Toward a better understanding of RRAM

May 01, 2012

Editor’s note: This work was first presented at the 2011 IEEE International Electron Devices Meeting (IEDM) and appears here courtesy of the IEEE. For more information about IEDM 2012 (San Francisco, CA; December 10-12), click here.

Abstract
The origin of switching parameter variations in metal-oxide resistive-switching random access memory (RRAM) is studied. The stochastic formation/rupture of the conductive filaments (CFs) is modeled and incorporated with a trap-assisted-tunneling (TAT) current solver. The experimental DC I-V characteristics and pulse transient waveform featuring the current fluctuation during the reset process are reproduced by Monte Carlo simulations.

It is found that the wide spread of high resistance states (HRS) are due to the variation of tunneling gap distances, and the tail bits of the HRS are due to the newly generated traps near the electrode at the end of the reset process. To solve the over-reset and tail bits problems, a device structure with active/buffer bi-layer oxides combined with the reset-verify technique is proposed. Our model is corroborated by measured experimental data of hafnium oxide (HfO$_x$) based RRAM.

Introduction
Metal oxide RRAM is one of the most promising candidates for future non-volatile memory applications. Among the RRAMs reported, HfOx-based devices have shown excellent performance such as scalability (better than 30 nm), switching speed (on the order of nanoseconds), endurance (approximately $10^{10}$ cycles), and data retention (better than 10 years extrapolated at 200º C) [1]. A 4-Mb array macro circuit was demonstrated [2]. However, poor uniformity of the switching parameters has been the major challenge for the large-scale manufacturing [3]. The physical origin of the variation is still not well understood so far. In this study, we aim to obtain new insights on this problem through modeling and comparison with experiment. The device fabrication and principal electrical characteristics of the TiN/HfO$_x$/Pt devices are reported in our previous work [4].

Model description
In the widely recognized bipolar switching mechanism, the set process from HRS to low resistance state (LRS) is the formation of CFs by generation of oxygen vacancies (Vo), while the reset process from LRS to HRS is the rupture of CFs by recombining Vo with the oxygen ions that migrate from the oxygen reservoir at the electrode/oxide interface [5] (see figure 1). The insensitivity of the measured current to temperature suggests that TAT is the dominate conduction mechanism [6]. Recent measurements and analysis of the lowfrequency noise and AC conductance provide additional evidence of the TAT conduction process [7].
Figure 1: In the set, oxygen atoms are pulled out of the lattice, generating oxygen vacancies (Vo) and leaving behind CFs in the bulk oxide. $O^{2-}$ drift to TiN, and are stored at the interface, where TiN acts as an oxygen reservoir. In the reset under a reversed bias, $O^{2-}$ migrate from the interface back to the bulk oxide and recombine with Vo, leading to the rupture of CFs near TiN; $O^{2-}$ migration is field- and temperature-assisted.

TAT includes three steps electrode to trap tunneling, trap-to-trap tunneling, and trap-to-electrode tunneling (see figure 2). We developed a TAT solver that calculates the electron occupation probability of all the traps based on equation 1, which accounts for all the possible tunneling paths between the traps, and between the traps and the electrodes.

$$\frac{df_n}{dt} = (1 - f_n)\sum_{m} r_{mn} f_m - f_n \sum_{m} (1 - f_m) r_{mn} + R_{cn}(1 - f_n) - R_{na}f_n \quad [1]$$

where $f_n$ is the $n^{th}$ trap's electron occupancy, $r$ is the hopping rate between traps, and $R$ is the tunneling rate between the electrodes and the traps.

Mott hopping (equation 2) [8] and WKB approximation (equation 3) are used for calculating the transition rates between the traps, and between the electrodes and the traps, respectively.

$$r_{hopping} = \nu_0 \exp\left(-\frac{2d}{\Sigma} + q\Delta V/kT\right) \quad [2]$$

where $d$ is the distance between traps, $\Sigma$ is the electron wavefunction localization length $\sim 0.3$ nm, $\nu_0 \sim 10^{13}$ Hz, $q$ is the electronic charge, and $\Delta V$ is the voltage difference.

$$R_{tunneling} = \nu_0 \int_{E_F}^{E_{Fermi}} f_{\text{norm}}(E) \exp\left(-\frac{1}{\hbar^2} \int \sqrt{2m^*(E - E - qV)} dx\right) dE \quad [3]$$

$L$ is the distance between trap and electrode, $m^*$ is the effective mass for HfO$_2$ $\sim 0.1 m_e$ [15], $\nu_0 \sim 10^{14}$ Hz, and $V$ is the applied voltage.

The steady state current is given by equation 4:

$$I = -q\sum f_n (R_{cn}(1 - f_n) - R_{na}f_n) \quad [4]$$
Figure 2: Schematic of the trap-assisted-tunneling (TAT) conduction mechanism shows (1) electrode to trap tunneling, (2) trap to trap tunneling, (3) trap to electrode tunneling. Current insensitivity on temperature [6] suggests TAT dominates instead of thermal activation conduction processes.

Take a 1-D Vo chain with a tunneling gap between the electrode and the filament as an example. Figure 3 shows an exponential decrease of the tunneling current with the increasing gap distance. Thus any variation of the gap distance would cause a significant variation of the HRS resistance.

Figure 3: I-V of 1-D filament for different gap distances. The inset shows current exponentially decreases with increasing gap distance, which is the main cause of the HRS variation.

Figure 4 shows that increasing the applied voltage reduces the electron occupancy probability in the traps near the cathode, and the electrode to trap tunneling becomes the bottleneck of the conduction.
Figure 4: Electron occupancy probability along 1-D filament under different bias voltages. With the increase of voltage, the occupancy near the cathode decreases and the electrode to trap tunneling becomes the bottleneck of the whole conduction.

To simulate the randomness of Vo configuration, we extended our TAT solver to a 2-D case (see figure 5). The results of this paper are based on 2-D simulation.

Figure 5: An example of 2-D electron occupancy probability, and the following simulations are based on 2-D.

The stochastic nature CF formation/rupture is considered as the following: In the forming/set process, the probability of Vo generation is determined by the attempt-to-escape rate that oxygen jumps over its barrier as:

$$P_g(F, T, t) = t / t_0 \exp \left[ - \frac{(E_o - \Delta E)}{kT} \right]$$ \hspace{1cm} [5]

**Determining probability of Vo recombination**

During the reset process, the probability of Vo recombination is determined by the product of the attempt-to-escape rate and the concentration of the oxygen ion at that position as given in equation 6:

$$P_r(F, T, t) = C \cdot P_g$$ \hspace{1cm} [6]

The velocity of the field and temperature-assisted ion migration from the interface obeys equation 7
where $t_0$ is the lattice vibration time $\sim 10^{13}$ s, $t$ is the simulation time step, $a$ is the lattice constant (also the mesh size) $\sim 0.25$ nm, $E_a$ is the Vo formation energy $\sim 1$ eV, $E_m$ is the oxygen ion migration barrier $\sim 1$ eV, $E$ is the electric field, and $C$ is the ion concentration.

At each grid point, a random number is generated and compared with the probability calculated above to determine success of generation (if that grid has no Vo) or recombination (if that grid has Vo). In each iteration step, the electric field and temperature are updated: the voltage is assumed to be dropped on the gap region since the highly conductive filaments can be viewed as a virtual electrode [3]. For the sake of simplicity, an average temperature inside the simulated cell is calculated from the macroscopic Joule heating following the analysis in [10]. A more rigorous treatment is to calculate the local temperature profile from the Fourier heat equation by considering the power dissipation due to the inelastic tunneling [11]. Figure 6 shows the established stochastic simulation flow. To handle the current overshoot problem with a drastic increase of the Vo density during the set process, a trial-and-error strategy to optimize the simulation time step is used.

Figure 6: The established stochastic simulation flow.

Figure 7 shows the simulated switching I-V curves of forming/reset/set and the corresponding Vo configuration with the percolation paths highlighted. It is seen that a larger reset stop voltage would result in a larger gap thus a higher HRS. A larger set compliance would result in more percolation paths thus a lower LRS. The abrupt set transition is due to a positive feedback of the temperature and field enhancement of the Vo generation probability. The gradual reset transition is due to the oxygen ion migration from the interface and gradual recombination of Vo and consequent increase of the gap distance. For a better appreciation of the switching dynamics, an animation video

\[ v_{ion} = a/t_0 \exp\left(-\frac{E_m}{kT}\right) \sinh\left(\frac{ca}{kT}\right) \]
showing the stochastic Vo evolution during the DC cycling is provided in [12].

Figure 7: Cross-section view of the simulated cell (left electrode: positive bias for forming/set and negative bias for reset, pink points are Vo), simulated cell (10 nm ×10 nm) corresponds to the weak spot region of a RRAM cell, e.g. the grain boundary. (a) initially randomly distributed Vo in as-fabricated cell; (b) forming process; (c) percolation paths after forming; (d) reset process with different stop voltages; (e) & (f) smaller/larger gap due to smaller/larger reset stop voltage; (g) set process with different compliance current; the current overshoot during forming/set is also shown; (h) & (i) fewer/more percolation paths due to smaller/larger compliance current. Percolation paths are found by the Dijkstra algorithm and color-coded in grey scale according to the conducting strength (darker means stronger).

Experimental corroboration
We developed DC I-V curves for the experimental (see figure 8) and simulated cases (see figure 9).

Figure 8: Experimental DC I-V characteristics of HfO_{x} memory for different reset stop voltages demonstrate abrupt set and gradual reset.
Figure 9: Simulated I-V characteristics of HfO$_x$ memory for different reset stop voltages also demonstrate abrupt set and gradual reset.

Reset pulse transient waveforms are shown for the experimental (see figure 10) and simulated (see figure 11) cases.

Figure 10: Experimental pulse transient current in the reset process. Current fluctuation is observed, and before the end of pulse, current jumps due to new Vo generation.

Figure 11: Simulated pulse transient current in the reset process. Current fluctuation is
For both DC and pulse programming, the reset process is gradual and a significant current fluctuation is observed. By tracking the Vo evolution in the simulation, we found that a current jump corresponds to a new Vo generation in the gap region at that moment. At the beginning of the reset process with a short gap region, the current fluctuation is caused by the competition between Vo generation process due to the presence of a large field and recombination process due to the presence of mobile oxygen ions migrated from the interface. It is suggested [13] that multilevel HRS states can be achieved (a) by linearly increasing the reset pulse amplitude (see figure 12 and 13) or (b) equivalently by exponentially increasing reset pulse width (see figure 14 and 15).

![Figure 12: Experimental multilevel states achieved by linearly increasing reset pulse amplitude.](image12.png)

![Figure 13: Simulated multilevel states achieved by linearly increasing reset pulse amplitude.](image13.png)
Examination of the simulated Vo configuration confirms that similar gap distances can be achieved by these two programming schemes, thus explaining the exponential voltage-time relationship [3] of the switching dynamics.

In Figure 16, the simulated HRS distribution after 1000-times pulse cycling is correlated with their Vo configuration. It is revealed that the lognormal spread of HRS is due to the variation of the average gap distances, while the tail bits of HRS are due to Vo generation near the electrode at the end of a programming pulse (see the last current jump of figure 10 as an experimental example).
Figure 16: Simulated HRS distribution after 1000-times pulse cycling. The lognormal distribution is due to the Gaussian distribution of the average gap distances. The tail bits are due to new Vo near the electrode generated at the end of the pulse (see Figure 10 for an experimental example of tail bit creation).

To reduce the tail bits of HRS, the reset-verify technique [1] can be used (see figure 17), but this may introduce more over-reset bits of HRS and consequently more tail bits of LRS [1]. Therefore, controlling the gap distance in the reset process is crucial to overcoming the over-reset problem.

Figure 17: Simulated LRS and HRS distribution with reset-verify technique. The purpose of multiple reset pulses is to decrease the new Vo generation probability at the end of a single pulse. With a maximum of three trials, above 99.9% HRS meet the criterion (1MΩ). The
over-reset bits may in turn cause more tail bits of LRS, however, which is undesired.

We propose using an additional buffer oxide layer that has a larger oxygen ion migration barrier (Em) to confine the switching in the active oxide. By tuning the active/buffer oxide thickness, the over-reset bits of HRS are reduced (see figure 18). This leads to a smaller number of tail bits of LRS as well; however, the tail bits of HRS still remain.

![Diagram of bi-layer structure](image1)

**Figure 18**: Simulated LRS and HRS distribution of a bi-layer structure. The buffer layer has $Em=1.2$ eV, the active layer has $Em=1$ eV, and the total oxide thickness remains the same (10 nm). We see that 1.5 nm active layer is effective for reducing the over-reset bits and the tail bits of LRS, but it does not help eliminate the tail bits of HRS.

Combining the bi-layer structure with the reset-verify technique is thus expected to achieve both uniform LRS and HRS distribution (see figure 19). Experimentally, uniformity improvement has been observed in the bi-layer structure such as HfO$_2$/AlO$_x$ [4] and HfO$_2$/ZrO$_x$ [14].

![Diagram of cumulative probability](image2)

**Figure 19**: Simulated both uniform LRS and HRS distribution are achieved by combining a bi-layer structure with the reset-verify technique.
Conclusion
The main achievements of this work include:

- A self-consistent TAT solver was developed to quantify the electronic conduction
- A stochastic oxygen ion generation/recombination/migration model was established to quantify the ion processes
- The variations of DC I-V characteristics and pulse transient waveform were reproduced
- Switching features such as current overshoot, multilevel capability and the exponential voltage-time relationship were revisited
- The current fluctuation during the reset process was attributed to the competing forces between Vo generation and recombination
- The origin of HRS tail bits is clarified
- A device structure with active/buffer bi-layer oxides combined with reset-verify was proposed to solve the over-reset and tail bits problems.

This paper provides new understanding of the physical origin of the switching parameter variations and provides device design guidelines to improve the switching uniformity in a large memory array. The development of the simulation tool opens up an opportunity to systematically study the variability of the metal oxide RRAM technology.

Acknowledgements
We thank Yi Wu, Yang Chai, J. Provine and Cambridge Nanotech for the device fabrication. This work is supported in part by DARPA SyNAPSE, NSF ECCS 0950305, the NRI of the SRC through the NSF/NRI Supplement to the NSF NSEC Center for Probing the Nanoscale, the member companies of the Stanford NMTRI, and the C2S2 Center of the FCRP, an SRC subsidiary. S. Yu is additionally supported by the Stanford Graduate Fellowship.

References
12. Animation video available @ http://nanowiz.stanford.edu/
username: stanfordnano; password: iedm2011

About the authors
Ximeng Guan received his Ph.D. degree and B.E. degree from Tsinghua University, China in 2010 and 2005, both with honor. He is now a postdoctoral scholar at the Stanford Nanoelectronics Group. His research interests are in the modeling of nanoscale electronic devices. Currently he is working on the modeling of resistive non-volatile memory, energy efficient electronic devices, and MOSFETs with high-mobility channels. He has authored or co-authored 24 journal and conference papers.

Shimeng Yu received the B.S. degree from Department of Microelectronics, Peking University, China, in 2009, and the M. S. degree from Department of Electrical Engineering, Stanford University, USA, in 2011, where he is currently working toward the Ph.D. degree. His past research activities include the simulation of the parameters fluctuation in nanoscale transistors and SRAM cells. He has been working on the fabrication, characterization, and modeling of the emerging resistive switching memory devices and their applications for neuromorphic computation system since 2008.

H.-S. Philip Wong received the B.Sc. (Hons.) in 1982 from the University of Hong Kong, the M.S. in 1983 from the State University of New York at Stony Brook, and the Ph.D. in 1988 from Lehigh University, all in electrical engineering. He joined the IBM T. J. Watson Research Center, Yorktown Heights, New York, in 1988. In September, 2004, he joined Stanford University as Professor of Electrical Engineering. His research interests are in nanoscale science and technology, semiconductor technology, solid state devices, and electronic imaging.