Theoretical Simulation of the Conductive Filament in the Resistive Switching Memory

Xu Xu

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

M. P. Anantram, Chair

Scott Dunham

Arka Majumdar

Program Authorized to Offer Degree:

Electrical and Computer Engineering

University of Washington

Abstract

Theoretical Simulation of the Conductive Filament in the Resistive Switching Memory

Xu Xu

Chair of the Supervisory Committee:
M. P. Anantram
Electrical and Computer Engineering

In this thesis, resistive switching properties of resistive memory (RRAM) is extensively modeled and investigated. The RRAM is a promising candidate to replace the Flash ram nowadays because of its performance, endurance and cost. The resistive switching behavior in RRAM is explored by solving a set of equations that represent the important elements of the device physics.

The non-equilibrium Green's function method is used to model charge transport for system size scale at atomistic scale. The Cu dopant in the α-alumina are calculated using ab initio density functional theory to give the Hamiltonian. The current saturation and negative differential resistance can be found in the system. The current saturation occurs due to the narrow bandwidth on the transmission coefficient. The negative differential resistance is analyzed based on the match and mismatch of local density of states. The current difference for different filament configuration makes multi-level memory cell possible.

Prior work has had difficulty in modeling the processes of set, reset and retention in a comprehensive manner. In this thesis, a more unified picture of these processes using Kinetic Monte Carlo simulations is provided. The Kinetic Monte Carlo simulations are based on atom/vacancy motion to model filament growth and dissolution in an oxide, in multiple scenarios. The kinetic Monte Carlo simulations are based on the processes of formation, diffusion and recombination. The classical heat and Poisson equations are solved to give the local electric field and temperature. In the surface generation model, oxygen vacancy is generated in Hafnia near the interface, with the corresponding oxygen atom entering the metal electrode. These oxygen atoms form a thin insulating oxide layer at the Hafnia-active electrode interface. This interfacial layer can change the direction of the electric field and help to thicken the filament. This thickening of the conducting filament is captured by my model and it explains the trend of resistance decrease with an increase in compliance current found in some experiments. In RRAM with negative thermophoresis materials, and two inert electrodes, the forming and unipolar reset processes are simulated. The negative thermophoresis precludes the traditional mechanism of RESET where the vacancy moves away from the filament as a result of high filament temperature. Simulations reveal a new RESET mechanism which involves the diffusion of oxygen interstitials to break the vacancy-based filament in a region close to the top electrode. With the mechanism of oxygen interstitial diffusion, the filament can be RESET at a higher temperature than in SET.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I want to express sincere appreciation to my PhD advisor Prof. M.P. Anantram for his help during my PhD research. Also, I appreciate Prof. Scott Dunham, Prof. Lucien Brush, Prof. Arka Majumdar and Prof. Xiaodong Xu for their precious time to serve in the committee. Also, I want to thank Winbond Inc. for funding and Fred Chen and J.J. Hsu of Winbond Inc. for regular technical discussions. I would like to particularly thank Jie Liu, Yunqi Zhao, Zhenni Wan and Hashem Mohammad, for the help in UW as group mates and as friends. Most of all, I would like to thank my parents.

Chapter 1. INTRODUCTION

The resistive random-access memory (RRAM) is a promising candidate for next-generation non-volatile memory (NVM). The RRAM devices can provide a high-density and fast-access memory cell array, with low power consumption and low fabrication cost. The memory cell uses resistive switching behavior to store information: when a conductive filament is formed in the insulating layer, the cell is set to be "1" or low resistance state (LRS); and when the conductive filament is ruptured, the information is reset to be "0" or high resistance state (HRS). The forming and rupture of atomic-scale conductive filament, which involves the migration of atoms in the insulating layer, can be controlled by applying external voltage. In order to offer a deep understanding of the electron transport through, the formation of, and the rupture of, a filament, this thesis focuses on the numerical simulation and theoretical modeling of RRAM devices.

1.1 BACKGROUND

The non-volatile memory (NVM) is a vital part of the digital storage system. Because the NVM can retrieve the stored information after having been powered off, the NVM is mainly used for the task of long-term persistence storage, such as storage of program and data. Nowadays, flash memory is prevailing in the NVM, because of its high reliability and performance. The flash memory is widely used in portable storage, such as the flash drive, the cache of magnetic disks, such as the hybrid hard drive, and also the stand also secondary storage such as the solid-state drive(SSD). In the flash memory cell, as shown in Fig 1.1, there is a floating gate, and the information is stored in the form of charge in the floating gate. During the programming process,

a large voltage is applied to the controlling gate, and a large current is applied from drain to source. Then the hot-electrons are injected from the channel to the floating gate. The charge in the floating gate can change the threshold voltage, which can be read from the peripheral sensing circuit. During the erase process, the polarity of the applied voltage is reversed, and the electrons in the floating gate can escape by quantum tunneling.

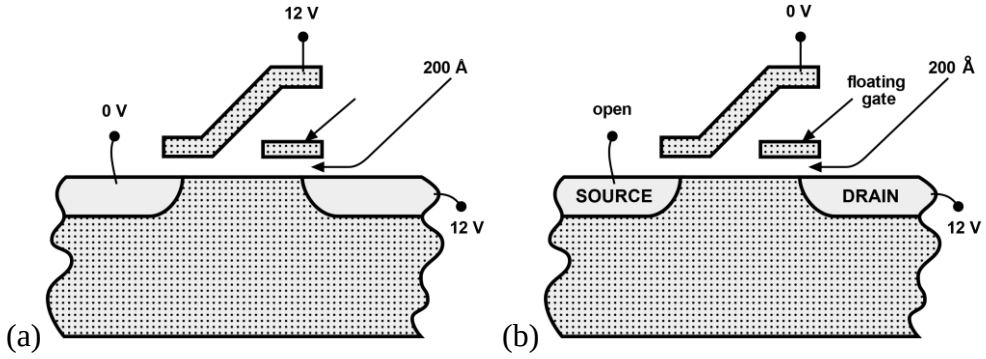


Fig. 1.1 Typical structure of Flash memory. In one memory cell, there is a floating gate to store the information in the form of charge, and a control gate is used to control the operations of program, erase and read.

However, the flash memory is facing problems in access time, retention time, scaling and also cost. In the CMOS technology today, the reading time for single floating gate MOSFET can be achieved within 100ns, but the writing time (in the program and erase both) is usually on the order of milliseconds. Good circuit design, such as NAND Flash, as shown in Fig. 1.2, can improve the average writing time by parallelization; however, the latency cannot be improved because of the physical limitation. Also, the tradeoff on memory cell size will delay the reading response time to the order of 10us. With the scaling of feature size of the device, the volume to hold the charge in the floating gate is decreasing, and therefore a smaller number of electrons are used to store the information. As reported from [1], only ~16 electrons are stored in the floating gate in 16nm

technology node. As a result, the memory can easily lose its information during the long-time storage [2]-[34].

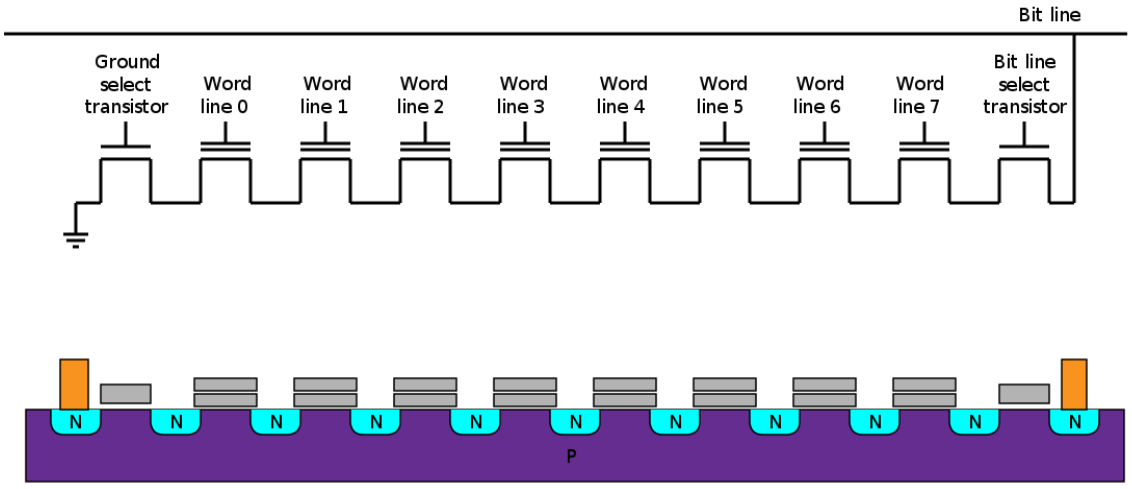


Fig. 1.2 Typical structure of NAND Flash memory. Memory cells are connected in series, which can shrink the average cell size, in penalty of block erasing.

RRAM is a promising candidate to replace flash memory. The RRAM device has a simple metal-insulator-metal structure, as shown in Fig 1.3, which can reduce the cost of fabrication. The insulating layer materials have the resistive switching characteristics: when a large voltage applied, the resistivity of the material will decrease, and when a reversed voltage applied, the resistivity can go back to a high value. The underlying mechanism of the switching characteristics is widely accepted as that the voltage induced atomic position changes. When a large voltage is applied to the switching layer, the atomic position of the defects in the switching layer can change. Then, the conductive filament is formed in the switching layer which results in different resistance. For the RRAM, experiments show that access time is expected to be shorter than flash memory. The memory cell is compatible with the high-density crossbar structure, whose size is the same as NAND Flash memory as $4F^2$. The fabrication process of RRAM is much simpler than NAND

Flash. The size of filament is in the nanometers scale, which provides the capability of the high-density memory array. Since the information is recorded by the position of atoms, the storage is much stable in the environment, especially insensitive to light and heat.

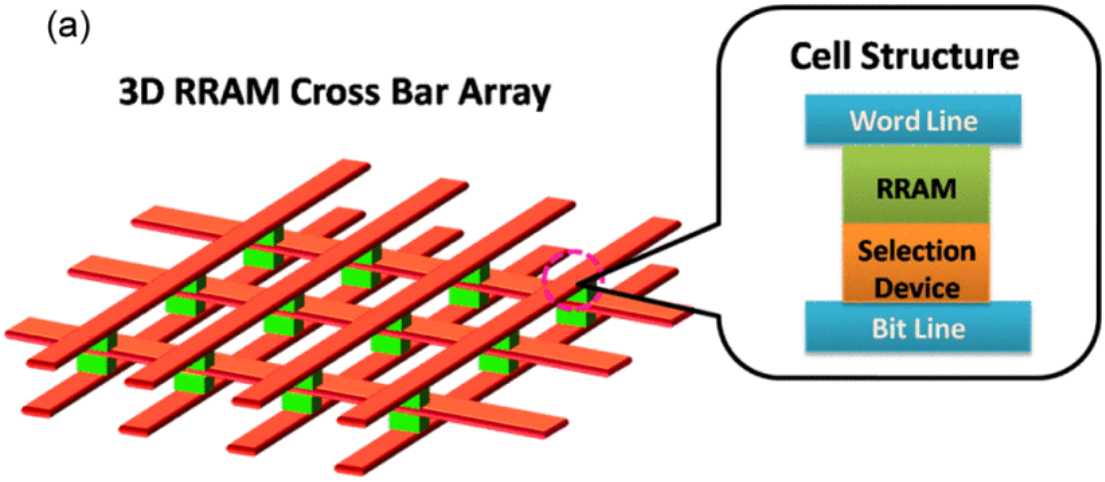


Fig. 1.2 Typical structure of RRAM [2]. Memory cells consist of two electrodes and a resistive switching layer. The selection device is usually needed for a memory array. RRAM has a simple structure which reduces the fabrication process and cost.

In the experiments of RRAM, various switching materials are investigated, such as perovskite oxides of $SrTiO_3$, $SrZrO_3$, or binary metal oxides such as NiO, TiO_2, HfOx or Al_2O_3. These materials can show good insulator properties in the pristine state, and they are compatible with the silicon fabrication today. Besides the switching material, switching characteristics are also sensitive to the choice of the electrode material. The metal, such as Cu, Pt or W, and other electrically conductive materials, such as TiN or ITO are widely used in the experiment. Today's experiments show that a single RRAM cell: can be smaller than 10nm by 10 nm in size, can be programmed in less than 100 nanoseconds, has an endurance of up to 10^{12} cycles, and has a retention time that is longer than 3000 hours at 150C°. Also, some experiments are done with the

3D integration of memory cells, which offers a promising packaging process for even higher storage density.

In the pristine state of an RRAM cell, the switching layer is sandwiched by two electrodes, and the memory cell shows a high resistance state. In the operation of setting the RRAM cell, a large voltage is applied between the electrodes. Then the conductive filament can be formed in the switching layer. The conductive filament can consist of metal atoms, via the process of the metal atoms immigration, or oxygen vacancies in the insulating layer, via the process of oxygen atoms emigration. The conductive filament has a low resistivity and when the filament bridges two electrodes, the resistance of the memory cell is switched to be a low value. The memory cell is therefore SET from “0” to “1”. To read out the stored information, a small voltage is applied to the cell, and then the values of current through the cell can tell the existence of the conductive filament. The rupture of filament will RESET the memory cell to “0”. The rupture process of conductive filament can be achieved in a bipolar way: apply a voltage on the memory cell with reversed polarity. Also, the unipolar rapture process is possible: applying a large current through the memory cell with the same polarity as the forming process. After the conductive filament raptures, the memory cell can go back to the high resistance state. After the conductive filament has ruptured, the resistance of memory cell changes back to a high value, which RESETs the cell to “0”. To understand the electron transmission through the conductive filament, the Green’s function calculation is used to extract the current-voltage characteristics. Kinetic Monte Carlo method is used to investigate the formation and rupturing processes of the conductive filament in the insulating layer.

1.2 OUTLINE

In Chapter 2, a brief review of some of the modeling and simulation approaches used in this thesis is given. The approaches include the ab initio density functional theory, the Green's function method, and the kinetic Monte Carlo method. These methods are used to simulate the different properties of resistive memory at different scales. In chapter 3, the Green's function method is used to investigate the electron transport properties through Copper filaments in alumina. Chapters 4 and 5 use the kinetic Monte Carlo method to investigate the migration of atoms and vacancies and the formation and recombination of oxygen vacancies and interstitials. The classical heat and Poisson equations are coupled with the kinetic Monte Carlo method; the details of my implementation are described in chapter 4.

In chapter 3, we will apply the steady-state calculation to give the current-voltage characteristics of the RRAM. This calculation assumes that the atomic positions are fixed, which gives the current in the reading process. The NEGF method in the frequency domain is used, which supposes the electrons are relaxed to the steady states. Then the transmission coefficient can be given, and then the current is computed. Negative differential resistance can be found in the filament, which can be explained as the match and mismatch of the local density of states. The current computation shows that the conductance of filament can change by more than 10 times due to the different positions of the atoms in the filament. The large change in the conductance will be promising to fabricate a multilevel memory cell. The migration barrier of Cu atom in Al_2O_3 is also calculated in the nudged elastic band method, which is based on the density functional theory.

Chapter 4 aims to study the filament formation and rupture process using the kinetic Monte Carlo method. Since the filament is formed in a switching layer around 10 nm thick, the ab initio method

described in chapter 3 can hardly handle this problem due to the computational resource requirements. Because the kinetic Monte Carlo method is highly efficient to compute the large-scale system, this method is used to simulate the formation and rupture processes of the conductive filament in the oxide. The RRAM cell consists of a switching layer, an inert electrode, and an active electrode. Under the simulation, the oxygen vacancy is generated near the active electrode and diffuse inside the switching layer to form the conductive filament. In the reset operation, the reversed voltage is applied on the cell, and filament can rapture. Since the voltage polarity is reversed in the reading and reset operations, the reading operation will not degrade the information stored in the cell. Various processes in operation of RRAM, including the clustering of vacancies, the surface formation process, the effect of the electric field, are studied in the kinetic Monte Carlo method. Also, the corresponding oxygen atoms are considered: these atoms can migrate into the metal directly and partially oxidized electrode. The partially oxidized electrode can form an insulating layer and then change the electric field near the filament tip. Therefore, after the filament has bridged two electrodes, the electric field is enhanced, and the direction of the electric field is pointing to the filament tip. The relationship between the current compliance and resistance is explained by the enhanced electric field.

Chapter 5 studies the resistive switching in the system with bulk generation. Although the device with surface generation studied in chapter 4 can produce a stable memory cell, the memory cell with inert electrodes on both sides is also studied. The oxygen vacancies are generated in the bulk oxide without the help of the active electrode. The reset of the memory cell is to apply a large current through the memory cell, which is also known as unipolar reset. Since the polarity of the voltage is the same in the set and reset process, a simple diode in series with the memory cell can be used selector in the memory array, which decreases the device area and reduces the circuit

complexity. In the experiment, the phenomenon of negative thermophoresis of the vacancy diffusion is also observed in the memory cell. The kinetic Monte Carlo method simulates the formation under the crystal enhanced electric field as well as the effect of the grain boundary. The negative thermophoresis of vacancy diffusion, i.e. the vacancy diffuse from the site with a lower temperature to the sites with higher temperature, is observed in the experiments. This negative thermophoresis is modeled by modification of the event rate. The simulation shows that power dissipation in the unipolar RESET process is higher than the SET process, and the unipolar RESET is induced by the diffusion of interstitial oxygen atom inside the oxide.

Chapter 6 discusses possible future developments for the RRAM including the problem of the sneak path for bipolar reset device, the microscopic mechanism of heat generation in a thin filament and the time-dependent current through the thin filament. A brief conclusion is given.

Chapter 2. BASIC THEORY BACKGROUND

In this chapter, the basic theoretical background for this thesis is presented. The main results of the thesis involve the electrical conductivity of nanoscale memory elements. An important component in modeling the resistance is prior knowledge of the location of atoms in the structure. To determine the location of the atoms in nanostructures of interest, we use the density function theory formalism and the kinetic Monte Carlo method. The DFT methods are used in Chapter 3, where we study the location of Cu atoms in alumina, which form an atomic scale filament involving a small number of Cu atoms. The DFT calculations are used to minimize the energy of the atomic-scale Cu filaments and study the energy barriers for diffusion of Cu atoms. Once the structures are energy minimized, we use the localized orbital based DFT method to calculate the Hamiltonian, followed by the use of the Green's function method to calculate the current through the atomic-scale Cu filaments.

In this thesis, we also study the formation of resistance of larger nanoscale filaments that are formed by oxygen vacancies in Hafnia. To model these systems, we use kinetic Monte Carlo method to model the movement and rearrangement of atoms and vacancies because the number of atoms in the system are too large to be modeled by a DFT based methods. The input to the KMC simulations is derived from DFT calculations in literature (and sometimes from experiments). We then follow this with a rather electrical engineering oriented resistive network between vacancies to calculate the current flowing in the nanodevice.

2.1 DENSITY FUNCTIONAL THEORY

Density functional theory helps us to find the electronic structure from first principles. DFT method is widely used in computing material properties since it avoids solving the many-electron Schrodinger equation [35]. In this thesis, DFT method is used to minimize the energy of atomic structure and give the Hamiltonian for further calculation in chapter 3. We provide a brief summary of DFT here. The time-independent many-electron Schrodinger equation is

$$\mathrm{H(r,R)}\psi(r,R) = E(R)\psi(r,R). \tag{2.1}$$

Here $\psi(\mathrm{r,R})$ is the many-electron wave-function and $\mathrm{H(r,R)}$ is the Hamiltonian for the many-electron system. The electron coordinates are r and nuclei coordinates are R; the nuclei are assumed to be fixed. The Hamiltonian can be decomposed as

$$\mathrm{H(r,R)} = \frac{\hbar}{2m}\sum_i -\nabla_i^2 + \sum_i V_{ext}(r_i,R) + \frac{1}{2}\sum_{i\neq j} V_{ij}(r_{ij}). \tag{2.2}$$

Here, the first term on the r.h.s. is the kinetic energy of an electron, the second term is the electrostatic potential which consist of the contribution due to the nuclei and external field. r_i represents the coordinate of i-th electron. The last term represents the interactions between electrons at r_i and r_j. $V_{ext}(r_i,R)$ is the electrostatic potential felt by the i-th electron due to the set of all nuclei in the system which is collectively represented by R. Note that the symbol R is used to represent the set of all locations of the nuclei $\{R_k\}$.

To solve the many-electron system in equation (2.2) without approximations is considered impossible since the degrees of freedom is on the order of 3n, where n is the number of electrons in the system. However, the Hohenberg-Kohn theorem indicates that for the ground state of the n-

electron system, the density of electrons can determine the ground state property uniquely. Therefore, the energy is a functional of electron density, and can be written as

$$\mathrm{E}[\rho] = T[\rho] + V_{ext}[\rho] + V_{ee}[\rho]. \tag{2.3}$$

Here, $\mathrm{T}[\rho]$ is the kinetic energy functional. The functional due to an external field can be written as $\mathrm{V_{ext}}[\rho] = \int dr\, \rho(r) V_{ext}(r, R)$. The interaction between electrons can be further decomposed as two terms the Hartree term for Columbic interaction and the exchange-correlation ($E_{xc}[\rho]$) term,

$$\mathrm{V_{ee}}[\rho] = \frac{1}{2}\sum_{i \neq j} \int\int dr_i dr_j \frac{\rho(r_i)\rho(r_j)}{|r_i - r_j|} + E_{xc}[\rho] \tag{2.4}$$

The Kohn-Sham (KS) approximation further reduces the complex $\mathrm{V_{ee}}[\rho]$ term to an "effective" single particle potential which depends on the electron density$[\rho]$. The single particle Hamiltonian solved for the Kohn-Sham method is

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V_{eff}(r)\right]\phi_i(r) = \epsilon_i \phi_i(r), \tag{2.5}$$

where, the effective potential $V_{eff}(r)$ consists of both the external potential and an effective single particle potential to represent $\mathrm{V_{ee}}[\rho]$. The electron density at zero temperature is

$$\rho = \sum_{\mathrm{i}} |\phi_i(r)|^2 \tag{2.6}$$

where the summation is performed over all occupied states. The kinetic energy term in KS system is

$$\mathrm{T}[\rho] = \sum_i \int \phi_i^*(r)\left(-\frac{\hbar^2}{2m}\nabla^2\right)\phi_i(r)dr. \tag{2.7}$$

In DFT computations, the potential and electron density are solved for a self consistently given exchange-correlation functional. The exchange-correlation functionals usually employ density

functionals such as the local-density approximation (LDA) or generalized gradient approximation (GGA) for accuracy and speed balance.

DFT computations have proven to be useful in studying a number of materials. For example, the IP and EA can be defined in terms of the ground state energy (E_0) of a system with a constrained number of electrons (shown in the brackets below):

$$\mathrm{E_{EA}} = E_0(n+1) - \mathrm{E_0}(n), \tag{2.8}$$

$$\mathrm{E_{IP}} = E_0(n) - E_0(n-1), \tag{2.9}$$

$$\mathrm{E_{BG}} = E_{IP} - E_{EA}. \tag{2.10}$$

$\mathrm{E_{BG}}$ is the bandgap. It should however be noted that for most materials, density functionals based on LDA and GGA do not perform an adequate job of calculating the bandgap.

The formation energy of defects in a system (such as vacancy and interstitial) can be estimated from DFT calculations. For example, an oxygen interstitial or vacancy formation energy in HfO2 can be calculated as

$$\mathrm{E(I_O)} = E(HfO_2{:}I_O) - E(HfO_2) - \mu_O, \tag{2.11}$$

$$\mathrm{E(V_O)} = E(HfO_2{:}V_O) - E(HfO_2) + \mu_o. \tag{2.12}$$

Here, $\mathrm{E(HfO_2{:}I_O)}$ and $\mathrm{E(HfO_2{:}V_O)}$ are the energy of HfO_2 (search for this in other parts of the document) with oxygen interstitial and vacancy respectively. $\mathrm{E(HfO_2)}$ is the energy of pure HfO_2 and μ_O is the chemical potential of oxygen. The chemical potential of oxygen is related to the environment. For example, in the case of oxygen rich environment, the oxygen will be absorbed or released in the form of gas (O_2 molecules). Therefore, the chemical potential is half of the energy of an oxygen molecule:

$$\mu_\mathrm{O} = \frac{1}{2}E(O_2). \tag{2.13}$$

On the other hand, in Hafnium environment, the oxygen generation or annihilation will involve HfO2, and the chemical potential of oxygen will be

$$\mu_O = \frac{1}{2}[E(HfO_2) - E(Hf)]. \quad (2.14)$$

The system may be more complex than oxygen rich or Hafnium rich cases. Then, we need to analyze the possible location for the final state of oxygen atom and give the correct formation energy. Mixture of these cases is also possible under certain conditions. The metal electrodes in RRAM devices are known to affect the value of the formation energy. For example, near an oxygen reactive metal surface, the value of the formation energy next to a metal contact can be lower than the formation energy in a defect-free bulk of the oxide [36]. If the oxide is next to an inert metal like Platinum, the surface formation energy can be higher than the bulk formation energy [36]. In addition, if there are defects in the bulk of the oxide, the formation energy there can be smaller than the formation energy near a metal surface [37].These features are used in chapters 4 and 5 to model RRAM devices.

The force on atoms is another quantity can be calculated form DFT computations. In general, all the first derivatives of total energy of the ground state can be calculated from DFT. The force on atom i in direction α is,

$$\mathrm{F_{i\alpha}} = -\frac{\partial E(r,R)}{\partial R_{i\alpha}}. \quad (2.15)$$

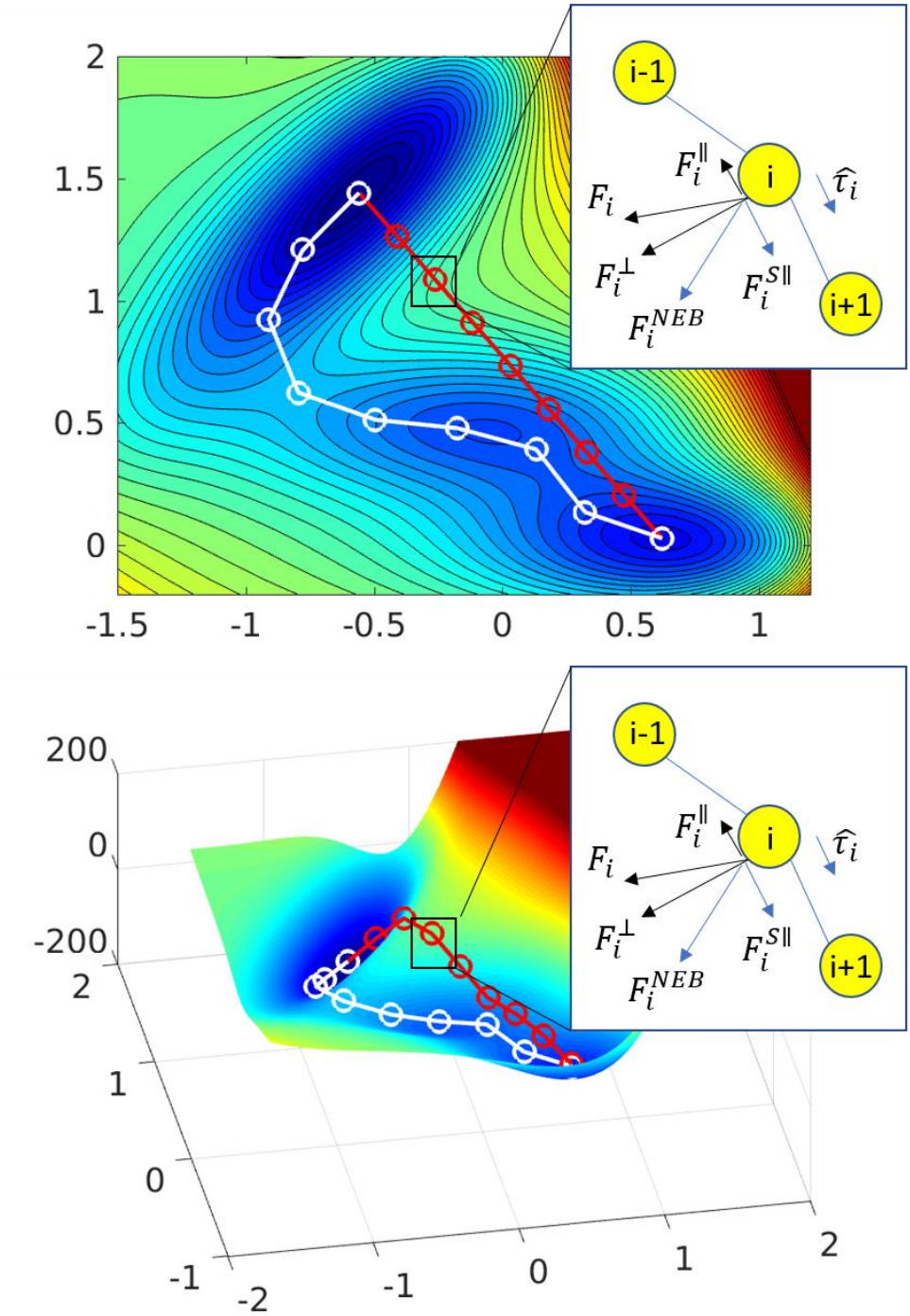


Fig. 2.1 (top) Illustration of NEB method. The image on PES can feel the force F_i, which can be decomposed to $F_i^{\perp}$ and $F_i^{\parallel}$. The image force parallel to the string is remove and another fictional force $F_i^{S\parallel}$ is added to keep images separated evenly on the MEP. (bottom) The energy surface is shown in 3D.

Here, the $R_{i\alpha}$ is the nuclear coordinate of the *i*-th atom along direction α. Since the system is in the ground state (for electrons), we can write the force as

$$F_{i\alpha} = -\frac{\partial}{\partial R_{i\alpha}}\langle\psi(r,R)|H(R)|\psi(r,R)\rangle$$

$$= -\frac{\partial\langle\psi|}{\partial R_{i\alpha}}H(R)|\psi\rangle - \left\langle\psi\left|\frac{\partial H(R)}{\partial R_{i\alpha}}\right|\psi\right\rangle - \langle\psi|H(R)\frac{\partial|\psi\rangle}{\partial R_{i\alpha}}. \quad (2.16)$$

Noticing that the ground state is an eigen state of the Hamiltonian,

$$H(R)|\psi\rangle = E|\psi\rangle \quad (2.17)$$

and the normalization condition is $\langle\psi|\psi\rangle = 1$, the force can be written as,

$$F_{i\alpha} = \left\langle\psi\left|-\frac{\partial H(R)}{\partial R_{i\alpha}}\right|\psi\right\rangle - E\frac{\partial\langle\psi|}{\partial R_{i\alpha}}|\psi\rangle - E\langle\psi|\frac{\partial|\psi\rangle}{\partial R_{i\alpha}} = \left\langle\psi\left|-\frac{\partial H(R)}{\partial R_{i\alpha}}\right|\psi\right\rangle. \quad (2.18)$$

In other words, given the self-consistent ground state, a calculation of the expectation value of $\partial H/\partial R_{i\alpha}$ in state $|\psi\rangle$ will yield the force. There is no need to perform another second self-consistent computation. With the force computed, molecular dynamics simulations including those that involve the melt-quench process, or molecular mechanics simulations to yield an optimized structure, can be performed.

The migration energy and minimum energy path (MEP) can also be derived from DFT calculations. By employing the nudged elastic band (NEB) algorithm, the transition state and migration barrier can be identified. The NEB method starts with initial, final and several intermediate states of a reaction. These states are optimized with the force calculated from the DFT computation together with the fictional repulsive force from other states in the coordinate space. As shown in Fig. 2.1, we can prepare 7 images between the initial and final states. Every image feels the force F_i on PES, with components $F_i^{\perp}$ and $F_i^{\parallel}$ perpendicular and parallel to the tangential direction $\hat{\tau}_i$ (I approximate this as pointing from image i to image $i+1$). In the NEB method,

while the force parallel to the string $F_i^{\parallel}$ is removed, another fictional force $F_i^{S\parallel}$ determined by the distance of images is added, and the image moves in the direction of $F_i^{NEB} = F_i^{\perp} + F_i^{S\parallel}$. As a result, the intermediate state will spread on the MEP evenly, and the migration energy is given from the maximum energy on the MEP.

2.2 NEGF AND CURRENT IN STEADY STATE

In order to calculate the current through devices (Fig. 2.2), the non-equilibrium Green's function (NEGF) method is often employed in nanostructures. The Schrodinger equation for a single electron in a potential H obeys,

$$H|\psi\rangle = \epsilon|\psi\rangle. \tag{2.19}$$

More explicitly, we can decompose the wave function as,

$$|\psi\rangle = |\psi_i\rangle + |\psi_s\rangle = \begin{pmatrix} |\psi_i^L\rangle \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} |\psi_s^L\rangle \\ |\psi_s^C\rangle \\ |\psi_s^R\rangle \end{pmatrix}. \tag{2.20}$$

Here $|\psi_i\rangle$ is the incident wave function, and $|\psi_s\rangle$ is the scattered one. The Hamiltonian can be decomposed in space as,

$$H = \begin{bmatrix} H_L & H_{LC} & 0 \\ H_{CL} & H_C & H_{CR} \\ 0 & H_{RC} & H_R \end{bmatrix}. \tag{2.21}$$

Here H_C is the Hamiltonian in the central (device) scattering region, H_L and H_R describe the semi-infinite electrodes, and $H_{LC}, H_{CL}, H_{CR}, H_{RC}$ are the coupling between the central region and leads respectively.

Inserting eqn (2.20) and eqn (2.21) to eqn (2.19), we have

$$(H - \epsilon I)|\psi_s\rangle = (\epsilon I - H)|\psi_i\rangle = -\begin{pmatrix} 0 \\ H_{CL}|\psi_i^L\rangle \\ 0 \end{pmatrix}. \tag{2.22}$$

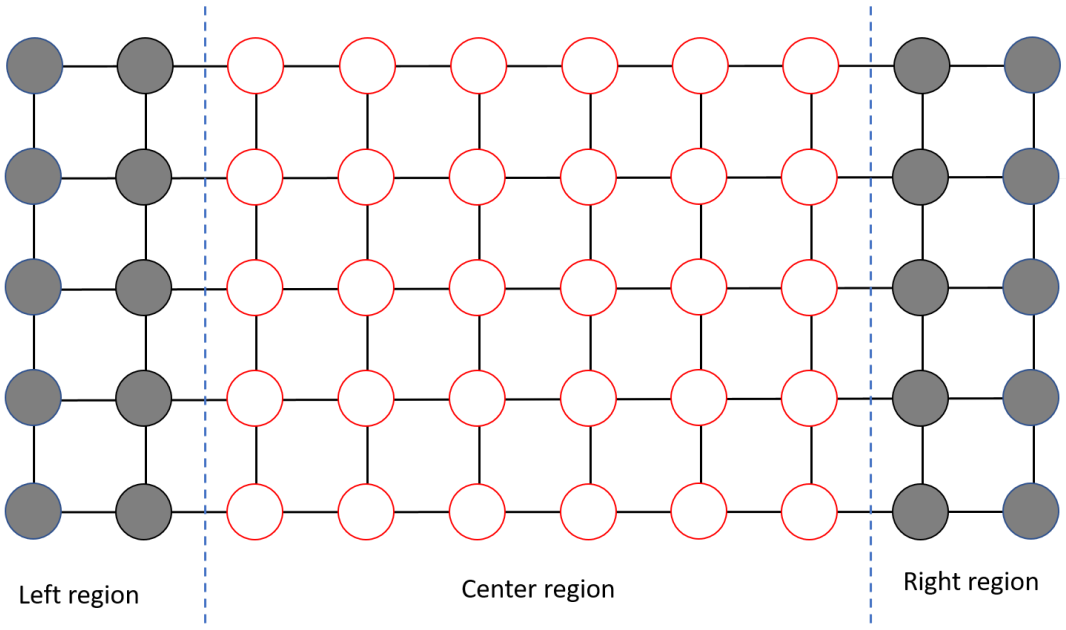


Fig. 2.2 The figure show the semi-infinite left and right regions (grey circles) and the central device region (red circles). The semi-infinite regions mimic contacts / open boundary conditions and the central region is where scattering occurs.

Then, the scattered wave function can be rewritten as

$$|\psi_s\rangle = G^r(\epsilon)\begin{pmatrix} 0 \\ H_{CL}|\psi_i^L\rangle \\ 0 \end{pmatrix}, \tag{2.23}$$

where $G^r(\epsilon)$ is the retarded Green's function in frequency domain, which is given by

$$G^r(\epsilon) = (\epsilon I - H - i\eta I)^{-1}. \tag{2.24}$$

The size of the Hamiltonian H (in matrix form) is infinity when considering the semi-infinite electrodes (Left (L) and Right (R) regions in Figure 2.2). The retarded Green's function can be written in matrix form as

$$\mathrm{G^r}(\epsilon) = \begin{pmatrix} G_L & G_{LC} & G_{LR} \\ G_{CL} & G_C & G_{CR} \\ G_{RL} & G_{RC} & G_R \end{pmatrix}. \tag{2.25}$$

Noticing that $\begin{pmatrix} 0 & H_{CL}|\psi_i^L\rangle & 0 \end{pmatrix}^{\mathrm{T}}$ is non-zero in the central region only, the scattered wave function is

$$|\psi_s\rangle = \begin{pmatrix} |\psi_s^L\rangle \\ |\psi_s^C\rangle \\ |\psi_s^R\rangle \end{pmatrix} = \begin{pmatrix} G_{LC}H_{CL}|\psi_i^L\rangle \\ G_C H_{CL}|\psi_i^L\rangle \\ G_{RC}H_{CL}|\psi_i^L\rangle \end{pmatrix}. \tag{2.26}$$

Then, the probability current carried by the transmitted state $|\psi_s^R\rangle$ can be written as, [38, 39]

$$\begin{aligned} \mathrm{i_t} &= -\frac{i}{\hbar}\left(\langle\psi_s^R|H_{RC}|\psi_s^C\rangle - \langle\psi_s^C|H_{CR}|\psi_s^R\rangle\right) \\ &= \frac{1}{\hbar}\left(-i\langle\psi_i^L|H_{CL}^\dagger G_{RC}^\dagger H_{RC} G_C H_{CL}|\psi_i^L\rangle + i\langle\psi_i^L|H_{CL}^\dagger G_C^\dagger H_{CR} G_{RC} H_{CL}|\psi_i^L\rangle\right) \\ &= \frac{1}{\hbar}\langle\psi_i^L|H_{LC}G_C\Gamma_R G_C H_{CL}|\psi_i^L\rangle. \end{aligned} \tag{2.27}$$

Here $\Gamma_\mathrm{R} = -iH_{CR}\left(g_R^\dagger - g_R\right)H_{RC}$, and $\mathrm{g_R} = (E - H_R)^{-1}$ is the surface Green's function. Then, the transmission coefficient is

$$\begin{aligned} \mathrm{T(E)} &= \sum_i \delta(\epsilon_i - E)\langle\psi_i^L|H_{LC}G_C\Gamma_R G_C H_{CL}|\psi_i^L\rangle \\ &= \sum_{i,j} \delta(\epsilon_i - E)\langle\psi_i^L|H_{LC}G_C\Gamma_R G_C|\phi_j\rangle\langle\phi_j|H_{CL}|\psi_i^L\rangle \\ &= \sum_j \langle\phi_j|\sum_i H_{CL}|\psi_j^L\rangle\,\delta(\epsilon_i - E)\langle\psi_i^L|H_{LC}G_C\Gamma_R G_C|\phi_j\rangle \\ &= \frac{1}{2\pi}\sum_j \langle\phi_j|\Gamma_L G_C\Gamma_R G_C|\phi_j\rangle = \frac{1}{2\pi}Tr[\Gamma_L G_C\Gamma_R G_C]. \end{aligned} \tag{2.28}$$

Here $|\phi_\mathrm{i}\rangle$ is a complete set of states and $|\psi_\mathrm{i}^\mathrm{L}\rangle$ are the eigenstates with energy $\mathrm{E_i}$ on the left region. And the current is the charge flow integrated for all possible energy, which can be written as

$$\mathrm{I} = \frac{2\mathrm{e}}{\hbar}\int dE\,[f_L(E) - f_R(E)]T(E).$$

Here,$\mathrm{f_L}(E)$ and $\mathrm{f_R(E)}$ are the Fermi functions on the left and right electrodes respectively.

In the calculation of current, we need the transmission coefficient, which relies on G_C and $\Gamma_{L/R}$. In order to compute the inverse of the Green's function in eqn. (2.24), Gaussian elimination method can be employed. Let $A = (EI - H)$, and the matrix A can be decomposed in L, C and R region as matrix H in eqn(2.21). Then multiply an elementary matrix P1 to A,

$$P_1 = \begin{bmatrix} I & 0 & 0 \\ -A_{CL}A_L^{-1} & I & -A_{CR}A_R^{-1} \\ 0 & 0 & I \end{bmatrix}$$

$$P_1 A = \begin{bmatrix} I & 0 & 0 \\ -A_{CL}A_L^{-1} & I & -A_{CR}A_R^{-1} \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} A_L & A_{LC} & 0 \\ A_{CL} & A_C & A_{CR} \\ 0 & A_{RC} & A_R \end{bmatrix}$$

$$= \begin{bmatrix} A_L & A_{LC} & 0 \\ 0 & A_C{}' & 0 \\ 0 & A_{RC} & A_R \end{bmatrix}, \tag{2.30}$$

Here $A_C' = A_C - A_{CL}A_L^{-1}A_{LC} - A_{CR}A_R^{-1}A_{RC}$. Then we can apply the second elementary matrix

$$P_2 = \begin{bmatrix} A_L^{-1} & 0 & 0 \\ 0 & A_C'^{-1} & 0 \\ 0 & 0 & A_R^{-1} \end{bmatrix}$$

$$P_2 P_1 A = \begin{bmatrix} A_L^{-1} & 0 & 0 \\ 0 & A_C'^{-1} & 0 \\ 0 & 0 & A_R^{-1} \end{bmatrix} \begin{bmatrix} A_L & A_{LC} & 0 \\ 0 & A_C{}' & 0 \\ 0 & A_{RC} & A_R \end{bmatrix} = \begin{bmatrix} I & A_L^{-1}A_{LC} & 0 \\ 0 & I & 0 \\ 0 & A_R^{-1}A_{RC} & I \end{bmatrix}. \tag{2.31}$$

The third elementary matrix is applied that

$$P_3 = \begin{bmatrix} I & -A_L^{-1}A_{LC} & 0 \\ 0 & I & 0 \\ 0 & -A_R^{-1}A_{RC} & I \end{bmatrix}$$

$$P_3 P_2 P_1 A = \begin{bmatrix} I & -A_L^{-1}A_{LC} & 0 \\ 0 & I & 0 \\ 0 & -A_R^{-1}A_{RC} & I \end{bmatrix} \begin{bmatrix} I & A_L^{-1}A_{LC} & 0 \\ 0 & I & 0 \\ 0 & A_R^{-1}A_{RC} & I \end{bmatrix} = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}. \tag{2.32}$$

By the uniqueness of matrix inversion, we can conclude that

$$A^{-1} = P_3 P_2 P_1$$

$$= \begin{bmatrix} I & -A_L^{-1}A_{LC} & 0 \\ 0 & I & 0 \\ 0 & -A_R^{-1}A_{RC} & I \end{bmatrix} \begin{bmatrix} A_L^{-1} & 0 & 0 \\ 0 & A_C'^{-1} & 0 \\ 0 & 0 & A_R^{-1} \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ -A_{CL}A_L^{-1} & I & -A_{CR}A_R^{-1} \\ 0 & 0 & I \end{bmatrix}$$

$$= \begin{bmatrix} * & * & * \\ * & A_C'^{-1} & * \\ * & * & * \end{bmatrix}. \tag{2.33}$$

Here star consists of nonzero elements that we do not require to calculate the current. So, we do not calculate them. Noticing the decomposition in equation (2.25), we have

$$G_c = A_C'^{-1} = (A_C - A_{CL}A_L^{-1}A_{LC} - A_{CR}A_R^{-1}A_{RC})^{-1}. \tag{2.34}$$

Usually, we denote $\Sigma_L = A_{CL}A_L^{-1}A_{LC} = H_{CL}(EI - H_L)^{-1}H_{LC}$ and $\Sigma_R = A_{CR}A_R^{-1}A_{RC}$ as the self-energies from left and right electrodes respectively. In the eqn (2.34), the dimensions of A_L and A_R are infinity, because we have assumed semi-infinite left and right electrodes, as shown in Fig. 2.2. But we can also observe that we only need the elements of A_L^{-1} that are connected to the center region (as shown in Fig. 2.2) to compute $A_C'^{-1}$. From eqn (2.21), A_L can be written as,

$$A_L = \begin{bmatrix} \ddots & \ddots & 0 \\ \ddots & a_{Ld} & a_{Lu} \\ 0 & a_{Ll} & a_{Ld} \end{bmatrix} = \begin{bmatrix} A_L & A_{Lu} \\ A_{Ll} & a_{Ld} \end{bmatrix}. \tag{2.35}$$

Here $A_{Ll} = A_{Lu}^{\dagger} = [\cdots \quad 0 \quad a_{Ll}\,]$. Then the inversion of this matrix can be presented as

$$A_L^{-1} = \begin{bmatrix} * & * \\ * & a_{Ld}'^{-1} \end{bmatrix}. \tag{2.36}$$

Here $a_{Ld}' = a_{Ld} - A_{Lu}A_L^{-1}A_{Ll}$. Note that because of the sparse property of A_{Ll} and A_{Lu},

$$a_{Ld}' = a_{Ld} - a_{Lu}(A_L^{-1})_{nn}a_{Ll}. \tag{2.37}$$

Here $(A_L^{-1})_{nn}$ refers to the sub-matrix of the left contact that is connected to the C-region, and we should note that this element is $a_{Ld}'^{-1}$. Thus, the element a_{Ld} satisfies the equation

$$a_{Ld}' = a_{Ld} - a_{Lu}a_{Ld}'^{-1}a_{Ll}. \tag{2.38}$$

Usually, we solve this equation iteratively as follows

$$(a'_{Ld})_{i+1} = a_{Ld} - a_{Lu}(a'^{-1}_{Ld})_i a_{Ll}. \quad (2.39)$$

Here $(a'_{Ld})_i$ means the approximated value of a'_{Ld} at the i-th iteration step. After the surface Green's function of A_L^{-1} is obtained, the Green's function can be obtained using matrix inversion packages. The NEGF method is used in chapter 3 to calculate the current through the RRAM device.

2.3 KINETIC MONTE CARLO

Using DFT calculations, the energy and force can be calculated at the atomistic level. However, DFT simulations of nanostructures considered in this thesis are very time consuming, if not impossible at this time. In order to accelerate the simulation and focus on the events we are primarily interested in, the kinetic Monte Carlo (KMC) method is employed. In the KMC method, the atoms are assumed to stay near a local energy minimum. It is implicitly assumed that an atom can vibrate in the potential of the solid, and these vibrations are not explicitly modeled. This vibration can result in atomic hopping with a probability that depends on an energy barrier. The KMC method takes this energy barrier as an input. Therefore, the KMC method saves plenty of computational resources compared to the ab initio molecular dynamics or force field molecular dynamics, which makes it viable to simulate a large system. The KMC method is used in chapter 4 and 5 to model the kinetics of defects in the oxide.

Consider an atom in a PES as shown in Fig. 2.3(a). It can hop from the initial state ($x = 0$) to the final state ($x = 1$). The energy along the MEP is plotted in Fig. 2.3(b). The probability of staying in the saddle point with velocity v, is

$$p_s = \frac{\exp\left(-\frac{mv^2}{2kT}\right)\exp\left[-\frac{V_s(x)}{kT}\right]}{Z_t} \approx \frac{\exp\left(-\frac{mv^2}{2kT}\right)\exp\left[-\frac{V_s(x)}{kT}\right]}{\int dx^{3N}\, dv^{3N} \exp\left(-\frac{mv^2}{2kT}\right)\exp\left[-\frac{V_m(x)}{kT}\right]}. \quad (2.40)$$

Here Z_t is the partition function in the trap, which is approximated as the partition function near the bottom of trap, and $V_s(x)$ and $V_m(x)$ are the potentials in the saddle and bottom of energy valley respectively ($x = 0$ and $x = 0.4$ in Fig. 2.3). Assuming the saddle point has a thickness of $dx_\parallel$, the transition rate is

$$r_i = \frac{\bar{v}_\parallel}{dx_\parallel} = \frac{\int dv\, v \exp\left(-\frac{mv^2}{2kT}\right)}{dx_\parallel} \frac{\int dx^{3N}\, dv^{3N-1} \exp\left(-\frac{mv^2}{2kT}\right)\exp\left[-\frac{V_s(x)}{kT}\right]}{\int dx^{3N}\, dv^{3N} \exp\left(-\frac{mv^2}{2kT}\right)\exp\left[-\frac{V_m(x)}{kT}\right]} \quad (2.41)$$

As $V_s(x)$ is a constant in the direction of $x_\parallel$, we can rewrite the rate

$$r_i = \frac{\int dv\, v \exp\left(-\frac{mv^2}{2kT}\right)}{dx_\parallel} \frac{\int dx^{3N-1}\, dv^{3N-1} \exp\left(-\frac{mv^2}{2kT}\right)\exp\left[-\frac{V_s(x)}{kT}\right]}{\int dx^{3N}\, dv^{3N} \exp\left(-\frac{mv^2}{2kT}\right)\exp\left[-\frac{V_m(x)}{kT}\right]} dx_\parallel$$

$$= \frac{\int dv\, v \exp\left(-\frac{mv^2}{2kT}\right)}{\int dv\, \exp\left(-\frac{mv^2}{2kT}\right)} \frac{\int dx^{3N-1}\, \exp\left[-\frac{V_s(x)}{kT}\right]}{\int dx^{3N}\, \exp\left[-\frac{V_m(x)}{kT}\right]}$$

$$= \sqrt{\frac{kT}{2\pi m}} \frac{\int dx^{3N-1}\, \exp\left[-\frac{V_s(x)}{kT}\right]}{\int dx^{3N}\, \exp\left[-\frac{V_m(x)}{kT}\right]} \quad (2.42)$$

The traps are assumed to be harmonic,

$$V_m(x) = \frac{1}{2}\sum m\omega_{mj}^2 x_j^2, \quad V_s(x) = E_{Barr} + \frac{1}{2}\sum m\omega_{sj}^2 x_j^2. \quad (2.43)$$

Here ω_{mj} are the harmonic frequencies in the energy minimum, and ω_{sj} are the harmonic frequencies at the saddle point. Then, the harmonic transition state theory gives, the rate of transition as [40]

$$r_i = r_0 \exp\left(-\frac{E_{barr}}{kT}\right). \tag{2.44}$$

Here, E_{barr} is the barrier energy, k is Boltzmann's constant and T is the temperature. The prefactor which is also known as the attempt frequency is given by,

$$r_0 = \frac{\prod_1^{3N} \omega_{mj}}{\prod_1^{3N-1} \omega_{sj}}. \tag{2.45}$$

In the KMC calculation, the Poisson process is assumed so that the total rate for all possible events is simply the sum of all rates,

$$r_{total} = \sum r_i. \tag{2.46}$$

The events considered in the system include the diffusion of vacancy, the diffusion of vacancy, the generation of Frenkel pair and recombination of Frenkel pair. The probability for event i to happen is then,

$$p_i = \frac{r_i}{r_{total}}. \tag{2.47}$$

For one single event, the elapsed time obeys an exponential distribution with the parameter of total rate (r_{total}), which means that the probability that the event happens in the time between T_1 and $T_1 + dT$ is

$$p_T(T_1)dT = r_{total} \exp(-r_{total} T_1)\, dT. \tag{2.48}$$

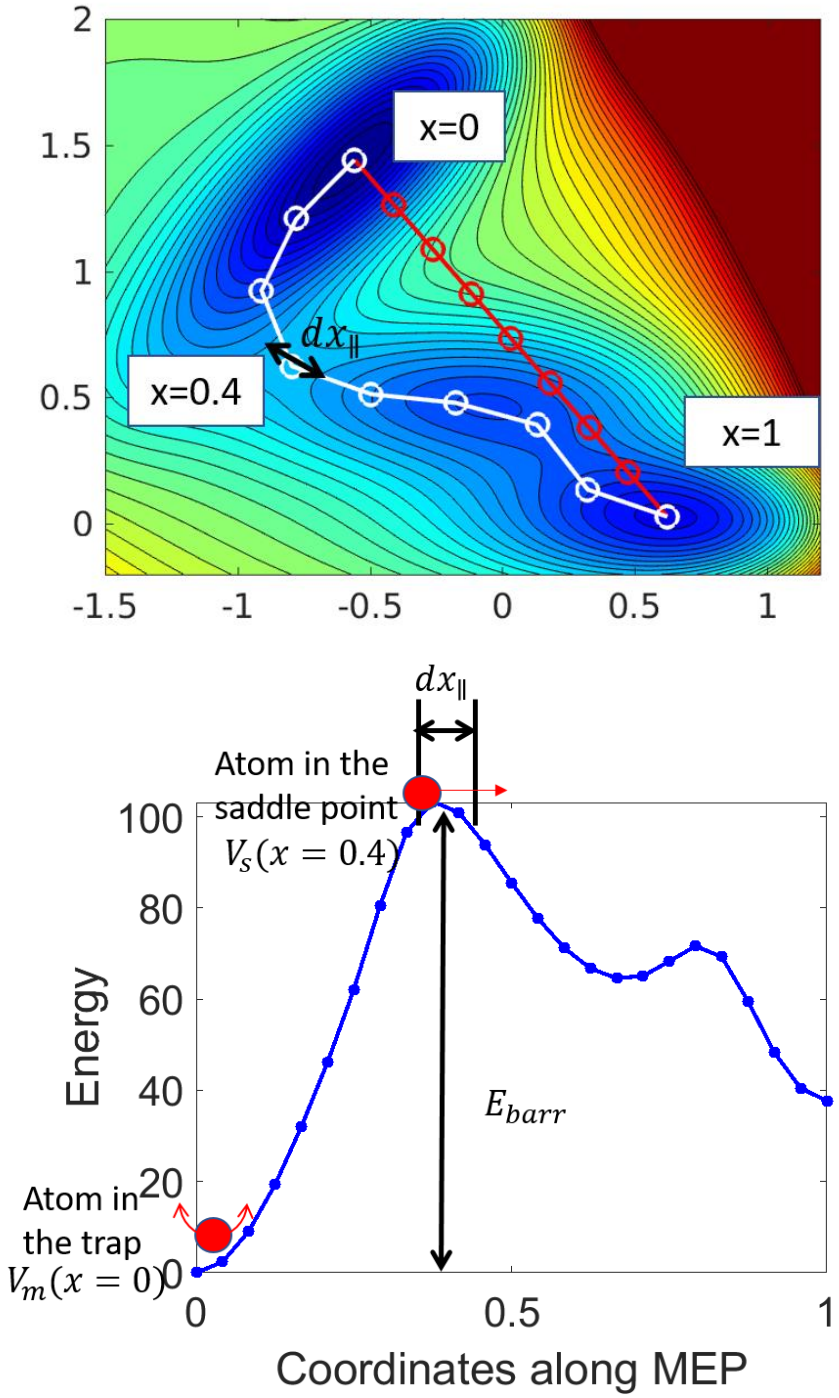


Fig 2.3. (top) An illustration of PES; atom in the initial state ($x = 0$) can overcome the energy barrier ($x = 0.4$) and reaches the final state ($x = 1$); (bottom) An illustration of energy along the MEP The height of the energy for an atom to leave the trap is E_{barr}. The atom can escape the trap when it reaches the saddle point with velocity v.

In the KMC simulation, the steps are:

1) Determine which event will occur in the simulation at the next time step using eqn (2.47).

2) Calculate the time elapsed before the next event that was determined in step 1) as in eqn (2.48).

3) Change the database where the location atoms are stored to account for the new location of all atoms as a result of the event in step 1. Recalculate all the transition rates in eqn (2.44) and the Frenkel pair generation rate. We will need these new transition rates when we iterate back to Step 1.

For step 1 above, we can use a random variable $0 < \chi_1 < 1$ to determine the event i, which should satisfy

$$\sum_{1}^{i} r_i/r_{total} < \chi_1 < \sum_{1}^{i+1} r_i/r_{total}, \tag{2.49}$$

For step 2), the elapsed time is calculated as

$$\mathrm{t} = \frac{\ln\chi_2}{r_{total}}. \tag{2.50}$$

Here $0 < \chi_2 < 1$ is another random variable uniformly distributed between 0 and 1.

For step 3), after one event happens, we should update the location of all related atoms and vacancies in the system (e.g. change the vacancy location in the system) and update all the related transition rates. Then the next iteration of simulation is performed by iterating back to step 1). A more detailed description can be found in Chapter 4 and 5.

Chapter 3. ELECTRON TRANSPORT THROUGH THIN FILAMENT[1]

In this chapter, the electron transport through an atomic scale filament is studied under steady state conditions. The assumption of steady state condition implies that the relaxation time for electrons is much smaller than the time to measure current. In the RRAM system, the conductive filaments are embedded in an insulating layer, which is sandwiched between metallic electrodes. The Green's functions in frequency domain are used to calculate the steady state current.

3.1 AN INTRODUCTION OF CURRENT THROUGH FILAMENT IN RRAM

A number of experimental studies use alumina in resistive switching [41-51]. Alumina is a desirable material because it has a large band gap and as a result the leakage current is small. It has been shown that the leakage current in a 1.2 nm alumina film is smaller than 10^{-4}A/cm^2 in experiment [49]. Also, the deposition of alumina using atomic layer deposition (ALD) is well studied and compatible with the advanced silicon CMOS technology. Unipolar RRAMs based on alumina have been reported in [52], and bipolar RRAM with a crossbar structure has also been reported in [45]. The RESET current reported in experiments can be as low as 1μA. Several groups have also studied a switching layer consisting of alumina stacked with other oxides [53, 54]. Therefore, alumina is a promising candidate as the switching layer in RRAM. Metallic filaments can be formed in alumina by interstitial Cu defects. In this chapter, the device physics of electron

[1] Most of the work in this chapter was published as: Xu Xu, Jie Liu, and M. P. Anantram, Conduction in alumina with atomic scale copper filaments, Journal of Applied Physics 116, 163701 (2014).

transport in metal-insulator-metal (MIM) configuration is studied, where the insulator is alumina and the metal electrodes are copper. The focus of this chapter is to understand the conductance with and without Copper atoms in the interstitials of alumina. Because the transport property is dominated by the insulating layer, both electrodes are chosen as copper to simplify the study.

The study of this chapter will yield insights into the smallest filaments that are capable to change the resistance of alumina appreciably and contribute to the understanding of the scaling properties of resistive memory devices. The results reveal that a few Cu atoms in Al_2O_3 can change the resistivity by about a thousand times at low biases. The saturation of current and negative differential resistance (NDR) that emerge with increase of bias are also discussed.

3.2 RRAM SYSTEM SETUP

The atomic positions of Al_2O_3 are obtained by variable-cell relaxation of the hexagonal 2×2×1 α-Al_2O_3 lattice, which has 120 atoms per unit cell. Using the conjugate gradient (CG) algorithm with periodic boundary condition, the Feynman-Hellmann forces on all atoms in the unit cell are less than 0.02 eV/Å, and the pressure on the unit cell is less than 1 GPa. We obtain lattice constants of a=4.7572 Å, c=13.0834 Å, which are close to the experimental values of a=4.7591 Å, c=12.9894 Å [55]. The calculated bandgap of Al_2O_3 is 6.0 eV, which is smaller than the experimental bandgap of 8.8 eV [56].

We first determine the interstitial sites in Al_2O_3 where a Cu atom can reside. The crystal structure has a number of hollow positions which can serve as an interstitial for Cu atoms. Since alumina is a material with high atomic density (1.16×10^{23} per cm^3), it is hard to find locations that the interstitial defects can reside. We have searched for locations larger than the covalent radius of a Cu atom (1.2 Å) in crystalline Al_2O_3, which can possibly hold interstitial Cu atoms. Then, we test

the three most possible locations for interstitial atoms, which are the three farthest positions from Al and O atoms. These positions are shown in Fig. 3.1.

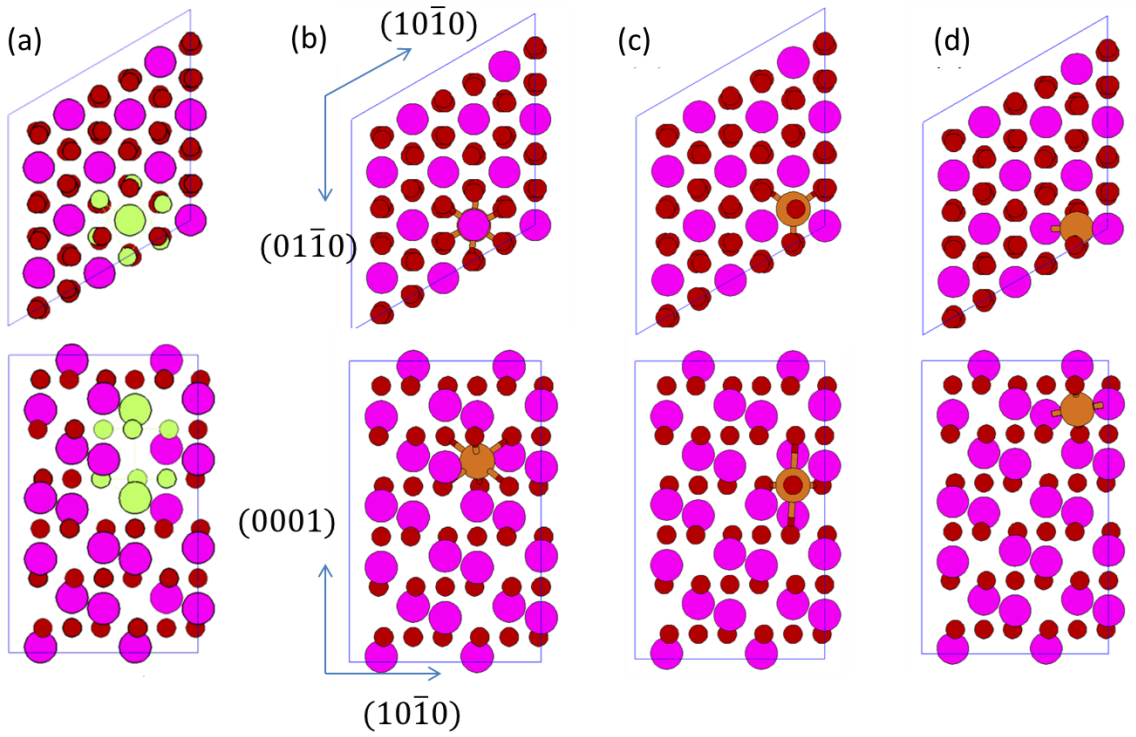


Fig. 3.1 (a) The structure of α-Al_2O_3 in a hexagonal unit cell viewed along (0001) and (10$\bar{1}$0) directions. The two Al and six O atoms highlighted in green enclose an interstitial for a Cu atom. (b), (c) and (d) are three different possible interstitial positions in α-Al_2O_3. Cu atom in (b) has the lowest energy.

In the hexagonal lattice of α-Al_2O_3, Al atoms form chains in the (0001) direction. The separations between two Al atoms on a chain are 3.82 Å and 2.674 Å, as shown in Fig. 3.1(a). The first interstitial position is shown in Fig. 3.1(b). It is located at the center of the 3.82 Å separation between Al atoms along the chain in (0001) direction. This position is the hollow site formed in α-Al_2O_3. The nearest neighbors for this interstitial position are two Al atoms, which are located along the (0001) chain and are 1.91 Å away. The second nearest neighbors are the surrounding six O atoms located on two (0001) planes with a separation of 2.12 Å, with the Cu-O distance of 1.97 Å, as shown in the highlighted atoms in Fig. 3.1 (a). Fig. 3.1(c) shows the second interstitial

position, which is along the O atom chain which lies along the (0001) direction. Three O atoms are closest to the interstitial position at a distance of 1.65 Å. And, Fig. 3.1(c) shows the third interstitial position, which is located on the plane formed by two Al and two O atoms. For this interstitial location, the nearest atoms are O atoms at a separation of 1.31 Å, and the second nearest neighbor atoms are two Al atoms at a distance of 1.39 Å. These three positions are the locations farthest from the atoms in crystalline Al_2O_3 structure, and therefore they are the most probable locations for the interstitial Cu atoms to reside.

We have considered these locations for the Cu atom and have performed energy minimization calculations to determine the most stable location. The formation (E_{form}) energy for the Cu atom is given by

$$E_{form} = E(Al_2O_3{:}Cu) - E(Al_2O_3) - E(Cu_{bulk}). \qquad (3.1)$$

Here $E(Al_2O_3{:}Cu)$ is the total energy of 120 atoms in Al_2O_3 with one Cu atom in the interstitial, $E(Al_2O_3)$ is the total energy of the 120 atom super cell of crystalline Al_2O_3, and $E(Cu_{bulk})$ is averaged the energy of fcc Cu whose super cell contains 64 atoms. The formation energy for the Cu atom located at the first possible interstitial position described above is 7.92 eV. We find that Cu atoms in either the second or third positions of the interstitials are relaxed to the position of the first interstitial. The relaxation gives the energy minimized structure with the constraint that the lattice parameters are the same as pure Al_2O_3, and local strain within this constraint is naturally included.

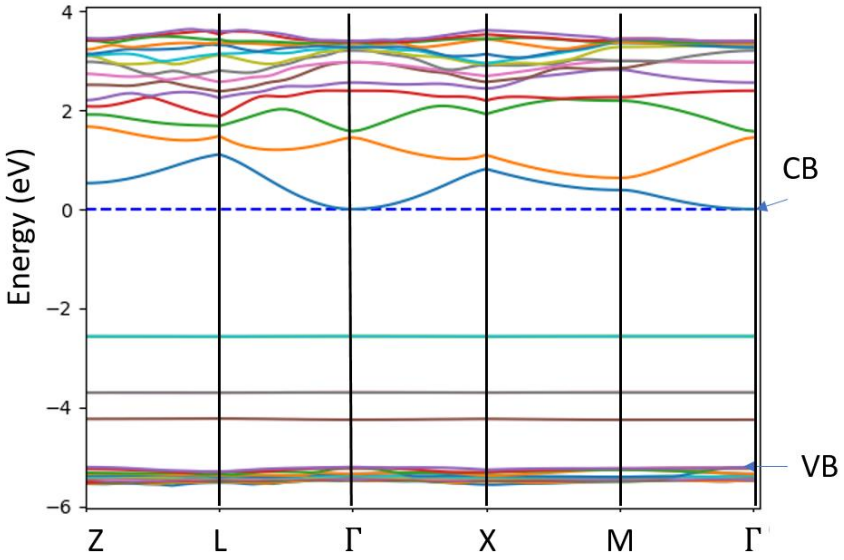


Fig. 3.2 The band structure and density of states of Al_2O_3 with Cu interstitial defect. Energy is shifted to the bottom of conduction band.

After the location of interstitial Cu atoms are determined, we perform DFT calculations with the Cu atoms to understand the change in electronic structure. In Fig 3.2, we observe that the Cu defects offer extra energy-states in the alumina band gap. These Cu-induced energy-states are mainly composed of 4s- and 3d- orbitals of Cu, which split into three energy levels. These states can either accept or donate electrons, and hence lead to an increase in conductivity when compared to pure alumina.

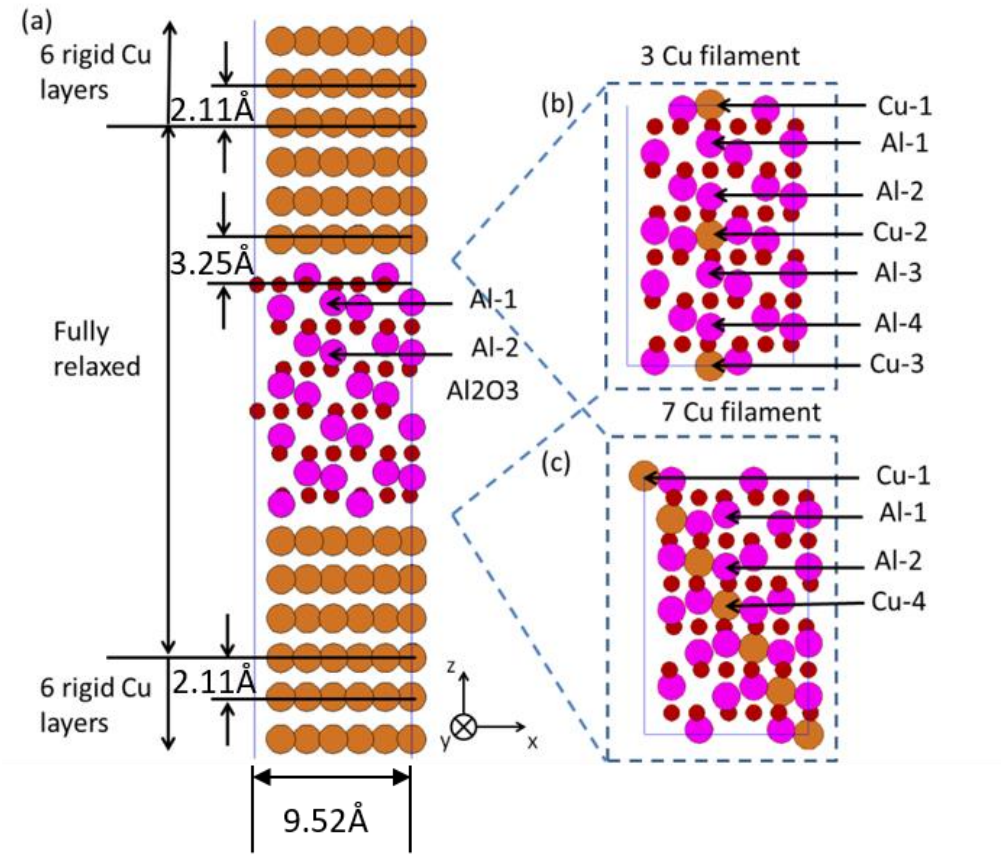


Fig. 3.3. Metal-Insulator-Metal (MIM) simulation models, in which alumina (a) without Cu filaments, (b) with 3 Cu atom filament along the (0001) direction, and (c) with 7 Cu atom filament along the $(1\bar{1}01)$ direction are shown. Periodic boundary conditions (PBC) are applied in the (x, y) plane. For clarity, only two out of the six rigid layers of Cu atom are shown in the figure.

Now, we consider two atomic filaments of Cu inside Al_2O_3. The first atomic filament considered has 3 Cu atoms that fill the interstitials along the (0001) direction, and the separation between Cu atoms equals to 6.549 Å (Fig. 3.3 (b)). Note that in this structure, there are two aluminum atoms between the consecutive copper atoms in the (0001) direction. The second filament considered has 7 Cu atoms in the $(1\bar{1}01)$ direction, with 3.25 Å separation between Cu atoms (Fig 3.3 (c)). The distances between two atomic filaments in the 3-Cu atom and 7-Cu atom cases are 9.514 Å and 6.958 Å respectively. We would like to emphasize that while we have relaxed these structures using energy minimization, these filament configurations do not emerge

from *ab initio* molecular dynamics simulations, which are extremely difficult for these large system sizes. Our intent here is to gain insight into the device physics for idealized filament configurations where the Cu atoms are placed in well-defined locations without any randomness, as shown in Fig 3.1 (c) and (d).

An ultrathin layer of Al_2O_3 that is one lattice constant thick (1.31 nm) is sandwiched by two Cu electrodes in our model. We note that such a calculation is computationally challenging because we include a total of 360 atoms (aluminum, oxygen and copper). However, there has been experimental work involving thin oxide layers. Reference [48] measured very small leakage currents in 1.2 nm thick alumina. Further, experiments on resistive memory have demonstrated successful reversible switching in oxides (GdO_x) of thickness of 2 nm [49] with a lateral feature size of 40 nm. Recently, reference [47, 57] has also demonstrated switching in alumina with thickness of 3 nm.

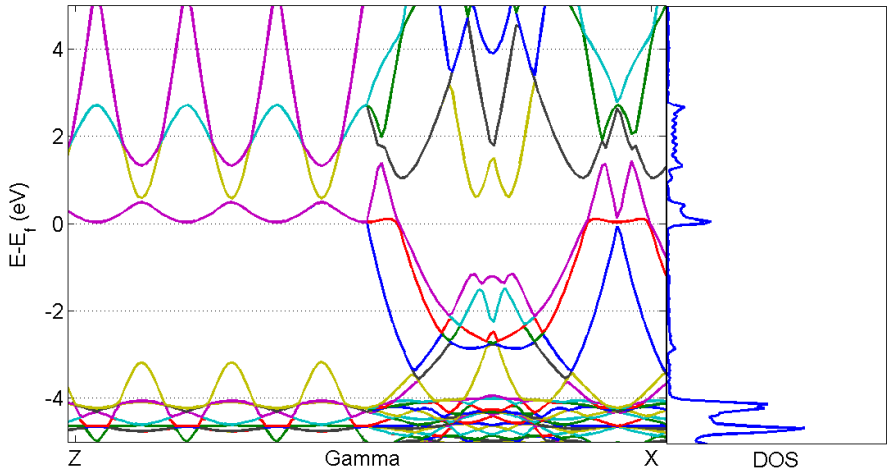


FIG. 3.4 Band structure and DOS of the 6.9% stretched Cu in the (111) plane. We can observe that the electronic structure is still metallic which expected give good conductance for the transport. The Fermi level is shifted to the zero of energy.

We can note that the surface of alumina on the (0001) planes forms a hexagonal lattice with a lattice constant of 9.52 Å (Fig. 3.3 (a)). On the other hand, fcc Copper with a lattice constant of 3.61 Å also has a hexagonal lattice in the (111) plane. The precise interface between Al_2O_3 and the copper contacts would depend on the experimental method for fabrication. This is a difficult problem in itself and here we are guided by reference [58, 59, 60, 61], which has studied the energy favorable contact between Al_2O_3 and Cu. We stretch the Cu lattice by 6.9% in the (111) plane to make the lattice constants of both Al_2O_3 and Cu at the metal-oxide interface identical [58]. Note that Cu remains metallic with no major changes, as shown in Fig. 3.4. For the electrode-alumina configuration, we first note that Al-ended Al_2O_3 is widely accepted as the most stable surface structure when exposed to vacuum. The O-top structure of Cu-Al_2O_3 contact is the most stable structure [58, 59, 60, 62]. In the O-top structure the Cu atoms sit directly on top of the first layer of O atoms near the interface. Our procedure for energy minimization of the O-top structure involves a full relaxation of a periodic structure consisting of four layers of Cu at each end of the Al_2O_3 as shown in Fig. 3.3 (a), with an initial Cu-O distance of 3.25 Å (distance d in Fig. 3.3(a)). We then attach a rigid structure consisting of six Cu layers (with an interlayer distance equal to 2.11Å) and perform a constrained energy relaxation where only the distance t can vary. The purpose of the outer six copper layers is to mimic crystalline copper, using which we will obtain the open boundary conditions to calculate the transport properties. Finally, a single point DFT calculation on the entire structure consisting of Al_2O_3 and a total of twenty Cu layers is performed to obtain the Hamiltonian (**H)** and overlap (**S**) matrices of the metal-insulator-metal (MIM) structures in the pseudo atomic orbital (PAO) basis.

The CG and DFT calculations are performed using the SIESTA package [63] with a Γ-point sampling since the unit cell is large; the plane wave cutoff is 300 Ry; the generalized gradient

approximation (GGA) of Perdew, Burke and Ernzerhof (PBE) [64] is used to estimate the exchange-correlation energy; the CG relaxation force threshold is set as 0.02 eV/Å. The single-zeta plus polarization (SZP) PAO is used for all atoms during relaxation since CG calculations with DZP is prohibitively high. However, all electronic structure calculations, including transmission, density of states, and current-voltage characteristics, are performed using the double-zeta polarized (DZP) PAO for the scattering region and SZP for the electrodes [65].

The transport properties of Al_2O_3 with and without Cu atoms are calculated using the equations described below. Note that while scattering due to various interfaces and the presence of Cu atoms in Al_2O_3 are included, our calculations do not account for decoherence. The transmission for electrons with energy E is given by

$$T(E,V) = Tr[\Gamma_T(E,V)G^r(E,V)\Gamma_B(E,V)G^a(E,V)], \tag{3.2}$$

Here the $G^r(E,V)$ is the retarded Green's function, which is given by

$$G^r(E,V) = [ES - V - H - \Sigma_T(E,V) - \Sigma_B(E,V)]^{-1}. \tag{3.3}$$

$\Sigma_{T/B}$ are self-energies of the top/bottom electrodes, which are constructed from the outer six Cu atom layers of Fig 3.3 (a). T and B refer to top and bottom electrodes respectively. In the presence of a voltage bias, the electrostatic potential (ϕ) is assumed to drop linearly along the z-direction in alumina. Thus, the term $V = q\phi$ corresponding to the electrostatic potential energy due to the external voltage drop is added to the Hamiltonian. The broadening matrices ($\Gamma_{T/B}$) are given by

$$\Gamma_{T/B}(E,V) = i\left[\Sigma_{T/B}(E,V) - \Sigma^+_{T/B}(E,V)\right], \tag{3.4}$$

The retarded Green's function naturally contains information that describes the scattering process. For example, the orbital-resolved local density of states (LDOS) can be calculated as, $LDOS_o(n,E,V) = -Im(G^r_{nn}(E,V))/\pi$ where $G^r_{nn}(E,V)$ is the nth diagonal element of matrix $G^r(E,V)$. And, the LDOS on atom is the summation of orbital-resolved LDOS over an

atom, $LDOS_A = \sum_{n\in\{atom\ i\}} LDOS_o(n,E,V)$. Here $\{atom\ i\}$ refers to the set of orbitals centered at the i-th atom. Moreover, the current-voltage relation is calculated using

$$I(V) = \frac{2q^2}{h}\int dE\ T(E,V)\left[f\left(E-\frac{V}{2}\right)-f\left(E+\frac{V}{2}\right)\right], \tag{3.5}$$

where $f(E)$ is the Fermi function at a temperature of 300 K.

3.3 TRANSMISSION AT ZERO BIAS

To understand how atomic scale Cu filaments affect the conductivity of Al_2O_3 film, we first present the transmission probability of electrons in the case of pure Al_2O_3, and the film with three and seven Cu atom filaments.

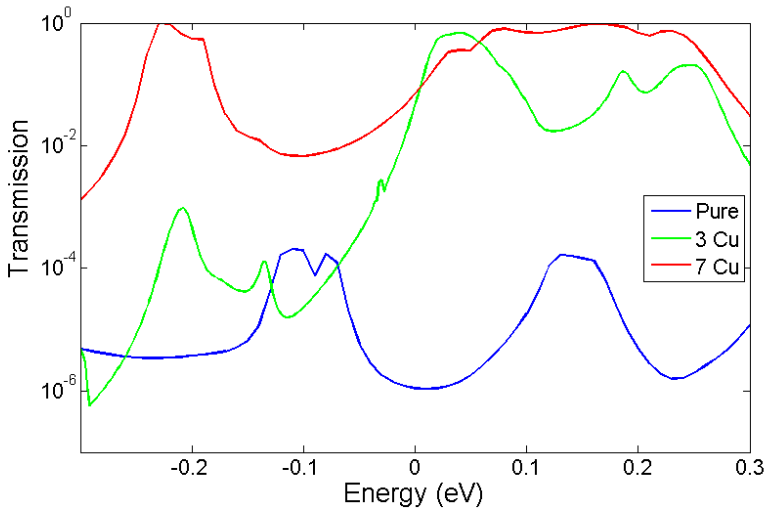


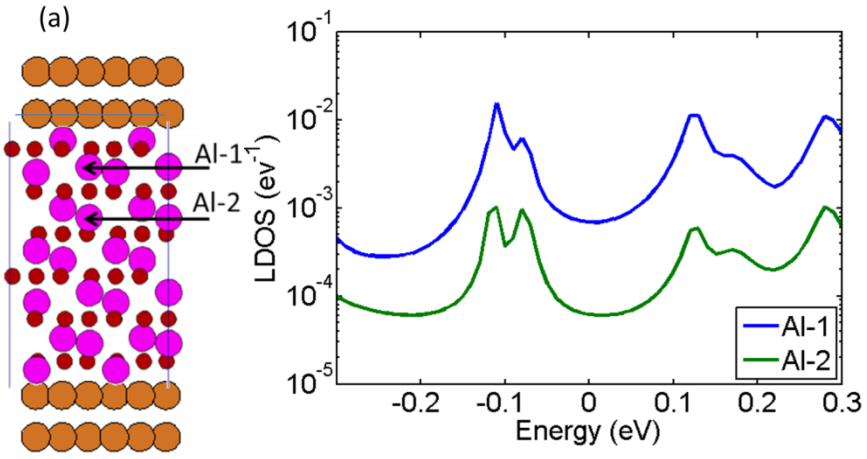
Fig. 3.5. Transmission spectrum for pure Al_2O_3 and the structures doped with 3 Cu atom and 7 Cu atom filaments. The energy axis is shifted so that the zero of energy is the Fermi energy at equilibrium.

For pure Al_2O_3, the tunneling probability through the 1.3 nm thick oxide is small at the Fermi energy (blue curve of Fig. 3.5) because of the large band gap. With the 3-Cu and 7-Cu atom filaments, the interstitial Cu atoms in the oxide create energy levels within the oxide band gap,

which increase the transmission significantly (red and green curves of Fig. 3.5). Our calculations show that both the 3-Cu atom and 7-Cu atom filaments change the transmission probability at the Fermi energy by over 10^3 times. The 3-Cu atom filament has two peaks in the transmission spectrum, and the transmission window width is approximately 0.12 eV while the 7-Cu atom filament has a wider transmission window approximately 0.2 eV wide. Note that while the peak transmission is close to unity as expected in a wide-narrow-wide geometry [66], the transmission at the Fermi energy is smaller than unity. As a result, the linear response conductance is smaller than the quantum of conductance $2q^2/h$, and low bias resistance is larger than 12.9 kΩ. The linear response resistances (R) calculated using

$$R = G^{-1} = \left(\int dE\, T(E) \frac{df(E)}{dE} \right)^{-1} \tag{3.6}$$

are 306.9MΩ, 78.37kΩ and 96.56kΩ for pure alumina, and the 3-Cu and 7-Cu atom filaments respectively. This implies that the diffusion of just a few Cu atoms can cause a huge change in the linear response resistance.



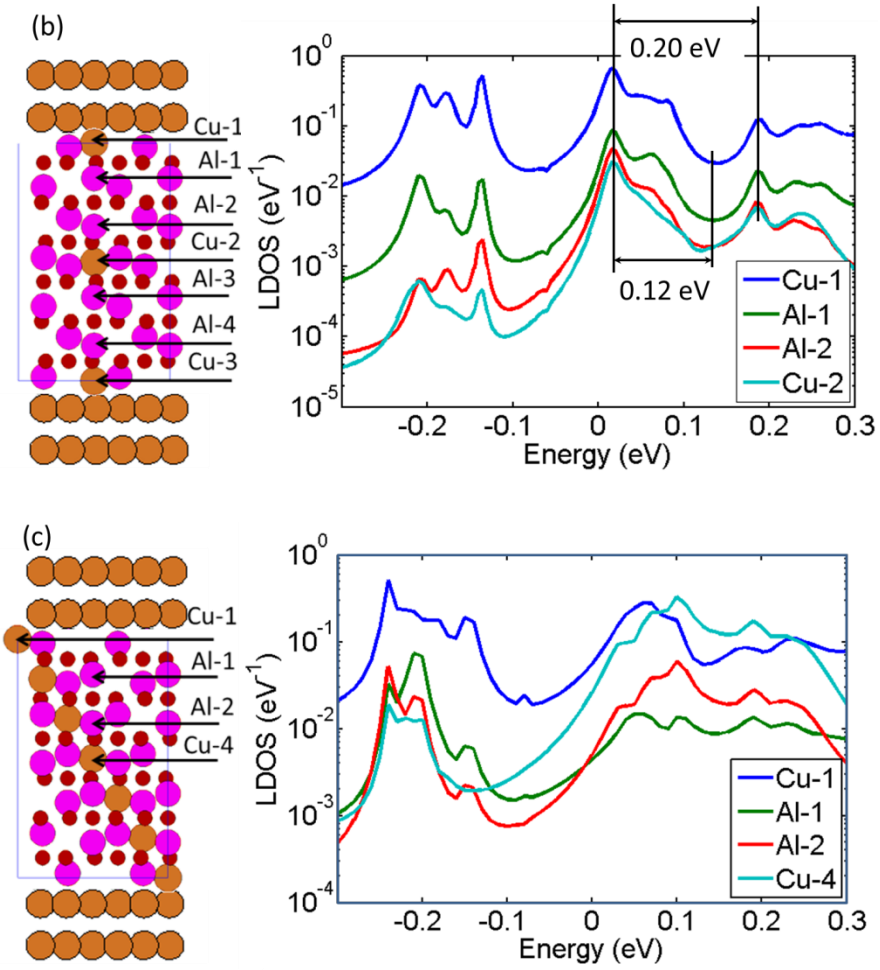


Fig. 3.6. LDOS for various atoms in the alumina layer for (a) pure alumina, (b) alumina with the 3-Cu atom filament and (c) alumina with the 7 Cu atom filament. LDOS is calculated with electrodes in the MIM structure. We find that the Cu atoms can increase the LDOS of neighboring Al atoms. Note that the energy axis is shifted such that the zero of energy is the Fermi energy at equilibrium.

The analysis of LDOS provides further insight. We plot the LDOS at five atoms to determine the conducting path (Fig. 3.6): Cu-1 (Cu atom near the left electrode), Al-1 and Al-2 (Al atoms in alumina), Cu-2 (Cu atom in the middle of the 3-Cu atom filament), and Cu-4 (Cu atom in the middle of the 7-Cu atom filament). As shown in Fig. 3.6(a), we observe that for pure alumina, the metal induced gap states penetrate into the pure alumina. However, the transmission is still low

near Fermi level because LDOS on Al-2, which is furthest away from the electrodes, has a small value. Thus, the low transmission results in high resistance in the OFF state, and therefore leads to high OFF/ON resistance ratio. On the other hand, atom Al-2 has a much higher LDOS in the 3-Cu atom filament because of close proximity to the filament (Fig. 3.6(b) and 3.6(c)), which causes the higher conductance. The conductive path for 3-Cu filament is along (0001) direction since atom Al-1 also has a large LDOS in the 3-Cu filament (Fig. 3.6(b)). In comparison, the conductive path in the 7-Cu filament is along the $(1\bar{1}01)$ direction. The LDOS on the aluminum atom (Al-2) is similar to the 3-Cu filament case since it is close to the Cu filament, while the LDOS is much smaller on Al-1 atom because it is located further away from the Cu-atom filament (Fig. 3.6(c)). To sum up, Cu atoms in Al_2O_3 increases the LDOS of atoms nearby, which aid to form the conductive path.

In order to understand the transmission of the filament in α-Al_2O_3 film better, we decompose the contribution to orbitals with different angular momentum along the z-direction and analyze their capability for carrying current. To carry out this analysis, we use the real space orbitals in SIESTA, where p_z and d_{z2} are the *m=0* components for p-and d-orbitals respectively. Besides the LDOS on different atoms, we can analyze the orbital resolved LDOS. Since the conductive filament of 3-Cu filament is along (0001) direction, we mainly focus on the central atoms, i.e. Cu-2 and Al-2, which are the bottleneck for conduction. For Cu-2, the LDOS contribution is mainly provided by 4s, $4p_z$ and $3d_{z2}$ orbitals near the Fermi level as in Fig. 3.7 (a). At the Al-2 atom, the LDOS contribution at the Fermi level is mainly from the 3s and $3p_z$ orbitals, as shown in Fig 3.7 (b). That is to say, the *m=0* orbitals on the central Cu atom and neighboring Al atoms along the (0001) direction form the conductive path for electrons. Physically, these orbitals contribute to current most, because

they are located near the Fermi level and these orbitals are naturally elongated along the transport direction.

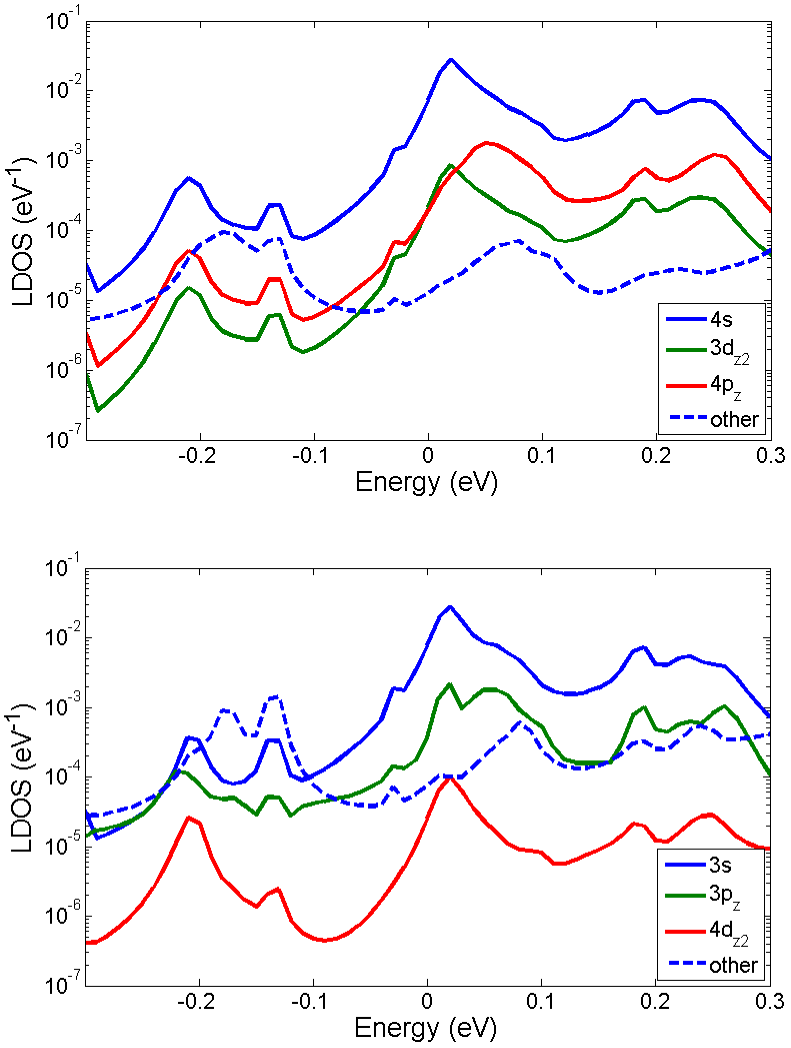


Fig. 3.7. LDOS on orbitals of the Cu-2 atom (upper) and Al-2 atom (lower) of the 3-Cu atom filament. *m=0* orbitals are shown by solid lines and the sum of other orbitals are shown by dashed lines. We find that the *m=0* orbitals provide the large LDOS near the Fermi level and form the conductive path in Al and Cu. The Fermi energy is at $E = 0$.

3.4 Current-Voltage Relation

We find that the current is larger for the 3-Cu atom chain when compared to the 7-Cu atom chain at applied biases smaller than 50 mV (Fig. 3.8), which agrees with the above results for linear

response resistance. For biases larger than 50 mV, there are three important features: (i) the current in the 3-Cu and 7-Cu atom filaments saturate at biases larger than 50 mV and 0.25V, (ii) negative differential resistance is observed in both filaments, with it being more pronounced in the 3-Cu atom filament and (iii) there is a large difference (approximately twenty times) in current between the 3-Cu and 7-Cu atom filaments when the applied bias is in the range of 0.4 V. This result is interesting because it shows that small changes in the filament can give rise to large changes in the features of the current-voltage characteristics. These results show that multi-level logic is feasible in the structures considered if the filament shape can be modified in a controlled manner. The reason for the above features is discussed now.

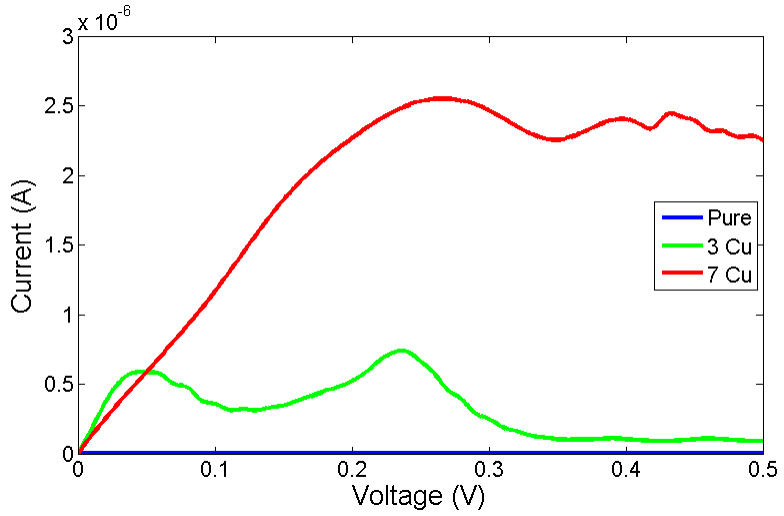


Fig. 3.8. Current-voltage characteristics of pure alumina and the structures with the 3-Cu and 7-Cu atom filaments. There is a large difference in current between the 3-Cu atom and 7-Cu atom filament cases when the bias is around 0.4 V. Features of negative differential resistance are seen in the current voltage characteristics, which occur due to a mismatch in the LDOS.

When a voltage is applied, the quasi-Fermi levels on left and right electrodes shift by $+eV/2$ and $-eV/2$ respectively, and the transmission at energies in between the two quasi-Fermi levels contributes to current flow. The reason for current saturation is the narrow transmission window

seen in Fig. 3.5. The narrower transmission peak in the case of the 3-Cu filament causes a current saturation at smaller biases, when compared to the 7-Cu atom filament. The negative differential resistance (NDR) is caused by the mismatch of LDOS along the conductive path. This notion is studied further by looking at the transmission spectrum and LDOS under finite bias for the 3-Cu filament. The NDR in the I-V curve emerges from a density of states effect and this should be observable at room temperature (300K) given that we have atomically narrow filaments. We find that the transmission versus energy has two peaks (green line of Fig. 3.5 and blue line of Fig. 3.9). Each peak split into two sub-peaks when a bias is applied as shown in Fig. 3.9(a) – while one sub-peak is shifted by $+eV/2$, the other sub peak is shifted by $-eV/2$. The reason for the split in transmission spectrum is that the LDOS of the Cu-1 and Cu-3 are shifted by energies approximately equal to $\pm eV/2$ (Fig. 3.9(b)) compared to the zero-bias case (Fig. 3.6(b)). The LDOS of Cu-2 has split into two peaks, one induced by the left electrode (shift by $+eV/2$) and the other induced by the right electrode (shift by $-eV/2$). When the LDOS on the different Cu atoms are matched in energy, there is a large current. The current-voltage characteristic in 3-Cu filament case shows a valley at 0.15 V and a second peak at 0.23 eV. Corresponding to these features, in the plot of LDOS (Fig. 3.6), the peak-to-valley distance is 0.12 eV and peak-to-peak distance is 0.20 eV. When the applied bias is 0.15 V, the LDOS peak on Cu-1 overlaps with the valley on Cu-3, resulting in small current, which explains the negative differential resistance. And, when the applied bias is 0.25 V, the LDOS peak on Cu-1 overlaps with the LDOS peak of Cu-3, which explains the second peak in current. Our physical picture here shows that the negative differential resistance and the second peak in the current-voltage curve correspond to the energy differences in the peak(Cu-1)-valley(Cu-3) and peak(Cu-1)-peak(Cu-3) features seen in the LDOS. In comparing the calculated result of current-voltage relation to the experiment, we can see that the

sublinear current behavior is observed in [41]. However, the exact match of current-voltage relation is not observed since the exact atomic configuration in the experiment may be different and the electron-phonon interaction is neglected in the calculation.

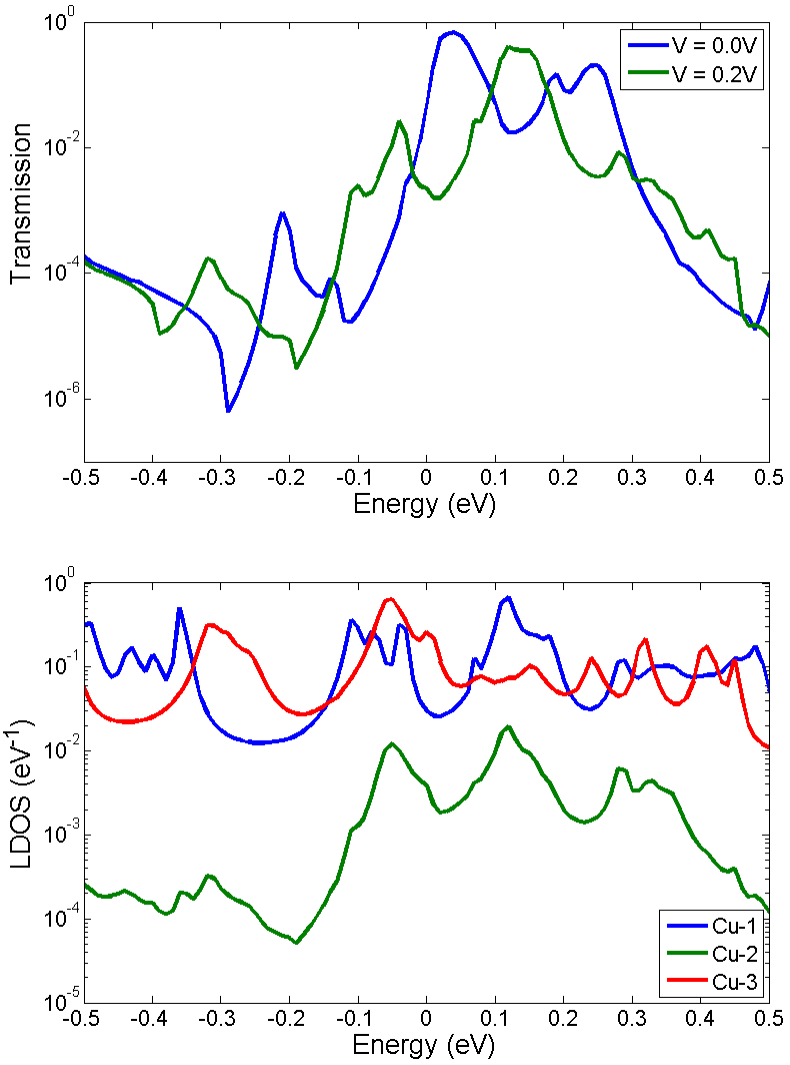


Fig. 3.9. (a) Transmission spectra of the 3-Cu atom filament under zero bias and a bias of 0.2 Volts; (b) LDOS of the 3-Cu atom filament when the bias is 0.2 V. One can see a split in LDOS by $\pm eV/2$, when a voltage of V is applied.

To summarize our results above again, we presented the electron transport properties of pure and doped Al_2O_3 using first principle calculations. Our thin film structure consisted of a 1.3 nm

thick α-alumina with crystalline copper electrodes. We find that the low bias conductance of crystalline alumina can be changed by more than a thousand times by incorporating few copper atoms in interstitials to form a filament. We show the current of the 3-Cu filament saturates at smaller voltages than the 7-Cu atom filament. The current saturates at much higher values in the 7-Cu atom filament, which is consistent with the greater hybridization between Cu atoms as a result of the smaller distance between Cu atoms in the 7-Cu atom filament. This shows that small changes in the filament structure can lead to large changes in the current-voltage characteristic in atomic scale filaments, a conclusion that is of relevance to multi-level logic. Finally, we find that the mismatch in the LDOS at Cu atoms along the filament in the presence of an applied bias leads to a negative differential resistance.

Chapter 4. KINETIC MONTE CARLO SIMULATION OF FILAMENT FORMATION[2]

In the previous chapter, we have investigated how the filament shape can change the IV characteristics of the memory cell. An asymmetric IV characteristics will help to solve the sneak path problem and enhance the force to RESET, as discussed in chapter 6. Thus, it is important to understand how to modulate the shape of filament to fabricate a memory cell array with high density and fast switching. Because of the computationally challenging nature of the problem, the ab initio method used in the previous chapter cannot be used to simulate memory cells, which are usually around 10 nm thick. Thus, in order to simulate the filament shape evolution under the forming voltage pulse, the kinetic Monte Carlo method is employed in this chapter since it is capable of modeling large filaments [67, 68, 69].

The material of the switching layer in RRAM can be chosen as Al_2O_3, HfO_2 or other types of oxides [70-79]. The function of the switching layer is to offer sites where defects can reside. The defects can be either metallic interstitials or oxygen vacancies, which are usually conductive. The defects in the oxide can carry a nonzero charge. Thus, these charged defects will move as a result of the applied voltage. For the case of simplicity, the defects are assumed to be positively charged vacancies, which is usually true for oxygen vacancies in the oxide. To conserve the number of oxygen atoms, the oxygen interstitial in the oxide should also be included. Other types of defects can be generalized without a significant modification of this model.

---

[2] Most of the work in this chapter was published as X. Xu, B. Rajendran & M. P. Anantram, "Kinetic Monte Carlo Simulation of Interface-Controlled Hafnia-Based Resistive Memory," in IEEE Transactions on Electron Devices, vol. 67, pp. 118-124, Jan. 2020.

In the metal oxide, the defects can form inside the oxide layer and increase the conductivity of the insulating layer. In the crystal lattice of oxide, the oxygen atoms and metal atoms are arranged in crystalline pattern. An *interstitial oxygen defect* is one where an Oxygen atom moves from its location in a perfect crystal to a location where there were no atoms in the perfect crystal (in between atoms). See Fig. 4.1. An oxygen atom that is missing from its location in a perfect crystal is referred to as the vacancy. If an oxygen atom moves from its original location in a crystal and moves to an interstitial location, an interstitial-vacancy pair is formed. This is known as a *Frenkel pair*. The energy required to forma Frenkel pair is referred to as the Formation energy in this thesis (see Fig. 4.1). The energy required for an interstitial (or vacancy) to diffuse is known as the migration barrier.

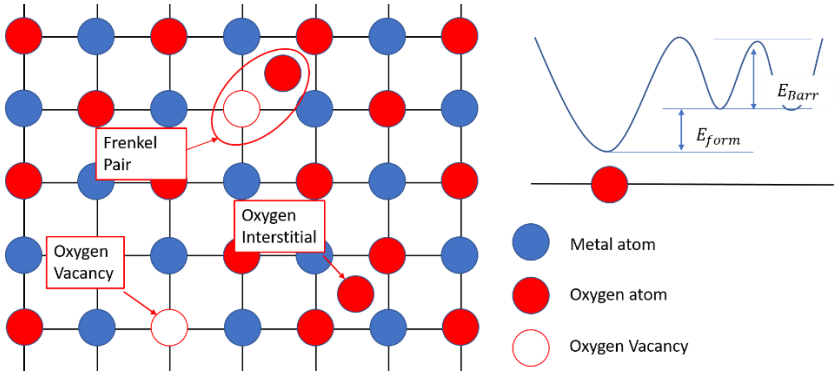


Fig. 4.1 The illustration of oxygen interstitial, vacancy and Frenkel pair. Formation energy (E_{form}) refers to the energy penalty to generate one vacancy-interstitial pair. Migration energy E_{barr} is the energy barrier for Diffusion.

4.1 Mathematical model of Monte Carlo simulation (independent of 2D / 3D)

The migration of vacancies is simulated using the Monte Carlo model. The migration processes include vacancy formation, recombination, and diffusion. The formation process involves the

generation of a Frenkel pair, consisting of an oxygen vacancy and an oxygen interstitial. In this thesis, I consider two types of formation processes (Fig. 4.2):

(a) Bulk formation (occurs in bulk of oxide)

(b) Surface formation (occurs at oxide-metal interface)

The formation process can take place in the bulk oxide. Under the assumption of thermal equilibrium, the rate of Frenkel pair generation is

$$R_G = r_0 \exp\left(-\frac{E_{form} + E_{barr}}{kT}\right). \tag{4.1}$$

Here, r_0 is the attempt frequency (described in Chapter 2) is not calculated explicitly in this work. r_0 usually ranges from $10^{13}\ s^{-1}$ to $\sim 10^{15}\ s^{-1}$in crystalline solids [40]. E_{barr} is the energy barrier for diffusion (migration barrier) [69, 80], k is the Boltzmann constant, and T is the temperature. The illustration of energy is shown in Fig. 4.1. The recombination process is the reverse of formation. When an oxygen vacancy meets an interstitial, they can recombine to form the native lattice with a rate

$$R_R = r_0 \exp\left(-\frac{-E_{form} + E_{barr}}{kT}\right). \tag{4.2}$$

The Frenkel pair generation can also occur at the oxide-metal interface and this is called surface formation. When the Frenkel pair is generated, the oxygen vacancy stays in the oxide while the interstitial oxygen goes into the metal. Such a process was first discussed in [36]. Also, the experiments, such as [81, 82], show that the material of electrodes can change the switching behavior. The underlying reason for a reduced formation energy is that the oxygen can form chemical bond with the metal in the electrode and therefore lower the total energy.

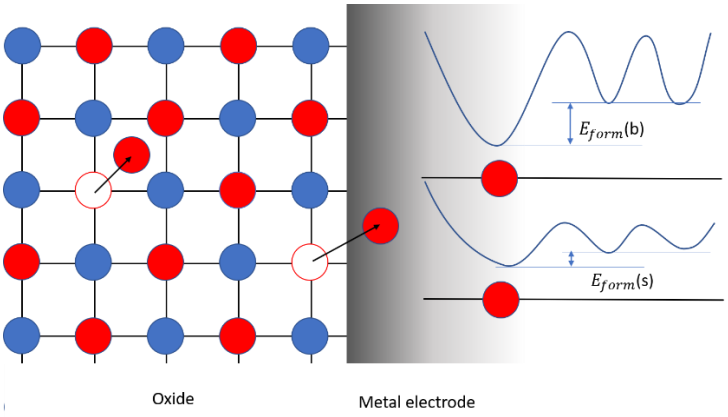


Fig. 4.2 The illustration of formation in the bulk and surface formation. In the bulk formation, the oxygen vacancy is generated with an oxygen interstitial in the oxide. In surface formation, the oxygen vacancy is generated at the interface of metal and oxide, with an oxygen moving into the metal directly. The formation energy for bulk generation and surface generation are $\mathrm{E_{form}}(b)$ and $\mathrm{E_{form}}(s)$ respectively.

The formula for the rate of surface formation/recombination is similar to bulk formation, except that the bulk formation energy should be replaced by the surface formation energy in equation (4.2) to give,

$$R_G = r_0 \exp\left(-\frac{E_{form} + E_{barr}}{kT}\right). \tag{4.3}$$

The diffusion process for oxygen vacancy consists of migration of the *on-lattice* oxygen to fill the oxygen vacancy, which effectively moves the oxygen vacancy. The rate of the diffusion is given by:

$$R_D = r_0 \exp\left(-\frac{\Delta E + E_{barr}}{kT}\right). \tag{4.4}$$

Here, ΔE refers to the energy change due to vacancy movement. Two factors that determine this energy change: the interaction between vacancies and the electrostatic potential energy. Thus,

$$\Delta E = \Delta\, E_{int} + \Delta\, E_{ele}/2. \tag{4.5}$$

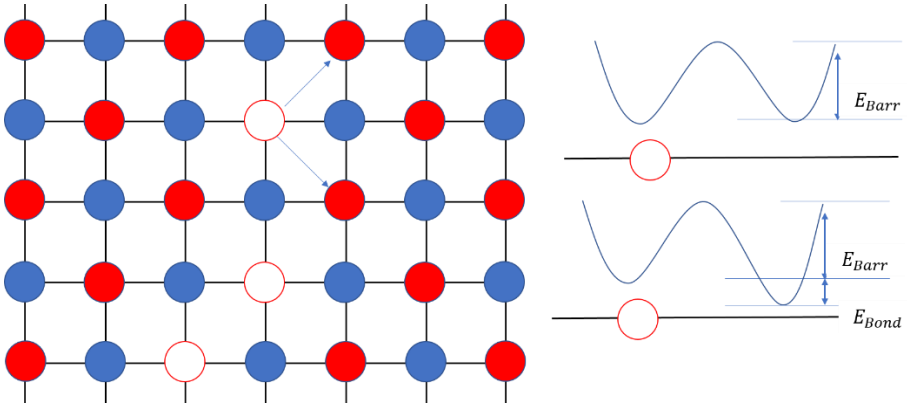


Fig. 4.3 The illustration of bond energy. The vacancy can move in direction of upper-right and down-right. After it moves to the down-right, the rate it moves again is smaller since it has to break the bond to move further.

The interaction energy can be approximated as the bonding energy between vacancies that occupy adjacent sites (show these sites in a figure), which gives

$$\Delta E_{int} = \Delta n \, E_{bond}. \tag{4.6}$$

Here, Δn is the change of bond number due to diffusion, and E_{bond} is the energy contribution per vacancy-vacancy bond. And the interaction energy is only considered when the difference is positive. The underlying reason for bond energy includes Coulomb and exchange energy between vacancies. The calculation in [83] shows that the energy to generate a vacancy is decreased when a vacancy is already existing. In other words, the bond formed between vacancies can lower the total energy and results in a more stable filament. For example, as shown in Fig. 4.3, the vacancy can move in the direction towards the upper-right or down-right. If the vacancy moved to upper-right the rate it moves again will not change, because the bond number is not changed before and after the moving. However, if the vacancy moves in the direction of downright, the total energy will be lowered by E_{bond}, and the rate of subsequent movement of this vacancy will be lower.

In estimation of the hopping time of vacancies and interstitials, the charge state of defects should be considered. The main reason that the charge state is important is that the migration barrier can

change due to charge state. The migration energy of neutral vacancies is around 1.1 eV, which means that it takes $r_0^{-1}\exp(E_{barr}/kT) \approx 1.285 \times 10^4\ s$ on average for one hopping to occur at room temperature. This hopping time is too long compared to experiments; the experimental time for switching is within milliseconds. On the other hand, the charged vacancy can migrate faster if the energy barrier is lower. Ab initio computations suggests that the migration barrier is around 0.5 eV [81, 82] for a charged vacancy, which means that hopping can occur in $r_0^{-1}\exp(E_{barr}/kT) \approx 4.85 \times 10^{-7}s$.

The electrostatic potential energy difference in a vacancy-move event is

$$\Delta E_{ele} = q\ \Delta\phi. \tag{4.7}$$

q is the charge of a vacancy, which is taken to be $+2e$ in the oxide for the case of Hafnia and titanium electrode. In the general case (different oxide or metal electrode), the charge can be different from $+2e$. $\Delta\phi$ is the change in electric potential between the initial and final locations of the vacancy. In the formation process, the formed Frenkel pair is surrounded by oxide, which is an insulating environment. Therefore, the net charge of the whole Frenkel pair has to be zero. Calculation in [84, 36, 85] shows that the donor state of Vo in HfO2 is higher than the acceptor state of Io. So, the electrons in the donor state will fall into the acceptor in Io, resulting in +2|e| charged Vo and -2|e| charged Io (Fig. 4.4). In the case of surface formation, the charge of Vo is determined by the Fermi level of the electrode. In case of Ti electrode and HfO2, the Fermi level is lower than the donor energy level of Vo, which will again result in a +2|e| charged Vo (Fig. 4.5). If the Fermi level of the electrode is near the acceptor level, the charge state of Vo is likely to be +1|e| or zero.

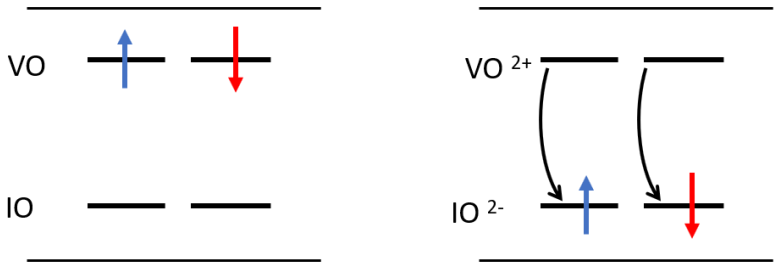


Fig. 4.4 The illustration of defects states in the bulk oxide (no metal). The vacancy and interstitial have energy levels in the bandgap of the oxide. In both figures, the Frenkel pair is neutral. In the state on the left, both the vacancy and interstitial are neutral (Io has ~ eight electrons; In the state on the right, the electrons fall down in energy to make the interstitial negatively charged (Io has ~ ten electrons).

In addition, the charge state of a vacancy is also determined by the voltage applied. The vacancies near the cathode are likely to be positively charged and the ones near the anode are likely to be neutral (it can even be negative if the bias is very large), when a bias is applied. However, if the tunneling current becomes significant, the charge state should be determined by the steady state current continuity equation. In this thesis, I assume *isolated vacancies* (not connected to cathode) to have $+2e$ charge, if a voltage is applied to the system. If a vacancy is connected to the anode, the vacancy charge is assumed to be zero.

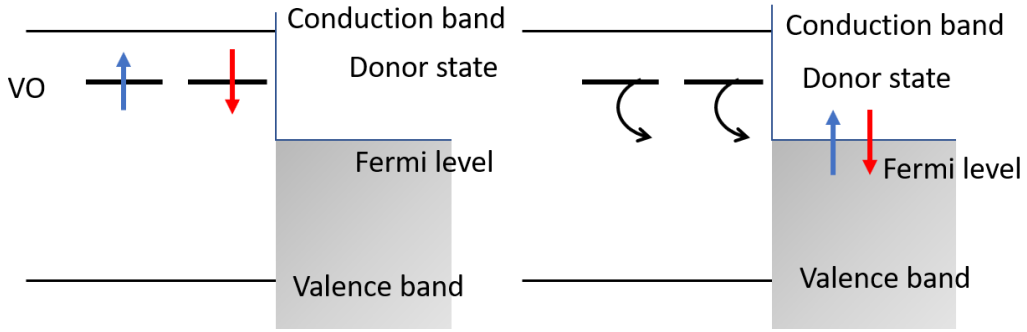


Fig. 4.5 The illustration of defects states in surface generation. The vacancies have energy levels in the bandgap of the oxide. In the state on the left, the vacancy is neutral and in the state on the right, the electrons fall down to the Fermi level of metal electrode to make the vacancy positively charged.

The electric potential satisfies Poisson's equation

$$\nabla \cdot (\epsilon_r \nabla \phi) = \frac{q(r)}{\epsilon_0}. \tag{4.8}$$

On the boundary of oxide, Dirichlet condition is applied, $\phi(x = 0) = 0$, and $\phi(x = t_{ox}) = V_{app}$, and periodic condition is applied on the y-direction. In the diffusion process, the vacancies and interstitials can leave the oxide at the right boundary (oxide/active electrode interface), but not the left boundary, as shown in Fig. 4.6.

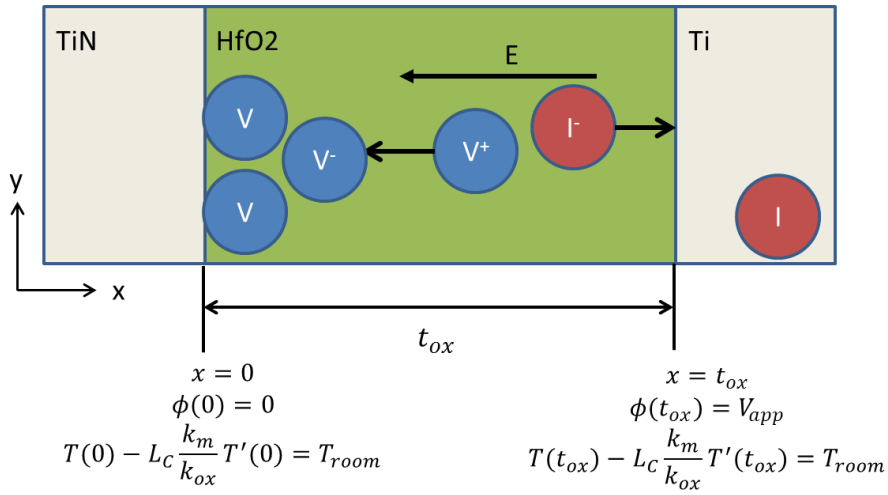


Fig 4.6 The system models the oxide region with thickness t_{ox}. Vacancies and interstitials can leave the oxide through the right boundary only. The Dirichlet boundary condition is applied for solving Poisson's equation, and mixed boundary condition is applied for heat equation.

The spatial variation of temperature is also taken into consideration. By solving Fourier's heat equation,

$$C \frac{\partial T}{\partial t} = \nabla(k \nabla T) + J, \tag{4.9}$$

we get the temperature distribution. Here C is the specific heat, k is the thermal conductivity, and J is the power density induced by Joule heating. This equation should be applied to the regions of oxide as well as metal electrode. We work with the approximation that the current is obtained from the current continuity equation (KCL) and KVL under the assumption that the current has a linear relation with the voltage. We can build a resistor network with the assumption that every vacancy

in the oxide is a node in resistor network. The effective resistance between nearest neighboring vacancies in the same island is (Fig. 4.7)

$$r_{ij} = r_N. \tag{4.10}$$

Here, r_N is the resistance between nearest neighbor vacancies in the same island. Also, in the model, note that each vacancy can only have a maximum of 6 nearest neighbor vacancies in Hafnia (only 4 nearest neighbors in 2D).

In order to understand the tunneling current between islands, we can consider the wave functions on the defects state. Because of the restricted computational resources, it is too time consuming to calculate the current by the method in Chapter 3, which requires a large Hamiltonian size from DFT calculations. However, the NEGF formalism can give us some information about the current. As stated from eqn. (2.27), the current can be expressed as $i_{ij} = i/\hbar(\langle\psi_i|H_{ij}|\psi_j\rangle - \langle\psi_j|H_{ji}|\psi_i\rangle)$, from the defect *i* to the defect *j*. Assuming that the defect states has a Slater type orbital $|\psi_i\rangle \propto \exp(-\alpha'/2\,|r - r_i|)$, the current is expected to depend exponentially on the distance $|r_i - r_j|$). Then, the effective resistance between two vacancies on separated islands is treated by an exponential dependence on the distance between the vacancies (Fig. 4.7)

$$r_{ij} = r_T = r_{T0}\exp(-\alpha\, d_{ij}). \tag{4.11}$$

Here d_{ij} is the minimum distance between vacancy islands *i* and *j* (detailed algorithm is in the Appendix). For example, in Fig 4.7, the vacancies can be separated into 5 islands as marked 1~5. Inside the islands the vacancies are staying at the nearest neighbor site with distance of lattice constant, as shown in the solid line of Fig 4.7. The resistance between vacancies in the same island is taken to be r_N. The shortest distance between islands are represented by d_{ij}, where $i, j = 1\sim5$. The resistance between islands are computed from (4.11), which gives the $r_{ij}, i, j = 1\sim5$. In the

simulations, the parameters of r_N, r_{T0}, and α are chosen to fit the low resistance of a memory cell. After computing the resistance between vacancies, we build the resistor network, and solve the current continuity equation

$$\sum_j I_{ij} = \sum_j (V_i - V_j)/r_{ij} = 0. \tag{4.12}$$

Here, I_{ij} is the current between vacancy i and j, and V_i and V_j are the voltage on vacancy i and j respectively. In the simulation, the values are fitted from experiment.

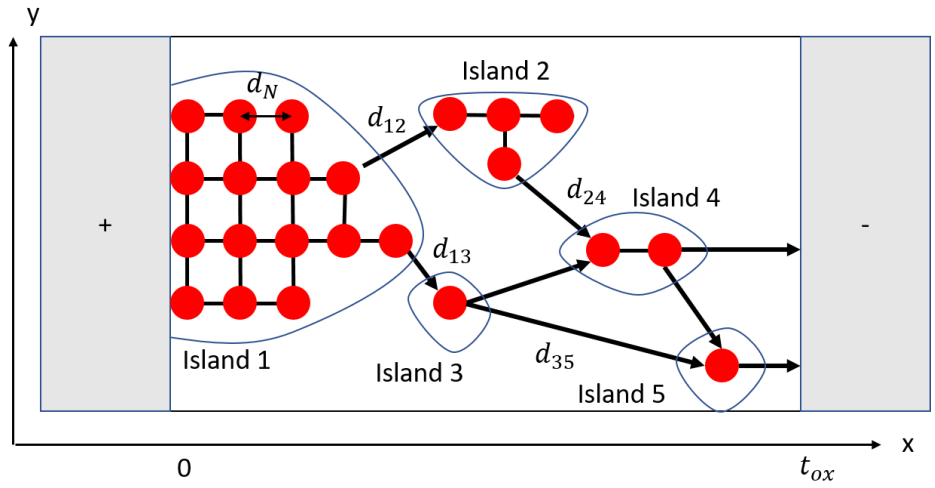


Fig. 4.7 The illustration of resistor network. The resistance between vacancies in the same island (represented by solid lines) is r_N. The resistance between vacancies that are more than a distance d_N apart in a single island are ignored. The resistance between islands (represented by arrows) is given by $r_T \exp(-\alpha\, d_{ij})$ [eqn. (4.11)]. Two islands are connected by only a single resistance r_{ij} which is determined by the shortest distance d_{ij} between islands i and j. The KVL and KCL are solved for such a resistive network to obtain the current and power density map in the structure.

The power density due to Joule heating is

$$J = \rho\, i^2 \tag{4.13}$$

Here ρ is the resistivity, which equals to r_N in the intra-island case. And in the tunneling case, $\rho = r_T \exp(-\alpha\, d_{ij})/d_{ij}$. Restricting ourselves to the adiabatic approximation, the temperature is found by solving the heat equation in steady state. That is, $\partial T/\partial t = 0$. This assumption means that we

let the temperature reach a new steady state value before determining how the vacancies and interstitials move using equations (4.1) – (4.4). That is, the time scale for thermal equilibrium is much shorter than the time scale involved in vacancy / interstitial hopping. This may not always be the case in a real experiment, but this is the assumption we use. Thus, the heat equation becomes

$$k_{ox}\nabla^2 T + J = 0. \tag{4.14}$$

On the boundary, we assume that the temperature decreases to room temperature (T_{room}) at a length L_c from the oxide-metal interface. A mixed boundary condition will is applied to the oxide/metal interface:

$$T_{room} = T - \frac{k_{ox}}{k_m}\frac{\partial T}{\partial x} L_c \, . \tag{4.15}$$

Then, from equations (4.1) to (4.15), we can solve the temperature distribution and electric potential to compute the rates of each process.

Independent stochastic Poisson processes are assumed for forming, recombining and diffusing, which means that these events will happen in an independent and memory-less manner. The total rate of all possible events is

$$R_{tot} = \sum R_i. \tag{4.16}$$

Here R_{tot} is the total rate, and R_i is the rate of i-th event. The probability that the next event is the i-th event is

$$p(i) = \frac{R_i}{R_{tot}}. \tag{4.17}$$

In the simulation, we will first calculate all the rates of possible events, and then we can use a random number generator to decide which event will happen. After that the rate should be updated

depending on the event that was chosen. Thus, Monte Carlo process can generate a series of events to simulate the process of resistive switching. The flow chart of Monte Carlo simulation is illustrated in Fig. 4.8.

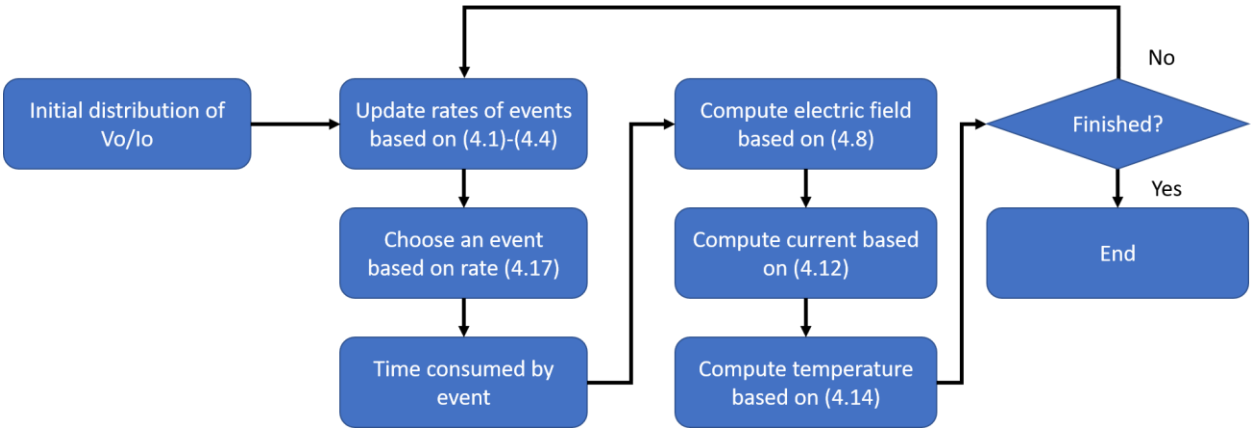


Fig. 4.8 Flow chart of Monte Carlo simulation.

I have numerically implemented the equations (4.1) to (4.15) first in two-dimensions (2D) to obtain an intuition for the growth of the filament and to understand the essential device physics. Then, I implemented a numerical solver in three-dimensions (3D), where the studies are performed in greater detail.

4.2 CODE AND MODEL DEVELOPMENT

In this section, the details of the KMC processes as implemented in our calculations, the solving of Poisson's equation and the model used for Joule heating are discussed.

4.2.1 Kinetic Monte Carlo processes

The KMC code for the 2D system is implemented as follows. Consider a 2D grid with $N_x \times N_y$ grid points (as shown in Fig. 4.9). At each of these grid points (sites) the hopping rates for the following events are calculated:

1) If the site is occupied by a vacancy, the hopping rate of vacancy to top, bottom, left and right (4 events). If there is a vacancy in the destination site, the hopping rate to that site is zero.

2) If the site is occupied by an interstitial, hopping rate of interstitial to top, bottom, left and right (4 events). If there is an interstitial in the destination site, the hopping rate to that site is zero.

3) if the site is occupied by a native oxygen atom, four types of Frenkel pair generation events are possible except if the top or right site is occupied by a vacancy or interstitial: type I, generate one vacancy on the site and the corresponding interstitial on the top neighbor; type II, generate one vacancy on the site (original oxygen atom site) and corresponding interstitial on the right neighbor; type III, generate one interstitial on the site and the corresponding vacancy on the top neighbor; and type IV, generate one interstitial on the site and corresponding vacancy on the right neighbor. Note that some of these events are zero and do not require a calculation.

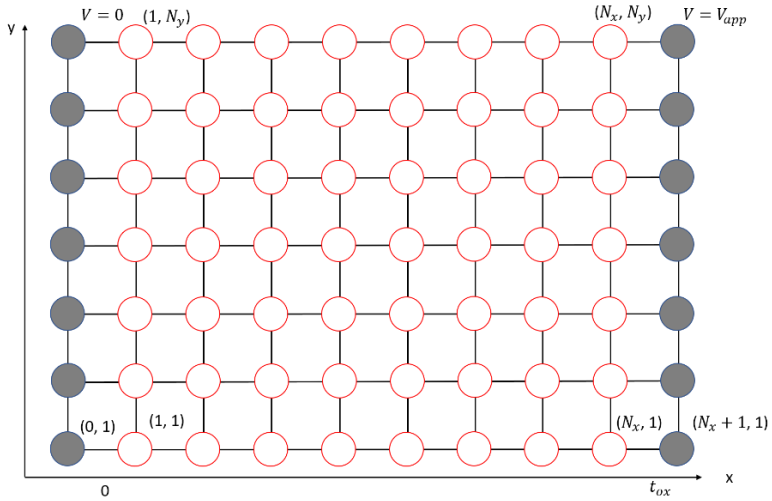


Fig. 4.9 The illustration simulation grid. The oxide is discretized to $N_x \times N_y$ grid and electrodes on the left and right are applied with Dirichlet boundary condition. Periodic condition in the y-axis is applied.

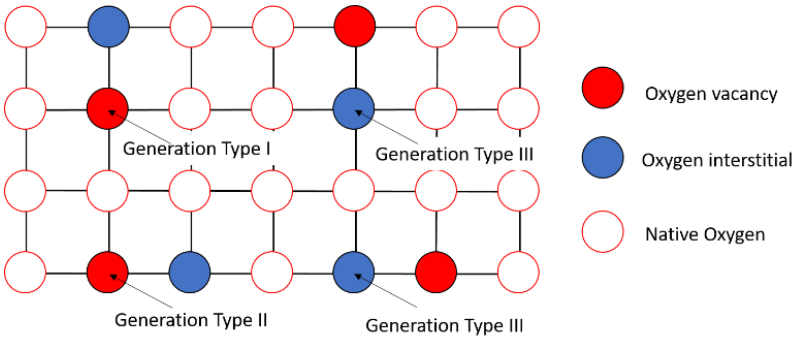


Fig. 4.10 The illustration of four types of generation. The vacancy or interstitial can generate on site with the corresponding interstitial or vacancy generate on the right or top neighbor respectively.

These events correspond to the index i in equation (4.16). So, the rates for a total up to $12 \times N_x \times N_y$ events are calculated in 2D. The rates at each grid point are calculated as per equations (4.1) to (4.4). A random event is chosen using equation (4.17), and the location of vacancies and interstitials are updated at all grid points. In the next step the rates identified above are recalculated for the new configuration (of vacancies and interstitials) and the process continues for n_1 simulation steps. After this both the Poisson and heat equations are solved once, and the above process of updating the location of vacancies and interstitials continues. This process is performed $N_{PoiHeat}$ number of times. In our simulations, we have used a variety of values for n_1 and $N_{PoiHeat}$. The total simulation step $\mathrm{n_1} \times N_{poiHeat}$ is chosen long enough to reach the desired resistance state (low resistance state in form and set, while high resistance state in reset). When the value of $\mathrm{n_1}$ varies from 10 to 100 steps, the switching process doesn't change significantly. Therefore, $\mathrm{n_i} = 100$ is chosen as the default value to shorten the simulation time.

In the 3D system, the number of grid points in the x, y, and z dimensions are taken to be N_x, N_y and N_z. The simulation procedure otherwise is identical to that described for the 2D case. The repeated evaluation of the partial differential equations (Poisson equation, Fourier equation and KCL) is time consuming. We adopt a procedure for the acceleration of the solvers for these

equations. As an example, the Poisson equation with Dirichlet boundary condition is solved with the backslash solver in MATLAB.

4.2.2 Poisson equation

The discrete form of Poisson's equation in (4.8) in 2D is:

$$c_{mn}\,\phi_{m,n} - c_{mn,1}\phi_{m,n-1} - c_{mn,2}\phi_{m,n+1} - c_{mn,2}\phi_{m-1,n} - c_{mn,4}\phi_{m+1,n} = 0 \qquad (4.18)$$

Here the coefficients $c_{mn,1} = (\epsilon_{m,n} + \epsilon_{m,n-1})/2, c_{mn,2} = (\epsilon_{m,n} + \epsilon_{m,n+1})/2, c_{mn,3} = (\epsilon_{m,n} + \epsilon_{m-1,n})/2, c_{mn,4} = (\epsilon_{m,n} + \epsilon_{m+1,n})/2$, and $c_{mn} = \sum_{i=1}^{4} c_{mn,i}$.

Here m and n are the indices on the x and y axis, $\phi_{\mathrm{m,n}}$ is the electrostatic potential at the grid (m, n). $\epsilon_{\mathrm{m,n}}$ is the dielectric constant and q in eqn (4.8) is set to be zero.

The electrostatic boundary condition at the grid points in the metal closest to the electrode is taken to be (Dirichlet boundary condition) zero at the left-hand side and $\mathrm{V_{app}}$ at the right hand side.

$$\phi_{0,\mathrm{n}} = 0, \text{ where } \quad n \in \{1, 2, 3, \ldots Ny\} \qquad (4.19)$$

$$\phi_{\mathrm{Nx+1,n}} = V_{app}, \text{ where } \quad n \in \{1, 2, 3, \ldots Ny\} \qquad (4.20)$$

Here n belongs to the set of all grid points $\{1, 2, 3, \ldots Ny\}$ along the y-direction. The electrostatic boundary condition for top and bottom are periodic,

$$\phi_{\mathrm{m},1} = \phi_{m,Ny+1} \quad \text{where} \quad m \in \{1, 2, 3, \ldots Nx\} \qquad (4.21)$$

The discretized version of Poisson's equation (4.8) results in a linear system of equations given by

$$A\phi = b. \qquad (4.22)$$

Here A is a $(N_x \times N_y + 2N_x) \times (N_x \times N_y + 2N_x)$ matrix for an $N_x \times N_y$ grid. Here $\mathrm{b} = \left(0 \cdot \mathrm{J}_{Nx}, 0 \cdot J_{Nx \times Ny}, \ \mathrm{V_{app}} \cdot J_{Nx}\right)^{\mathrm{T}}$. J_{m} is unit vector of size $1 \times \mathrm{m}$.

The matrix A is

$$A = \begin{pmatrix} I_{Nx} & 0 & 0 \\ A_{lc} & A_{cc} & A_{cr} \\ 0 & 0 & I_{Nx} \end{pmatrix}. \tag{4.23}$$

Here I_{Nx} is the identity matrix with dimensions of N_x. The elements of $\mathrm{A_{lc}}$, A_{cc} and $\mathrm{A_{cr}}$ are given in eqn. (4.22). As we can easily see, the matrix is not symmetric. The backslash solver in MATLAB is not optimized for an asymmetric matrix. We can transform the matrix equation to a symmetric form by the following operation:

$$\begin{pmatrix} I_l & 0 & 0 \\ -A_{lc} & I_c & -A_{cr} \\ 0 & 0 & I_r \end{pmatrix}\begin{pmatrix} I_l & 0 & 0 \\ A_{lc} & A_{cc} & A_{cr} \\ 0 & 0 & I_r \end{pmatrix}\phi = \begin{pmatrix} I_l & 0 & 0 \\ 0 & A_{cc} & 0 \\ 0 & 0 & I_r \end{pmatrix}\phi = \begin{pmatrix} I_l & 0 & 0 \\ -A_{lc} & I_c & -A_{cr} \\ 0 & 0 & I_r \end{pmatrix} b.$$

Multiplying out this matrix, we have one of the three equation to be

$$A_{cc}\phi_c = (-A_{lc} \quad I_c \quad -A_{cr})b. \tag{4.24}$$

Where ϕ_c corresponds the potential at the device nodes only and so is a vector of length equal to $Nx \times Ny$. The symmetric form A_{cc} is useful because the numerical solver can now use the fast Cholesky decomposition instead of full LU decomposition. It is also useful to remark that using A_{cc} is faster than using $-A_{cc}$ because A_{cc} is positively definite, and so the Cholesky decomposition can be applied. My tests show that for the Poisson solver with symmetric positive definite solver (equation (4.24)) is nearly 10 times faster than the asymmetric one (equation (4.23). For the 3D system, we can simply extend the 2D system in the z-direction with different layers. Then, eqn (4.18) becomes

$$c_{mnl}\,\phi_{m,n,l} - c_{mnl,1}\phi_{m,n-1,l} - c_{mnl,2}\phi_{m,n+1,l} - c_{mnl,3}\phi_{m-1,n,l} - c_{mnl,4}\phi_{m+1,n,l} - c_{mnl,5}\phi_{m,n,l-1} - c_{mnl,6}\phi_{m,n,l+1} = 0. \tag{4.25}$$

The boundary conditions applied are Dirichlet in x-direction:

$$\phi_{0,n,l} = 0, \text{ where } \quad n \in \{1, 2, 3, \ldots Ny\} \; and \; l \in \{1, 2, 3, \ldots Nz\} \tag{4.26}$$

$$\phi_{Nx+1,n,l} = V_{app}, \text{ where } \quad n \in \{1, 2, 3, \ldots Ny\} \; and \; l \in \{1, 2, 3, \ldots Nz\} \tag{4.27}$$

and periodic in y- and z- directions:

$$\phi_{m,1,l} = \phi_{m,Ny+1,l} \quad \text{where} \quad m \in \{1, 2, 3, \dots Nx\} \; and \; l \in \{1, 2, 3, \dots Nz\} \tag{4.28}$$

$$\phi_{m,n,1} = \phi_{m,n,Nz+1} \quad \text{where} \quad m \in \{1, 2, 3, \dots Nx\} \; and \; n \in \{1, 2, 3, \dots Ny\}. \tag{4.29}$$

4.2.3 Joule heating in system

In solving the heat equation (eq. 4.14), the discretized version of the 3D heat equation is

$$c_{mnl}\,\phi_{m,n,l} - c_{mnl,1}\phi_{m,n-1,l} - c_{mnl,2}\phi_{m,n+1,l} - c_{mnl,3}\phi_{m-1,n,l} - c_{mnl,4}\phi_{m+1,n,l} - c_{mnl,5}\phi_{m,n,l-1} - c_{mnl,6}\phi_{m,n,l+1} = xyz. \tag{4.30}$$

The boundary conditions applied are Dirichlet in x-direction:

$$\phi_{0,n,l} = 0, \text{ where} \quad n \in \{1, 2, 3, \dots Ny\} \; and \; l \in \{1, 2, 3, \dots Nz\} \tag{4.31}$$

$$\phi_{Nx+1,n,l} = V_{app}, \text{ where} \quad n \in \{1, 2, 3, \dots Ny\} \; and \; l \in \{1, 2, 3, \dots Nz\} \tag{4.32}$$

As the mathematical for heat equation is similar to the Poisson equation above. To solve this, we first symmetrized as in equation (earlier). Then the same numerical solver as Poisson is used to solve the heat equation.

The main discussion of this subsection is around the models form heat dissipation used when Fourier equation is solved. These models are essential to prevent unphysical increase in temperature in the nanoscale filaments.

The source of heating is the energy loss from the current flowing through the device. In the Joule heating model, electrons lose energy continuously – that is, the amount of energy loss equals to the electric potential change multiplied by charge of an electron (q). The underlying assumption in the Joule heating model is that the electrochemical potential difference is equal to q times the electrostatic potential difference. Therefore, the power source for Joule heating is

$$P = \frac{Q\Delta\phi}{t} = IV. \tag{4.33}$$

Here P is the heating power, Q is the total charge runs through the filament in time interval t and $\Delta\phi$ is the potential difference between electrodes.

We should notice that the Joule heating assumes two conditions: 1) the energy lost by electrons as heat is equal to q times the change in electric potential; 2) the energy dissipated due to the change in potential energy of the electron instantaneously becomes heat in a continuous manner as the electron travels through the system. In other words, the phonons are re-thermalized at an infinitesimally small length scale (significantly smaller than any physical dimension relevant to the system size).

When we applied the straightforward Joule heating model above with an infinitesimally small length scale for heat dissipation, we observe that the temperature in the oxide can easily rise over $3500K$, which is higher than the melting point of HfO_2.

For example, given that the current compliance is $100\mu A$, and a low device resistance of $2k\Omega$, the power dissipated in the system due to Joule heating is $20\mu W$. Assuming a a filament cross sectional area of $4nm^2$ and a length of $10nm$, the power density in the filament is $500nW/nm^3$. The thermal conductivity of HfO2 is around $1.1W/(m \cdot K)$ or $1.1nW/(nm \cdot K)$. I estimate the temperature using the simple 1D heat equation,

$$\kappa \frac{d^2}{dx^2} T = P_{heat}. \tag{4.34}$$

The boundary condition is $T = 300K$ at the oxide-electrode interfaces at $x = 0nm$ and $x = 10nm$. I calculate the maximum temperature (at the center of the filament $x = 5nm$) to be $T_{max} \approx 5982K$. Solving the 3D heat equation with the boundary condition that the temperature at the metal-electrode interface is 300 K give a maximum temperature of 2182K which is half of the 1D

result, but still too high versus the result of experiment [68, 86] also estimates the maximum temperature in the filament could be higher than the melting point in the given conditions. However, experiments do not report melting in HfO2 based resistive memory with a current compliance of less than 1mA. Therefore, I conclude that the conditions assumed for Joule heating are not valid in the atomic scale filaments.

There are two relevant scenarios to obtain a lower temperature:

1) Slow phonon relaxation model: The length scale assumed with phonon thermalization is finite (the Fourier equation assumes immediate phonon thermalization). The phonon can then on an average travel a mean free path before thermalization. In the Fourier equation, the source term for heat generation at any particular site will have contributions from phonons emitted in the surrounding sites.

2) Slow electron relaxation model: The electrons are slowly relaxed in the filament, and therefore when the electron enters the anode, they still carry a high energy. This energy is dissipated inside the metal.

In KMC simulation, both models are applied, and the temperature is lower in the oxide with the model in Scenarios 1 and 2 compared with the case of applying Fourier equation directly.

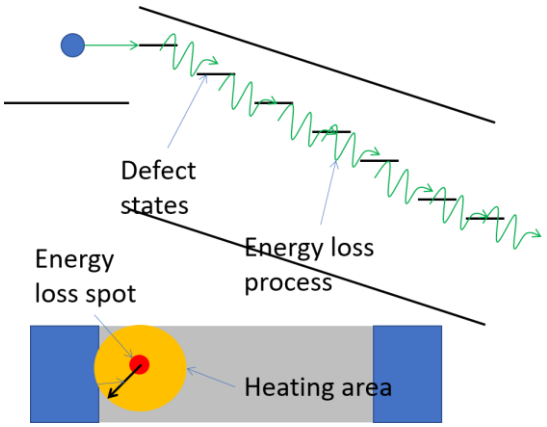


Fig.4.11. Illustration of Scenario 2, the slow phonon relaxation model: Energy loss occurs in conductive filament. But the phonon takes a long length (longer than filament) to reach thermal equilibrium. Thus, the heating area is larger than the energy loss spot (red). Results in a large diffused heating area (orange).

Slow phonon relaxation model: In general, a phonon will thermalize over a length scale equal to the energy mean free path of the phonon. The phonon mean free path in an oxide can be estimated from the thermal conductivity [87]

$$L_{mfp} = 3\kappa/Cv \tag{4.35}$$

Here κ is the thermal conductivity of the material, C is the heat capacity and v is the speed of sound. With the parameters of HfO2, $\kappa = 1.1\mathrm{W/(m \cdot K)}$, $\mathrm{C} = 1.16\mathrm{J/(cm^3 \cdot K)}$ and $\mathrm{v} = 2815\mathrm{m/s}$, we can conclude that the mean free path of phonons is approximately 1.12nm. Also, as suggested in Ref. [87], the mean free path in an insulator is larger than the above value, when the dispersion of phonons is considered. The contribution of thermal conductivity due to phonons of various frequencies is given by:

$$\kappa = \frac{1}{3} L_{mfp} \int_0^{\omega_m} C(\omega) v_g(\omega) d\omega \tag{4.36}$$

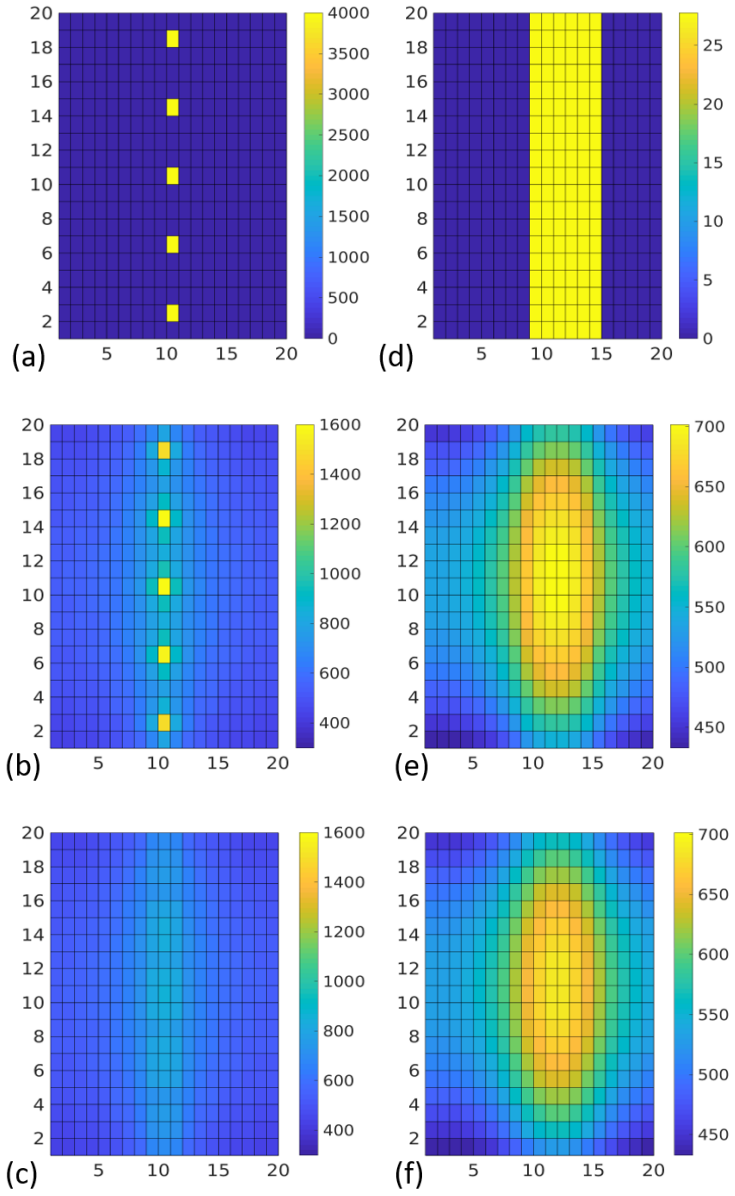


Fig.4.12. a) The power distributed in the yellow regions is assumed to be $32\mu W/nm^3$. The temperature in the filament obtained by solving the Fourier equation (b) without and (c) with the slow phonon relaxation model. While the maximum temperature in b) is over 1600K, it is below 800K in c). d) The same total power of 20μW as in (a) is distributed in the yellow regions of (d), which is much larger in size. e) and f) are temperature distributions with and without slow phonon relaxation models by assuming a power dissipation as shown in (d). The maximum temperature of e) and f) are nearly the same at slightly over 700K.

Here, $C(\omega)$ is the heat capacity, $v_g(\omega)$ is the group velocity and ω_m is the maximum frequency of acoustic phonons. It should be noted that the heat capacity has contributions due to both acoustic and optical phonons,

$$C = \int_0^{\omega_m} C(\omega)d\omega + \frac{3\kappa}{V}\left(\frac{\hbar\omega_o}{kT}\right)^2 e^{\frac{\hbar\omega_o}{kT}}\left(e^{\frac{\hbar\omega_o}{kT}} - 1\right)^{-2} \quad (4.37)$$

Here, ω_0 is the optical phonon frequency. The reason for the underestimation of the phonon mean free path when equation (4.36) is used arises because the optical phonon contributes to the heat capacity, but not to the thermal conductance. In other words, the energy of the electron can be transferred to phonon, but the phonon takes on average one mean free path to deposit the energy in the thermal form.

I will now discuss the implementation of Scenario 2, the Slow Phonon Relaxation Model, in my code. Let P_{mn} be the power density dissipated at grid point (m, n) as given by eqn. (4.33) (That is Local power density is proportional to the current density at that point as given by Joule's law). In Scenario 2, we assume that P_{mn} is dissipated evenly over all grid points that lie within a phonon mean free path of L_{mfp} from grid point (m, n). Then, on grid point (i, j), the power density contribution due to P_{mn} is

$$\rho_{\text{ij}} = P_{mn}\, f(i - m, j - n). \quad (4.38)$$

In general, the distribution function f(i, j) should vary with distance from grid point (m, n) and energy conservation requires that

$$\sum_{\text{ij}} f(i - m, j - n) = 1. \quad (4.39)$$

To simplify the problem, we will not account for a spatial dependence in the contribution to ρ_{ij} from P_{mn} as along as grid point (i, j), is within L_{mfp} from grid point (m, n). That is, we will assume a uniform distribution given by

$$\text{f}(\text{i} - \text{m}, \text{j} - \text{n}) = \frac{1}{N_{mfp}}, \qquad \text{when } |\text{l} - \text{i}| < \frac{\text{L}_{\text{mfp}}}{a} \; or \; |\text{m} - \text{j}| < \frac{\text{L}_{\text{mfp}}}{a} \quad (4.40)$$

$$\text{f}(\text{i} - \text{m}, \text{j} - \text{n}) = 0. \;\; \text{when } |\text{l} - \text{i}| > \frac{\text{L}_{\text{mfp}}}{a} \; or \; |\text{m} - \text{j}| > \frac{\text{L}_{\text{mfp}}}{a} \quad (4.41)$$

where $N_{mfp} = 4\left(\frac{L_{mfp}}{a}\right)^2$.The power density used on the r.h.s. of Fourier equation (eqn. (4.8)) is

$$\rho_{ij} = \sum_{lm} \frac{P_{lm}}{N_{mfp}}, \tag{4.42}$$

where the summation is performed over all grid points.

As shown in Fig.4.12 (a), the power density is distributed in several regions (shown in yellow) inside the oxide layer. If slow phonon relaxation (scenario 1) is neglected, the maximum temperature can be 1600K, as in Fig. 4.12(b). On the other hand, if the slow phonon relaxation is taken into consideration (we assume $L_{mfp} = 1.5nm$), the maximum temperature in the oxide can be reduced significantly to less than 800K, as shown in Fig. 4.12(c). We also show the temperature distribution if the power density is uniformly distributed over a larger volume inside a filament ($3x3x10nm^3$), as shown in Fig. 4.12 (d). The total power dissipated is taken to be the same as in 4.14 (a). Then, the temperature distribution can be computed without considering slow phonon relaxation, as in Fig. 4.12 (e) and with slow phonon relaxation as in Fig. 4.12 (f). As these figures show, the difference between the cases with and without slow phonon relaxation is small if the same power is distributed over a larger volume. Figure (4.12) shows that with the slow phonon relaxation model, the temperature can be a lot lower even if the power density is large in a small localized region.

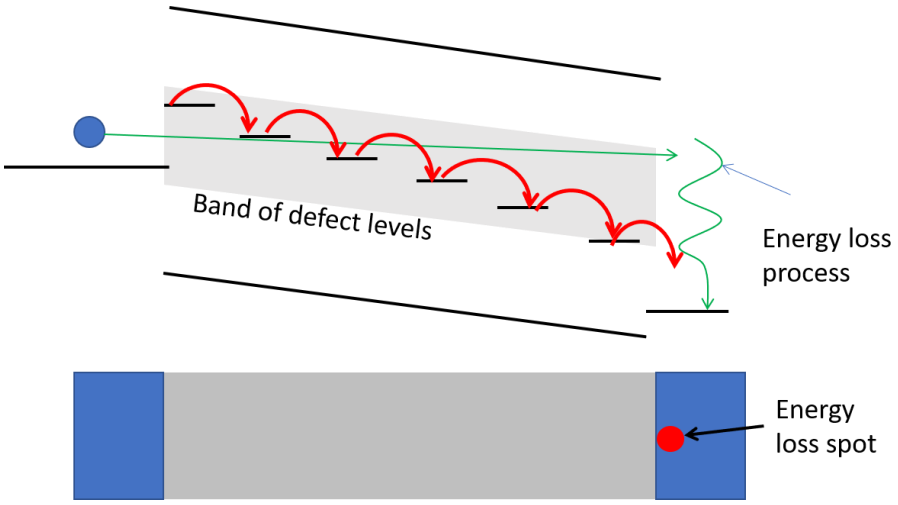


Fig. 4.13. Illustration of Slow electron relaxation: Electrons can pass through the oxide in one extreme by dissipating energy as they hop between defect levels (red arrows), and, in the other extreme they can pass through the device without energy loss through the band of defect levels (grey region) and only lose heat in the right contact. The reality is expected to be somewhere in between in well-formed filaments. If the electron mean free path is longer than the oxide/filament length, I expect the heating to occur primarily in the contact. Note that we do not model this scenario using a detailed model that accounts for the phonons and defect energy levels shown in this figure. This is beyond my scope here. Instead, I simply estimate the temperature by assuming a fraction of the heat to be dissipated in the filament and the rest in the right contact (solid red circle in the right contact).

Slow electron relaxation model (Scenario 2): Another factor that should be considered is that in the tunneling of electrons through the oxide, the energy is not entirely transferred to the phonons as the electron passes through the oxide. Instead, electrons can keep part of their energy as they travel through the filament. These electrons, which keep their energy, together with those electrons which lose energy to phonons, contribute to the total current. The electrons that keep their energy become hotter during transit and have a high energy when they enter the cathode (right electrode on Fig. 4.13). The high-energy electron then releases the excess energy in the electrode. In this model, the energy dissipated in the oxide is less than in the Joule heating model. As the metal is a good thermal conductor, it is expected that the heat dissipated in the metal will quickly dissipate away. Therefore, the temperature in the oxide is expected to be less than that obtained from the Joule heating model. Note that I do not model this scenario using a detailed model that accounts for the phonons and defect energy levels. This is not the intent here and is beyond the scope of this

thesis. Instead, I simply estimate the temperature by assuming a fraction of the heat to be dissipated in the filament and the rest in the right contact.

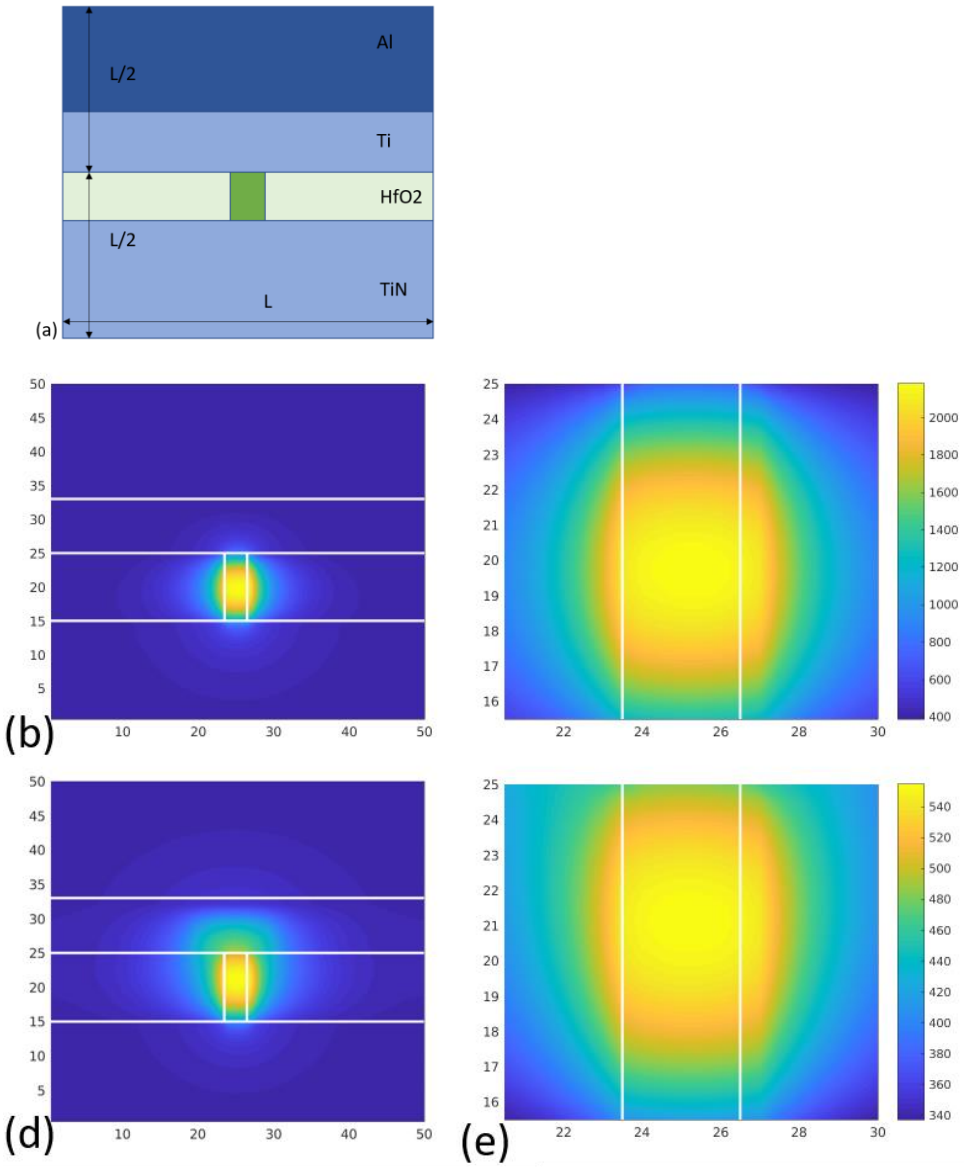


Fig. 4.14. *Slow electron relaxation model (Scenario 2):* a) Device structure is shown. The thickness of HfO2 is 10nm, thickness of Ti is 8nm, cross sectional dimension of the filament is $3 \times 3\, nm^2$. The total volume of the simulation box is 50x50x50nm ($L = 50\, nm$). b) Temperature when all heat is dissipated inside the oxide; (c) a zoomed view of the filament region in (b); d) Temperature when 10% of the heat is dissipated inside oxide and 90% is dissipated in the Ti/Al metal; (e) a zoomed view of the filament region in (d).

In the simulation of the *slow electron relaxation model (Scenario 2)*, part of the heat is dissipated in the oxide and the remaining is dissipated in the metal. To model this, I solve the Fourier equation with the Dirichlet boundary condition applied deeper inside the metal. We assume metal contacts with dimensions of $L/2 \times L$ as shown in Fig. 4.14 (a). We choose $L = 50\ nm$, and assume that the temperature reaches $300\ K$ at the edges of the box. Note that the oxide is $10\ nm$ thick and the cross-section area of the filament is $3 \times 3\ nm^2$. If all the power is dissipated inside the filament, the temperature can be as high as 2000K, as shown in Fig. 4.14(b) and enlarged in (c). On the other hand, if we consider the electron relaxation inside the filament is slow, so that when the electron enters the metal, it still carries a high energy. Then the high energy electron will relax inside the metal and generate heat. Because the thermal conductivity of the metal is much higher than the oxide, the temperature will not be high. The result is shown in the Fig. 4.14 (d) and enlarged in (e). In obtaining this temperature distribution, I have assumed that 10% of the heat is dissipated in the oxide and 90% of the heat is dissipated in the metal electrode. We remark that the $600\ K$ maximum temperature is close to the experimental results in [86].

I now make the following general remarks. A detailed analysis of the heating process based on an atomistic model, which takes electron-phonon interaction into account will be useful in understanding heat dissipation at these length scales. The electron-phonon interaction may affect the vibration modes of a few atoms resulting in the loss of energy due to absorption / dissipation of localized vibrational modes.

4.2.4 Bulk and surface generation

A vacancy should be generated together with an interstitial in order to conserve the number of oxygen atoms. From ab initio computations, the formation energy for a vacancy/interstitial pair

E_{form} ranges from 5eV to 8eV. Such large values means that it takes a time period of $r_0^{-1}\exp((E_{form}+E_{barr})/kT)/\mathrm{N}$ to form one vacancy in the system (N is the number of sites for vacancy generation). Suppose $kT = 0.025eV, E_{form} = 5eV, E_{barr} = 1.1eV$, and $r_0^{-1} = 10^{-15}s$, the time to form one vacancy should be $7.22 \times 10^{71}s$ (2.29e64 years). This time is much longer than the values from experiments, which means that the real formation energy should be much smaller than the value from ab initio computation. However, a small bulk formation energy leads to a short retention time in the high resistance state because of vacancy generation. For example, if the formation energy is $\mathrm{E_{form}} = 1.4eV$, the time to form one vacancy should be $2.092 \times 10^{24}\mathrm{s}$. Considering the huge number of sites in the oxide (HfO_2 oxide cell occupies $0.125\mathrm{nm}^3$ volume), $100\mathrm{x}100\mathrm{nm}^2$ cell takes less than 290.5 hours to form a conductive chain in the 10nm thick layer oxide. Experiments however suggest a short set time and a long retention time in the HRS (high resistance state). This is not possible if there is a single energy barrier corresponding to the formation energy E_{form} (equation (4.1)) during both the set and HRS retention processes. The surface generation is a candidate that satisfies the requirement of short switching time and a long retention time in the HRS. In surface generation, the vacancy can generate at the metal-oxide interface with the rate associated with $\mathrm{E_{form(s)}}$, which is different from the bulk formation energy of $\mathrm{E_{form}}$. Therefore the generation rate on the surface can be faster than the rate in the bulk. Then, a filament can grow under an applied electric field since the oxide interface provide a source for the vacancy generation. The growth rate is mainly determined by the generation rate of vacancy which is associated with surface formation energy $\exp(-\mathrm{E_{form}(s)}/\mathrm{kT})$. On the other hand, in the HRS, while vacancies can still form near the interface without the driving electric field, vacancies will occupy all the sites with a low formation energy on the surface. Further generation of vacancies at the surface is not possible unless the previously generated vacancies at the surface

move away (but there is no electric field to move the vacancies). Therefore, the filament can hardly form, and the high resistance state is maintained for a long time. For example, if the $E_{form(s)} = 0.3eV$, then the generation time is $1.627 \times 10^{-10}s$. However, the diffusion of one vacancy takes $1.128 \times 10^{4}s$, and the direction is random without an applied electric field. Therefore, the HRS can be retained for a long time. The ability of maintain an HRS is also accentuated by the bonding energy. In the 2D case, the vacancies on the filament has 1~3 bonds. Compared with the condition $E_{bond} = 0eV$, the diffusion time for one vacancy with $E_{bond} = 0.03\text{eV}$ increases by $\exp\left(\Delta n \frac{E_{bond}}{kT}\right) \approx 3.32\sim36.59$ times. This further helps increase the retention time.

However, a small formation energy leads to a short retention time in the high resistance state because of vacancy generation. Experiments however suggest a short set time and a long retention time. This is not possible if there is a single energy barrier corresponding to the formation energy E_{form} (equation (4.1)) during both the set and retention processes. The surface generation is a candidate that satisfies the requirement of short switching time and a long retention time.

4.3 RESULTS AND ANALYSIS IN 2D SYSTEM

The simulation is started in a 128x128 two-dimensional grid for illustration purpose. This grid describes the switching layer in the RRAM. The left edge of simulation box is the interface between inert electrode and switching layer; the inert electrode does not react with defects or permit their permeation into them. The right edge is the interface with the active electrode, where the defect can be generated. In the simulation, the inert electrode is grounded, and the voltage on the active electrode varies. When the applied voltage is positive, the electric field will push the positively charged vacancy toward the left inert electrode, and negatively charged oxygen

interstitial will diffuse to the right active electrode. Because oxygen can migrate into the metal and form defects, the right edge is an open boundary for the interstitial oxygen. In other words, when the interstitial meets the right edge, it can disappear from the simulation box (oxide). For the vacancy, the chance to migrate out of the system is proportional to the concentration of oxygen in the active electrode. Thus the rate of a vacancy migration (R_{Vout}) out of the simulation box from the right edge is

$$R_{Vout} = R_{diff}\, C_O. \tag{4.43}$$

Here, R_{diff} is the rate of vacancy diffusion, and C_O is a factor proportional to the probability that there is an oxygen at a site on the surface (oxide-metal interface). For R_{diff}, I use the hopping rate that is assumed in the oxide as in equation (4.4) but a user of my code can easily change this. In my code, for simplicity, C_O is calculated by counting the number of oxygen interstitials that leave the oxide and multiply this by a user tunable parameter.

We show the formation of filaments in this chapter. Unless otherwise stated only the oxide region from $x = 0$ to $x = tox$ in figure 4.15 is shown.

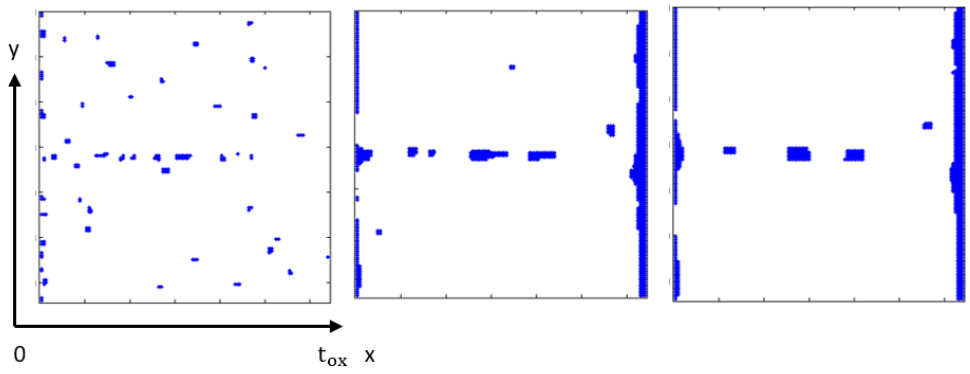


Fig. 4.15 Evolution of vacancies in 2D system with parameters $E_{bond} = 0.3eV, E_{form} = 6.15eV, E_{form(s)} = 0.15eV$. We can observe that the clustering of vacancies, and the vacancy islands are hard to move. The left and right boundaries of each of the three figures corresponds $x = 0$ to $x = tox$ respectively.

We first perform a simulation study to understand how the bonding energy (E_{bond}) can change the clustering of vacancies in the filament. To recapitulate, when two defects (vacancy/vacancy or interstitial/interstitial) meet and form a bond, the total energy will be lowered by one bond energy (E_{bond}). A larger bond energy means that there is a larger chance that the vacancies will stay together to form a cluster rather than breaking up. A large bond energy can offer stronger binding between vacancies, improving the stability of the filament. However, the large bond energy can also cluster the vacancy to form islands inside the switching layer. Because of the large bond energy, it is difficult for a vacancy to detach from these islands and diffuse along the electric field. In other words, the large bond energy will prevent the vacancy from moving along the electric field and forming a filament. In our simulation, we can see that when the $E_{bond} > qaV_{app}/L_x$ and $E_{bond} \gg kT$, the binding energy will be strong enough to overcome the electric

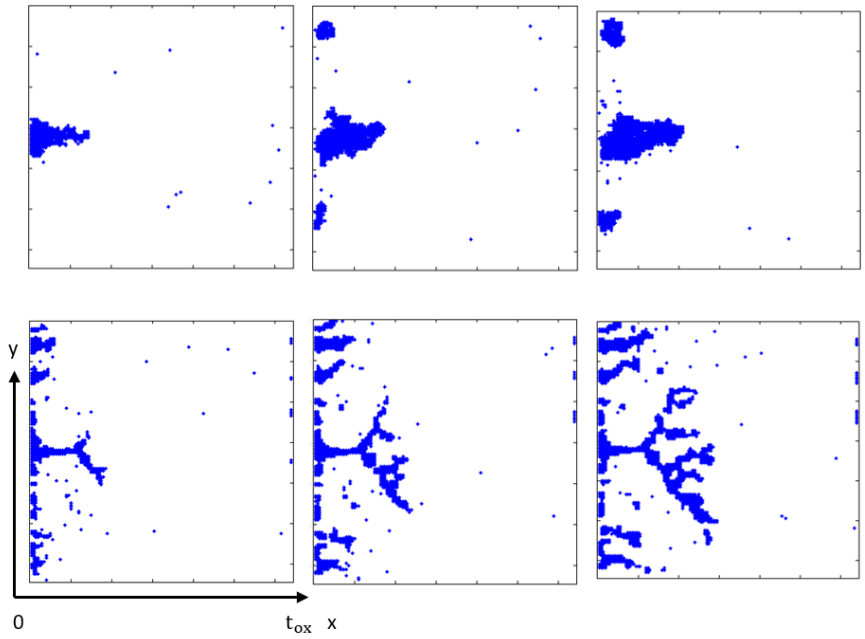


Fig 4.16. (a) (b) and (c) are simulations with surface generation at 10k, 1M and 2M steps. The parameters are $E_{bond} = 0.03\ eV$ $E_{form} = 6.2\ eV$ and $E_{form(s)} = 0.3\ eV$, $kT = 0.025\ eV$. The applied voltage $V_{app} = 6\ V$. (d), (e) and (f) correspond to a lower formation energy $E_{form(s)} = 0.15\ eV$, which results in a faster generation of Frenkel pairs. We can clearly see the branching of the filament. The oxide thickness is 64nm, and surface generation occurs over the entire oxide-active electrode interface (right interface). The left and right boundaries of each of the three figures corresponds $x = 0$ to $x = tox$ respectively.

field. Here V_{app} is the applied voltage, L_x is the thickness of switching layer and q is the charge of one vacancy. If the binding energy is large, we can only observe the growth of vacancy islands, and these islands cannot move, which means the memory cell cannot switch back to the high resistance state. To minimize the effect from the bond energy, we will set the value of $E_{bond} = 0.03eV$ as a default. We can observe that at room temperature $kT = 0.025eV$, this value of bond energy can offer enough stability to keep a filament shape.

We perform simulations with the parameters of $E_{form} = 6.2eV$ and $E_{form(s)} = 0.3eV$. Note that the value of $E_{form(s)} = 0.3eV$ corresponds to an activation energy of $1.4eV$ for formation (eqn 4.1) since the energy barrier for migration is taken as $1.1eV$. The temperature is set uniformly at $kT = 0.025eV$. The applied voltage, $V_{app} = 6V$. We observe that the generation rate with these parameters is low because of the limited region for generation. At the end of two million Monte Carlo simulation steps, the number of vacancies generated is too small to form a filament connecting the left and right electrodes, as shown in Fig. 4.16 (a)-(c). To generate more vacancies, we decrease the surface formation energy to $0.15\ eV$. We observe that the filament now grows with the same applied voltage of 6V. The filament tends to branch rather than stay in a thin cylindrical region, as shown in Fig. 4.16 (d)-(f).

The reason for branching is that the filament growth is determined by both the electric field and random walk of vacancies. The electric field guides vacancies to the other electrode and the vacancies form a partial filament. As filament tip is sharp, the enhanced electric field there causes the tip to grow faster than other parts. However, the diffusion of vacancies is also determined a random walk that depends on the temperature. As a result, the vacancies have a chance to attach to regions other than the tip (where the electric field is not a maximum). The vacancy attachment in regions other than the tip of the partial filament lead to protrusions where the electric field is

large. This local electric field can enhance the growth of the protruded part and lead to the filament branching. Thus under the driving of both electric field and thermal motion, the filament tends to branch as it grows if the generation rate is large enough, rather than stay in a confined area. The branching area at the oxide thickness tells the physical limit of a memory cell. And the large cell area will decrease the density of the memory array. We can roughly estimate the ratio of growth and also notice that the growth process is away from the equilibrium, which means the equilibrium assumption cannot be used. For example, if the oxide thickness is $\mathrm{L} = 5\mathrm{nm}$ and $\mathrm{V_{app}} = 2\mathrm{V}$ voltage is applied on the memory cell. The rate enhancement from the one-atom thick protruded filament is

$$\frac{\mathrm{R_1}}{R_0} = \exp\left(qd\frac{E_1 - E_0}{2kT}\right) = \exp\left(\frac{qd}{2kT}\left(\frac{V}{L-d} - \frac{V}{L}\right)\right) \approx 2.366.$$

Here we take the lattice constant $\mathrm{d} = 0.5\mathrm{nm}$. In other words, the enhancement for the protruded atom is 2.366 times larger than the un-protruded part. In general, the one protruded filament has 0~4 un-protruded part. Therefore, the growth rates on the filament on the protruded part and un-protruded part are comparable if the vacancies stay in the space with same probability. The comparable growth rate will result in large chance of growth in the un-protruded part, and lead to a flat tip of filament. And the flat tip provides another stage for the filament branching.

4.4 RESULTS AND ANALYSIS IN 3D SYSTEM

To understand the realistic switching process better, we need to simulate the 3D system with kinetic Monte Carlo method. The simulation is started in a 32x32x16 three-dimensional grid, which corresponds to an 8nm thick HfO_2 with a $16 \times 16nm^2$ cross section. The top surface of the simulation box is the interface between Ti and HfO_2, where surface generation can occur. The

bottom surface is the interface with TiN. Neither vacancies nor interstitials can migrate into the TiN electrode. In the simulation, the inert electrode is grounded, and the voltage on the active electrode varies. When the applied voltage is positive, the electric field will help the generation of vacancies with oxygen interstitials in the active electrode and vacancies in the oxide. Since the generation happens near the electrode, the electrons in the vacancy energy level can easily tunnel into the electrode, and therefore these vacancies are likely to be the positively charged (Fig. 4.10). Then, the charged vacancy will be pushed by the electric field towards the bottom inert electrode. The vacancy can diffuse under the electric field and attach to the bottom electrode. Because the bottom electrode is grounded, it can inject electrons into the defect states in vacancy, and result in neutralization of the charged vacancy. The neutralized defect can have a very different migration barrier, and therefore the time scale difference between forming and retention is significantly more than that the enhancement factor from the electric field (the $\exp\left(+\frac{q\mathcal{E}d}{2kT}\right)$ term in eqn (4.44)). The time scale difference is explained as following. For the diffusion of defects, the modified hopping rate under electric field is

$$D(\mathcal{E}) = D_0 \exp\left(-\frac{E_0}{kT} + \frac{q\mathcal{E}d}{2kT}\right). \tag{4.44}$$

Here E_0 is the migration barrier, q is the charge of vacancy, $\mathcal{E}$ is the electric field and d is the distance of migration. Therefore, the ratio of hopping rate with and without electric field is $D(\mathcal{E})/D(0) = \exp(q\mathcal{E}d/2kT)$, which should be correlated to the time ratio of forming and retention. Given a set of typical values in experiments: $applied\ Voltage\ (V) = 2V,\ t_{ox} = 8nm,$ $q = 2e, d = 0.5nm,$ and $kT = 25.8meV$, we find that the ratio of the hopping rate with and without electric field is near 1.6×10^4. However, in experiments the ratio of retention and forming time is more than 10^9, given that the forming time is near $10^{-6}s$, and the retention time is more than $10^3 s$. In other words, the enhancement of diffusion by electric field is not large enough to

warrant the long retention times that are experimentally observed. Ref [81, 82] suggested that there is a large migration barrier difference of vacancies and interstitials with different charge states. The equation before could be rewritten as

$$D(\mathrm{q},\mathcal{E}) = D_0 \exp\left(-\frac{E_0(q)}{kT} + \frac{q\mathcal{E}d}{2kT}\right). \quad (4.45)$$

Here, the migration barrier $E_0(q)$ is a function of vacancy charge. The ratio of hopping rate with and without the charge dependent energy barrier.

$$\frac{D(q,\mathcal{E})}{D(0,0)} = \exp\left(-\frac{E_0(q)-E_0(0)}{kT} + \frac{q\mathcal{E}d}{2kT}\right). \quad (4.46)$$

Once the difference of migration barrier with and without charge $E_0(q) - E_0(0) \approx 0.6eV$ is accounted for, the enhancement of hopping rate could be more than 1.2×10^{10}, which is enough to support the time scale difference in forming and retention.

I notice that oxygen can react with the metal in the active electrode. As a result, the surface generation process can create one vacancy in the oxide and the corresponding oxygen atom migrates into the metal Ti electrode directly. More attention should be paid to the oxygen in the Ti electrode. The migrated oxygen may oxidize the metal electrode and result in a thin insulating layer. This layer can enhance the electric field near the tip which helps yield the scaling of resistance with increase in current compliance. The details are explained in section below.

4.4.1 3D Model: Scaling of low resistance and current compliance

In this section, a 3D model for HfO_2-based RRAM devices is developed to show the switching property of RRAM. The underlying physics is based on oxygen vacancy-interstitial generation, recombination and diffusion. The model solves for the kinetics of oxygen (O) and vacancy (Vo) hopping rate determined by Arrhenius law coupled with the local temperature and electric field. The kinetics of the filament formation is mainly dictated by the Vo generation rate near the Ti

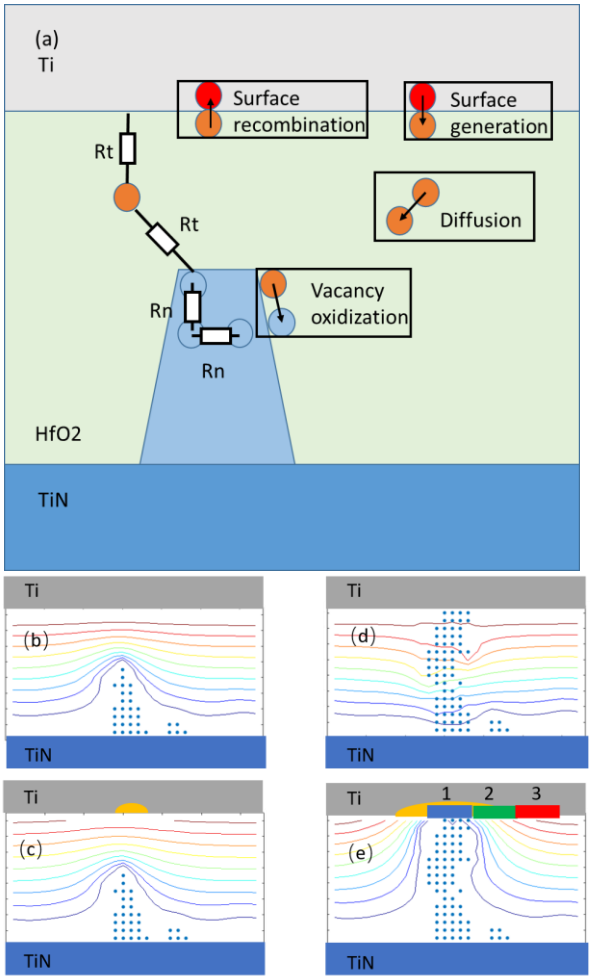


Fig. 4.17. Main processes in the RRAM switching are shown in (a). (b) and (c) show the contour plot of electric potential with and without the TiOx layer when the filament has not fully bridged the two electrodes. (d) and (e) show the electrical potential after the fully bridged filament forms. The electrical field is enhanced near the tip of filament when a TiOx layer is present.

electrode, which is much larger than the bulk generation rate [36]. The current is calculated using a resistive network discussed in section 4.1.

The device structure considered is shown as in Fig.4.17(a), a thin HfO_2 layer is sandwiched by two electrodes. The bottom electrode is inert, and is usually made of TiN or Pt. The top electrode is an active electrode, which is made of a metal such as Ti. Oxygen vacancies can be generated in the interior of the oxide, as well as at the oxide-active electrode interface. The vacancy generation inside the oxide forms a Frenkel pair, where one vacancy is generated together with its paired interstitial oxygen inside HfO_2. On the other hand, the generation process on the HfO_2/Ti

interface involves an oxygen atom migrating into the Ti electrode. Ab initio calculations shows that the surface formation energy for the process, where the oxygen atom resides in the Ti electrode and the corresponding interstitial resides in the oxide, is much lower than the formation energy of Frenkel pairs [82, 36]. Thus, most vacancies are generated near the Ti electrode. Since the Fermi level of the Ti electrode is lower than the energy level of the oxygen vacancy, the generated vacancies are likely to be positively charged, and therefore their diffusion will be guided by the electric field. After the vacancies attach to the inert electrode, electrons can be injected into these vacancies. The neutral vacancies have a higher migration barrier than the charged one. As a result, these vacancies cannot move and a stable conducting filament of vacancies bridging the two electrodes forms.

Since the generation is mainly at the interface, where the formation energy is much lower than the one in bulk, the time to generate one vacancy is much shorter. As a result, the resistive switching can be fast. The small formation energy at the interface can also lead to faster vacancy generation even if there is no bias. Note that this does not affect the high resistance state because of the localized nature of generation. Most sites in the bulk oxide will not be affected by the surface generation, and the resistance state can be retained for a long time. More attention should be paid to the oxygen interstitial formed in the Ti layer. The metal in the active electrode can be easily oxidized and therefore a thin insulating layer of TiOx is formed on the oxide/Ti interface, which is calculated as tunneling current in the current solver. Fig. 4.17(b) and (c) show that the electric field distributions are nearly the same for the cases with and without the TiOx layer, when the filament does not fully bridge the two electrodes. However, after the filament has fully bridged the two electrodes, the presence of the TiOx layer results in a much stronger electric field near the active electrode with components along the oxide-metal interface (Fig. 4.17 (d) and (e)). This

component along the oxide-metal interface helps the filament become thicker. On the other hand, without the TiOx layer, the electric field is perpendicular to the metal-oxide interface, as shown in Fig. 4.17 (c). As a result, the generated vacancy is likely to diffuse along the filament, rather than perpendicular to it. This makes it difficult for the filament to grow thicker, and the current does not increase sufficiently even when the applied voltage is kept on, which is contrary to experimental observations. The presence of the TiOx layer, as shown in Fig 4.17(e), makes the electric field near the active electrode large and highly non-uniform even after the filament forms in region 1. As a result, vacancies are formed more easily in region 2, rather than in the region 3, where the electric field strength is lower. The diffusion of the generated vacancies in region 2 is guided by the large electric field in Fig. 4.17(e). Therefore, the filament will be thickened after switching and the resistance of the switched device, which is roughly inversely proportional to the thickness of the filament, decreases with the increase in current compliance.

In this model, the surface formation energy is lower by 0.05eV in a small surface area of $2 \times 2\text{nm}^2$ at the Ti-Hafnia interface located at the center of the oxide–active electrode interface. The site spacing is 0.5nm, which is length of HfO_2 unit cell. Denser vacancy distribution is not considered. This might be due structural variations, such as grain boundary intersections or mechanical stress. When the vacancy is attached to the inert electrode or other neutral Vo which are electrically connected to the inert electrode, electrons can fill the vacancy energy states. Therefore, the vacancy can become neutral as a result of the applied voltage. Since the electrons are injected from the inert electrode, vacancies near the electrode have a small net charge and their activation energy for diffusion is high. On the other hand, because of the voltage drop on the filament, Vo near Ti have a higher chance to be charged and hence a smaller activation barrier, which helps vacancy diffusion. In Fig. 4.18, $\mathrm{E}_{\mathrm{barrier}} = 1.1eV$ (diffusion barrier for vacancies);

$E_{form(s)} = 1.4eV$, $E_{form} = 6eV$ for the bulk formation energy, $r_N = 1k\Omega$, $r_{T0} = 1k\Omega$ and $\alpha = 1.5$ have been discussed earlier in equations (4.1) to (4.6) and figures 4.1 to 4.3. These parameters are fitted from experiment data. In the simulation with surface generation, the vacancy is considered only since the oxygen atom has entered the metal electrode.

A typical switching current and voltage versus time are shown in Fig. 4.18, when a $2V$ bias is applied on the $8nm$ thick oxide. The voltage is applied with the constant numerical value until current compliance is reached, then the voltage on memory cell is decreased to comply with the

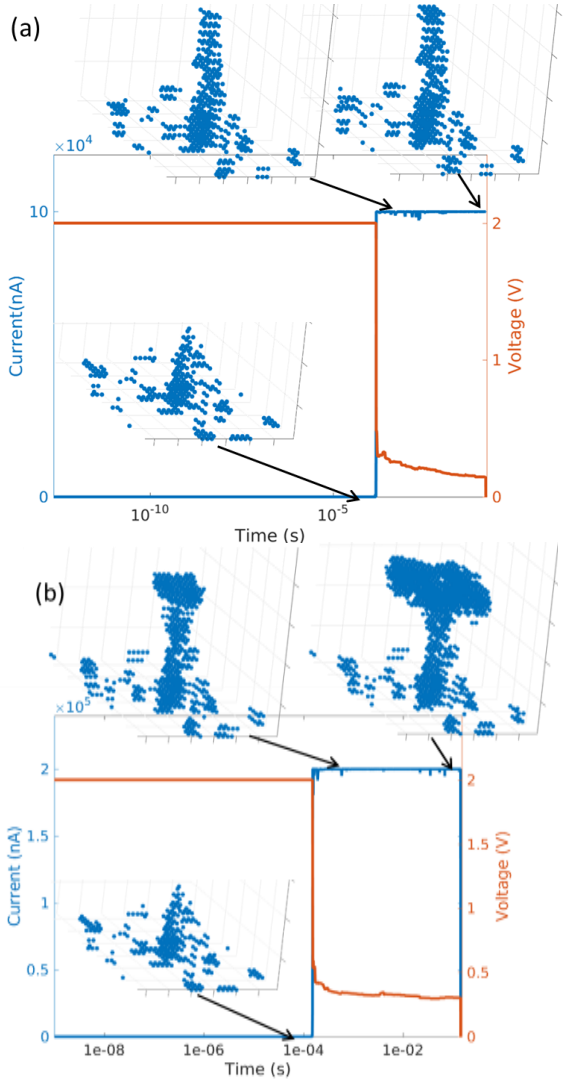


Fig. 4.18. The current (blue) and voltage (orange) versus time for (a) 100 µA and (b) 200 µA compliance respectively. Insets show the shape of filament at different stages of formation.

specified current compliance value. The current through the memory cell suddenly rises at $200\ \mu s$ after the switch is turned on, and because of current compliance, the voltage drops. At low current compliance ($100\ \mu A$), we find that the shape of the filament is conical as seen in insets of Fig. 4.18(a). In contrast at high current compliance (200 μA), the filament develops a dumbbell shape, with over all thickening of the filament as shown in Fig.4.18 (b). The reason for the dumbbell shape in the higher current compliance case is the deposition of vacancies near the active electrode due to the existence of a high voltage for a longer time. The generated vacancies diffuse towards the inert electrode and the filament grows gradually.

An important characteristic of ReRAM is that it can offer multi-level resistances in a single cell by applying different current compliances. We can test the forming process in kinetic Monte Carlo method with different initial random seeds, and then compute the resistance after current compliance is reached. The mean resistance versus current compliance is shown in Fig. 4.19 (a). At small current compliances (< 100μA), we can observe that for both cases of $\mathrm{E_{bond}} = 0.03eV$ and $\mathrm{E_{bond}} = 0.1eV$, the mean resistance decreases with increase in current compliance. On the other hand, in the large current compliance region (> 100μA), the resistance continues to decrease with current compliance in the small bond energy case $\mathrm{E_{bond}} = 0.03eV$, while for the large bond energy case, $\mathrm{E_{bond}} = 0.1eV$, the resistance decrease is negligible. We also observe that in the case of a larger bond energy (0.1eV), the mean resistance is larger (from 100 μA to 500 μA).

The resistance decrease with an increase in current compliance is related to vacancy generation and filament growth. We should notice that most vacancies are generated near the tip of the filament due to electric field enhancement. Then, the electric field guides the vacancy towards the filament and thickens the filament. With a larger current compliance, the voltage across the device is large for a longer period of time during which the thickening process occurs.

This leads to the smaller resistance. As shown in the Fig. 4.19, with the TiOx on top of the filament,

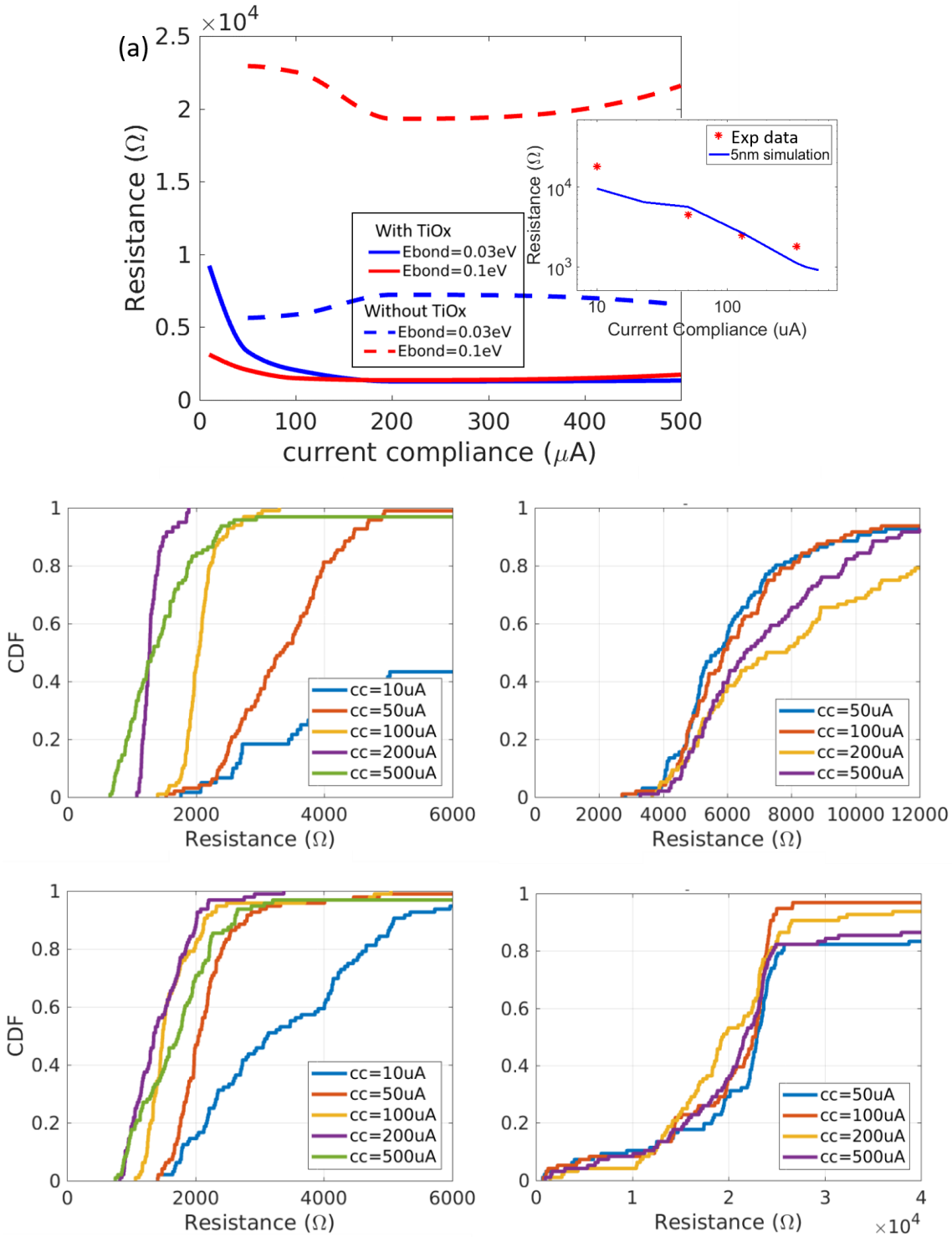


Fig. 4.19. (a). Filament resistance vs. compliance current, inset is the simulation of 5nm oxide comparing with experiment data from [88]. (b) and (c) are the cumulative distribution functions (CDF) of forming for Ebond=0.03eV and Ebond=0.1eV respectively with the TiOx part on top of the filament. (d) and (e) are the cumulative distribution functions (CDF) of forming for Ebond=0.03eV and Ebond=0.1eV respectively without the TiOx part on top of the filament.

the resistance of filament after switching is decreasing with increase in compliance current. On the other hand, if we remove the top TiOx part, we can notice that the filament resistance is kept at the same value regardless of the value of compliance current. The reason for this is that after switching, the electric field is distributed uniformly as shown in Fig. 4. 17(d), and therefore can hardly help the thickening process of filament. Once the filament has bridged the two electrodes, the filament growth stops (or becomes very slow) because the electric field is not strong enough to guide the vacancy toward an existing filament. Therefore, the resistance of the filament is determined by the filament thickness when it bridges the two electrodes – the resulting distribution of resistance values is found to be quite large.

The resistance decrease with increase in current compliance is different at high and low bond energies. When the bond energy is high, the electric field is not effective in pushing the vacancies because it takes a large energy to break the bond between vacancies. On the other hand, with a smaller bond energy, the vacancies can diffuse more freely, which helps lead to the thickening of the filament. Therefore, the resistance stops decreasing in the large current

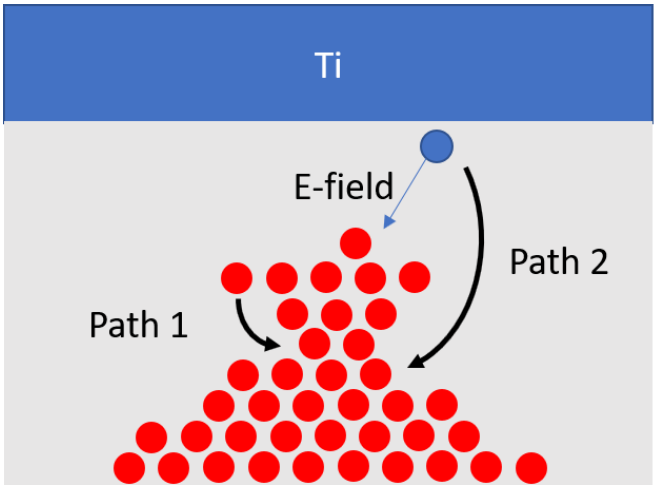


Fig. 4.20. An illustration of thin spot in the filament. The thin spot can hardly thicken via path 1 since the vacancy attached to the filament and so can hard to move because of the high migration barrier of neutral vacancies; the thin spot can hardly get thicker via path 2 since the generated vacancy near top electrode can hardly reach the thin spot because of electric field screening due to the broad region on top of the filament. See the dashed oval box for the region where electric field screening occurs.

compliance region. To understand the resistance distribution, we can also plot the cumulative distribution function (CDF) of resistance under different current compliances. Fig. 4.19(b) and (c) shows the CDF of 100 samples for small and large bonding energy. We observe that at current compliances of 100 μA and 200 μA, the lower 90% of the resistances distribute with a small deviation. This distribution is different from the results obtained from free energy calculations.

In the kinetic Monte Carlo simulation, a narrow region in the filament is often found that has wider filamentary regions both above and below (both towards the top and bottom electrodes, see Fig. 4.20). It is observed that these narrow regions can hardly thicken during the filament growth process. The reason for this is discussed now using Fig. 4.20. The neutral vacancies that are already attached to the filament in regions towards the bottom electrode find it difficult to diffuse to the narrow region through path 1 because neutral vacancies have a large diffusion barrier. The vacancies are neutral because they are connected to the bottom electrode. Lone vacancies in between the filament and the bottom electrode also find it difficult to diffuse towards the narrow region because the electric field is screened by the wider filamentary regions (See path 2 of Fig. 4.20). As a result, these narrow regions cannot grow and they determine the resistance of the filament and a long tail of CDF shows up in the distribution of resistance values.

4.4.2 RESET process

Different from the hard breakdown of oxide, the resistive switching in RRAM is reversible. In experiments with an oxygen reactive metal as an electrode, the bipolar RESET is commonly used. In bipolar RESET, a reverse (opposite to SET) bias voltage is applied to the memory cell with low resistance. As a result, the filament is ruptured, and the memory cell is RESET to a high resistance state.

The RESET operation is modeled with a reverse bias of -2V, as shown in Fig.4.21. The device is RESET in a short time and current decrease smoothly with time. In the case of RESET, the electric field is reversed and therefore the vacancy tends to diffuse to the active electrode, which has a lower electric potential now. The TiOx can serve as a reservoir of oxygen, and when an oxygen vacancy is near the surface, the recombination of vacancy and oxygen occurs; oxygen in the TiOx reservoir diffuses back to the oxide to fill the oxygen vacancy. The TiOx layer can

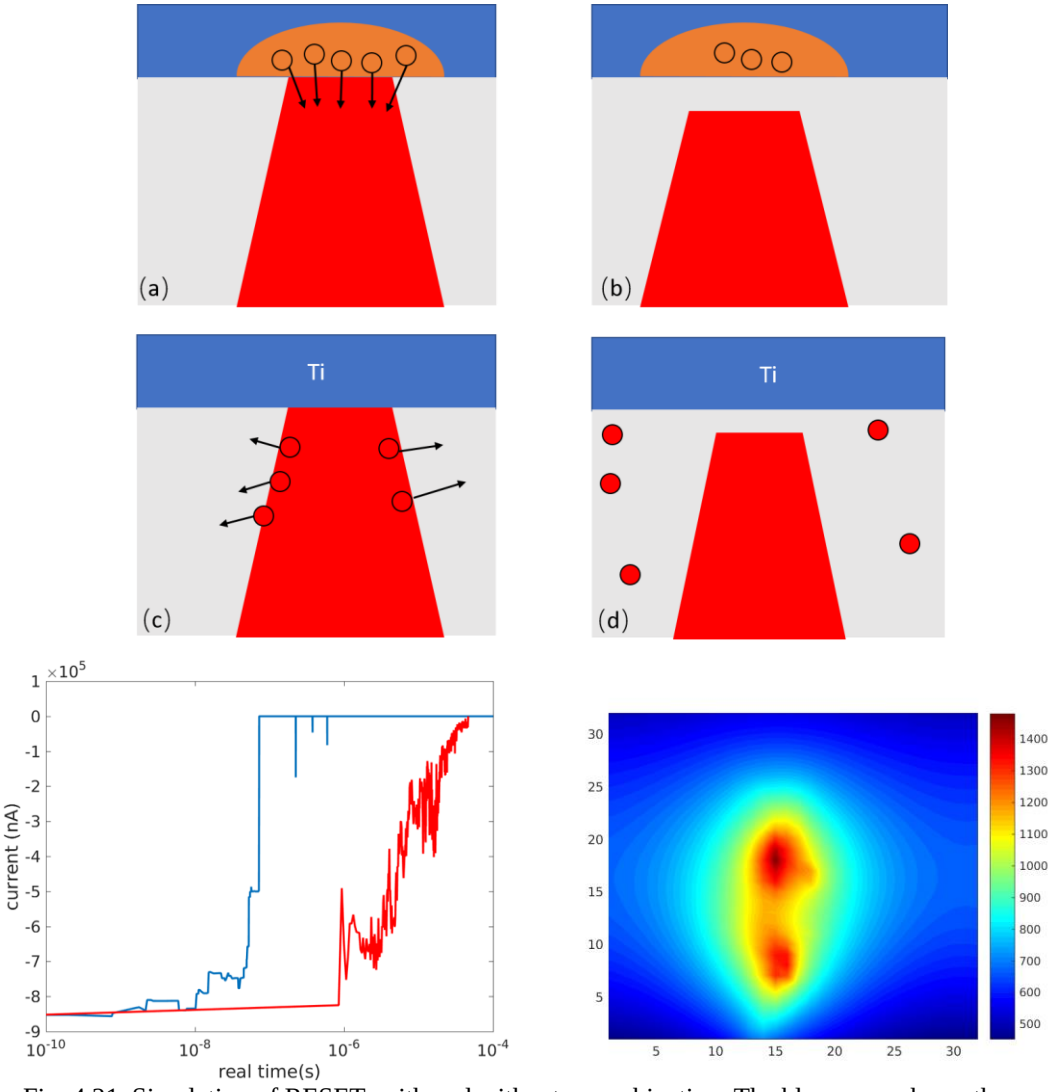


Fig. 4.21. Simulation of RESET, with and without recombination. The blue curve shows the current in the RESET process where oxygen in Ti diffuses back to the oxide and recombine with the vacancy, which leads to a RESET time of ~ 3×10^{-8}s. If the oxygen in Ti cannot diffuse back, Joule heating can increase the temperature of the filament above 1400K, as shown in (f). Then vacancies diffuse away and the memory cell can RESET. The RESET with Joule heating is shown by the red curve of (e).

also enhance the electric field in the RESET process. The diffusion process is fast because recombination is energy favorable and the electric field direction is helping the recombination happen. The creation of a region without oxygen vacancy forms resistive regions, leading to the high resistance.

The prospect of RESET occurring when the process of oxygen vacancy recombination is removed is also studied. It is found that the RESET process can still occur but now it is driven by Joule heating, which can break the filament. As shown in the Fig. 4.21 (c) and (d), the Joule heating can build up a temperature gradient and vacancies can diffuse away from the filament, leading to RESET. But I notice that the RESET time now is longer than the case with surface recombination (Fig. 4.21 (e)). Further, for the REST to occur, a positive thermophoresis coefficient is required. The role of thermophoresis is further discussed in Chapter 5.

4.4.3 Retention Modeling

Since the resistance of filament is determined by the configuration of vacancy, the information of memory cell can be kept for a long time after the cell is set to a low resistance state. KMC simulations are performed to study the retention of the memory cell, as shown in Fig. 4.22. In the retention, the voltage is not applied to the memory cell and the vacancies are neutral. Therefore, the migration barrier is higher than the charged vacancy, as stated in Eqn (4.46). In the simulation, the migration barrier is taken as $\mathrm{E_{barr}} = 1.1eV$ [83]. After the KMC simulation is finished, a test voltage of 0.3V is applied on the vacancy configuration to compute the current through the filament and calculate the filament resistance ($0.3V/Current$). As shown in Fig 4.22 (e), when $\mathrm{E_{bond}} = 0.03eV$, the low resistance state can be maintained for around 10^6s (~10 days) at 300K. On the other hand, if the vacancies have a high bonding energy, $\mathrm{E_{bond}} = 0.1eV$, the

resistance of filament can last longer. As shown in Fig. 4.22 (j), the low resistance state can last for $10^8 s$ (~3years). Because of the high bonding energy, the vacancies can hardly diffuse away from the filament region and therefore the filament can persist for a longer time.

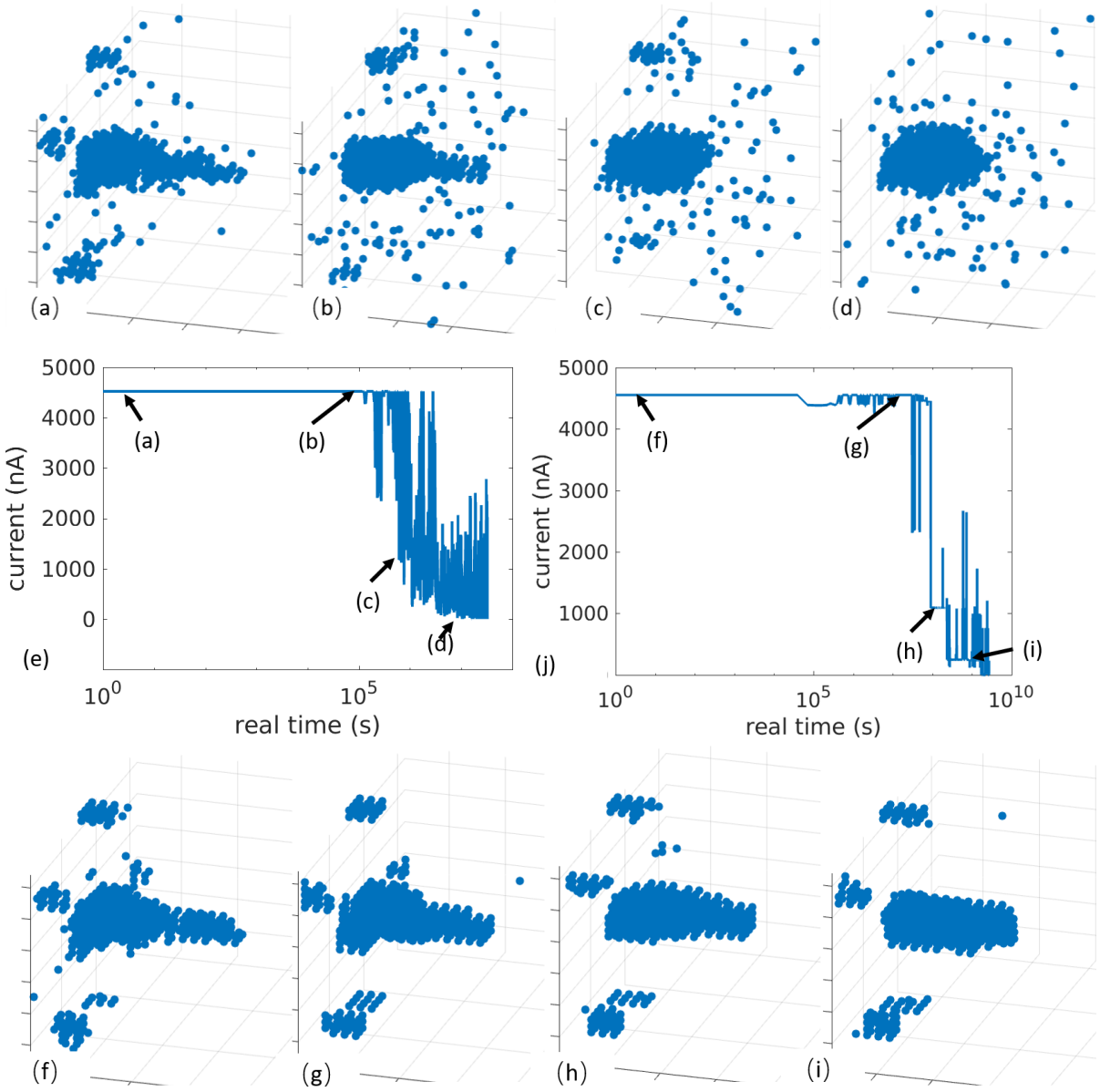


Fig. 4.22 Retention of filament cell for low (a-e) and high (f-j) bonding energy. The ambient temperature is 300K. The migration barrier is $E_{barr} = 1.1$eV. Current is computed from a test voltage of 0.3V after the KMC simulation. In the low bond energy (a-d), $E_{bond} = 0.03eV$, the vacancies are diffuse away from the central region and current drops, as shown in (e). In the high bond energy (f-i), $E_{bond} = 0.1eV$, the vacancies are staying in the central region and current can be kept for a long time, as shown in (j).

4.4.4 Discussion on forming time

We can use this simulator to discuss the forming time of resistive switching. In the experiment, the forming time of a memory cell depends on the electric field and temperature. [88, 89] have found that the forming time can be faster than the intuitive electric field enhancement of $\exp(eFd/kT)$, with electric field F and site spacing d. The explanation of the fast switching is attributed to the crystal structure enhancement [90], which can accelerate the switching time with a factor of $\exp(\alpha\, eFd/kT)$ with $\alpha = (\epsilon_r + 2)/3$, and ϵ_r is the relative permittivity of switching layer. More detailed discussion on the electric field enhancement is in Chapter 5. Here we are fitting the data from [89] by using the parameters that the formation energy $E_{form(s)} = E_{form(s)}(0) - \alpha\, eFd \approx E_{form(s)}(0) - \alpha e \frac{V_{app} d}{t_{ox}} = 2.2\text{eV} - V_{app} \times 0.4e$. The result of switching is shown as in Fig. 4.23. The switching time is decreased by 8 orders of magnitude when

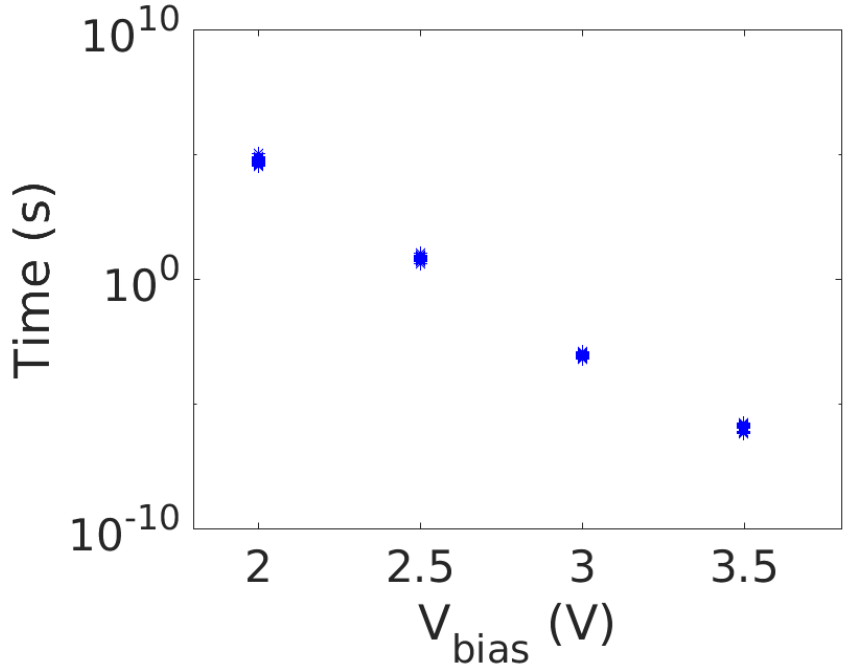


Fig. 4.23 RRAM forming time versus applied voltage under the assumption that the formation energy is depending on the applied voltage $E_{form(s)} = E_{form(s)}(0) - \alpha\, eV_{app}\, d\, /t_{ox}$. $E_{form(s)}(0) = 2.2\text{eV}$ and $\alpha\, d/t_{ox} = 0.4$.

the applied voltage increases by 1.5V. This result approximately matches the data from [71] and data from [88]. However, we should notice that the experimental data does not match the prediction from Arrhenius law over the entire voltage range. For example, the switching time decrease between 1.7V to 1.8V in [89] is much slower than at other voltage ranges. Also, in [88], switching time changes slowly between 2.5 to 3V applied voltage. The slow change in certain voltage ranges could be a measurement error, but it also can come from the defect charge state changing with voltage. An illustration of the charge states is shown in Fig. 4.24a. In the surface generation process, the vacancy can form with different charge states under different electric fields. And the different charge states are formed with different formation energies, which can result in the switching time changing rapidly within certain voltage ranges. For example, under the large voltage applied (>2V), the electric field is high, and +2e vacancy is more likely to be generated. The +2e vacancy formation energy is lower than the +1e vacancy and neutral vacancy formation energies [82], so the switching time is short. And when the applied voltage is lower, the electric field is smaller and therefore more +1e vacancies are generated. The formation energy and migration barrier for +1e vacancy is higher, which result in a longer switching time. In the simulation, migration barrier is 0.5eV if applied voltage is equal to or larger than 2V, and the migration barrier is modified to 0.7eV if the applied voltage is less than 2V. The simulation result is shown as in Fig. 4.24(b) that the switching time can change by around 3 orders of magnitude between 1.5V to 2V, while the switching time change is less than 1 order of magnitude between 2V and 3V. Two possible mechanisms to explain the switching time are simulated and the results

are shown in Fig 4. 23 and Fig. 4.24. However, without further experimental data, it is hard to make a conclusive statement of the underlying physics.

To sum up, a kinetic Monte Carlo model is developed to simulate filament formation, SET and RESET processes in HfO2-based resistive memory devices. The memory cell consists a switching layer, an inert electrode and an active electrode. The active electrode, which is usually made by Ti in experiment, can easily react with oxygen. Most oxygen vacancies are generated near the active electrode interface. The generated vacancy can diffuse under the electric field and form the conductive filament. The generated oxygen interstitial directly goes into the active electrode and partially oxidizes the Ti, which enhance the electric field near the filament tip. The electric field enhancement helps the thickening of the filament. The thickening process during formation is controlled by current compliance, and therefore the filament resistance is related to the value of current compliance. The RESET process happens when a reversed voltage is applied. The oxygen

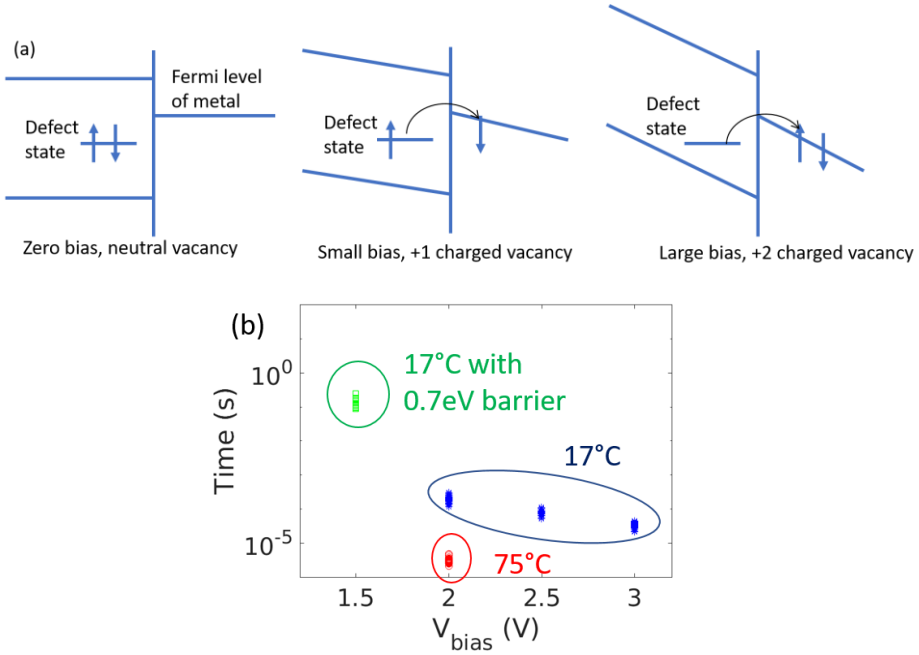


Fig. 4.24 (a) Vacancy charge state can change under different bias voltage. (b) RRAM forming time versus applied voltage under different migration barrier due to charge state of vacancy.

in the active electrode diffuse back to the oxide and recombine with the vacancy in the filament.

Because the current compliance can be easily controlled in experiment, the multi-level storage of memory cell can be implemented by changing the value of current compliance.

Chapter 5. FILAMENT FORMATIONA DN DISSOLUSION IN UNIPOLAR DEVICES[3]

We discussed in Chapter 1 that there are two different kinds of RRAM devices (1) Bipolar devices and (2) Unipolar devices and discussed the relative advantages of them. To recall, a unipolar device is both SET and RESET when the voltage across the device has the same polarity. That is, unlike bipolar devices, two different polarities ($+V_e$ and $-V_e$ voltage drops) are not required to SET and RESET the devices. The unipolar device is attractive because it can offer a solution to the sneak path problem in RRAM arrays.

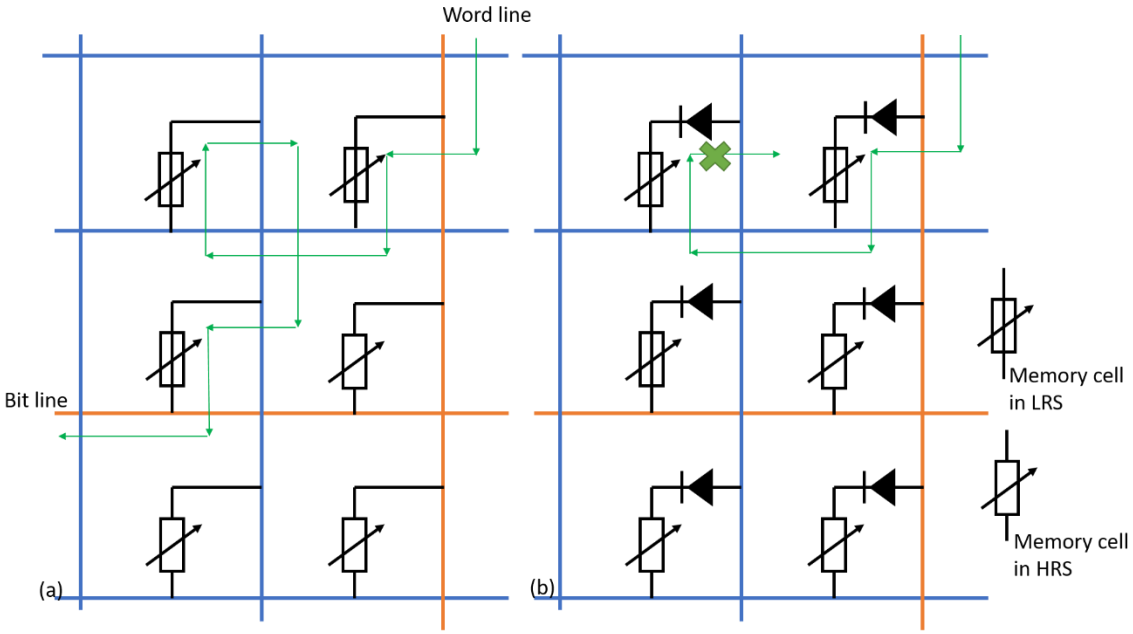


Fig. 5.1 (a) shows the sneak path problem of memory cell array without diodes. When reading a memory cell by activating a bit line and a word line, the current can run through surrounding memory cells in LRS and result a false reading. (b) shows that the memory array with one diode (1D1R) can avoid the sneak path problem. The additional diodes prevent the current to flow through the sneak path in (a), and the desired memory can be read.

[3] Most of the work in this chapter has been submitted for publication

One of the issues in developing a cross-wire based resistive memory array is ensuring that the resistances can be properly read. Primarily, the issue is that if a voltage is applied to one horizontal line and the current is measured from one vertical line to read a single element, then there may be other possible pathways that can contribute to the overall current (see Fig. 5.1). If the element has a high resistance, then these other pathways can dominate the current, and result in a misread value. From an architectural perspective, this issue is circumvented by connecting a diode in series with the memory element lying between the vertical and horizontal nanowires.

To RESET the memory state in unipolar RRAM, a large current is applied through the memory cell with the same polarity as the SET process. As the voltage is applied in the same direction for both set and reset operation, a diode connected in series with the memory cell can avoid the sneak path problem

In experiment, the unipolar reset is usually observed in a memory cell fabricated with inert metal electrodes on both sides of the switching layer. With inert metals, the preferential surface generation of vacancy and interstitial near the oxide-metal interface is difficult due to the high formation energy of oxygen in the inert metal. The Frenkel pairs are likely generated in the bulk switching layer, especially in regions with lower formation energy, such as in grain boundaries. Since the electric field direction is the same in SET and RESET processes, the temperature is accepted to be the main reason for resistive switching. The physics of the SET/RESET process is very sensitive to the sign of the thermophoresis coefficient because a single vacancy will be driven from high to low temperature regions if the thermophoresis coefficient is positive, while when the thermophoresis coefficient is negative, the vacancies diffuse to the sites with high temperature. In the case of positive thermophoresis, unipolar RESET is traditionally explained by vacancies radially diffusing out to break the filament (the filament is made up of vacancies) [91]. The high

temperature, which is induced by the high current running through the filament, helps this process. Recently, a negative thermophoresis effect of oxygen vacancy has been reported in Hafnia based RRAM devices [92, 93, 94]. Here oxygen vacancies tend to diffuse to the sites with higher temperature. A strong temperature gradient in the RESET process will cause vacancies to diffuse inward towards the core of the filament, making the filament stronger, thus preventing RESET. One idea in the literature is that temperature may be lower in RESET than in SET, which makes the negative thermophoresis effect weak. Then, the vacancies can diffuse out radially driven by the concentration gradient, resulting in a RESET due to partial filament dissolution. However, experiments involving unipolar devices show that power dissipation during RESET is higher than in SET, which conflicts with the explanation in [95, 42].

In this chapter, we use the kinetic Monte Carlo method to model the generation and diffusion of defects with bulk generation and propose a mechanism for RESET in devices with a negative thermophoresis coefficient. The results show that RESET can occur if power dissipation is higher than in SET. The underlying mechanism involves diffusion of interstitials to break the vacancy-based filament in a region close to the top electrode.

5.1 PHYSICAL MODEL OF RRAM IN BULK GENERATION

We will first start out with an explanation of crystal field enhancement and hopping rates. Then we will discuss Positive and Negative thermophoresis. Finally, we will present results of form, set and reset processes under the negative thermophoresis.

5.1.1 Grain boundary, McPherson's crystal field enhancement

The HfO_2-based RRAM cell is modeled in 3D with physics underlying oxygen vacancy (Vo) and interstitial (Io) generation, recombination and diffusion. In this model, Vo and Io kinetic rates are determined by Arrhenius law, coupled with the heat and Poisson's equations. The current is calculated using a resistive network consisting of vacancy cluster islands. The structure considered consists of a thin HfO_2 layer that is sandwiched by two electrodes (Fig. 5.2). We assume that both electrodes are inert with respect to reaction with oxygen. Examples include platinum (Pt) and titanium nitride (TiN). Vo can be generated in the interior of oxide along with an associated Io, which is referred to as a Frenkel pair. Ab initio calculations suggest that the formation energy of a Frenkel pair is around 5~8eV [36, 96]. Then, Arrhenius law gives the average generation time for one Frenkel pair, $\mathrm{t_G} = \tau_0 \exp(E_{form}/kT) \approx 10^{19}s$, given 300K with the pre-factor of $\tau_0 =$ 1fs. We note that this time is too long when compared with experiments. In experiments, the resistance switching can happen within 1us [95]. Therefore, to match results from experiments, we need to assume a region with a lower formation energy in the oxide. The intersection between two grain boundaries might be an explanation for the low formation energy, as indicated by [68]. Mechanical stress introduced during the fabrication process might also be responsible for a decrease in the formation energy as per [67] though a microscopic reason for this is not given. For example, mechanical stress can change the lattice constant and lead to a lower formation energy. In this chapter, we assume that there is a region with a formation energy of 1.5eV, to match the order of magnitude results from experiments. Note that the generation time for a Frenkel-pair with this formation energy is ~10^{10} s seconds at 300K, which is again too long for the formation of a Frenkel-pair.

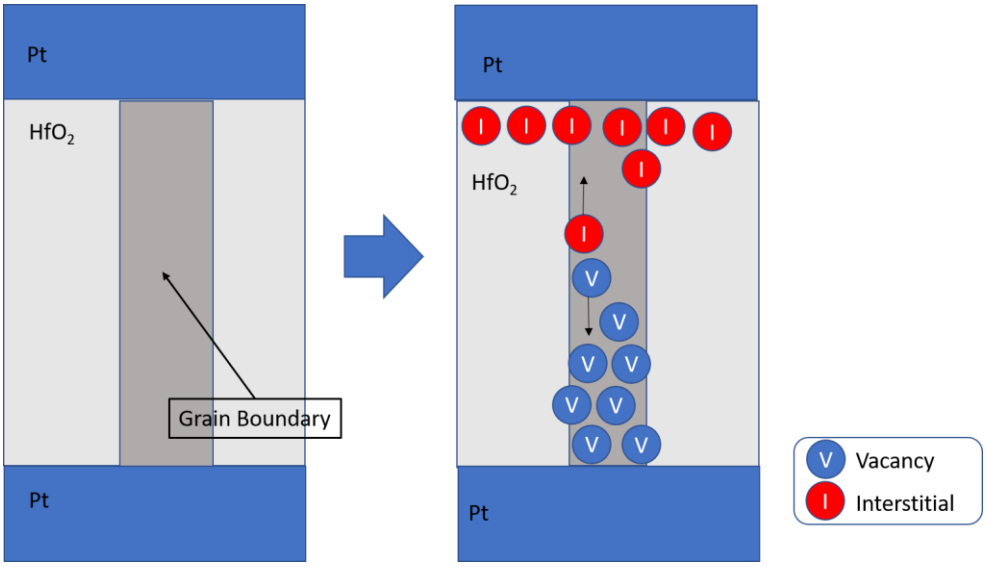


Fig. 5.2 System of bulk generation. The vacancy and interstitial can be generated in the grain boundary region of bulk switching layer.

It should be noticed that outside the grain boundary region, the formation energy for a Frenkel pair is large. Therefore, once a Frenkel pair is generated inside the grain boundary region, there is an extra migration barrier for one or both defects to diffuse outside the grain boundary region. In other words, at least one type of defect stays inside the grain boundary region. In the switching process, vacancies stay in the grain boundary region while interstitials can more easily leave. This extra migration barrier for vacancies to diffuse outside the grain boundary region is neglected in the simulations of this chapter for simplicity. Note that in the simulation results with negative thermophoresis and bond energy between vacancies, vacancies stay inside the grain boundary region even without the neglected higher migration barrier across the regions with two different formation energies. The results presented will be similar if we account for the extra migration barrier as this only promotes the vacancies to stay in the region with the lower formation energy.

The low formation energy region is insufficient to explain all experimental observations. With the low formation energy, the defect generation rate is large compared to bulk regions, when the

current is both small (e.g. in the SET process) and high (e.g. in the RESET process). In the high current case, the temperature is also high, which results in a larger defect generation rate. When the generation rate for defects is high, the memory cell is unable to reset, which conflicts with the unipolar reset experiment. McPherson proposed that the enhancement of the applied electric field due to crystal structure can significantly speed up defect formation in hi-K materials [90], and this enhancement can explain the RESET process. The following expression is used to determine the generation rate,

$$\mathrm{R_G} = R_0 \exp\left(-\frac{E_{form} - qFd/2}{kT}\right) \tag{5.1}$$

Here, $\mathrm{E_{form}}$ is the formation energy, q is the charge of defect and d is the lattice constant. The electric field F is enhanced by crystal structure, where $\mathrm{F} = \frac{\epsilon_r+2}{3}\mathrm{V/L}$. Here ϵ_r is the relative dielectric constant, V is the applied voltage and L is the thickness of oxide.

If the applied voltage is V and the oxide thickness is L, the local electric field in the oxide should be $(\varepsilon_r + 2)\mathrm{V}/3\mathrm{L}$ [90]. In case of HfO_2, the dielectric constant $\varepsilon_r = 25$ makes the local electric field nine times larger than the macroscopic field. The enhanced electric field from the crystal structure adopted in the simulation lowers the formation barrier by $9\mathrm{qVd}/2\mathrm{L}$ (d is the lattice spacing). This factor is important for the set process and crucial to prevent re-generation in the unipolar reset process as discussed later. The charge of interstitials and vacancies also play an important role in the filament formation and rupture. From ab initio calculations, the acceptor level of an interstitial is 0.3eV above the valence band maximum (VBM), while the donor level from a vacancy is 3.1 eV above VBM [84]. Therefore, in the process of Frenkel pair generation, Io is

likely to be negatively charged, while Vo is positively charged. In this chapter, we assume that the charge on the defects are +2e and -2e for Vo and Io respectively.

5.1.2 Basic Concept of Positive and Negative thermophoresis

So far in this thesis, we have assumed that both vacancies and interstitials move from a lattice site at a higher temperature to a lattice site at a lower temperature, provided that there are no energetic differences between the two sites. The usual process to model this phenomenon consists of two diffusion processes between the sites as shown in Fig. 5.3. In both processes, the hopping rate out of a site depends on the source site where the vacancy resides.

The hopping rate R_1, from the left-site to right-site depends on the temperature of the left-site (T_1). Note that here the left-site is the source site from where the vacancy diffuses, and the right-site is the destination site for the vacancy. Similarly, the hopping rate from the right-site (source) to left-site (destination) R_2, depends on the temperature of the right-site (T_2). The net hopping rate from left-site to right-site, $R = R_1 - R_2$ is given by,

$$R = R_1 - R_2 = R_0 \exp\left(-\frac{E_{bar}}{k}\left(\frac{1}{T_1} - \frac{1}{T_2}\right)\right). \tag{5.3}$$

Under the above conditions, the net diffusion will be from left-site to right-site if $T_1 > T_2$. This is known as *positive thermophoresis*. Note that there are a number of lattice sites in the system but to understand the basic phenomena, the simplified two-site description conveys the essential point of the underlying physics.

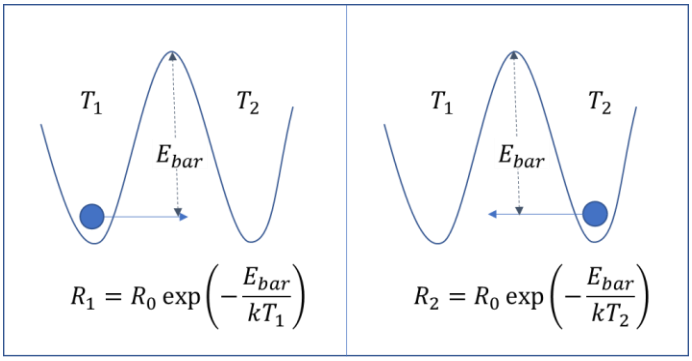


Fig. 5.3 Illustration of positive thermophoresis. When the temperature $T_1 > T_2$, the hopping rate $R_1 > R_2$. As a result, the defects in the high temperature region are likely to diffuse into the low temperature region.

However, it has been observed that the net flow of particles can be from a lattice site with a lower temperature to a lattice site with a higher temperature, and this phenomenon is known as *negative thermophoresis*. To model negative thermophoresis, the usual procedure is to change the temperature in the formulae for hopping rates to depend on the temperature of the destination site as shown in Fig. 5.4. The net hopping rate from left-site to right-site, $R = R_1 - R_2$ is then given by,

$$R = R_1 - R_2 = R_0 \exp\left(-\frac{E_{bar}}{k}\left(\frac{1}{T_2} - \frac{1}{T_1}\right)\right). \tag{5.4}$$

It can be verified that now the net flow of particles is from the right-site (lower temperature T_2) to the left-site (higher temperature T_1).

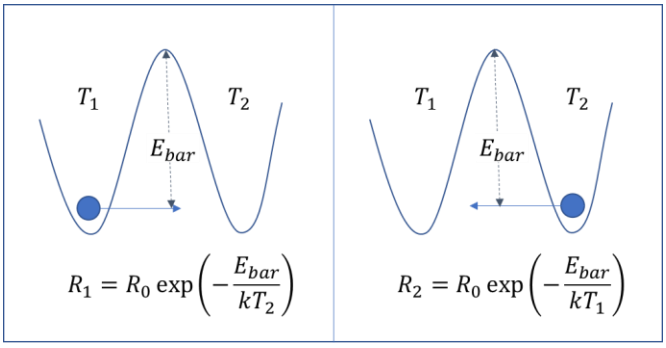


Fig 5.4 Illustration of negative thermophoresis. When the temperature $T_1 > T_2$, the hopping rate $R_1 < R_2$. Therefore, the defects in the low temperature region are likely to diffuse into the high temperature region.

In the discussion below, we model the filament growth in a material with a negative thermophoresis for the vacancies. To summarize, the negative thermophoresis is modeled by the formula

$$\mathrm{R}_{\mathrm{diff(V)}} = R_0 \exp(-E_{diff(V)} / kT_d), \tag{5.5}$$

where, $\mathrm{T_d}$ is the temperature at the destination site. (V) stands for vacancy.

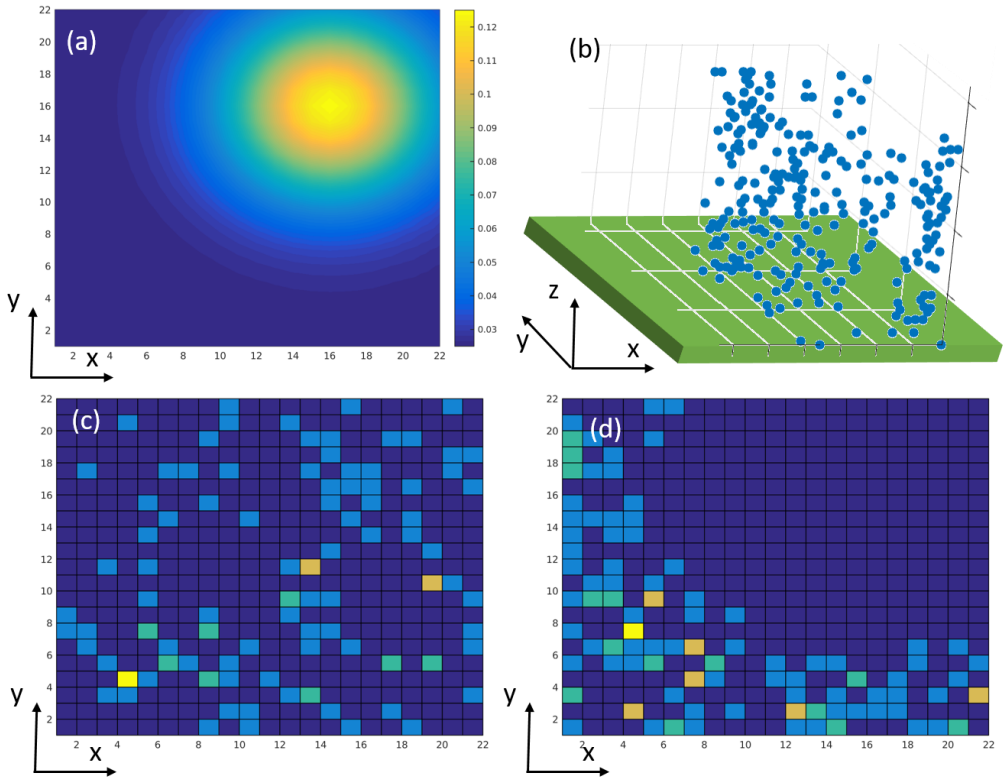


Fig. 5.5 An evolution of defects with positive thermophoresis coefficient. (a) shows the temperature distribution in xy-plane, and temperature is the same alone z-axis. (b) shows the 3d distribution of vacancy after diffusion of vacancies from the initial distribution. (c) is the initial distribution from top view (along z-axis); and (d) is the final state from top view (along z-axis).

We test the Monte Carlo code in the cases of both positive and negative thermophoresis with an ideal temperature distribution as shown in Fig. 5.5 and 5.6. The role of bond energy and generation are neglected in this test. The temperature is assumed to change radially from 1000 K at the hottest

region to 300 K at the coldest region. There is an initial random distribution of vacancies (generation and recombination are switched off for this test). We then look at the distribution of vacancies after a fixed simulation time. In Fig. 5.5, which is the positive thermophoresis case, we see that the vacancies have moved away from the hot regions. In contrast, in Fig 5.6, which considers negative thermophoresis, the vacancies have moved inwards to cluster at the hottest region, at the end of the simulation. The microscopic reason is that the defects in the hot region takes a short time to diffuse to the cold region, while it takes a longer time to diffuse back to the hot region, in the case of negative thermophoresis.

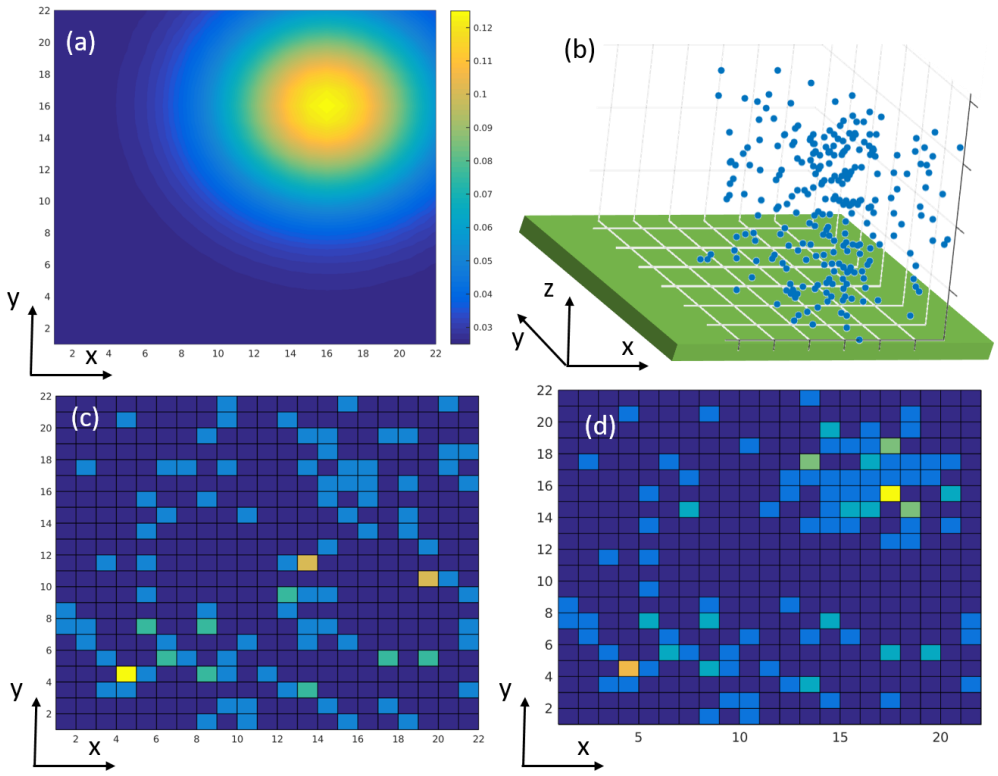
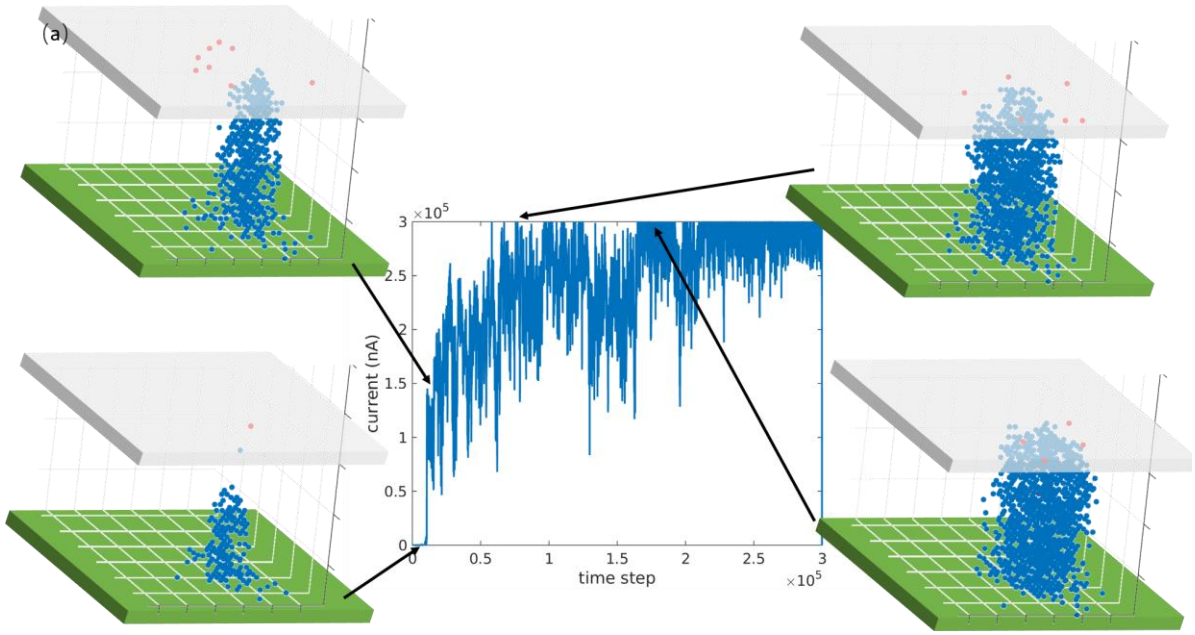
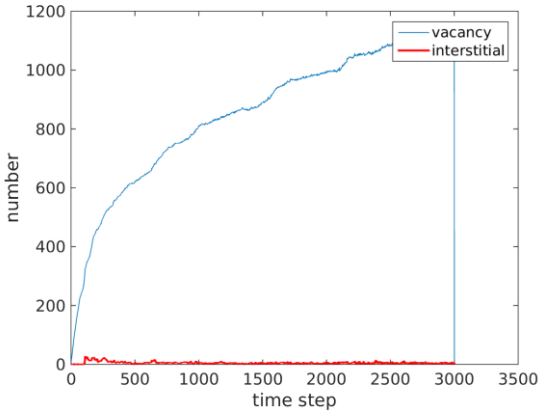


Fig. 5.6 An Evolution of defects with negative thermophoresis coefficient. (a) shows the temperature distribution in xy-plane, and temperature is the same alone z-axis. (b) shows the 3d distribution of vacancy after diffusion from the initial distribution. (c) is the initial distribution from top view (along z-axis); and (d) is the final state from top view (along z-axis).

5.2 RESULTS AND ANALYSIS IN 3D SYSTEM WITH BULK GENERATION

We model the formation process and demonstrate results that qualitatively agree with experimental observations and prior modeling work [97]. Specifically, we find a filament that is primarily in





(b)

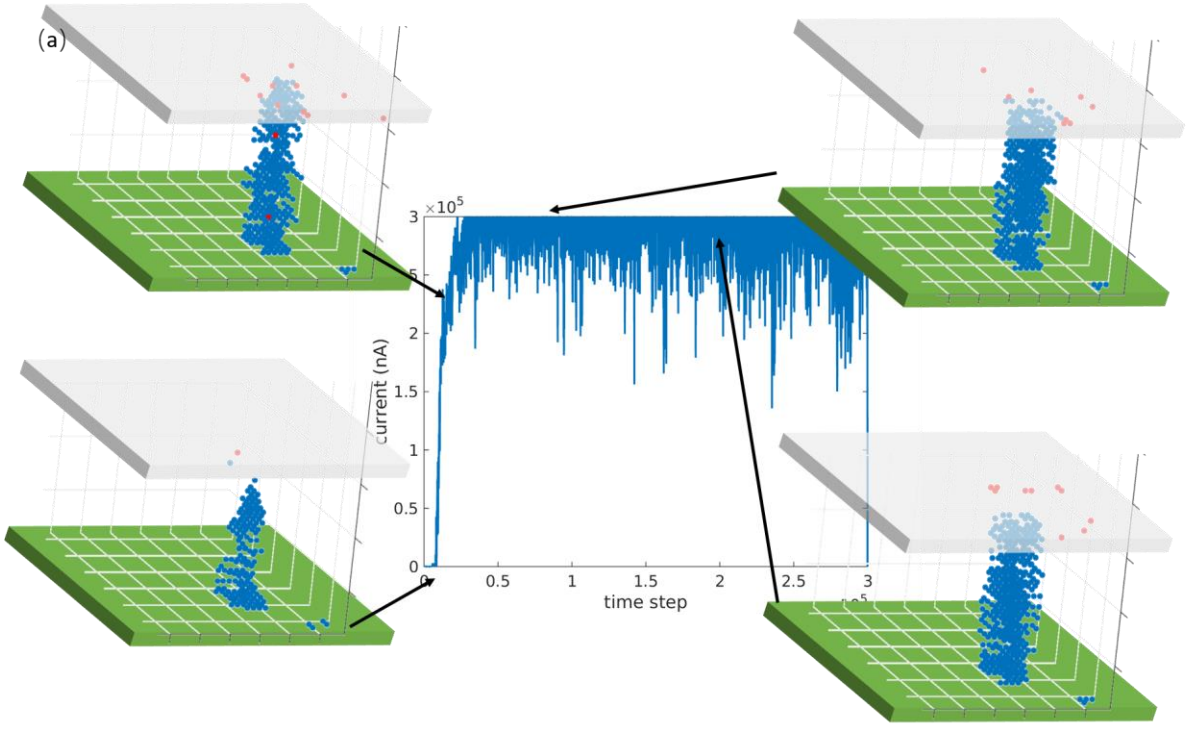
Fig. 5.7 (a) shows the filament formation with bulk generation, and (b) shows the number of interstitials and vacancies. The interstitial can migrate out of system and $E_{bond} = 0.0eV$. $E_{barr}(V) = 0.6eV$ for isolated vacancy, $E_{barr}(Va) = 1.2eV$ for attached vacancy, $E_{barr}(I) = 0.3eV$ for interstitial and $E_{barr}(Io) = 0.3eV$ for the interstitial to migrate out of oxide region. The formation energy for the bulk oxide is $E_{form}(b) = 6eV$ and in the grain boundary region, the formation energy is $E_{form}(GB) = 1.4eV$.

regions with a low formation energy. Since the defect generation location is inside the bulk switching layer, the filament formation process is different from the surface generation dominated mechanism discussed in Chapter 4. We discuss two cases: (i) Metal electrode can absorb Oxygen (Oxygen migrates into the metal) and (ii) Metal electrode is inert to Oxygen (Oxygen remains in the oxide). These two cases are important because of the ability of electrodes to absorb oxygen versus being inert to oxygen is important in determining both the type of switching (unipolar and bipolar) and endurance (ability to repeatedly switch between on and off states)

After the vacancy and interstitial are generated in the oxide, they diffuse, guided by the electric field. As a result of the externally applied voltage, electrons can flow in or out of the defect levels if the defects are connected to the electrodes. As a result, a vacancy will be neutralized after it attaches to the electrode [82]. We have assumed that the neutral Vo has 0.6eV higher migration energy than the charged one in the numerical simulations presented in all sections of this chapter.

(i) **Oxygen migrates into the metal:** We present simulation results for different values of bond energy, which we define as *low* and *high* bond energies for reasons that will become clear. When the bond energy is low ($E_{bond} = 0.0eV$), as shown in Fig. 5.7, the filament grows in the lateral direction as a result of generation in the grain boundary region. The extent of the lateral dimensions of the filament can be larger than the diameter of the grain boundary. This is because in the low bond energy case, the interaction between vacancies are weak, so that the

vacancy can easily diffuse out of the grain boundary region and make space for further generation. Recall that the hopping rate depends on the diffusion barrier and bond energy as,



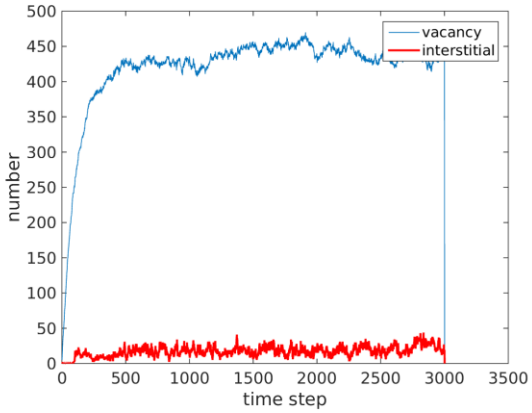


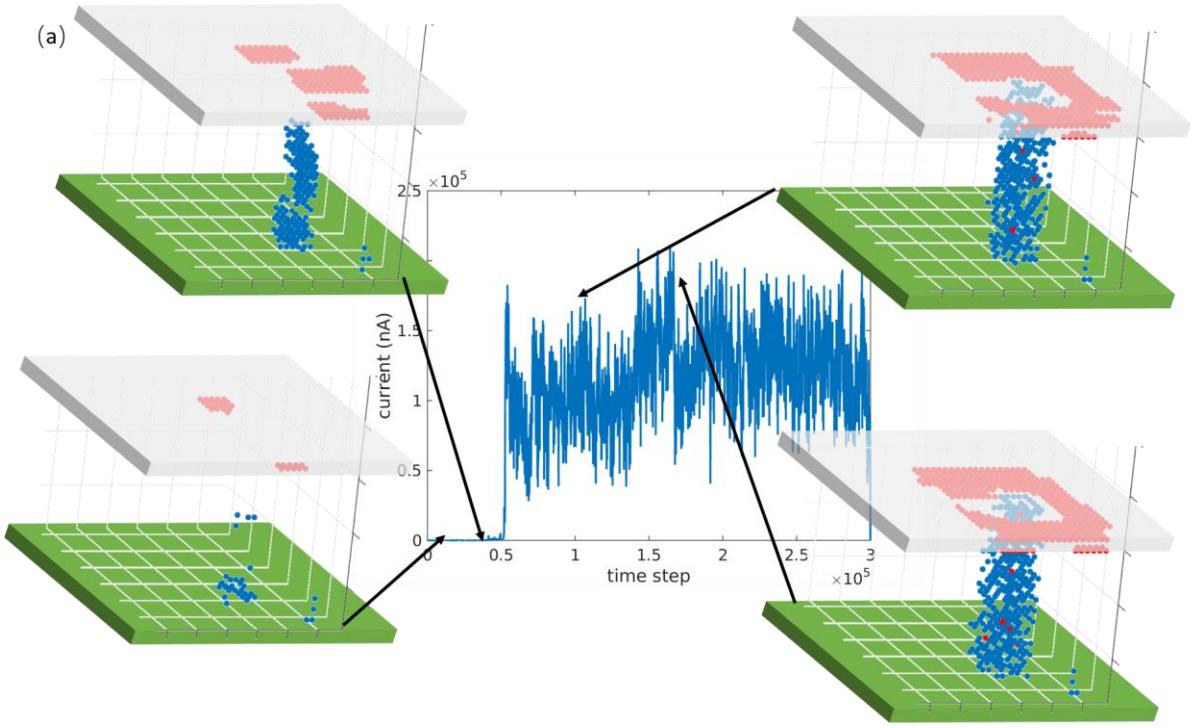
Fig. 5.8 (a) shows the filament formation with bulk generation and (b) shows the number of interstitials and vacancies. The interstitial can migrate out of system and $E_{bond} = 0.1eV$. Parameters: $E_{barr}(V) = 0.6V$ for isolated vacancy, $E_{barr}(Va) = 1.2eV$ for attached vacancy, $E_{barr}(I) = 0.3eV$ for interstitial and $E_{barr}(Io) = 0.3eV$ for the interstitial to migrate out of oxide region. The formation energy for the bulk oxide is $E_{form}(b) = 6eV$ and in the grain boundary region, the formation energy is $E_{form}(GB) = 1.4eV$.

$$R_{Diff} = R_0 \exp\left(-\frac{Ebarr - \Delta n\, E_{bond}}{kT}\right) \quad (5.2)$$

It is also interesting to contrast the shape of the filament here to the surface generation case where the filament got thickened at the tip of filament (Fig. 4.18) as opposed to the cylindrical thickening in Fig 5.7, $\Delta n = n_f - n_i$, where n_f and n_i are the final (after diffusion) and initial number (before diffusion) of vacancies in the vacancy-cluster.

On the other hand, when the bond energy between vacancies is *high (*Fig. 5.8*)*, the lateral growth of the filament is slow, because the vacancy needs to pay a larger energy penalty to diffuse out as given by the equation (5.2). Then, the sites with low formation energy are occupied by vacancies and additional generation of interstitial-vacancy pair is prohibited. Moreover, we can observe that the filament with a *high* bond energy is denser than the one

with a *low* bond energy. The reason is that the bond energy helps the vacancy to stay together and cluster. This also leads to a more stable current, as shown in Fig. 5.8.



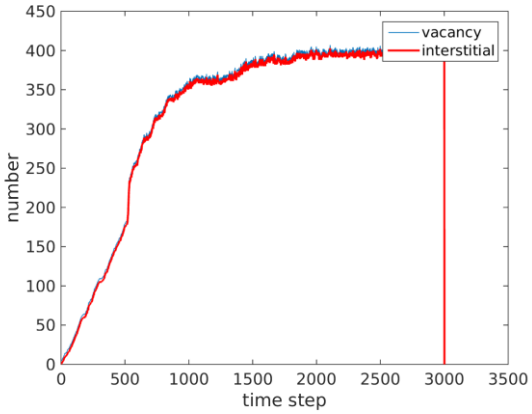


Fig. 5.9 (a) shows the simulation of filament formation with low migration barrier for interstitial within the oxide, and interstitial cannot migrate out of the system into the metal and (b) shows the number of interstitials and vacancies. The current compliance of 300µA is not reached in the simulation. Parameters: with $E_{bond} = 0.1eV$, $E_{barr}(V) = 0.6eV$ for isolated vacancy, $E_{barr}(Va) = 1.2eV$ for attached vacancy, $E_{barr}(I) = 0.3eV$ for interstitial. The formation energy for the bulk oxide is $E_{form}(b) = 6eV$ and in the grain boundary region, the formation energy is $E_{form}(GB) = 1.4eV$.

(ii) **Oxygen remains in the oxide:** Filament formation is determined by the (a) vacancy generation rate and (b) vacancy hopping rate. However, a vacancy cannot be generated in an interstitial rich region because a newly generated vacancy will immediately recombine. Vacancy diffusion into an interstitial rich region will also lead to a loss/recombination of the vacancy. Interstitial rich regions are formed when interstitial diffusion process is slow. Interstitial diffusion is affected by both the migration barrier and the interstitial bond energy as per equation (5.2).

So, (a) the interstitial cannot move out radially and (b) additional vacancy generation is not possible because a vacancy cannot be generated on top of the existing partial filament. If the interstitial diffusion in the oxide is slow, the interstitial cannot diffuse out laterally. This leads to interstitial rich regions at the tip of an existing filament even though some amount of diffusion

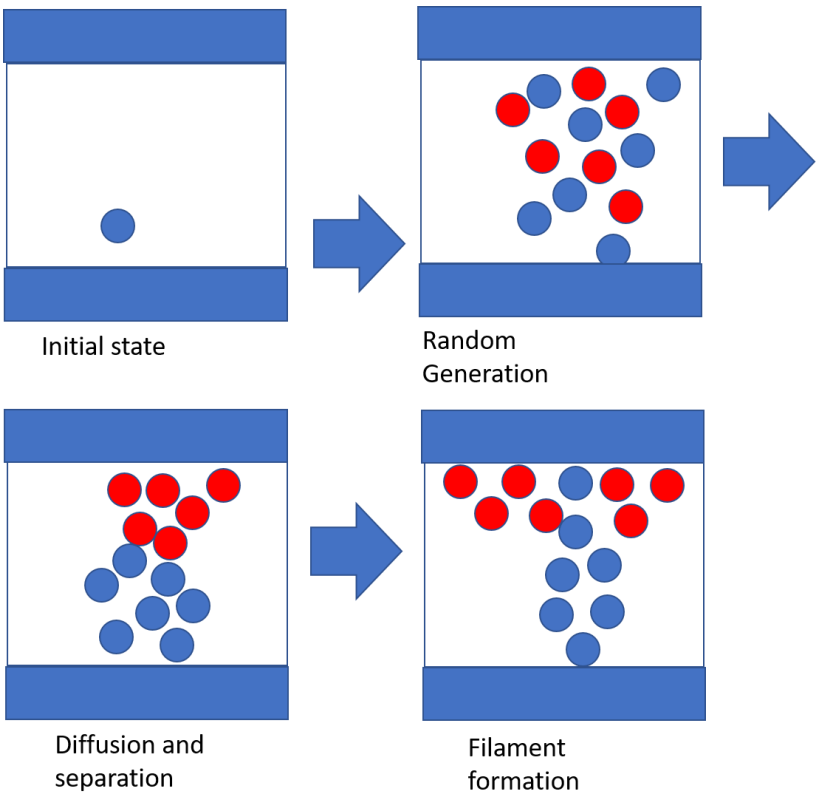


Fig. 5.10 Pattern of filament formation with bulk generation when the interstitial cannot migrate out.

along the filament is possible because of the applied electric field. See the third box titled "Diffusion and separation" in the Fig. 5.10

When the interstitial cannot migrate into the metal, the diffusivity of interstitials has a significant impact on the growth of the filament. When interstitial hopping rate is large (small migration barrier and bond energy for interstitials), the current through the memory cell cannot rise to the current compliance level as seen in Fig. 5.8. The formation process is illustrated as in Fig 5.10. The vacancy and interstitial are generated in the grain boundary region, and they are diffusing in different direction under the same electric field. The vacancy and interstitial will be separated in the oxide. Because of the constantly applied voltage, the generation continues and filament forms with the lateral diffusion of interstitial.

We can tune the migration barrier for the interstitial to a lower value so that it helps diffusion. For a 0.3eV migration barrier for interstitials and a bond energy of 0.1eV, the switching process as a function of time is shown in Fig. 5.9. We observe that the current rises in a short time but does not reach current compliance. If the hopping rate of interstitials is high, the interstitials can diffuse back to the vacancy region and recombine with the vacancy. Therefore, the current can hardly rise to a large value. In other words, the fast diffusion of interstitial stops the growth of filament and limits the current that can run through a single filament. This is especially true for a memory cell with a small lateral dimension such as in our simulation.

To sum up, under the condition of positive thermophoresis coefficient, the vacancy tend to form a sparse structure because of the lateral temperature gradient. As a result, the current through the filament can hardly reach the compliance, and the resistance is limited to a high value.

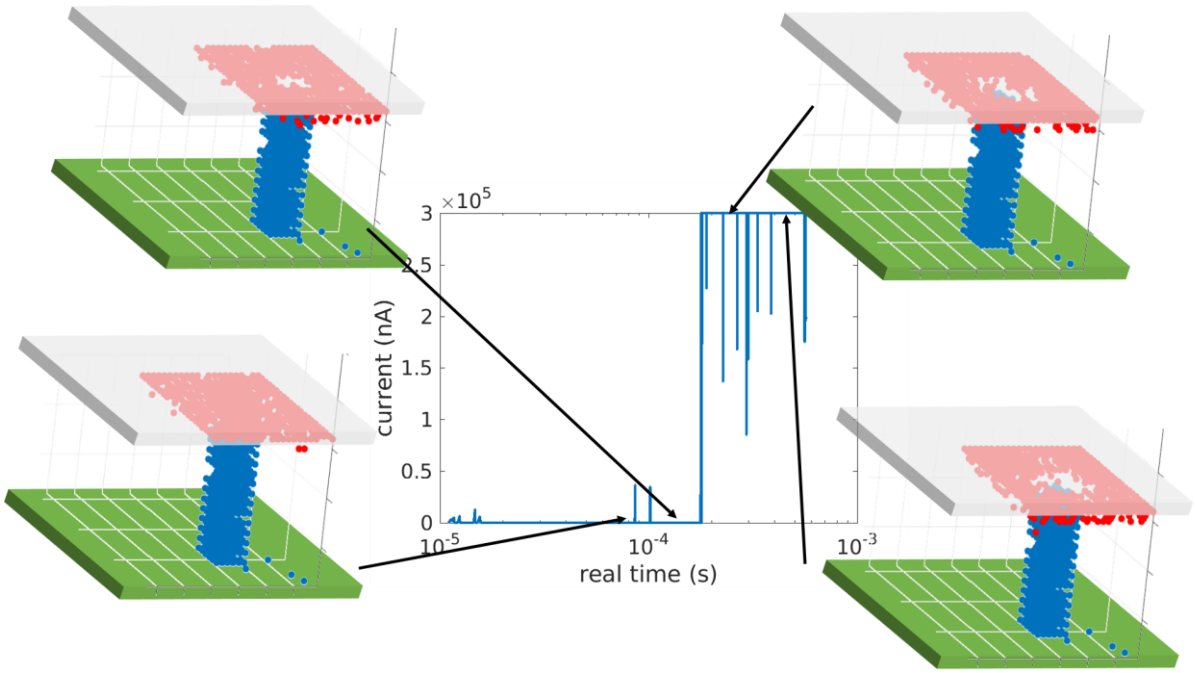
5.2.1 Results with Negative thermophoresis

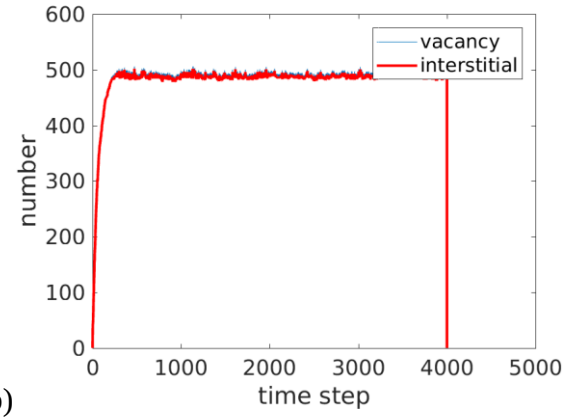
Experiments show that the diffusion of oxygen vacancy in Hafnia has a negative thermophoresis coefficient because the diffusion of vacancy arises from the diffusion of lattice oxygen in the oxide [92, 93]. In Fig. 5.9, we saw that it was not possible to achieve current compliance. Here, we show that with negative thermophoresis, current compliance can be reached.

A typical switching current versus time is shown in Fig. 5.11, when a voltage of 4V is applied on the 8nm thick oxide. The central 9 nm^2 region has a formation energy of $E_{form} = 1.4eV$ while the formation energy is 6eV outside. As mentioned before, the lower formation energy may be due to grain boundary [68] or mechanical stress [67]. When there is a region of low formation energy enclosing a region with a higher formation energy, thermodynamics requires there to be a higher diffusion barrier for vacancies and/or interstitials to move from the region of low to high formation energy. In the simulations presented in this chapter, we have neglected this higher diffusion barrier and instead assume a uniform diffusion barrier throughout the simulation region. The results of the simulation show that the vacancies are clustered in the low formation energy region. A more accurate picture that accounts for the higher diffusion barrier across the regions with two different formation energies will only strengthen the picture of vacancies clustering in the low formation energy region and will not change our conclusions on switching if the barrier for interstitial to escape is small and thus the barrier for vacancy to escape is large. References [92] and [98] report experimentally measuring filaments of comparable cross-sectional area. The formation energy in the rest of the oxide is 6eV. The thermophoresis coefficient for interstitials is positive and for vacancy is negative, as discussed before. The current runs through the memory cell for 2ms after the voltage is turned on. When the current compliance of 300μA is reached, the current ceases to rise, and the voltage across the memory cell decreases. The Vo-Io pairs are generated in the region

with low formation energy first, as in Fig. 5.11 (a). Then, the vacancies and interstitials diffuse guided by the electric field respectively, which makes space for further generation, as in Fig 5.11 (b). Since the electrode is inert to oxygen, the interstitials cannot leave the oxide to enter the electrode. Then, an interstitial layer is formed near the electrode, as in Fig 5.11 (c). Because a large voltage is constantly applied to the memory cell before the current reaches compliance, the bonds between interstitials can break, and the vacancies can align to form the filament, as in Fig 5.11 (d). The resistance drops after the filament has bridged two electrodes. The migration barrier for neutral interstitials (value is 1.2 eV) helps to prevent interstitial diffusion back to the filament region and therefore a long retention time is observed after resistive switching. Note that in contrast to the case with positive thermophoresis (Fig. 5.9), in this case, a filament that achieves current compliance is possible.

Under the condition of negative thermophoresis, the temperature gradient helps the vacancy to diffuse inward. As a result, a solid filament can form in the system as shown in Fig 5.11. Comparing with the case of positive thermophoresis, as shown in Fig 5.9, the density of vacancy in the filament region is much higher. The solid filament can carry a large current through it and therefore the current can reach compliance.



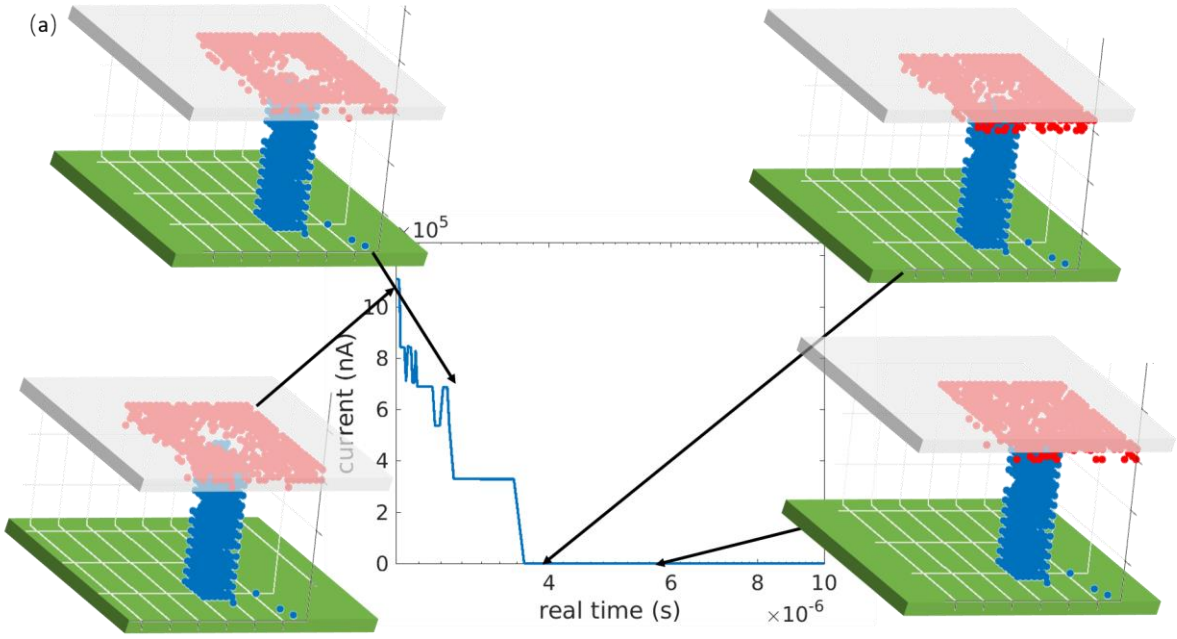


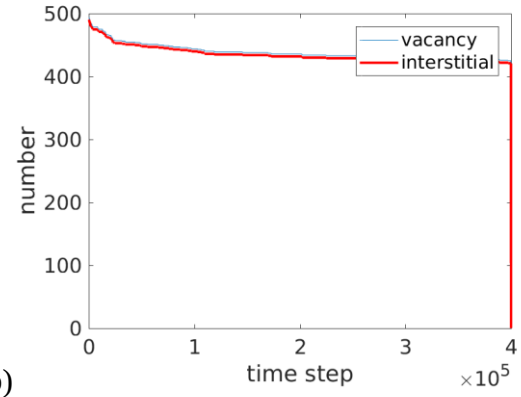
(b)

Fig. 5.11 (a) Filament formation with negative thermophoresis coefficient and (b) shows the number of interstitials and vacancies. Parameters: $E_{bond} = 0.1eV$, $E_{barr}(V) = 0.5eV$ for isolated vacancy, $E_{barr}(Va) = 1.2eV$ for attached vacancy, $E_{barr}(I) = 0.3eV$ for interstitial. The formation energy for the bulk oxide is $E_{form}(b) = 6eV$ and in the grain boundary region, the formation energy is $E_{form}(GB) = 1.4eV$. Current compliance is 300µA for this simulation.

5.2.2 RESET Process

The process of reset is not well understood – with differing claims on where the filament breaks. Using experimental techniques, references [72] claimed that in the bipolar reset, the filament breaks at the top end, which is near the top electrode. A recent experimental work [99] claims that in the bipolar reset, the filament breaks close to the bottom end of the tunneling region.





(b)

Fig. 5.12 (a) shows the process of RESET of filament cell and (b) shows the number of interstitials and vacancies. Parameters: $E_{bond} = 0.1eV$, $E_{barr}(V) = 0.5eV$ for isolated vacancy, $E_{barr}(Va) = 1.2eV$ for attached vacancy, $E_{barr}(I) = 0.3eV$ for interstitial. The formation energy for the bulk oxide is $E_{form}(b) = 6eV$ and in the grain boundary region, the formation energy is $E_{form}(GB) = 1.4eV$.

An important behavior of unipolar RRAM is that it is reset after applying a large current with the same polarity as in the set process. Usually, unipolar reset is observed when the electrode is not reactive with oxygen and there is a large power dissipation. The large power dissipation can result in a high temperature in the filament, which helps to increase the hopping rates of defects. Also, the power dissipated on the conductive path builds a temperature gradient in the lateral direction. The filament breakage was modeled with positive thermophoresis coefficient. When the vacancy has a positive thermophoresis coefficient, as claimed in [91], they can diffuse in the lateral direction along the temperature gradient. The lateral diffusion results in the vacancy disconnection, and the resistance can switch to a high value.

Recently, experiments show that the oxygen vacancy in HfO_2 has a negative thermophoresis coefficient, which means that the vacancy will diffuse to the region with high temperature. Then, the dissolution of the filament due to a high temperature cannot occur. When a large current flow in the filament, the central region has the highest temperature, and vacancies diffuse towards the center, making the filament stronger (lower resistance), rather than weaker. As shown in the Fig. 5.11, with the negative thermophoresis coefficient, the vacancies tend to diffuse inward to the central region, where the temperature is the highest. Then the filament can hardly break, since the large temperature gradient prevents the RESET process from happening. Thus, another physical mechanism is necessary for unipolar RESET in devices with a negative thermophoresis coefficient for vacancies.

With the negative thermophoresis coefficient, the RESET can occur if the interstitials diffuse due to the high temperature and recombine with vacancies in the filament. When the current is high during unipolar reset, the temperature of the memory cell is over 700K because of Joule

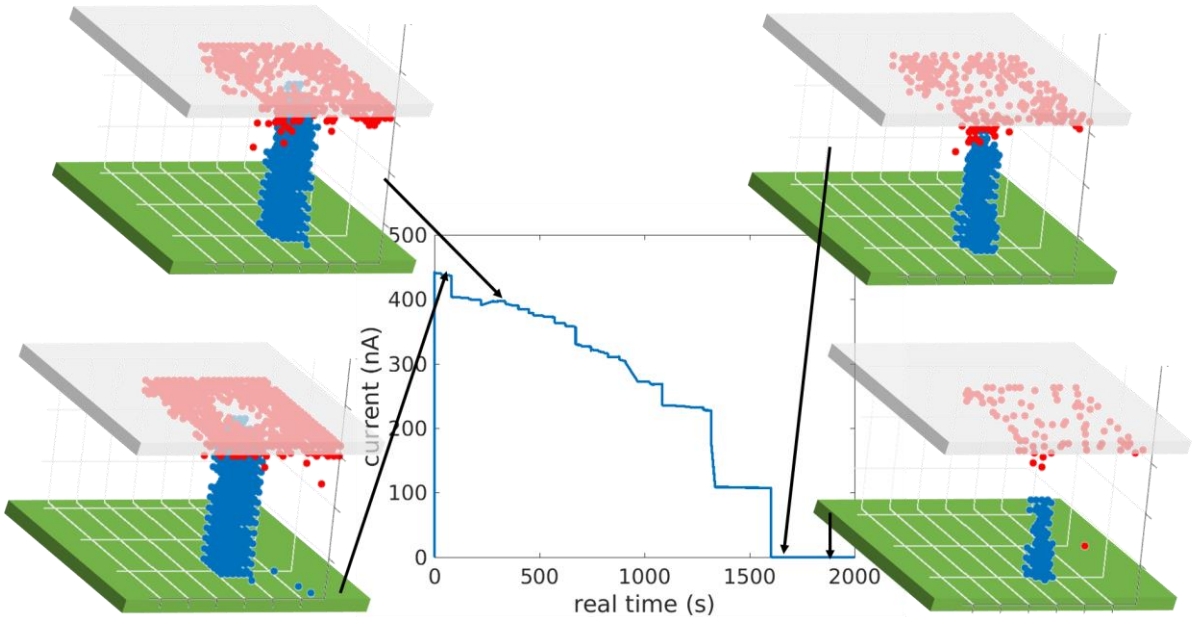
heating. The high temperature activates the diffusion of vacancies and interstitials, which makes the RESET process fast enough to match the data in the experiment. If the Joule heating is turned off, diffusion of a vacancy or interstitial will take more than 10^3s, which is too long when compared to experimental data. During reset, we find that the interstitials, which have accumulated near the electrode diffuse actively. The interstitials diffuse back towards the filament because of the concentration gradient, and they recombine with the vacancies. Then, an insulating gap is formed at the tip of the filament (Fig. 5.12). This figure clearly shows that we find that the reset process does not follow the hourglass model in unipolar devices. Breakage occurs close to the top electrode. This should be experimentally verifiable using techniques such as the novel method in reference [99].

We now discuss the importance of crystal field enhancement in the process of reset. If the electric field used in the generation rate lacks crystal field enhancement, the timescales of reset and retention cannot match experiments [95]. When the normal electric field is applied to determine the generation and hopping rates, we see that the memory cell is switched to LRS in a short time. However, we cannot observe reset of the memory cell without the crystal field enhancement. In reset also, the electric field accelerates the generation process, as in the forming process. The applied voltage difference in the forming, set and reset processes results in different electric field strengths playing a role in determining the generation rate. Without the crystal field enhancement, the electric field difference is approximately $\Delta \mathrm{E} = \Delta \mathrm{V}/\mathrm{L}$ (L is the thickness of oxide and $\Delta \mathrm{V}$ is the difference between voltages used for set and reset); then, the time scale difference can be estimated to be $\exp(\mathrm{qd}\,\Delta\,\mathrm{V}/2\mathrm{kTL}) \approx 12.2$ ($q = 2e, d = 0.5nm, \Delta V = 1V$). In other words, when the reset voltage is applied, it takes only 12.2 times longer for a Vo-Io generation compared to the set process. As a result, the cell should be in a low resistance state

immediately after reset; that is, reset is not possible. But if we take the crystal field enhancement into consideration, the electric field difference becomes $(\varepsilon_r + 2)/3 \ \Delta V/L \approx 9 \ \Delta V/L$. Then, the time scale difference is increased to $\exp(9qd \Delta V/2kTL) \approx 2.94 \times 10^9$, which effectively prevents generation under a 1V applied voltage. Therefore, the memory cell could be reset in a unipolar fashion.

5.2.3 Retention of memory cell

For both the interstitial and vacancy, the migration barrier depends on the charge state. For vacancy, the charge is assumed to be zero ($+2e$) if it is (is not) connected to the bottom electrode. For the interstitial, the charge depends on the applied voltage. If the voltage is zero (or small) the charge is assumed to be zero (neutral), and if the voltage is large such as in set or reset, the charge has been assumed to be $+2$ and -2 for vacancy and interstitial respectively. Note that with voltages corresponding to set and reset, the Io and Vo is always assumed to be charged in this thesis. We will see in this section that the charge of the interstitial is important to include correctly for a long retention time.



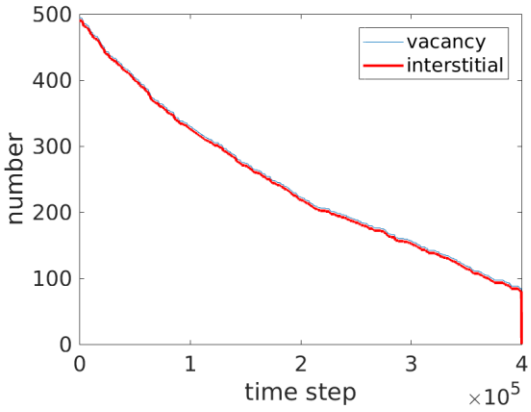


Fig. 5.13 (a) shows the retention of filament cell and (b) shows the number of interstitials and vacancies. Parameters: $E_{bond} = 0.1eV$, $E_{barr}(V) = 0.5eV$ for isolated vacancy, $E_{barr}(Va) = 1.2eV$ for attached vacancy, $E_{barr}(I) = 0.3eV$ for interstitial. The formation energy for the bulk oxide is $E_{form}(b) = 6eV$ and in the grain boundary region, the formation energy is $E_{form}(GB) = 1.4eV$

A good memory cell can retain information for an extremely long time (10+ years). The simulation of retention is shown in Fig. 5.13. The change in charge state with applied voltage should be noticed in the retention process. In the retention process, the diffusion of defects is not affected by the electric field because there are no applied biases. The time ratio of the retention and set times ($t_{retention}/t_{set}$) can be roughly estimated to be $\exp\left(-\frac{\mathrm{qEd}}{2\mathrm{kT}}\right)$. Taking the following

parameters for HfO_2, $q = 2e, d = 0.5nm$ and electric field $E \approx \frac{2V}{8nm} = 0.25V/nm$ the retention time is only about 148 times larger than the set time. However, experiments show a $t_{retention} > 10^5 s$ and a $t_{set} < 10^{-5} s$. We believe that besides the rate difference induced by electric field, the large difference in timescales between $t_{retention}$ and t_{set} is due to the large difference in

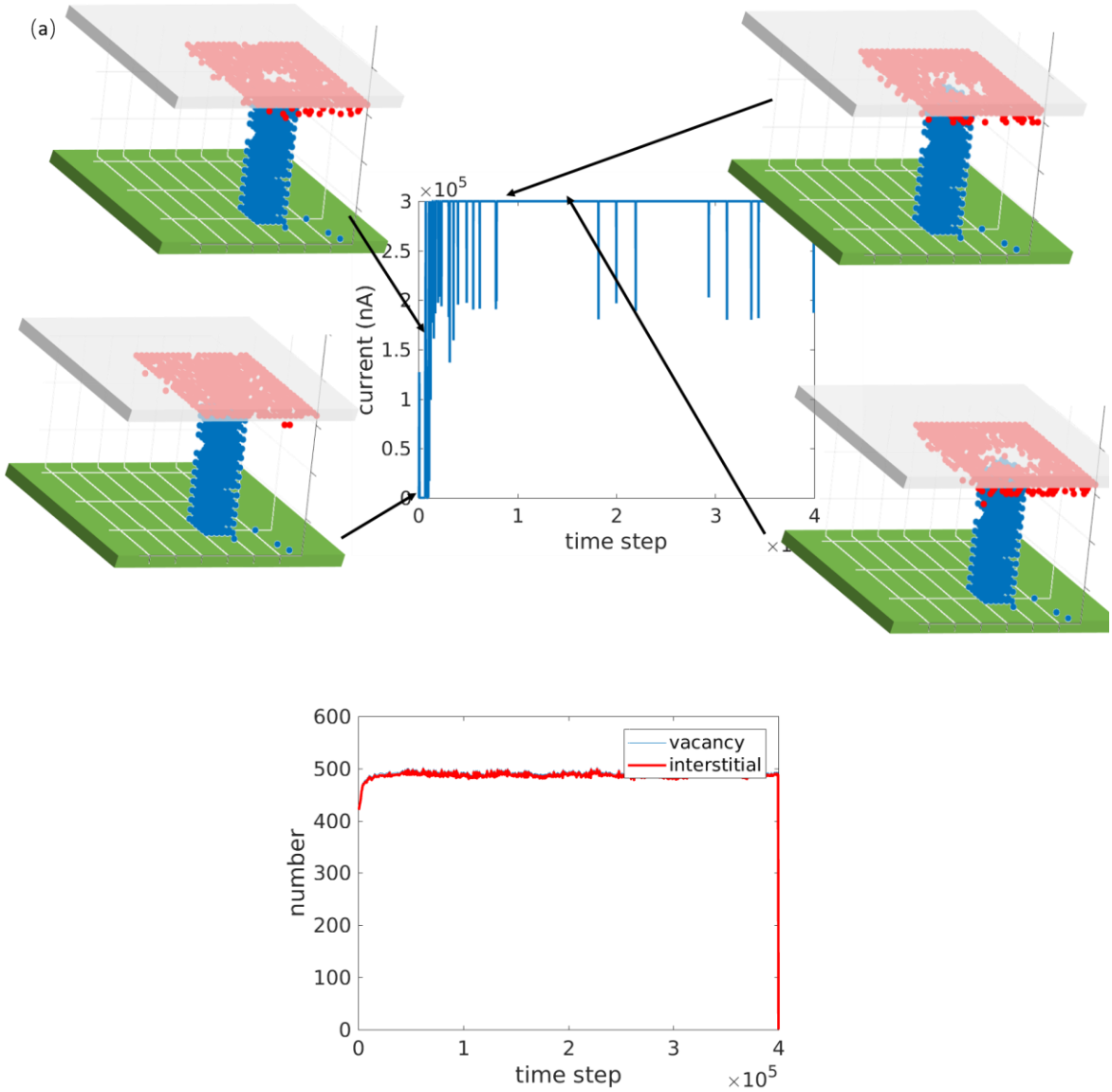


Fig. 5.14 (a) shows the second set of the filament. Figure (a) shows the filament after the first RESET and this is the initial condition and (b) shows the number of interstitials and vacancies. A second SET voltage pulse is applied, and the filament evolution is show in Figures (b) and (c). Figures (d), (e) and (f), where the filament is easier to visualize show the side top view of (a), (b) and (c) respectively. Parameters: $E_{bond} = 0.1eV$, $E_{barr}(V) = 0.5eV$ for isolated vacancy, $E_{barr}(Va) = 1.2eV$ for attached vacancy, $E_{barr}(I) = 0.3eV$ for interstitial. The formation energy for the bulk oxide is $E_{form}(b) = 6eV$ and in the grain boundary region, the formation energy is $E_{form}(GB) = 1.4eV$.

migration barrier of the interstitial with and without an applied bias. Since there is no voltage applied in retention, the interstitials are likely to be charge neutral, and therefore have a higher migration barrier as shown in [82]. A larger migration barrier prevents the diffusion of interstitials in the retention process, which helps to maintain the filament. For a migration barrier of 1.0 eV in the neutral interstitial, the memory state can be retained for more than 10^3 s in our simulation.

After the memory cell is reset to HRS, a second set could be performed. After a high voltage is applied to set the filament again, vacancy-interstitial pairs can be generated. Fig. 5.14 (a) shows the result when the initial filament configuration is the one from Fig. 5.12 (f). Now, applying a set voltage of 2 V causes generation of vacancy-interstitial pairs, which shows a growth of the filament (blue) in Fig. 5.14 (b) to (c). One can see that as the filament forms again, the interstitials move away laterally as shown by the arrow in Fig. 5.14 (f).

We have developed a 3D model to simulate filament formation and reset processes in HfO_2-based resistive memory devices with a negative Soret coefficient. The Frenkel pair consisting of oxygen vacancy and interstitial is assumed to be formed in regions such as the grain boundary of the oxide, where the formation energy for Frenkel pair is lower than the defect free bulk region. The applied voltage separates the vacancy and interstitial and guides the vacancy and interstitial to the top and bottom electrodes respectively. The vacancies may neutralize after it is attached to the bottom electrode. As a result, a solid filament can form in the system due to the high migration barrier for neutral vacancies. As the filament grows, when it bridges the two electrodes, the resistance of the memory cell drops. The negative Soret coefficient means that the vacancies tend to stay near the region with a high temperature, which makes the vacancies stay tightly in the filament region. The unipolar RESET found in experiment is analyzed in the case with a negative Soret coefficient for vacancies. With a negative Soret coefficient, the vacancies are likely to stay

in the high temperature region, which makes the heat induced dissolution of filament impossible. A novel mechanism involving the diffusion of interstitials activated by high temperature and voltage, can help the reset occur in a short time, yielding a unipolar RESET. The vacancies are neutralized if the applied voltage is turned off. However, because of the existence of interstitials in the system, the retention time of memory cell is shorter than the bipolar case.

Chapter 6. FUTURE DEVELOPMENT AND CONCLUSION

In chapter 3, the current through conductive filament was given under the steady-state condition. In chapter 4 and chapter 5, the formation and rupture of the filament are simulated with surface and bulk generation by the kinetic Monte Carlo method. This chapter will discuss possible developments of RRAM in the future and also give a brief conclusion.

6.1 MORE CALCULATIONS ON FILAMENT CONFIGURATION

In chapter 5, we have discussed the sneak path problem in RRAM array. An external diode and the unipolar RESET can solve the sneak path problem, but unipolar RESET relies on the mechanism of Joule heating. A better solution is to embed the diode into the memory cell, which gives a self-rectifying current-voltage characteristic. If a memory cell has a rectifying current-voltage characteristic, the sneak path problem will be solved since little current can pass through the memory cell with a reverse bias. Also, the rectifying current- voltage characteristics can reduce energy consumption in the RESET process of the memory cell because when the voltage bias is reversed, the current through the memory cell is small, and therefore no significant heat will be dissipated.

A possible approach to implement the rectifying characteristic is to modulate the conductive filament shape by specific preparation of the switching layer [92] or by controlling the electrical pulse in the SET process [66]. For example, the pyramid-shaped filament with a small gap to the electrode, as shown in Fig. 6.1 (a), is a possible configuration for ReRAM rectifying characteristics. Comparing to other configurations with the same number of Cu defects and a large conductance, such as the linear configuration, as shown in Fig 3.3 and zigzag configuration, as shown in Fig. 6.1 (b), the number of Cu-Cu bonds is large. Because of the exchange interaction

between electrons near the defects, the Cu defects will attract each other. Thus, the large number of Cu-Cu bonds can reduce the total formation energy, and the pyramid configuration is easier to form spontaneously.

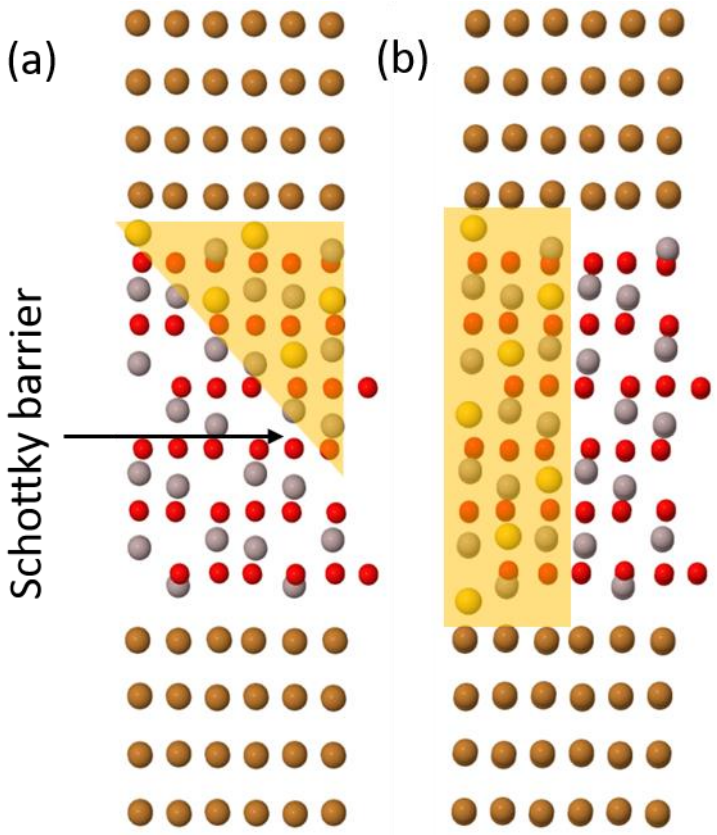


Fig. 6.1. Metal-Insulator-Metal (MIM) simulation models of pure alumina with different Cu filaments embedded. (a) contains pyramid Cu filament and (b) contains zigzag filament. Region with Cu atoms in alumina is highlighted in yellow.

Next, we can calculate the current-voltage characteristics of these configurations using Green's function formalism. Because the pyramid configuration is not symmetric, I expect that the current-voltage characteristics will also not be symmetric. The conductive but diode-like behavior can be understood by a small band formed by defects. Because the Cu ions are so dense in the switching layer, the defect states induced by Cu can form a small band. Then, the region doped by Cu behaves as a semiconductor rather than a large bandgap insulator. The Schottky junction will be formed between the inert electrode and the doped region, which results in the rectifying current-voltage

characteristics (Fig. 6.1). The rectifying current-voltage characteristics will help solve the sneak path problem in the reading operation of the RRAM array.

The pyramid structure also helps the RESET process. Because the Schottky barrier between the inert electrode and doped region forbids large current, the voltage is mainly dropped on the tip of the pyramid. Because the Schottky barrier is reverse-biased, the current through the memory cell can be low in the RESET process. The energy is mainly used to move the ion rather than dissipate energy via Joule heating.

6.2 MICROSCOPIC HEATING MECHANISM IN CONDUCTION

In the RESET process of RRAM operation, the heat generated by current flow plays an important role. Experiments show that the temperature of the filament in the RESET process can reach 900K [4]. The high temperature can increase the probability of ion hopping. An accurate calculation of the temperature (or more accurately, the diffusion and generation rates) is necessary. In chapter 3, the current is calculated using the NEGF formalism without considering the heating effect, and in chapters 4 and 5, the current is calculated using classical Ohm's law, and Joule heating is assumed. Future work to model energy dissipation that goes beyond Joule heating is necessary because of the nanoscale size of the filament, which is comparable to both the mean free path for the electron and mean free path of the phonon. The physics that we have failed to capture in this thesis is that electron relaxation and energy dissipation, which will more accurately reflect the probability for diffusion and generation. One approach that can go beyond the Joule heating model is the Buttiker probe model. In this model, the two-terminal device is generalized to a n-terminal device. The extra terminals are introduced to represent the energy exchange between electrons and phonons. These interactions also result in decoherence in electron transport. The current between two terminals can be calculated in the Green's function formalism as

$$I_{ij} = \int d\epsilon \left[f(\epsilon - \mu_i) - f\left(\epsilon - \mu_j\right) \right] T_{ij}(\epsilon) \, .$$

As the Buttiker probes are fictitious, the total electrical current flowing in them should be zero. That is, for all Buttiker probes i, the net current should be zero as given by:

$$\sum_j I_{ij} = 0$$

Meanwhile, we can calculate the power delivered into probe i using the following expression,

$$P_i = \sum_j \int d\epsilon \, \epsilon \left[f(\epsilon - \mu_i) - f\left(\epsilon - \mu_j\right) \right] T_{ij}(\epsilon).$$

The rate equation for the local phonon number is given by [100]

$$\dot{n} = \frac{P_i}{\hbar\omega} - \gamma_i \, (n - n_0).$$

Here n is the number of phonons, ω is the frequency of the phonon, γ_i is the relaxation rate of the phonon, and the n_0 is the number of phonons at equilibrium. The energy loss of electron can result in an increase in the number of phonons, and the time-dependent phonon number can be solved numerically.

6.3 TIME DEPENDENT CALCULATION OF ELECTRON TRANSPORT

In chapter 3, the assumptions made in calculating the current through atomic-scale filaments are: 1) the Hamiltonian is varying slowly; 2) we only looked at the electronic state of the system after steady state has been reached. This method was used successfully to interpret the mechanism of current flow. However, it cannot handle a rapid change in the applied voltage (i.e. not steady state) or significant levels of decoherence. In order to understand the detailed interaction between the electron and atom, the Buttiker probe method, which assumes equilibrium between the phonon reservoir and electron system, is not enough.. A more precise time-dependent method must be used

to calculate current flow [101, 102]. The derivation of the time-dependent calculation is briefly described in Appendix C . In general, the calculation needs the time evolution of the density matrix to calculate the instantaneous current.

We can use a simple example to validate the time-dependent code. In this example, the device contains 8 sites, whose on-site energy is $h_{nn} = 0\, eV$. The hopping can happen only between the nearest neighbor sites; the off-diagonal term in the Hamiltonian is $t = 50\, meV$. The broadening function Γ is assumed to be 1 eV. The current at 200 fs approaches the steady state as shown in Fig. 6.2.

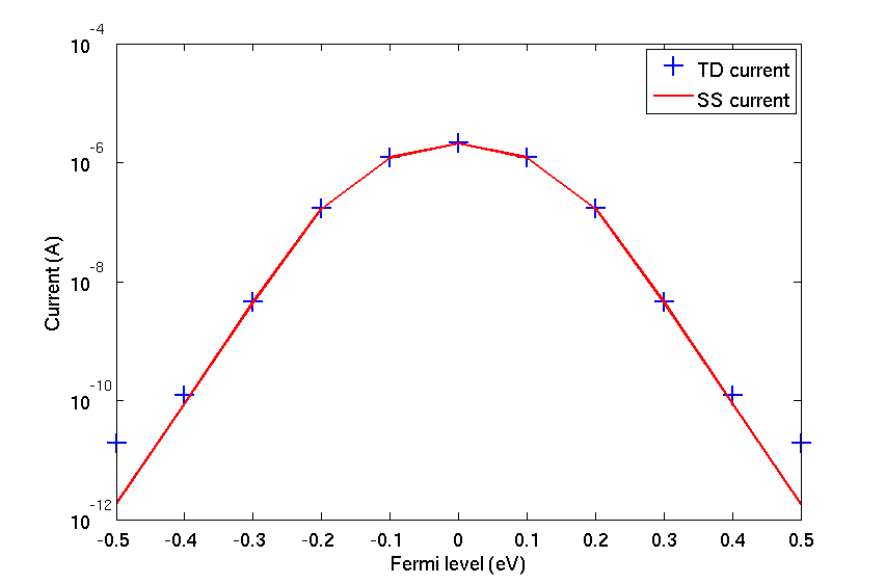


Fig. 6.2 The time-dependent current at 200 fs (blue) and steady state current for 50 meV hopping energy.

In this thesis, we have studied the properties of conductive filament in the RRAM. In chapter 3, the Green's function in the frequency domain is used to calculate the transmission of RRAM, and then give the steady-state current The current-voltage characteristics are investigated, and the mechanism for electron transmission is analyzed using the local density of states. In chapter 4, the form, SET and RESET processes of filaments are studied using the kinetic Monte Carlo method.

Under conditions when one electrode of RRAM is made of an oxygen reactive metal, the surface generation process will dominate the vacancy generation. The generated vacancy can diffuse along the electric field and form the conductive filament. The oxygen generated can partially oxidize the metal electrode and form an insulating layer. This layer helps to enhance the electric field near the filament tip and therefore accelerate the filament forming process. As a result, the resistance in LRS can be controlled by current compliance, which makes the multi-level memory cell feasible. In chapter 5, the bulk generation process in the RRAM is studied. Under conditions when the insulating layer is sandwiched by two oxygen inert metals, bulk generation is likely to dominate the generation process. The vacancies are generated in the grain boundary region and the crystal field enhancement helps the Frenkel pair generation process. The diffusion of vacancy is studied when the oxide has a negative thermophoresis coefficient. I find that the RRAM can be RESET in the unipolar manner, which helps to solve the sneak path problem in the RRAM array. This thesis focused on the underlying mechanisms for current flow, SET, RESET and retention of prototype RRAM devices.

REFERENCES

[1] J. Hutchby and M. Garner, "Assessment of the Potential Maturity of Selected Emerging Research," in *Memory Technologies Workshop ERD/ERM Working Group Meeting*, 2010.

[2] H. Akinaga and H. Shima, "Resistive Random Access Memory (ReRAM) Based on Metal Oxides," *Proceedings of the IEEE,* vol. 98, no. 12, p. 2237, 2010.

[3] H.-S. P. Wong, H.-Y. Lee, S. Yu, Y.-S. Chen, Y. Wu and P.-S. Chen, "Metal–Oxide RRAM," *Proceedings of the IEEE,* vol. 100, p. 1951, 2012.

[4] A. Sawa, "Resistive switching in transition metal oxides," *Materials Today,* p. 28, 2008.

[5] A. Chen, S. Haddad, Y. C. Wu, Z. Lan, T. N. Fang and S. Kaza, "Switching characteristics of Cu2O metal-insulator-metal resistive memory," *Applied Physics Letters,* vol. 91, p. 123517, 2007.

[6] D. C. Kim, S. Seo, S. E. Ahn, D.-S. Suh, M. J. Lee and e. al, "Electrical observations of filamentary conductions for the resistive memory switching in NiO films," *Appl. Phys. Lett.*, vol. 88, p. 202102, 2006.

[7] X. Guo, C. Schindler, S. Menzel and R. Waser, "Understanding the switching-off mechanism in Ag+ migration based resistively switching model systems," *Appl. Phys. Lett.*, vol. 91, p. 133513, 2007.

[8] V. V. Zhirnov, R. K. I. Cavin, S. Menzel, E. Linn, S. Schmelzer, D. Brauhaus, C. Schindler and R. Waser, "Memory Devices: Energy–Space–Time Tradeoffs," *Proceedings of the IEEE,* vol. 98, no. 12, p. 2185, 2010.

[9] R. Waser, R. Dittmann, G. Staikov and K. Szot, "Redox-Based Resistive Switching Memories – Nanoionic Mechanisms, Prospects, and Challenges," *Advanced Materials,* vol. 21, p. 2632, 2009.

[10] D. B. Strukov, G. S. Snider, D. R. Stewart and R. S. Williams, "The missing memristor found," *nature,* vol. 453, p. 80, 2008.

[11] Z. Wang, H. Yu, X. A. Tran, Z. Fang, J. Wang and H. Su, "Transport properties of HfO2-x based resistive-switching memories," *Phys. Rev. B,* vol. 85, p. 195322, 2012.

[12] H. D. Lee, B. Magyari-Kope and Y. Nishi, "Model of metallic filament formation and rupture in NiO for unipolar switching," *Phys. Rev. B,* vol. 81, p. 193202, 2010.

[13] J. H. Hur, M.-J. Lee, C. B. Lee, Y.-B. Kim and C.-J. Kim, "Modeling for bipolar resistive memory switching in transition-metal oxides," *Phys. Rev. B,* vol. 82, p. 155321, 2010.

[14] S.-G. Park, B. Magyari-Köpe and Y. Nishi, "Electronic correlation effects in reduced rutile TiO2 within the LDA+U method," *Phys. Rev. B,* no. 82, p. 115109, 2010.

[15] S. Tsui, Y. Y. Xue, N. Das, Y. Q. Wang and C. W. Chu, "Interfacial resistive oxide switch induced by reversible modification of defect structures," *Phys. Rev. B,* vol. 80, p. 165415, 2009.

[16] S. Yu, R. Jeyasingh, Y. Wu and H.-S. P. Wong, "Characterization of low-frequency noise in the resistive switching of transition metal oxide HfO2," *Phys. Rev. B,* vol. 85, p. 045324, 2012.

[17] J. Son and S. Stemmer, "Resistive switching and resonant tunneling in epitaxial perovskite tunnel barriers," *Phys. Rev. B,* vol. 80, p. 035105, 2009.

[18] T. Gu, Z. Wang, T. Tada and S. Watanabe, "First-principles simulations on bulk Ta2O5 and Cu/Ta2O5/Pt heterojunction: Electronic structures and transport properties," *Journal of Applied Physics,* vol. 106, p. 103713, 2009.

[19] H. Akinaga, H. Shima, F. Takano, I. H. Inoue and H. Takagi, "Resistive Switching Effect in Metal/Insulator/Metal Heterostructures and Its Application for Non-volatile Memory," *IEEJ Trans,* vol. 2, p. 453, 2007.

[20] S. Ambrogio, S. Balatti, V. Milo, R. Carboni, Z. Q. Wang, A. Calderoni, N. Ramaswamy and D. Ielmini, "Neuromorphic Learning and Recognition With One-Transistor-One-Resistor Synapses and Bistable Metal Oxide RRAM," *IEEE Transactions on Electron Devices 63,* p. 1508, 2016.

[21] Y. Bai, H. Wu, R. Wu, Y. Zhang, N. Deng, Z. Yu and H. Qian, "Study of Multi-level Characteristics for 3D Vertical Resistive Switching Memory," *Scientific Reports 4,* p. 5780, 2014.

[22] M. Coll, J. Fontcuberta, M. Althammer, M. Bibes, H. Boschker, A. Calleja, G. Cheng, M. Cuoco and e. al, "Towards Oxide Electronics: a Roadmap," *Applied Surface Science,* vol. 482, p. 1, 2019.

[23] D. Berco and T.-Y. Tseng, "A comprehensive study of bipolar operation in resistive switching memory devices," *Journal of Computational Electronics,* vol. 15, no. 2, p. 577, 2016.

[24] T. Hasegawa, K. Terabe, T. Tsuruoka and M. Aono, "Atomic Switch: Atom/Ion Movement Controlled Devices for Beyond Von-Neumann Computers," *Adv. Mater.,* vol. 24, p. 252, 2012.

[25] D. Ielmini, "Modeling the Universal Set/Reset Characteristics of Bipolar RRAM by Field- and Temperature-Driven Filament Growth," *IEEE Transactions on Electron Devices,* p. 4309, 2011.

[26] S. H. Jo, T. Kumar, S. Narayanan and H. Nazarian, "Cross-Point Resistive RAM Based on Field-Assisted Superlinear Threshold Selector," *IEEE Transactions on Electron Devices,* vol. 62, p. 3477, 2015.

[27] M. Lanza, H. P. Wong, E. Pop, D. Ielmini, D. Strukov, B. C. Regan, L. Larcher, M. A. Villena, J. J. Yang, L. Goux and e. al, "Recommended Methods to Study Resistive Switching Devices," *Adv. Electron. Mater.,* vol. 5, p. 1800143, 2019.

[28] S. Kim, J. Zhou and W. D. Lu, "Crossbar RRAM Arrays: Selector Device Requirements During Write Operation," *IEEE Transactions on Electron Devices,* vol. 61, p. 2820, 2014.

[29] D. Panda, P. P. Sahu and T. Y. Tseng, "A Collective Study on Modeling and Simulation of Resistive Random Access Memory," *Nano Research Letter,* vol. 13, p. 8, 2018.

[30] T. Tsuruoka, K. Terabe, T. Hasegawa and M. Aono, "Forming and switching mechanisms of a cation-migration-based oxide resistive memory," *Nanotechnology,,* vol. 21, p. 425205, 2010.

[31] Y. C. Yang, F. Pan, Q. Liu, M. Liu and F. Zeng, "Fully Room-Temperature-Fabricated Nonvolatile Resistive Memory for Ultrafast and High-Density Memory Application," *Nano Lett. 9,* p. 1636, 2009.

[32] 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 and M. Oshima, "Highly reliable TaOx ReRAM and direct evidence of redox reaction mechanism," in *2008 IEEE International Electron Devices Meeting*, San Francisco, CA, 2008.

[33] J. J. Yang, M. D. Pickett, X. Li, D. A. A. Ohlberg, D. R. Stewart and R. S. Williams, "Memristive switching mechanism for metal/oxide/metal nanodevices," *Nature Nanotechnology*, vol. 3, p. 429, 2008.

[34] R. Waser and M. Aono, "Nanoionics-based resistive switching memories," *Nature Mater.*, vol. 6, p. 833, 2007.

[35] R. G. Parr and Y. Weitao, Density-Functional Theory of Atoms and Molecules, New York, NY: Oxford University Press, 1994.

[36] B. Traore, P. Blaise, E. Vianello, E. Jalaguier, G. Molas, J. F. Nodin, L. Perniola, B. D. Salvo and Y. Nishi, "Impact of electrode nature on the filament formation and variability in HfO2RRAM," in *2014 IEEE International Reliability Physics Symposium*, Waikoloa, HI, 2014.

[37] S. R. Bradley, A. L. Shluger and G. Bersuker, "Electron-Injection-Assisted Generation of Oxygen Vacancies in Monoclinic HfO2," *Phys. Rev. Applied,* vol. 4, p. 064008, 2015.

[38] A.-P. Jauho, N. S. Wingreen and Y. Meir, "Time-dependent transport in interacting and noninteracting resonant-tunneling systems," *Phys. Rev. B* , vol. 50, p. 5528 , 1994.

[39] J. Taylor, H. Guo and J. Wang, "Ab initio modeling of quantum transport properties of molecular electronic devices," *Phys. Rev. B,* vol. 63, p. 245407 , 2001.

[40] L. A. Girifalco, Statistical Physics of Materials, New York, NY: John Wiley & Sons Inc., 1973.

[41] D. S. Jeong, H. Schroeder and R. Waser, "Coexistence of Bipolar and Unipolar Resistive Switching Behaviors in a Pt/TiO2/Pt Stack," *Electrochem. Solid State Lett. 10,* p. G51, 2007.

[42] L. Goux, Y.-Y. Chen, L. Pantisano, X.-P. Wang, G. Groeseneken, M. Jurczak and D. J. Wouters, "On the Gradual Unipolar and Bipolar Resistive Switching of TiN\HfO2\Pt Memory Systems," *Electrochem. Solid State Lett. 13,* p. G54, 2010.

[43] L. Goux, K. Opsomer, R. Degraeve, R. Müller, C. Detavernier, D. J. Wouters, M. Jurczak, L. Altimime and J. A. Kittl, "Influence of the Cu-Te composition and microstructure on the resistive switching of Cu-Te/Al2O3/Si cells," *Appl. Phys. Lett.*, vol. 99, p. 053502, 2011.

[44] L. Goux, K. Opsomer, A. Franquet, G. Kar, N. Jossart, O. Richard, D. Wouters, R. Müller, C. Detavernier, M. Jurczak and J. Kittl, "Thermal-stability optimization of Al2O3/Cu–Te based conductive-bridging random access memory systems," *Thin Solid Films,* vol. 533, p. 29, 2013.

[45] Y. Wu, S. Yu, B. Lee and P. Wong, "Low-power TiN/Al2O3/Pt resistive switching device with sub-20 μA switching current and gradual resistance modulation," *Journal of Applied Physics,* vol. 110, p. 094104, 2011.

[46] L. Chen, H.-Y. Gou, Q.-Q. Sun, P. Zhou, H.-L. Lu, P.-F. Wang, S.-J. Ding and e. al, "Enhancement of Resistive Switching Characteristics in Al2O3-Based RRAM With

Embedded Ruthenium Nanocrystals," *IEEE Electron Device Letters,* vol. 32, p. 794, 2011.

[47] A. Belmonte, W. Kim, B. Chan, N. Heylen, A. Fantini, M. Houssa, M. Jurczak and L. Goux, "90nm W\Al2O3\TiW\Cu 1T1R CBRAM cell showing low-power, fast and disturb-free operation," in *2013 5th IEEE International Memory Workshop*, Monterey, CA, 2013.

[48] H. Lin, P. Ye and G. Wilk, "Current-transport properties of atomic-layer-deposited ultrathin Al2O3 on GaAs," *Solid-State Electronics,* vol. 50, p. 1012, 2006.

[49] K. Aratani, K. Ohba, T. Mizuguchi, S. Yasuda, T. Shiimoto, T. Tsushima and T. Sone, "A Novel Resistance Memory with High Scalability and Nanosecond Switching," in *International Electron Devices Meeting*, Washington, DC, 2007.

[50] D. S. Jeong, B.-k. Cheong and H. Kohlstedt, "Pt/Ti/Al2O3/Al tunnel junctions exhibiting electroforming-free bipolar resistive switching behavior," *Solid-State Electronics,* vol. 63, p. 1, 2011.

[51] K. Sankaran, L. Goux, S. Clima, M. Mees, J. A. Kittl, M. Jurczak, L. Altimime, G.-M. Rignanese and G. Pourtois, "Modeling of Copper Diffusion in Amorphous Aluminum Oxide in CBRAM Memory Stack," *ECS Trans.,* vol. 45, no. 3, p. 317, 2012.

[52] W. Zhu, T. P. Chen, Z. Liu, M. Yang, Y. Liu and S. Fung, "Resistive switching in aluminum/anodized aluminum film structure without forming process," *Journal of Applied Physics,* vol. 106, p. 093706, 2009.

[53] W. Banerjee, S. Maikap, S. Z. Rahaman, A. Prakash, T.-C. Tien, W.-C. Li and J.-R. Yang, "Improved Resistive Switching Memory Characteristics Using Core-Shell IrOx Nano-Dots in Al2O3/WOx Bilayer," *Journal of The Electrochemical Society,* vol. 159, p. H177, 2011.

[54] W. G. Kim, J. Y. Kim, J. W. Moon, M. S. Joo, H. J. Choi, S. G. Kim and e. al, "Effect of Inserting Al2O3 Layer and Device Structure in HfO2-Based ReRAM for Low Power Operation," in *2012 4th IEEE International Memory Workshop*, Milan, Italy, 2012.

[55] D. R. Lide, Ed., CRC Handbook of Chemistry and Physics, Boca Raton, FL: CRC Press, 2005.

[56] R. H. French, "Electronic Band Structure of Al2O3, with Comparison to Alon and AIN," *Journal of the American Ceramic Society,* vol. 73, p. 477, 1990.

[57] J. Y. Shin and Y.-H. Son, "Direct observation of conducting filaments on resistive switching of NiO thin films," *Appl. Phys. Lett. 92,* p. 222106, 2008.

[58] W. Zhang, J. Smith and A.G. Evans, "The connection between ab initio calculations and interface adhesion measurements on metal/oxide systems: Ni/Al2O3 and Cu/Al2O3,," *Acta Materialia,* vol. 50, p. 3803, 2002.

[59] I. G. Batyrev and L. Kleinman, "In-plane relaxation of Cu(111) and Al(111) α-Al2O3(0001) interfaces," *Phys. Rev. B,* vol. 64, p. 033410, 2001.

[60] S. Shi, S. Tanaka and M. Kohyama, "First-principles study of the tensile strength and failure of α-Al2O3(0001)/Ni(111) interfaces," *Phys. Rev. B,* vol. 76, p. 075431, 2007.

[61] S. Prada, M. Rosa, L. Giordano, C. Di Valentin and G. Pacchioni, "Density functional theory study of TiO2/Ag interfaces and their role in memristor devices," *Phys. Rev. B,* vol. 83, p. 245314, 2011.

[62] S. Eremeev, S. Schmauder, S. Hocker and S. Kulkova, "Investigation of the electronic structure of Me/Al2O3(0001) interfaces," *Physica B: Condensed Matter,*, vol. 404, p. 2065, 2009.

[63] J. M. Soler, E. Artacho, J. D. Gale, A. Garcia, J. Junquera, P. Ordejon and D. Sanchez-Portal, "The SIESTA method forab initio order-N materials simulation," *Journal of Physics: Condensed Matter*, vol. 14, p. 2745, 2002.

[64] J. P. Perdew, K. Burke and M. Ernzerhof, "Generalized Gradient Approximation Made Simple," *Phys. Rev. Lett.*, vol. 77, p. 3865, 1997.

[65] M. Strange, I. S. Kristensen, K. S. Thygesen and K. W. Jacobsen, "Benchmark density functional theory calculations for nanoscale conductance," *J. Chem. Phys.*, vol. 128, p. 114714, 2008.

[66] F. A. Ponce, T. Yamashita and S. Hahn, "Structure of thermally induced microdefects in Czochralski silicon after high-temperature annealing," *Appl. Phys. Lett*, vol. 43, p. 1051, 1983.

[67] S. Menzel, P. Kaupmann and R. Waser., "Understanding filamentary growth in electrochemical metallization memory cells using kinetic Monte Carlo simulations," *Nanoscale 7*, p. 12673, 2015.

[68] G. Bersuker, D. C. Gilmer, D. Veksler, P. Kirsch, L. Vandelli, A. Padovani, L. Larcher, K. McKenna, A. Shluger, V. Iglesias, M. Porti and M. Nafría, "Metal oxide resistive memory switching mechanism based on conductive filament properties," *Journal of Applied Physics*, vol. 110, p. 124518, 2011.

[69] P. Huang, X. Y. Liu, W. H. Li, Y. X. Deng, B. Chen, Y. Lu, B. Gao, L. Zeng, K. L. Wei, G. Du, X. Zhang and J. F. Kang, "A physical based analytic model of RRAM operation for circuit simulation," in *2012 International Electron Devices Meeting*, 2012.

[70] P. Gonon, M. Mougenot, C. Vallée, C. Jorel, V. Jousseaume, H. Grampeix and F. E. Kamel, "Resistance switching in HfO2 metal-insulator-metal devices," *Journal of Applied Physics 107*, p. 074507, 2010.

[71] S. Aldana, P. García-Fernández, A. Rodríguez-Fernández, R. Romero-Zaliz, M. B. González, F. Jiménez-Molinos, F. Campabadal, F. Gómez-Campos and J. B. Roldán, "A 3D kinetic Monte Carlo simulation study of resistive switching processes in Ni/HfO2/Si-n+-based RRAMs," *J. Phys. D: Appl. Phys. 50*, p. 335103, 2017.

[72] R. Degraeve, A. Fantini, S. Clima, B. Govoreanu, L. Goux, Y. Chen, D. Wouters, P. Roussel and e. al, "Dynamic ‘hour glass’ model for SET and RESET in HfO2 RRAM," in *2012 Symposium on VLSI Technology (VLSIT)*, Honolulu, HI,, 2012.

[73] X. Guan, S. Yu and H. -.. 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, p. 1172, 2012.

[74] J. Guy, G. Molas, P. Blaise, M. Bernard, A. Roule, G. L. Carval, V. Delaye, A. Toffoli, G. Ghibaudo, F. Clermidy, B. D. Salvo and L. Perniola, "Investigation of Forming, SET, and Data Retention of Conductive-Bridge Random-Access Memory for Stack Optimization," *IEEE Trans. Electron Devices 62*, p. 3482, 2015.

[75] D. D. Lee, Y. Sung, I. Lee, J. Kim, H. Sohn and D. Ko, "Enhanced bipolar resistive switching of HfO2 with a Ti interlayer," *Appl. Phys. A*, p. 997, 2011.

[76] E. Linn, R. Rosezin, C. Kugeler and R. Waser, "Complementary resistive switches for passive nanocrossbar memories," *Nature Matt. 9,* p. 403, 2010.

[77] M. Liu, Z. Abid, X. He, Q. Liu and W. Guan, "Multilevel resistive switching with ionic and metallic filaments," *Applied Physics Letters,* vol. 94, no. 23, p. 233106, 2009.

[78] A. Padovani, L. Larcher, O. Pirrotta, L. Vandelli and G. Bersuker, "Microscopic Modeling of HfOx RRAM Operations: From Forming to Switching,," *IEEE Transactions on Electron Devices 62,* p. 1998, 2015.

[79] S. Yu, X. Guan and H. -.. 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, p. 1183, 2012.

[80] P. Huang, Y. Deng, B. Gao, B. Chen, F. Zhang, D. Yu, L. Liu, G. Du, J. Kang and X. Liu, "Optimization of Conductive Filament of Oxide-Based Resistive-Switching Random Access Memory for Low Operation Current by Stochastic Simulation," *Jpn. J. Appl. Phys,* vol. 52, p. 04CD04, 2014.

[81] M. Y. Yang, K. Kamiya, B. Magyari-Köpe, M. Niwa, Y. Nishi and K. Shiraishi, "Charge-dependent oxygen vacancy diffusion in Al2O3-based resistive-random-access-memories," *Applied Physics Letters,* vol. 103, p. 093504, 2013.

[82] 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, p. 400, 2016.

[83] S. M. Aspera, H. Kasai, H. Kishi, N. Awaya, S. Ohnishi and Y. Tamai, "Realization of the Switching Mechanism in Resistance Random Access Memory Devices: Structural and Electronic Properties Affecting Electron Conductivity in a Hafnium Oxide–Electrode System Through First-Principles Calculations," *J. Elec. Mater. 42,* p. 143, 2013.

[84] 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, p. 183505, 2005.

[85] Y. Guo, S. J. Clark and J. Robertson, "Electronic and magnetic properties of Ti2O3, Cr2O3, and Fe2O3 calculated by the screened exchange hybrid density functional," *J. Phys.: Condens. Matter 24,* p. 325504, 2012.

[86] U. Russo, D. Ielmini, C. Cagli and A. L. Lacaita, "Self-Accelerated Thermal Dissolution Model for Reset Programming in Unipolar Resistive-Switching Memory (RRAM) Devices," *IEEE Transactions on Electron Devices,* vol. 56, p. 193, 2009.

[87] G. Chen, "Thermal conductivity and ballistic-phonon transport in the cross-plane direction of superlattices," *Phys. Rev. B,* vol. 57, p. 14958, 1998.

[88] D. Garbin, E. Vianello, O. Bichler, Q. Rafhay, C. Gamrat and e. al, "HfO2-Based OxRAM Devices as Synapses for Convolutional Neural Networks," *IEEE Transactions on Electron Devices,* vol. 68, p. 2494, 2015.

[89] M. Alayan, Investigation of HfO2 based Resistive Random Access Memory (RRAM), Grenoble, France: Micro and nano technoloies/Microelectronics Universite Grenoble Alpes, 2018.

[90] J. McPherson, J.-Y. Kim, A. Shanware and H. Mogul, "Thermochemical description of dielectric breakdown in high dielectric constant materials," *Appl. Phys. Lett.,* vol. 82, p. 2121, 2003.

[91] S. Larentis, C. Cagli, F. Nardi and D. Ielmini, "Filament diffusion model for simulating reset and retention processes in RRAM," *Microelectron. Eng. 88,* p. 1119, 2011.

[92] S. Kumar, Z. Wang, X. Huang, N. Kumari, N. Davila, J. P. Strachan, D. Vine, A. L. D. Kilcoyne, Y. Nishi and R. S. Williams, "Conduction Channel Formation and Dissolution Due to Oxygen Thermophoresis/Diffusion in Hafnium Oxide Memristors," *ACS Nano,* vol. 10, p. 11205, 2016.

[93] D. B. Strukov, F. Alibart and R. Stanley Williams, "Thermophoresis/diffusion as a plausible mechanism for unipolar resistive switching in metal--oxide--metal memristors," *Applied Physics A,* vol. 107, p. 509, 2012.

[94] S. U. Sharath, S. Vogel, L. Molina-Luna, E. Hildebrandt, C. Wenger, J. Kurian, M. Duerrschnabel, T. Niermann, G. Niu, P. Calka, M. Lehmann, H. Kleebe, T. Schroeder and L. Alff, "Control of Switching Modes and Conductance Quantization in Oxygen Engineered HfOx based Memristive Devices," *Adv. Funct. Mater.,* vol. 27, p. 1700432, 2017.

[95] H. W. Pan, K. P. Huang, S. Y. Chen, P. C. Peng, Z. S. Yang, C.-H. Kuo and e. al, "1Kbit FinFET Dielectric (FIND) RRAM in pure 16nm FinFET CMOS logic process," in *2015 IEEE International Electron Devices Meeting (IEDM)*, Washington, DC, 2015.

[96] C. Tang and R. Ramprasad, "Point defect chemistry in amorphous HfO2: Density functional theory calculations," *Phys. Rev. B,* vol. 81, p. 161201(R), 2010.

[97] M. Lanza, G. Bersuker, M. Porti, E. Miranda, M. Nafria and X. Aymerich, "Resistive switching in hafnium dioxide layers: Local phenomenon at grain boundaries," *Appl. Phys. Lett. 101*, p. 193502, 2012.

[98] J. Y. Son and Y.-H. Shin, "Direct observation of conducting filaments on resistive switching of NiO thin films," *Appl. Phys. Lett.,* vol. 92, p. 222106, 2008.

[99] Z. Chai, J. Ma, W. D. Zhang, B. Govoreanu, J. F. Zhang, Z. Ji and e. al, "Probing the Critical Region of Conductive Filament in Nanoscale HfO2 Resistive-Switching Device by Random Telegraph Signals," *IEEE Transactions on Electron Devices,* vol. 64, p. 4099, 2017.

[100] W. Liu, K. Sasaoka, T. Yamamoto, T. Tada and S. Watanabe, "Inelastic transient electrical currents and phonon heating in a single-level quantum dot system," *Journal of Applied Physics,* vol. 113, p. 123701, 2013.

[101] A. Croy and U. Saalmann, "Propagation scheme for nonequilibrium dynamics of electron transport in nanoscale devices," *Phys. Rev. B,* vol. 80, p. 245311, 2009.

[102] J. Maciejko, J. Wang and a. H. Guo, "Time-dependent quantum transport far from equilibrium: An exact nonlinear response theory," *Phys. Rev. B ,* vol. 74, p. 085324, 2006.

APPENDIX A CONNECTED-COMPONENT LABELING ALGORITHM

In the KMC simulation, we need to identify the linkage between vacancies: for the vacancies in the same vacancy island, Ohm's law will be applied, while for those on different islands, the tunneling current should be computed. Then we need to use the connected component labeling algorithm to mark the vacancies in the same island.

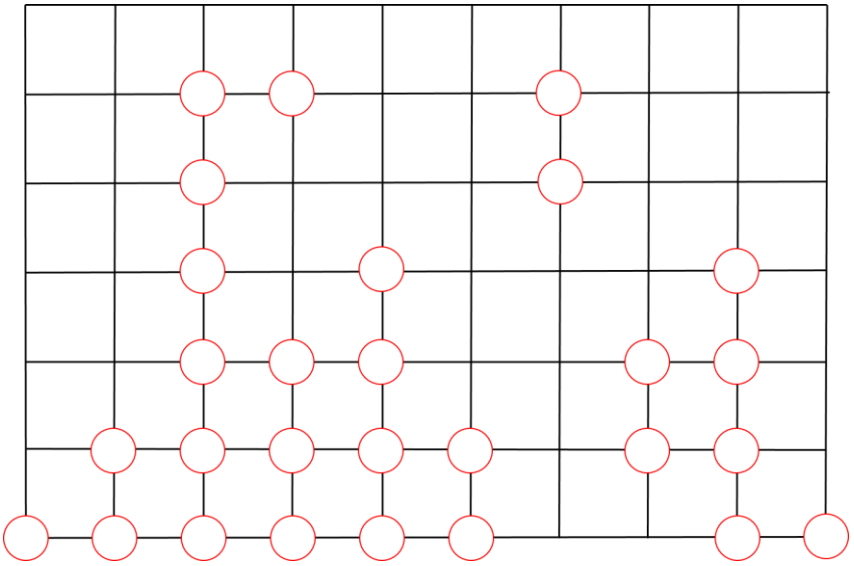


Fig. A.1. A picture of vacancies in the lattice. The red balls represent the vacancies. The algorithm in this appendix will identify the vacancies in different islands.

The connected component labeling algorithm is a two-pass process. In the first pass, every vacancy will be scanned and labeled with the same symbol as its neighbor. For example, we have a vacancy configuration as shown the Fig. A.1, where the red balls represent vacancies in the lattice.

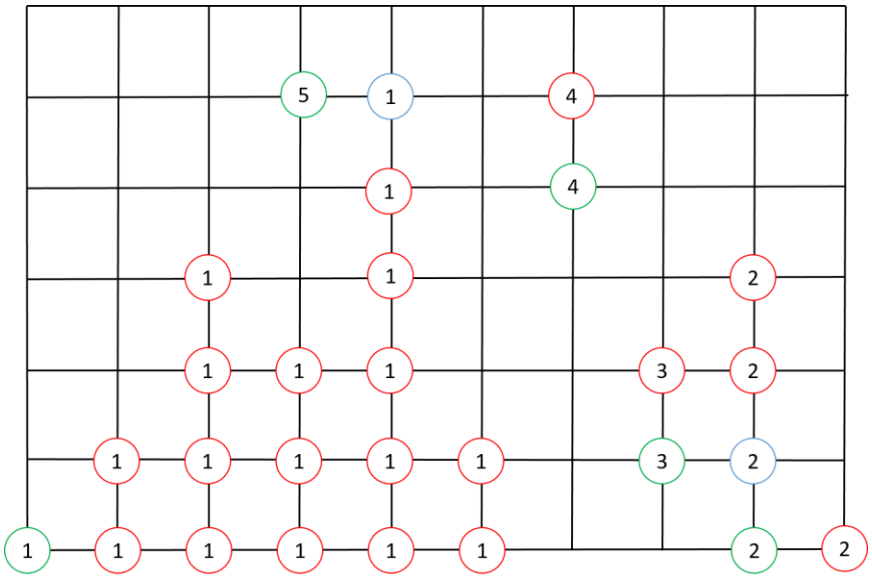


Fig. A.2. The first pass marks green vacancies with a new number since they don't have any neighbors marked when they are scanned. And mark the red vacancies with it neighbors' number because all the neighbors have the same number. The blue vacancies are marked with one of its neighbors' number and record that the neighbors should be in the same type.

The first pass is started as steps below: 1) scan the lattice in a certain sequence, for example, from left to right, then from bottom to top; 2) when a vacancy is found in the lattice, its neighbors will be checked. There are 3 cases for the checking results: case 1) if there are not marked neighbor, mark the vacancy with a new number; case 2) if there are neighbors with mark, and all the marked neighbors have the same number, mark the vacancy with its neighbors' number; and case 3) if there are neighbors with mark, and the mark numbers are different, mark the vacancy with one of the neighbors, and record that all the numbers of it neighbor should be sorted into the same type. For example, as shown in Fig. A.2, the vacancy in the green balls don't have any neighbor marked when they are scanned, as in case 1. Therefore, they are marked with a new number. The red balls in Fig. A.2 have at least one neighbor already marked when they are scanned, then they are marked as their neighbors' number, as in case 2. The blue balls have neighbors with different numbers. Therefore, they are marked with one of the neighbor's

numbers (the smaller number is taken here) and record that the neighbors' number should have the same type. For example, as in Fig. A.2, the blue balls show that number 1 and 5 should be in the same type, and number 2 and 3 should be in the same, as in case 3. In the second pass, the numbers with the same type will be assigned to the same island. For example, number 1 and 5 are assigned to the type of "color red", number 2 and 3 are assigned to the type of "color green" and number 4 is assigned to the type of "color blue". Then the islands of vacancies can be shown as in Fig. A.3.

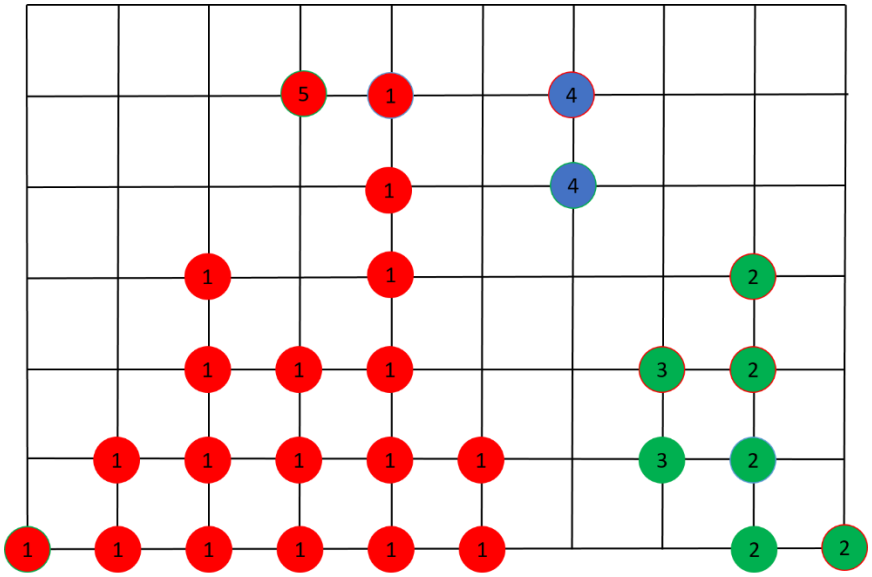


Fig. A.2. The second pass assign the vacancies in the same type into the same color, and the islands of vacancies are identified.

Then the pseudocode for the algorithm is

```
for site i is a vacancy site
      if neighbor[i] is empty
            label[i] = new(label)
            type[label[i]] = new(type)
      else
            label[i] = label[neighbor[i,0]]
            for j in neighbor
```

```
                type[label[j]] = type[label[i, 0]]
            end
        end
    end

for site i is a vacancy site
        label[i] = type[label[i]]
    end
```

APPENDIX B PERTURBATION OF PHONON AND ELECTROMIGRATION FORCE

In chapter 4 and 5, the attempt frequency is assumed as a constant, and in chapter 2, we can see that the attempt frequency is related to the phonon frequencies in the local minimum and saddle point. The phonon frequency can be given from the ab initio calculation. In the context of atomic filament in the crystal, the perturbation of atom mainly comes from the phonon interaction with the defect. Thus we need to know how the phonon distributed in the crystal and how they interact with the defect. The phonon energy and mode could be presented by force constant matrix that

$$\ddot{x}_\alpha = \sum_\beta F_{\alpha\beta}\, x_\beta\, M_\alpha^{-1}.$$

Here x_α is the displacement from the equilibrium position, M_α is the mass of atom, and the force constant matrix is defined as

$$F_{\alpha\beta} = \frac{\partial^2 E}{\partial x_\alpha \partial x_\beta}.$$

The calculation of force constant matrix could be finish in two methods: small displacement method (SDM) or density functional perturbation theory (DFPT). In the SDM, the atoms are moved by a small value from its equilibrium positions. Then the force constant matrix is approximated by

$$F_{\alpha\beta} \approx \left(\frac{\partial E(\Delta x_\alpha)}{\partial x_\beta} - \frac{\partial E(0)}{\partial x_\beta}\right)/\Delta x_\alpha \;\; = \big(F_\beta(\Delta x_\alpha) - \; F(0)\;\big)/\Delta x_\alpha.$$

In the DFPT, force constant matrix is given in the perturbation formalism of functional density theory [cite]

$$F_{\alpha\beta} = \sum_v \langle\psi_v| \frac{\partial^2 H}{\partial x_\alpha \partial x_\beta} |\psi_v\rangle.$$

In general, calculation of DFPT does not require large unit cell to give the long range interaction between atoms in different unit cell, while the SDM has to use a large unit cell to calculate the long range interaction. However, DFPT has to apply the periodic boundary condition in calculation, which makes it less versatile as the SDM method.

In order to understand the perturbation of current from phonon, we have to analyze the detailed interaction between phonon and defect. Here, we apply Born-Oppenheim approximation to study this interaction, which assumes the states of electron is separable from the states of ion. Because the mass of ion is much larger and thus move much slower than electron, and phonon is the vibration of ions. Then the Born-Oppenheim approximation is valid. Thus, we only need to calculate the electron states for different atomic configuration rather than solve a dynamical problem with coupled ion and electrons system. Under the phonon circumstance, we can treat the phonon induced displacement as a perturbation to the equilibrium stable state. Thus the Hamiltonian expanding to the first order of the displacement should be

$$H \approx H_0 + H_1 = H_0 + \sum_{\alpha} \frac{\partial H}{\partial x_\alpha}\, dx_\alpha$$

Here H_0 refers to the unperturbed Hamiltonian, which is the one when system is relaxed to the minimum energy state. The elements of perturbation matrix could be given by the partial derivative of the position of nuclei that

$$H_{1mn} = \sum_{\alpha} \langle m \left| \frac{\partial H}{\partial x_\alpha} \right| n \rangle \, dx_\alpha.$$

Actually, these elements are widely used in calculating the Hellmann-Feynman force that

$$F_\alpha = \sum_{n} \langle n \left| \frac{\partial H}{\partial x_\alpha} \right| n \rangle.$$

Then the perturbed Green's function is

$$G(E) = G_0(E) + G_0(E)H_1G_0(E)$$

The transmission becomes

$$T(E) = T_0(E) + 2\,Tr[G_0(E)H_1G_0(E)\Gamma_L G_0(E)\Gamma_R]$$

The current variation induced by a phonon mode is given that

$$var(T) = 2\pi\, Tr\left[G_0(E)\,\langle m\left|\frac{\partial H}{\partial x_\alpha}\right| n\rangle\, G_0(E)\Gamma_L G_0(E)\Gamma_R\right].$$

Thus we can give the DC noise of current induced by the phonon. Also the noise will change with the temperature as well as the configuration of Cu filament.

The atomic displacement can affect the current, and on the other hand, the current can apply force on the atoms. In the RESET process of ReRAM, the current density could be as high as xxx[cite], which is likely to induce the electromigration force on the atoms in filament.

The mechanism of electromigration force is that the atoms, especially defects, scatter the fast moving electron in the current. The Hellerman-Feynman theorem is still valid under the condition that the electrons are not in the equilibrium state [\cite]

$$F_\alpha = \sum_n \langle n\left|\frac{\partial H}{\partial x_\alpha}\right| n\rangle.$$

However, here we need to change the Hamiltonian as the functional of the density matrix given in the non-equilibrium state that

$$\rho = \int d\epsilon\; G^<(\epsilon).$$

Then the relationship between force and current could be given. When a force F_α is applied on an atom, it is equivalently to add an extra linear term to the potential energy surface that

$$V_{em} = -F_\alpha \cdot x_\alpha.$$

Then the energy barrier for ion hopping is lower because of the electromigration force. In general, larger current will induce a larger force on the atoms, meanwhile the energy consumption to

RESET a memory cell will be larger. Thus the tradeoff between the RESET speed and energy consumption must be made. In the RESET of ReRAM, the time is determined by the lifetime of defect diffusion, which is

$$\tau = \tau_0 \exp[-(E_b - F_{em} \cdot x)/kT].$$

With both electromigration force and heating effects, we can analyze the mechanism of RESET process in ReRAM.

APPENDIX C A SIMPLE DERIVATION OF TIME DEPENDENT GREEN'S FUNCTION

In the chapter 3, the assumptions made in calculating the current through atomic scale filaments are: 1) the Hamiltonian of the system due to the location of atoms is varying slowly; and 2) we only looked at the electronic state of the system after steady state has been reached. This method was used successfully to interpret the mechanism of current flow. However, it cannot handle a rapid change in applied voltage (not steady state) or significant levels of decoherence. A more precise time-dependent method must be used to calculate current flow [53-54].

The time dependent current can be derived from the time dependent Hamiltonian,

$$H = H_L + H_{LC} + H_C + H_{RC} + H_R$$

and the distribution function in the contacts. Here, H_L and H_R are the Hamiltonians of the isolated left and right electrodes, which are

$$H_{L/R} = \sum_{\alpha \in L/R} \epsilon_\alpha \, a_\alpha^\dagger \, a_\alpha .$$

Here ϵ_α is the energy on the electrodes, and a_α is the annihilation operator of electron on the electrodes. H_C is the Hamiltonian of the device, and is given by

$$H_C = \sum_i \epsilon_i(t) b_i^\dagger \, b_i + \sum_i h(t) b_i^\dagger b_{i+1} + H.c.$$

Here, $\epsilon_i(t)$ is the on-site energy in the device, $h(t)$ is the coupling between neighbor sites in the device, and b_i is the annihilation operator of electron on the i-th site of the device.

H_{LC} and H_{RC} are the coupling between device and electrodes,

$$H_{L/RC} = \sum_{\alpha \in L/R} V_{ij} \, a_i^\dagger \, b_j + H.c.$$

Then, the current in the left/right electrodes is

$$I_{L/R} = \frac{d}{dt} N_{L/R}.$$

Here $N_{L/R}$ are the number of electrons in the left and right electrodes respectively.

As suggested from [101, 102], the current in the time dependent system could be written as

$$J_\alpha(t) = \frac{2e}{\hbar}\ Re\left\{\int_{-\infty}^{\infty} dt'\ [G^<(t,t')\Sigma_\alpha^a(t',t) + G^r(t,t')\Sigma_\alpha^<(t',t)]\right\}.$$

Here, $J_\alpha(t)$ is the instantaneous current between the electrode α and the central region, $G^<$ and G^r are the lesser and retarded Green's function, and Σ_α^a and $\Sigma_\alpha^<$ are the self-energies for the electrode $\alpha \in L/R$. The greater and lesser self-energies could be explicitly given by

$$\Sigma_\alpha^>(t,t') = -i \int \frac{d\epsilon}{2\pi}\ (1 - f_\alpha(\epsilon))\ e^{-i\epsilon(t'-t)}\ \Gamma_\alpha(\epsilon,t,t'),$$

$$\Sigma_\alpha^<(t,t') = i \int \frac{d\epsilon}{2\pi}\ f_\alpha(\epsilon)\ e^{-i\epsilon(t'-t)}\Gamma_\alpha(\epsilon,t,t').$$

Here, $f_\alpha(\epsilon)$ is the Fermi function in electrode α, and $\Gamma_\alpha(\epsilon,t,t') = 2\pi\rho_\alpha(\epsilon)V_{\alpha i}(\epsilon,t)V_{\alpha j}^\dagger(\epsilon,t')$ with the density of state $\rho_\alpha(\epsilon)$ on the electrode $\alpha \in L/R$ is the level-width function on the electrode. The retarded and advanced self-energies have the relations to the lesser and greater self-energies as

$$\Sigma_\alpha^r(t,t') = \ u(t-t')\ [\Sigma_\alpha^>(t,t') - \Sigma_\alpha^<(t,t')],$$

$$\Sigma_\alpha^a(t,t') = \ -u(t'-t)[\Sigma_\alpha^>(t,t') - \Sigma_\alpha^<(t,t')].$$

Also this type of relation can be applied to Green's function too. Then we can obtain another expression for the instantaneous current

$$J_\alpha(t) = \frac{2e}{\hbar}\ Re\ Tr\{\Pi_\alpha(t)\}$$

Here, we defined the current matrix

$$\Pi_\alpha(t) = \int_{t_0}^{t} dt' [G^{>}(t,t')\Sigma_\alpha^{<}(t',t) - G^{<}(t,t')\Sigma_\alpha^{>}(t',t)]$$

The evolution of electron density function can be given by

$$i\,\dot{\rho}(t) = [H(t), \rho(t)] + i \sum_\alpha [\Pi_\alpha(t) + \Pi_\alpha^\dagger(t)]\,.$$

Please note that this is the Liouville equation that the first term on the r.h.s is the evolution of density matrix within the central system, and the second term describes the interaction between the central system and the electrodes.

We can expand the Fermi function on the poles of complex energy plane that

$$f_\alpha(\epsilon) = \frac{1}{1 - \exp\left(\frac{\epsilon - \mu_\alpha}{kT}\right)} \approx \frac{1}{2} - kT \sum_p \left(\frac{1}{\epsilon - \chi_{\alpha p}^{+}} + \frac{1}{\epsilon - \chi_{\alpha p}^{-}}\right).$$

Here $\chi_{\alpha p}^{\pm} = \mu_\alpha \pm kT\,\chi_p$ is the poles of Fermi function on the complex energy plane, and $Im\{\chi_p\} > 0$ is assumed. A well-known expansion of Fermi function is Matsubara expansion, which converges slowly at low temperature. Nevertheless, any other types of expansion which approximate the Fermi function can be used here. Here we are using the auxiliary expansion described in [101].

In the wide band limit, the level-width function is a constant that $\Gamma_\alpha(\epsilon) = \Gamma_\alpha$, then we can explicitly give the lesser self-energy that

$$\Sigma_\alpha^{<}(t',t) = \frac{i}{2}\Gamma_\alpha\,\delta(t-t') + \sum_p \Sigma_{\alpha p}(t',t)\,.$$

Here the self-energies on auxiliary poles are

$$\Sigma_{\alpha p}(t',t) = kT\,\Gamma_\alpha \exp\left[i\int_{t'}^{t} d\tau\,\chi_{\alpha p}^{+}(\tau)\right].$$

The greater self-energy can be expressed in a similar way. Then the current matrix becomes

$$\Pi_\alpha(t) = \frac{1}{4}[1 - 2\rho(t)]\Gamma_\alpha + \sum_p \Pi_{\alpha p}(t).$$

Here,

$$\Pi_{\alpha p}(t) = \int_{t_0}^{t} dt' \left[G^{>}(t,t')\Sigma_{\alpha p}^{<}(t',t) - G^{<}(t,t')\Sigma_{\alpha p}^{>}(t',t)\right]$$

We can find the equation of motion that

$$i\frac{\partial}{\partial t}\Pi_{\alpha p}(t) = kT\,\Gamma_\alpha + \left[H(t) - \frac{i}{2}\Gamma - \chi_{\alpha p}^{+}(t)\right]\,\Pi_{\alpha p}(t).$$

Here, $\Gamma = \sum_\alpha \Gamma_\alpha$. Solve the equations of motion for the current matrices and the density matrix, which depend on the current time only, then we can give instantaneous current.

In order to compute the time dependent current, we have to solve the density matrix equation fast. The equation for density matrix should be

$$\dot{\rho} \;=\; \rho[-\Gamma - iH] + [-\Gamma + iH]\rho + J^{in}(t)$$

Then the solution of the density matrix is

$$\rho(t + \Delta t) = \exp(-A\Delta t)\,\rho(t)\exp(-A^{\dagger}\Delta t) + \int_t^{t+\Delta t} \exp(A\tau)\; J^{in}(t-\tau)\exp(A^{\dagger}\tau)\,d\tau$$

The $J^{in}(t)$ is the current flow into the device, and it satisfies

$J^{in}(t) = J_L^{in}(t) + J_R^{in}(t)$.

The current matrix can be decomposed on poles of complex plane as,

$$J_L^{in}(t) = \frac{\Gamma_L}{4} - \sum_p \Pi_{Lp}(t) + H.c.$$

and

$$i\dot{\Pi}_{Lp} = \frac{\Gamma_L}{\beta} + \left(H(t) - \frac{i}{2}\Gamma - \chi_{\mathrm{Lp}}\right)\Pi_{Lp}$$

When Γ_L are non-zero only for injection into the left most device grid point, J_L^{in} has non-zero elements only on the first column and row only because the current matrix can be solved as

$$\Pi_{Lp}(t) = \frac{\Gamma_L}{\beta}\int_0^t \exp\left[\left(A + i\chi_{Lp}\right)\tau\right] d\tau.$$

The rank of J_L^{in} is at most 2. Then it can be decomposed as

$$J_L^{in}(t) = Z(t)Z^{\dagger}(t)$$

Here $Z(t)$ is a 2 by n matrix, which can be decomposed. For example,

$$J_L^{in}(t) = \begin{pmatrix} J_{11} & J_{1\alpha} \\ J_{\alpha 1} & 0 \end{pmatrix}$$

Here, $J_{\alpha 1}$ is an n-1 by 1 matrix and

$J_{\alpha 1} = J_{1\alpha}^{\dagger}$.

Then,

$$J_L^{in}(t) = \begin{pmatrix} J_{11} & J_{1\alpha} \\ J_{\alpha 1} & 0 \end{pmatrix} = P_1 \begin{pmatrix} 1 & 0 \\ 0 & -J_{\alpha 1} J_{11}^{-1} J_{1\alpha} \end{pmatrix} P_1^{\dagger}$$

Here

$$P_1 = \begin{pmatrix} J_{11}^{1/2} & 0 \\ J_{11}^{-1/2} J_{1\alpha} & I_{n-1} \end{pmatrix}$$

And we should notice that $-J_{\alpha 1} J_{11}^{-1} J_{1\alpha}$ remains Hermitian and

$$-J_{\alpha 1} J_{11}^{-1} J_{1\alpha} = \left(J_{\alpha 1} J_{11}^{-1/2}\right)\left(J_{\alpha 1} J_{11}^{-1/2}\right)^{\dagger}$$

Then we can conclude

$$Z = \begin{pmatrix} J_{11}^{1/2} & 0 \\ J_{11}^{-1/2} J_{1\alpha} & J_{\alpha 1} J_{11}^{-1/2} \end{pmatrix}$$

So the evolution of density matrix could be calculated as

$$\rho(t + \Delta t) = \exp(-A\Delta t)\, \rho(t) \exp(-A^{\dagger}\Delta t) + \int_t^{t+\Delta t} \exp(A\tau)\ Z(t-\tau)\, Z^{\dagger}(t-\tau) \exp(A^{\dagger}\tau)\, d\tau$$

To compute the terms of exponential, the Krylov subspace method can be employed, which takes $O(m^2N)$ time, where m is the bandwidth of Hamiltonian.