Monte Carlo Simulation of a Three-Terminal RRAM with Applications to Neuromorphic Computing

Akhilesh Balasingam
Department of Electrical Engineering
Stanford University
Stanford, California
avb03@stanford.edu

Akash Levy
Department of Electrical Engineering
Stanford University
Stanford, California
akashl@stanford.edu

Haitong Li
Department of Electrical Engineering
Stanford University
Stanford, California
haitongl@stanford.edu

Priyanka Raina
Department of Electrical Engineering
Stanford University
Stanford, California
praina@stanford.edu

***Abstract*—We developed a Monte Carlo simulator to compute the state-dependent I-V characteristics of three-terminal (3T) RRAM devices. State switching in these devices is modeled using a combination of vacancy migration and trap-assisted-tunneling mechanisms. We describe key elements of the simulator, compute hysteresis curves under typical voltage cycling conditions, and demonstrate agreement with experimental results. We then study the response of 2T- and 3T-RRAMs under pulsed operation and show that 3T-RRAM conductance values have both greater dynamic range than 2T-RRAMs and the potential to deliver superior inference accuracy in neuromorphic applications.**

Keywords—RRAM, Three-Terminal, Non-Volatile Memory, Monte Carlo Simulator, Neuromorphic

I. Introduction

Two-terminal Resistive Random-Access Memory (2T-RRAM) devices have been researched extensively for high-density memory and neuromorphic computing applications. Recently, a three-terminal variant (3T-RRAM) of this device family has been investigated experimentally, which, by separating the terminals used for *read* and *write* operations, seeks to enable efficient resistive state update and flexible neuromorphic architectures [1, 2]. State switching in these devices has been theorized to occur via a non-filamentary mechanism, involving a combination of (1) field-driven oxygen vacancy migration and (2) electron transport based on trap-assisted-tunneling [1]. We develop a Monte Carlo simulator for 3T-RRAMs which captures these two mechanisms and makes predictions that show broad agreement with published experimental results. Using this simulator, we also study the response of 3T-RRAMs under pulsed operation and investigate their potential for use as synapses in neuromorphic applications.

The device structure we model is motivated by the experimental structure from [1], which is excerpted in Fig. 1a. The simulated structure is an idealization, consisting of a top electrode (TE), a bottom electrode (BE) and a gate electrode (GE), all of which are in contact with the switching layer, as shown in Fig. 1b. The dots in Fig. 1b depict oxygen vacancies, whose spatial distribution is modulated during *write* by a gate voltage V_G applied on GE, with TE and BE grounded. During *read*, GE floats, BE is grounded, and a voltage V_D applied on TE draws a current I_D, which depends on the effective distance between the defect population and the channel region between TE and BE. While the discussion below focuses on the geometry shown in Fig. 1b, alternative multi-terminal configurations and biasing arrangements for *read* and *write* can also be studied using this simulator.

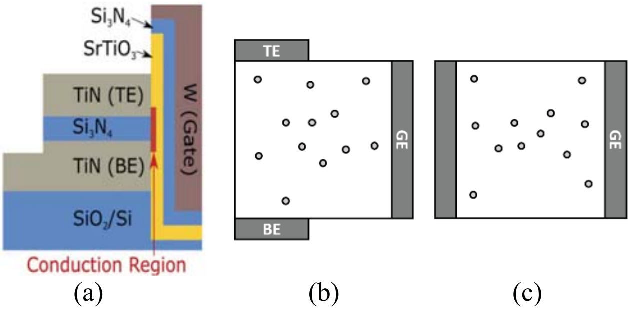


Fig. 1 (a) Schematic of 3T-RRAM structure, excerpted directly from [1], (b) 2D schematic of the 3D simulation domain, in which vacancy dynamics within the switching layer (e.g., TiO_2, $SrTiO_3$) are captured, and (c) the 2T-RRAM structure for comparison.

In Fig. 1c, we show, for comparison, a schematic of a 2T-RRAM device where the switching layer is sandwiched between two electrodes. In the 2T device, current is maximized when the centroid of the defect population is approximately half-way between the two electrodes. In the 3T configuration, conductance between TE and BE is maximized when the defect population is pushed closer to the TE-BE channel region. Thus, we can expect the 3T-RRAM to exhibit a greater dynamic range of conductance values than the 2T-RRAM. This is confirmed by the simulations presented in this paper.

II. Monte Carlo Simulator

Our model for this device extends an earlier two-terminal, non-filamentary vacancy modulation-based RRAM model by adding support for multiple terminals [3]. In this approach, the simulations are initialized with a population of N vacancies distributed at random within the switching layer. While the vacancies have the freedom to move within the simulation domain during device operation, their total number is assumed to be conserved. For the results reported below, the simulation domain is a rectangular prism with dimensions $10nm \times 10nm \times 12nm$, containing $N = 30$ vacancies, corresponding to a concentration of $n_v = 2.5 \times 10^{19} cm^{-3}$. The current flowing within the device is calculated using a fully-connected non-linear resistor network constructed by linking each vacancy m to each of the electrodes, and to every other vacancy n within the oxide, as illustrated in Fig. 2a. The tunneling current I_{mn} between vacancy m and vacancy n is approximated, as in [3], by an empirical function which is strongly nonlinear in the distance d_{mn} and the potential difference $V_{mn} = (V_m - V_n)$ between them:

$$I_{mn} = \quad g_0 V_{mn}[2 - e^{c_1(d_{mn}-c_2)}]e^{-\frac{c_3 d_{mn}}{V_{mn}}} \quad \text{(Eqn. 1)}$$

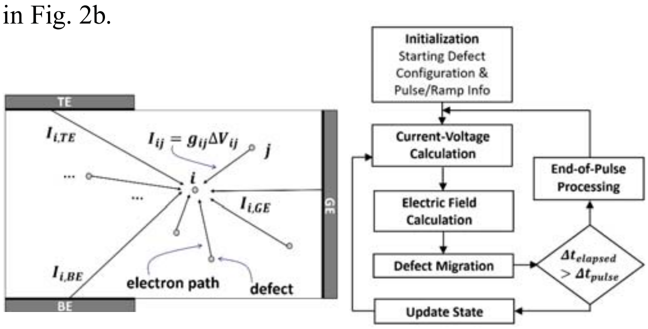
where $g_0 = 10nA.\text{V}^{-1}$, $c_1 \approx 0.069\, nm^{-1}$, $c_2 = 1\, nm$, $c_3 = 1\, V\, nm^{-1}$ are parameters of the model [3]. The potential V_m at the location of each vacancy m is then computed by iteratively solving a system of current continuity equations. The electric field $\vec{E}_m$, obtained from the local gradient of the potential distribution $\{V_m\}$, is used to calculate the vacancy migration rates:

$$r_m = \upsilon e^{-\Phi_b/k_B T} \sinh\left(\frac{\gamma q E_m}{k_B T}\right) \quad \text{(Eqn. 2)}$$

where k_B is the Boltzmann constant, $|q| = 1.6 \times 10^{-19} C$ is the charge on an electron, T is the absolute temperature, and $\upsilon = 10^{13} s^{-1}$ is the migration attempt frequency [4]. The energy barrier for defect movements Φ_b=0.8eV and the local field enhancement factor $\gamma = 12$ are empirical parameters of the model, which can depend both on the properties of the materials system under study and on process conditions [5]. The vacancy m that moves is selected with probability proportional to the rates r_m , and is moved a distance corresponding to a typical lattice spacing in transition metal oxides, $\delta = 0.2nm$, in the direction of the local electric field $\vec{E}_m$. The elapsed time for the event is then computed using the sum of all the defect migration rates:

$$\Delta t_i = -(\textstyle\sum_m r_m)^{-1} \ln(s_i) \quad \text{(Eqn. 3)}$$

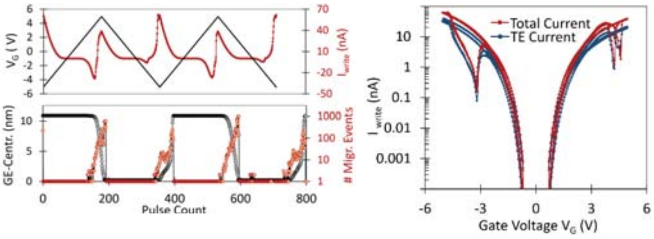
where s_i is a random number distributed uniformly on the unit interval [4]. The key elements of the simulator are shown in Fig. 2b.



(a) (b)

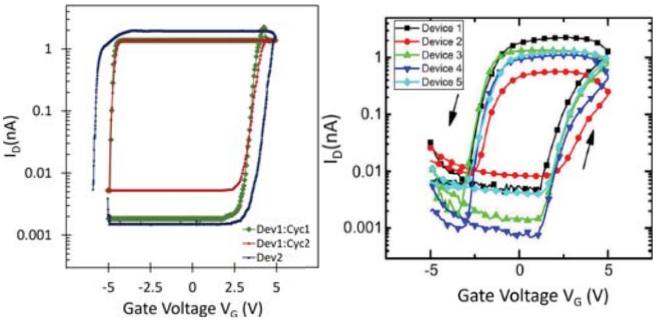
Fig. 2 (a) Continuity condition imposed at one of the vacancies forming the non-linear resistor network used to model current flow via trap-assisted tunneling (TAT), and (b) flowchart showing key functional units of the simulator.

During programming, TE and BE are grounded, and a voltage stimulus is applied on GE as a series of pulses, each of duration $\Delta t = 10\mu s$. The pulse height V_G is cycled twice between $-5V$ and $+5V$ in $|\Delta V_G| = 50mV$ increments per pulse, as depicted in the triangular profile in the upper graph of Fig. 3a, which also shows the corresponding write current I_{write}. As V_G is cycled, the defects within the switching layer migrate in response to the local electric field, which is oriented primarily in a direction perpendicular to GE. The distance between GE and the centroid of the defect distribution, and the number of defect migration events during each pulse are shown in the lower graph of Fig. 3a. The same *write* current I_{write} is shown as a function of V_G in Fig. 3b, which shows the standard hysteresis behavior.



(a) (b)

Fig. 3 (a) The upper graph shows the heights of the voltage pulses and the total 'write' current I_{write} drawn through the GE, and the lower graph shows the distance between the centroid of the defect population and GE and the number of defect migration events occurring during each pulse, (b) 3T-RRAM hysteresis loop, showing I_G from (a) as a function of V_G, which splits approximately evenly between the TE and BE. The TE current I_{TE} is also shown; the BE current $I_{BE}= I_G - I_{TE}$ is not shown to preserve clarity of figure.



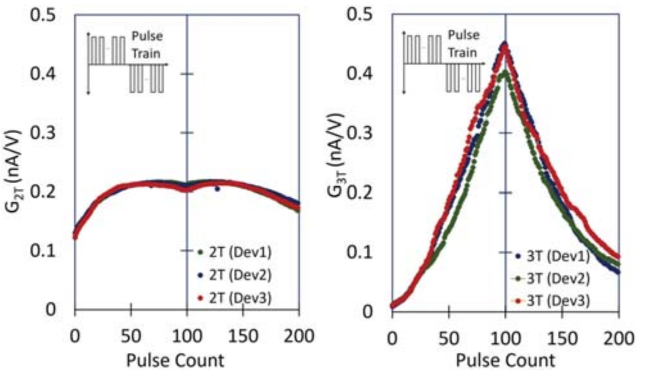
(a) (b)

Fig. 4. The curves labeled Dev1:Cyc1 and Dev1:Cyc2 show the 'read' current corresponding to the two 'write' cycles shown in Fig. 3. The curve Dev2 corresponds to a second device instance with V_G swept with a different pulsing schedule: $\Delta t = 10\mu s$, and $|\Delta V_G| = 10mV$. The 'read' is performed nondestructively with GE floating, BE grounded and TE biased to V_D = 2V, and (b) Graph directly excerpted from experimental paper [1] shows box-like hysteresis loops comparable to our simulation results. The upper and lower bounds of I_D in the simulation results shown in Fig. 6 are within the band of variability seen experimentally.

As the number of defects within reach of the electric field between the TE and BE rises or falls, the drain current increases or decreases, respectively. So, at the conclusion of each gate pulse, we apply a voltage pulse on TE, grounding BE and floating GE, to *read* the drain current I_D as a function of V_G, which is shown in Fig. 4a. The box-like hysteresis loop in this graph agrees qualitatively with the experimental results shown in Fig. 4b (excerpted from [1]), in shape and magnitude.

III. Neuromorphic Benchmarking

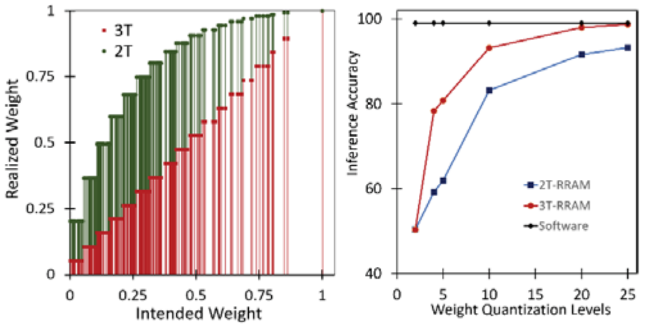
With this validation and demonstration of model capabilities, we explore the suitability of 3T-RRAMs for neuromorphic applications by computing their drain conductance as a function of the number of fixed-height pulses M applied on the gate. Fig. 5a and 5b show characteristics for the case of $M = 100$ for 2T- and 3T-RRAMs. It is evident that the 3T-RRAM exhibits better linearity and greater on/off ratio. Unlike the 2T-RRAM, in which the conductance peaks when the centroid of the defect population is roughly equidistant from the two electrodes, the conductance of the 3T-RRAM steadily increases as the defect population is pushed closer to the BE and TE.



(a) (b)

Fig. 5 Comparison of the response of 2T-RRAM and 3T-RRAM devices when a sequence of gate pulses of constant height and duration are applied during the 'write' phase. The first M=100 pulses have a height of $V_G = 3V$ (potentiation) and the next M=100 pulses have a height of $V_G = -3V$ (depression). (a) 2T-RRAM conductance with read voltage of $V_G = 2V$, for three devices (b) 3T-RRAM conductance with a read voltage of $V_D = 2V$, between TE and BE, for three devices.

Next, we model the mapping of an offline-trained fully-connected neural network (NN) to two crossbar circuits, with synapses based on 2T- and 3T-RRAMs, respectively, and simulate inference accuracy in MATLAB. The NN consists of 400 input neurons, 25 neurons in the hidden layer, and 10 output neurons, and it is trained in software on the MNIST data [6]. The weights of the NN are then mapped to discrete RRAM conductance levels attainable by applying a fixed number of gate pulses. With increasing M, it is clear that the 3T-RRAM will yield a greater number of discrete levels than the 2T-RRAM, leading to a mapping with higher fidelity. Fig. 6b shows the inference accuracy achieved when NN weights are mapped to 2T- and 3T-RRAM conductance levels, as well as the best accuracy of 98.86% attained in software (SW). This analysis shows that as M increases, the inference accuracy of the 3T-based circuit approaches levels achieved in SW more quickly than the 2T-based circuit.



(a) (b)

Fig. 6 Comparison of the neuromorphic performance of crossbar circuits configured with 2T- and 3T-RRAM synapses with characteristics shown in Fig. 5. (a) Weights realized with the conductances of 2T- and 3T-RRAM vs. scaled weights determined in SW (intended weights), with 20-level quantization, and (b) inference accuracy when NN weights are represented by 2T- and 3T-RRAMs at different quantization levels.

IV. Conclusion

We have developed a new Monte Carlo simulator for multi-terminal, non-filamentary RRAM devices. The model is simple, capturing only the dynamics within the switching layer of the device, with only a few fitting parameters for calibration. Yet it produces results that are in qualitative agreement with behavior observed experimentally in a 3T-RRAM device. Using the simulator, we demonstrated that 3T-RRAMs show promise for use as synapses in neuromorphic crossbar circuits, warranting further experimental study of multi-terminal RRAMs. Future work includes model enhancements such as the addition of vacancy generation and annihilation, support for additional material layers and interfaces and an exploration of the performance of multi-terminal RRAMs when used in NN training.

Acknowledgments

Akash Levy is supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1650114. We thank Weier Wan from Stanford University for insightful discussions.

References

[1] Herrmann, E., et al. (2018). Gate Controlled Three-Terminal Metal Oxide Memristor. *IEEE Electron Device Letters,* 39(4), 500–503. doi: 10.1109/led.2018.2806188

[2] Rush, A. J., Jones, A., Herrmann, E., & Jha, R. (2019). Gated-ReRAM Based Strategies for On-Chip Supervised Learning. *2019 IEEE National Aerospace and Electronics Conference (NAECON)*. doi:10.1109/naecon46414.2019.9057906

[3] Subhechha, S., et al. (2017). Kinetic defect distribution approach for modeling the transient, endurance and retention of a-VMCO RRAM. *2017 IEEE International Reliability Physics Symposium (IRPS)*. doi: 10.1109/irps.2017.7936322

[4] Voter A.F. (2007) Introduction to the Kinetic Monte Carlo Method. In: Sickafus K.E., Kotomin E.A., Uberuaga B.P. (eds) Radiation Effects in Solids. NATO Science Series, vol 235. Springer, Dordrecht

[5] Guan, X., Yu, S., & Wong, H. P. (2012). A SPICE Compact Model of Metal Oxide Resistive Switching Memory With Variations. *IEEE Electron Device Letters,* *33*(10), 1405-1407. doi:10.1109/led.2012.2210856

[6] Le Cun, Y., et al. (2013). THE MNIST DATABASE. Retrieved from http://yann.lecun.com/exdb/mnist/