark, "Defect energy levels in HfO2 high-dielectric-constant gate oxide," *Applied Physics Letters*, vol. 87, no. 18, p. 183505, 2005.

[122] T. Perevalov, V. Aliev, V. Gritsenko, A. Saraev, and V. Kaichev, "Electronic structure of oxygen vacancies in hafnium oxide," *Microelectronic Engineering*, vol. 109, pp. 21–23, 2013, insulating Films on Semiconductors 2013.

[123] K. J. Chang, B. Ryu, H.-K. Noh, J. Bang, and E.-A. Choi, "Electronic Structure of O-vacancy in High-k Dielectrics and Oxide Semiconductors," *MRS Proceedings*, vol. 1370, pp. mrss11–1370–yy01–01, 2011.

[124] D. Ielmini and Y. Zhang, "Analytical model for subthreshold conduction and threshold switching in chalcogenide-based memory devices," *Journal of Applied Physics*, vol. 102, no. 5, p. 054517, 2007.

[125] G. Jegert, A. Kersch, W. Weinreich, U. Schröder, and P. Lugli, "Modeling of leakage currents in high-k dielectrics: Three-dimensional approach via kinetic monte carlo," *Applied Physics Letters*, vol. 96, no. 6, p. 062113, 2010.

[126] S. Yu, X. Guan, and H.-S. P. Wong, "Conduction mechanism of TiN/HfOx/Pt resistive switching memory: A trap-assisted-tunneling model," *Applied Physics Letters*, vol. 99, no. 6, p. 063507, 2011.

[127] W. Zhu, T.-P. Ma, T. Tamagawa, J. Kim, and Y. Di, "Current transport in metal/hafnium oxide/silicon structure," *IEEE Electron Device Letters*, vol. 23, no. 2, pp. 97–99, 2002.

[128] W. Stehling, E. Abbaspour, C. Jungemann, and S. Menzel, "Kinetic Monte Carlo modeling of the charge transport in a HfO2-based ReRAM with a rough anode," 2017, pp. 1–4.

[129] L. Larcher, "Statistical simulation of leakage currents in mos and flash memory devices with a new multiphonon trap-assisted tunneling model," *IEEE Transactions on Electron Devices*, vol. 50, no. 5, pp. 1246–1253, 2003.

[130] G. Bersuker, D. Gilmer, D. Veksler, P. Kirsch, L. Vandelli, A. Padovani, L. Larcher, K. Mckenna, A. Shluger, V. Iglesias, M. Porti, and M. Nafría, "Metal Oxide RRAM Switching Mechanism Based on Conductive Filament Properties," *Journal of Applied Physics*, vol. 110, p. 124518, 2012.

[131] L. Vandelli, A. Padovani, L. Larcher, R. G. Southwick, W. B. Knowlton, and G. Bersuker, "A physical model of the temperature dependence of the current through SiO_2/HfO_2 stacks," *IEEE Transactions on Electron Devices*, vol. 58, no. 9, pp. 2878–2887, 2011.

[132] L. Vandelli, A. Padovani, G. Bersuker, D. Gilmer, P. Pavan, and L. Larcher, "Modeling of the Forming Operation in HfO2-Based Resistive Switching Memories," in *2011 3rd IEEE International Memory Workshop (IMW)*, 2011, pp. 1–4.

[133] O. Pirrotta, A. Padovani, L. Larcher, L. Zhao, B. Magyari-Köpe, and Y. Nishi, "Multi-scale modeling of oxygen vacancies assisted charge transport in sub-stoichiometric TiOx for RRAM application," in *2014 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, 2014, pp. 37–40.

[134] L. Larcher, A. Padovani, O. Pirrotta, L. Vandelli, and G. Bersuker, "Microscopic understanding and modeling of HfO2 RRAM device physics," in *2012 International Electron Devices Meeting*, 2012, pp. 20.1.1–20.1.4.

[135] X. Guan, S. Yu, and H.-S. P. Wong, "On the switching parameter variation of metal-oxide rram—part i: Physical modeling and simulation methodology," *IEEE Transactions on Electron Devices*, vol. 59, no. 4, pp. 1172–1182, 2012.

[136] B. Gao, B. Sun, H. Zhang, L. Liu, X. Liu, R. Han, J. Kang, and B. Yu, "Unified physical model of bipolar oxide-based resistive switching memory," *IEEE Electron Device Letters*, vol. 30, no. 12, pp. 1326–1328, 2009.

[137] A. Miller and E. Abrahams, "Impurity conduction at low concentrations," *Physical Review*, vol. 120, pp. 745–755, 1960.

[138] E. Abbaspour, "Modeling and simulation of valence-change based resistive switching," Ph.D. dissertation, Dissertation, Rheinisch-Westfälische Technische Hochschule Aachen, 2019.

[139] W. Zhou, "Master equation study on organic light-emitting diodes," Ph.D. dissertation, Dissertation, Rheinisch-Westfälische Technische Hochschule Aachen, 2020.

[140] I. Lundström and C. Svensson, "Tunneling to traps in insulators," *Journal of Applied Physics*, vol. 43, no. 12, pp. 5045–5047, 1972.

[141] D. Duncan, B. Magyari-Köpe, and Y. Nishi, "Filament-induced anisotropic oxygen vacancy diffusion and charge trapping effects in hafnium oxide rram," *IEEE Electron Device Letters*, vol. 37, no. 4, pp. 400–403, 2016.

[142] D. Ielmini, "Modeling the universal set/reset characteristics of bipolar rram by field- and temperature-driven filament growth," *IEEE Transactions on Electron Devices*, vol. 58, no. 12, pp. 4309–4317, 2011.

[143] P. Jiang, X. Qian, and R. Yang, "Tutorial: Time-domain thermoreflectance (TDTR) for thermal property characterization of bulk and thin film materials," *Journal of Applied Physics*, vol. 124, no. 16, p. 161103, 2018.

[144] M. A. Panzer, M. Shandalov, J. A. Rowlette, Y. Oshima, Y. W. Chen, P. C. McIntyre, and K. E. Goodson, "Thermal properties of ultrathin hafnium oxide gate dielectric films," *IEEE Electron Device Letters*, vol. 30, no. 12, pp. 1269–1271, 2009.

[145] K. Szot, W. Speier, G. Bihlmayer, and R. Waser, "Switching the electrical resistance of individual dislocations in single-crystalline srtio3," *Nature Materials*, vol. 5, no. 4, pp. 312–320, Apr 2006.

[146] C. G. Van de Walle and J. Neugebauer, "First-principles calculations for defects and impurities: Applications to III-nitrides," *Journal of Applied Physics*, vol. 95, no. 8, pp. 3851–3879, 2004.

[147] C. G. Van de Walle, S. Limpijumnong, and J. Neugebauer, "First-principles studies of beryllium doping of GaN," *Physical Review B*, vol. 63, p. 245205, 2001.

[148] W. Chen and A. Pasquarello, "First-principles determination of defect energy levels through hybrid density functionals and GW," *Journal of Physics: Condensed Matter*, vol. 27, no. 13, p. 133202, 2015.

[149] S. Anand, M. Y. Toriyama, C. Wolverton, S. M. Haile, and G. J. Snyder, "A convergent understanding of charged defects," *Accounts of Materials Research*, vol. 3, no. 7, pp. 685–696, 2022.

[150] J. Robertson and S. J. Clark, "Limits to doping in oxides," *Physical Review B*, vol. 83, p. 075205, 2011.

[151] A. O'Hara, G. Bersuker, and A. A. Demkov, "Assessing hafnium on hafnia as an oxygen getter," *Journal of Applied Physics*, vol. 115, no. 18, p. 183703, 2014.

[152] M. Schie, S. Menzel, J. Robertson, R. Waser, and R. A. De Souza, "Field-enhanced route to generating anti-frenkel pairs in HfO_2," *Physical Review Materials*, vol. 2, p. 035002, 2018.

[153] S. Clima, B. Govoreanu, M. Jurczak, and G. Pourtois, "HfOx as RRAM material – First principles insights on the working principles," *Microelectronic Engineering*, vol. 120, pp. 13–18, 2014, mAM2013, March 10-13, Leuven, Belgium.

[154] S. Hida, T. Morita, T. Yamasaki, J. Nara, T. Ohno, and K. Kinoshita, "Repeatable and reproducible formation/rupture of oxygen vacancy filaments in the vicinity of a polycrystalline HfO2 surface," *AIP Advances*, vol. 9, no. 3, p. 035309, 2019.

[155] S. Kumar, C. E. Graves, J. P. Strachan, A. L. D. Kilcoyne, T. Tyliszczak, Y. Nishi, and R. S. Williams, "In-operando synchronous time-multiplexed O K-edge x-ray absorption spectromicroscopy of functioning tantalum oxide memristors," *Journal of Applied Physics*, vol. 118, no. 3, p. 034502, 2015.

[156] S. R. Bradley, G. Bersuker, and A. L. Shluger, "Modelling of oxygen vacancy aggregates in monoclinic HfO2: can they contribute to conductive filament formation?" *Journal of Physics: Condensed Matter*, vol. 27, no. 41, p. 415401, 2015.

[157] S. R. Bradley, A. L. Shluger, and G. Bersuker, "Electron-injection-assisted generation of oxygen vacancies in monoclinic HfO_2," *Physical Review Applied*, vol. 4, p. 064008, 2015.

[158] D. Gao, J. Strand, M. Munde, and A. Shluger, "Mechanisms of Oxygen Vacancy Aggregation in SiO2 and HfO2," *Frontiers in Physics*, vol. 7, p. 43, 2019.

[159] K. Kamiya, M. Y. Yang, B. Magyari-Köpe, M. Niwa, Y. Nishi, and K. Shiraishi, "Vacancy cohesion-isolation phase transition upon charge injection and removal in binary oxide-based rram filamentary-type switching," *IEEE Transactions on Electron Devices*, vol. 60, no. 10, pp. 3400–3406, 2013.

[160] K. Kamiya, M. Y. Yang, T. Nagata, S.-G. Park, B. Magyari-Köpe, T. Chikyow, K. Yamada, M. Niwa, Y. Nishi, and K. Shiraishi, "Generalized mechanism of the resistance switching in binary-oxide-based resistive random-access memories," *Physical Review B*, vol. 87, p. 155201, 2013, URL: http://dx.doi.org/10.1103/PhysRevB.87.155201.

[161] K. Kamiya, M. Yang, S.-G. Park, B. Magyari-Kope, Y. Nishi, M. Niwa, and K. Shiraishi, "On-off switching mechanism of resistive-random-access-memories based on the formation and disruption of oxygen vacancy conducting channels," *Applied Physics Letters*, vol. 100, p. 073502, 2012.

[162] K.-H. Xue, P. Blaise, L. R. C. Fonseca, G. Molas, E. Vianello, B. Traoré, B. De Salvo, G. Ghibaudo, and Y. Nishi, "Grain boundary composition and conduction in HfO2: An ab initio study," *Applied Physics Letters*, vol. 102, no. 20, p. 201908, 2013.

[163] B. Butcher, G. Bersuker, L. Vandelli, A. Padovani, L. Larcher, A. Kalantarian, R. Geer, and D. Gilmer, "Modeling the effects of different forming conditions on RRAM conductive filament stability," in *2013 5th IEEE International Memory Workshop*, 2013, pp. 52–55.

[164] L. Vandelli, A. Padovani, L. Larcher, G. Broglia, G. Ori, M. Monia, G. Bersuker, and P. Pavan, "Comprehensive physical modeling of forming and switching operations in HfO2 RRAM devices," *Technical Digest - International Electron Devices Meeting, IEDM*, 2011.

[165] D. J. Dumin, "Oxide reliability : a summary of silicon oxide wearout, breakdown, and reliability," 2002.

[166] M. Kimura and H. Koyama, "Mechanism of time-dependent oxide breakdown in thin thermally grown SiO2 films," *Journal of Applied Physics*, vol. 85, no. 11, pp. 7671–7681, 1999.

[167] J. W. McPherson and H. C. Mogul, "Underlying physics of the thermochemical E model in describing low-field time-dependent dielectric breakdown in SiO2 thin films," *Journal of Applied Physics*, vol. 84, no. 3, pp. 1513–1523, 1998.

[168] J. R. Tessman, A. H. Kahn, and W. Shockley, "Electronic polarizabilities of ions in crystals," *Physical Review*, vol. 92, pp. 890–895, 1953.

[169] J. W. McPherson, "Lorentz factor determination for local electric fields in semiconductor devices utilizing hyper-thin dielectrics," *Journal of Applied Physics*, vol. 118, no. 20, p. 204106, 2015.

[170] C. Kittel, *Introduction to Solid State Physics*. Wiley, 1996.

[171] K. Park, K. Park, S. Im, S. Hong, K. Son, and J. Jeon, "Development of an Advanced TDDB Analysis Model for Temperature Dependency," *Electronics*, vol. 8, no. 9, 2019.

[172] J. McPherson, J.-Y. Kim, A. Shanware, and H. Mogul, "Thermochemical description of dielectric breakdown in high dielectric constant materials," *Applied Physics Letters*, vol. 82, no. 13, pp. 2121–2123, 2003.

[173] A. Padovani, L. Larcher, G. Bersuker, and P. Pavan, "Charge Transport and Degradation in HfO2 and HfOx Dielectrics," *IEEE Electron Device Letters*, vol. 34, no. 5, pp. 680–682, 2013.

[174] W. T. Doyle, "The clausius-mossotti problem for cubic arrays of spheres," *Journal of Applied Physics*, vol. 49, no. 2, pp. 795–797, 1978.

[175] H. Fröhlich, "General theory of the static dielectric constant," *Transactions of the Faraday Society*, vol. 44, pp. 238–243, 1948.

[176] L. Vandelli, A. Padovani, L. Larcher, and G. Bersuker, "Microscopic modeling of electrical stress-induced breakdown in poly-crystalline hafnium oxide dielectrics," *IEEE Transactions on Electron Devices*, vol. 60, no. 5, pp. 1754–1762, 2013.

[177] Y. Guo and J. Robertson, "Materials selection for oxide-based resistive random access memories," *Applied Physics Letters*, vol. 105, no. 22, p. 223516, 2014.

[178] N. Capron, P. Broqvist, and A. Pasquarello, "Migration of oxygen vacancy in HfO2 and across the HfO2/SiO2 interface: A first-principles investigation," *Applied Physics Letters*, vol. 91, no. 19, p. 192905, 2007.

[179] K. McKenna and A. Shluger, "The interaction of oxygen vacancies with grain boundaries in monoclinic HfO2," *Applied Physics Letters*, vol. 95, no. 22, p. 222111, 2009.

[180] C. Tang, B. Tuttle, and R. Ramprasad, "Diffusion of o vacancies near Si : HfO_2 interfaces: An ab initio investigation," *Physical Review B*, vol. 76, p. 073306, 2007.

[181] Y. Dai, Z. Pan, F. Wang, and X. Li, "Oxygen vacancy effects in hfo2-based resistive switching memory: First principle study," *AIP Advances*, vol. 6, no. 8, p. 085209, 2016.

[182] K. Rushchanskii, S. Blügel, and M. Lezaic, "Routes for increasing endurance and retention in HfO2-based resistive switching memories," *Physical Review Materials*, vol. 2, 2018.

[183] P. A. Leighton, "Electronic Processes in Ionic Crystals (Mott, N. F.; Gurney, R. W.)," *Journal of Chemical Education*, vol. 18, no. 5, p. 249, 1941.

[184] D. Strukov, F. Alibart, and S. Williams, "Thermophoresis/diffusion as a plausible mechanism for unipolar resistive switching in metal–oxide–metal memristors," *Applied Physics A*, vol. 107, 2012.

[185] A. R. Genreith-Schriever and R. A. De Souza, "Field-enhanced ion transport in solids: Reexamination with molecular dynamics simulations," *Physical Review B*, vol. 94, p. 224304, 2016.

[186] J. Janek, C. Korte, and A. B. Lidiard, *Thermodiffusion in Ionic Solids —Model Experiments and Theory*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 146–183.

[187] S. Yu, X. Guan, and H.-S. P. Wong, "On the Switching Parameter Variation of Metal Oxide RRAM—Part II: Model Corroboration and Device Design Strategy," *IEEE Transactions on Electron Devices*, vol. 59, no. 4, pp. 1183–1188, 2012.

[188] R. Degraeve, A. Fantini, N. Raghavan, L. Goux, S. Clima, B. Govoreanu, A. Belmonte, D. Linten, and M. Jurczak, "Causes and consequences of the stochastic aspect of filamentary RRAM," *Microelectronic Engineering*, vol. 147, pp. 171–175, 2015, insulating Films on Semiconductors 2015.

[189] D. Veksler and G. Bersuker, "Advances in RRAM Technology: Identifying and Mitigating Roadblocks," *International Journal of High Speed Electronics and Systems*, vol. 25, no. 01n02, p. 1640006, 2016.

[190] V. G. Karpov and D. Niraula, "Log-normal statistics in filamentary RRAM devices and related systems," *IEEE Electron Device Letters*, vol. 38, no. 9, pp. 1240–1243, 2017.

[191] F. M. Dekking, C. Kraaikamp, H. P. Lopuhaä, and L. E. Meester, *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media, 2006.

[192] D. T. Gillespie, "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions," *Journal of Computational Physics*, vol. 22, no. 4, pp. 403–434, 1976.

[193] J. J. Lukkien, J. P. L. Segers, P. A. J. Hilbers, R. J. Gelten, and A. P. J. Jansen, "Efficient Monte Carlo methods for the simulation of catalytic surface reactions," *Physical Review E*, vol. 58, pp. 2598–2610, 1998.

[194] J. Robertson, O. Sharia, and A. A. Demkov, "Fermi level pinning by defects in HfO2-metal gate stacks," *Applied Physics Letters*, vol. 91, no. 13, p. 132912, 2007.

[195] B. Meng and W. H. Weinberg, "Monte Carlo simulations of temperature programmed desorption spectra," *The Journal of Chemical Physics*, vol. 100, no. 7, pp. 5280–5289, 1994.

[196] K. A. Fichthorn and W. H. Weinberg, "Theoretical foundations of dynamical Monte Carlo simulations," *The Journal of Chemical Physics*, vol. 95, no. 2, pp. 1090–1096, 1991.

[197] R. Nieminen and A. Jansen, "Monte carlo simulations of surface reactions," *Applied Catalysis A: General*, vol. 160, no. 1, pp. 99–123, 1997, kinetic Methods in Heterogeneous Catalysis.

[198] A. U. Modak and M. T. Lusk, "Kinetic Monte Carlo simulation of a solid-oxide fuel cell: I. Open-circuit voltage and double layer structure," *Solid State Ionics*, vol. 176, no. 29, pp. 2181–2191, 2005.

[199] N. Sano, K. Matsuzawa, M. Mukai, and N. Nakayama, "On discrete random dopant modeling in drift-diffusion simulations: physical meaning of 'atomistic' dopants," *Microelectronics Reliability*, vol. 42, no. 2, pp. 189–199, 2002.

[200] J. J. M. van der Holst, M. A. Uijttewaal, B. Ramachandhran, R. Coehoorn, P. A. Bobbert, G. A. de Wijs, and R. A. de Groot, "Modeling and analysis of the three-dimensional current density in sandwich-type single-carrier devices of disordered organic semiconductors," *Physical Review B*, vol. 79, p. 085203, 2009.

[201] H. Takeuchi, D. Ha, and T.-J. King, "Observation of bulk HfO2 defects by spectroscopic ellipsometry," *Journal of Vacuum Science and Technology A*, vol. 22, no. 4, pp. 1337–1341, 2004.

[202] S. Monaghan, P. Hurley, K. Cherkaoui, M. Negara, and A. Schenk, "Determination of electron effective mass and electron affinity in HfO2 using MOS and MOSFET structures," *Solid-State Electronics*, vol. 53, no. 4, pp. 438–444, 2009, special Issue with papers selected from the Ultimate Integration on Silicon Conference, ULIS 2008.

[203] M. Hinz, O. Marti, B. Gotsmann, M. A. Lantz, and U. Dürig, "High resolution vacuum scanning thermal microscopy of HfO2 and SiO2," *Applied Physics Letters*, vol. 92, no. 4, p. 043122, 2008.

[204] N. D. Milošević and K. D. Maglić, "Thermophysical Properties of Solid Phase Hafnium at High Temperatures," *International Journal of Thermophysics*, vol. 27, no. 2, pp. 530–553, 2006.

[205] X. Wu, K. Li, N. Raghavan, M. Bosman, Q.-X. Wang, D. Cha, X.-X. Zhang, and K.-L. Pey, "Uncorrelated multiple conductive filament nucleation and rupture in ultra-thin high-κ dielectric based resistive random access memory," *Applied Physics Letters*, vol. 99, no. 9, p. 093502, 08 2011.

[206] X. Wu, K. Yu, D. Cha, M. Bosman, N. Raghavan, X. Zhang, K. Li, Q. Liu, L. Sun, and K. Pey, "Atomic scale modulation of self-rectifying resistive switching by interfacial defects," *Advanced Science*, vol. 5, no. 6, p. 1800096, 2018.

[207] X. Saura, J. Suñé, S. Monaghan, P. K. Hurley, and E. Miranda, "Analysis of the breakdown spot spatial distribution in Pt/HfO2/Pt capacitors using nearest neighbor statistics," *Journal of Applied Physics*, vol. 114, no. 15, p. 154112, 10 2013.

[208] F. Stellari, E. Y. Wu, T. Ando, E. Cartier, M. M. Frank, C. Cabral, P. Song, and D. Pfeiffer, "Resistive random access memory filament visualization and characterization using photon emission microscopy," *IEEE Electron Device Letters*, vol. 42, no. 6, pp. 828–831, 2021.

[209] E. Wu, F. Stellari, L. Ocola, M. Frank, P. Song, and T. Ando, "Gibbs spatial process for characterization of filament interaction in ReRAM devices via photon emission microscopy," *Applied Physics Letters*, vol. 120, no. 13, p. 132902, 03 2022.