Statistical Modeling of Metal-Oxide RRAM SET/RESET Behavior Using Deep Neural Networks

Aarav Wattal*
Dept. of Electrical Engineering
Stanford University
Stanford, CA
awattal@stanford.edu

Akash Levy*
Dept. of Electrical Engineering
Stanford University
Stanford, CA
akashl@stanford.edu

Zainab Faryal Khan
Dept. of Electrical Engineering
Stanford University
Stanford, CA
zainabk@stanford.edu

***Abstract*—We propose the use of end-to-end deep neural network (DNN) models to predict the statistical behavior of resistive RAM (RRAM) during SET/RESET programming. We demonstrate that such models may accurately generate the probability distributions describing the final conductance, without the need for fitting parameters by hand (R^2 = 0.983 for SET, R^2 = 0.817 for RESET). We also attempt to connect some of our models' qualitative behavior to physics-based understanding of RRAM SET/RESET processes. Finally, we describe a use case in which our models serve as key testing infrastructure in functional verification of a multiple-bits-per-cell RRAM controller.**

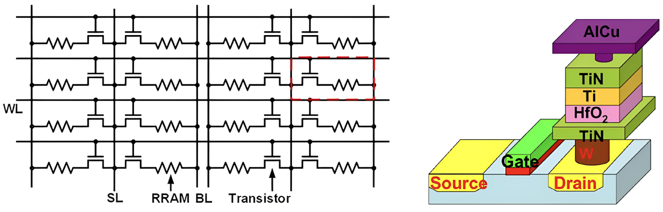
***Index Terms*—resistive RAM (RRAM), non-volatile memory, deep neural network, statistical device modeling**

I. Introduction

The recent proliferation of low-power system-on-chip (SoC) designs has led to a rise in demand for embedded non-volatile memories to reduce the overhead associated with moving data to and from off-chip memories [1], [2]. Metal-oxide Resistive RAM (RRAM) is one of the most promising candidates due to its low cost and compatibility with the CMOS back end of line [3]. Additionally, RRAM has the potential to store multiple bits per cell, thereby improving density for high-capacity on-chip storage [4]. However, a major challenge with RRAM is that its behaviors are highly stochastic and therefore difficult to understand and model [5], [6]. This paper proposes two end-to-end deep neural network (DNN) models to rapidly make well-fitted empirical predictions about RRAM's nonlinear and stochastic behavior. We connect our results with previously-developed physics-based explanations of RRAM resistive switching. Finally, we show how our models can serve as functional verification components for a digital multiple-bits-per-cell RRAM controller.

II. Background

To quantify SET/RESET variability, we start by collecting data on a 1T1R RRAM array (reported in [7]), shown in Fig. 1, by performing sweeps of SET/RESET pulses across different combinations of pulse parameters, followed by READ pulses. For the SET sweep, we start with cells in the high-resistance state (HRS), then apply different wordline voltages (V_{WL}), bitline voltages (V_{BL}), and pulse widths (t_{pw}). Conversely, for the RESET sweep, we start with cells in the low-resistance state (LRS), then apply different combinations of $\{V_{WL}, V_{SL}, t_{pw}\}$ as in the SET sweep (V_{SL} is the source-line voltage). The parameter ranges we sweep over in the SET/RESET sweeps are given in Table I. Note that pulse width is swept (roughly) exponentially, with six possible values: {20 ns, 40 ns, 100 ns, 200 ns, 400 ns, 1000 ns}. We measure the resistance before and after each pulse, and the SET/RESET sweep datasets each have 32,768 pulses logged in total.



(a) (b)

Fig. 1: (a) 1T1R RRAM cell array, showing word lines, bit lines, and a shared source line for every two bit-lines, excerpted from [8]. (b) Structure of HfO_2 1T1R RRAM, reproduced from [3] (dashed red rectangle in (a))

Parameter	Start	Stop	Step
Pulse Width (ns), t_{pw}	20	1000	×2 (or 2.5)
Wordline Voltage (V), V_{WL}	0.5	3.4	+0.1
Bitline Voltage (V), V_{BL}	0.5	3	+0.5

TABLE I: Table showing the sweep parameter ranges

III. Problem Formulation and Modeling Methodology

A. Problem Formulation

The goal is to estimate the final conductance (g_f) of a cell as a probabilistic function of the initial conductance (g_i), wordline voltage (V_{WL}), bitline voltage (V_{BL}), and pulse width (t_{pw}):

$$F_{SET}(g_i, V_{WL}, V_{BL}, t_{pw}) = P(g_f) \qquad (1)$$

Note that we choose to use conductance rather than resistance, because this value is more easily bounded and has behavior with less nonlinearity. Similarly, for RESET, we can write the function as:

$$F_{RST}(g_i, V_{WL}, V_{BL}, t_{pw}) = P(g_f) \qquad (2)$$

We can approximate both of these functions using neural networks that attempt to generate the probability distribution of the final conductance. In our dataset, the conductance values are scaled by a constant factor to normalize them between 0 and 1. The dataset is split into training data (60%), validation data (10%), and testing data (30%).

*Aarav Wattal and Akash Levy contributed equally to this work.

B. *Modeling Methodology*

We develop two different small machine learning models (shown in Fig. 2) to make predictions on the final conductance following a SET/RESET pulse. The first model is a multi-layer perceptron (MLP) (Fig. 2a), a simple feed-forward artificial neural network that makes predictions without estimating variability [9]. We train this network to predict final conductance values from the pulse width, wordline voltage, bitline/source-line voltage, and initial conductance. We train this network to predict final conductance values from the pulse width, wordline voltage, bitline/source-line voltage, and initial conductance. The output is a single number (g_f) for the predicted final conductance value, and $P(g_f) = \delta(g_f)$, δ(g) is the Dirac delta function. The second model is a Probability Density-based Network (PDN) (Fig. 2b) that predicts a Gaussian probability distribution [10], $P(g_f) = \mathcal{N}(g_f, \sigma_f)$, where $\mathcal{N}(\mu, \sigma)$ is the normal distribution with mean μ and standard deviation σ. Details of model architecture/training are provided in the caption of Fig. 2.

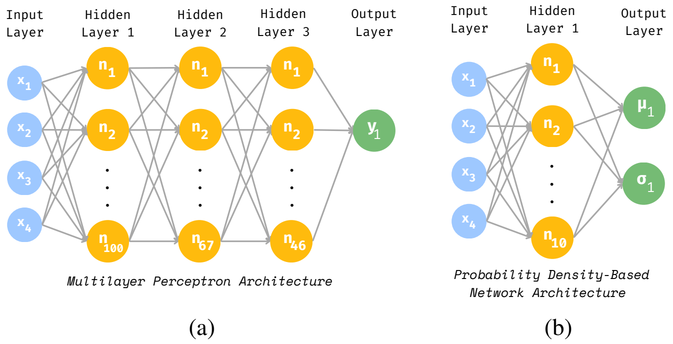
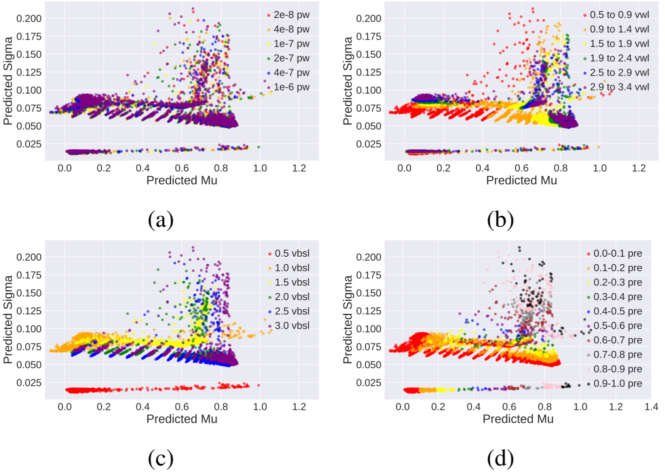


Fig. 2: (a) Model 1 is a multi-layer perceptron (MLP) with 3 hidden layers: the first with 100 neurons, the second with 67, and the third with 46. We use the Stochastic Gradient Descent (SGD) optimization algorithm with a mean-squared error (MSE) loss function, a learning rate of 0.01, and early stopping with validation data to avoid over-fitting. (b) Model 2 is a small probability density-based network (PDN), consisting of an input layer, a hidden layer with 10 neurons (tanh activation function), and two output neurons for the mean and standard deviation. The model is trained with 384 batches (64 samples each) and uses Adam optimization algorithm ($\gamma = 0.01, \beta_1 = 0.9, \beta_2 = 0.999$) with negative log-likelihood loss function: $\mathcal{L} = -\log P(y|x)$.

IV. Results and Observations

Both the MLP and PDN are trained and tested on the same set of preprocessed data. For the SET sweep data, the MLP makes final conductance predictions with a root mean squared error (RMSE) of 5.918%, while the PDN makes predictions with an RMSE of 6.557%. When tested on the subset of SET data that yields $>$20% changes in conductance in response, the PDN predicts with an RMSE of 4.971%. On the RESET sweep data, MLP yields an RMSE of 3.839%, PDN yields an RMSE of 4.420%. For both models, there is an overall strong correlation between the actual and predicted conductance values for SET (R^2 = 0.983 for MLP, R^2 = 0.976 for PDN) and moderate correlation for RESET (R^2 = 0.817 for MLP, R^2 = 0.742 for PDN), since RESET is a more stochastic process than SET [11].

As shown in Fig. 3, the SET data with intermediate bitline voltage has the greatest predicted standard deviation. Physically, this is because intermediate V_{BL} often yields a physical state in which the RRAM is transitioning between being thin filament(s) dominated by electron hopping to being thicker filament(s) dominated by metallic behavior [12]. A thin filament state means that the conductance is more sensitive to small changes in filament geometry. Wordline voltage is negatively correlated to predicted variability; as the wordline voltage increases, the conductive filament effective width increases and the filament becomes more metallic. While the pulse width can influence the resistive switching of the cell, the measured data and our models indicate that the variability is much more dependent on the voltages/currents in the device. As shown in Figs. 4 and 5, the lower bitline voltage values were predicted at a lower accuracy, which we theorize happens as the formation of new filaments introduces an additional source of randomness. Fig. 6 demonstrates that our RESET data models have better RMSE but worse correlation than our SET data models. This is because most RESETs are not difficult to predict, either resulting in a fully high-resistance state or in no change at all. However, some outliers do exist at the high wordline and high bitline voltages, likely because under these conditions, there exists rapid migration of ions/oxygen vacancies that constitute the conductive filament(s). Fig. 7 provides a visual depiction of the SET sweep data's predicted final conductance distribution compared to the measured final conductance distribution for a particular t_{pw} and g_i.



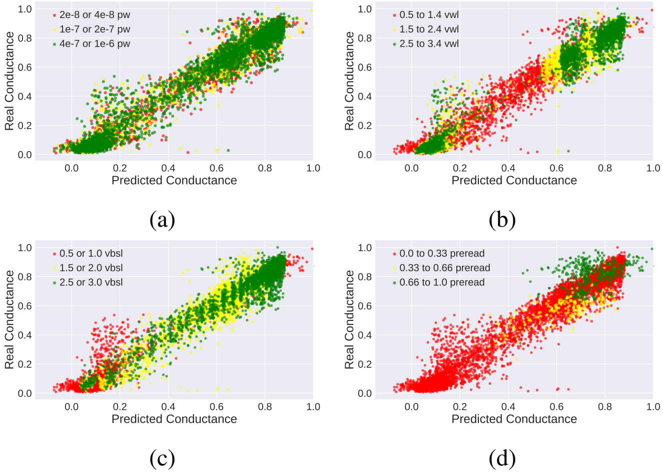
(a) (b)

(c) (d)

Fig. 3: The PDN is trained on the SET dataset and utilized to predict mean final conductance (mu) and standard deviation (sigma). Predicted mu vs. sigma plots are colored by input variables: (a) pulse width, (b) wordline voltage, (c) bitline voltage, and (d) pre-read conductance. (a) Pulse width does not demonstrate a significant effect on predicted sigma. (b) There is a negative correlation between wordline voltage and sigma. (c) Pulses with mid- and high-range bitline voltages are the hardest to make predictions about. (d) Positive correlation exists b/w pre-read conductance and predicted sigma.

A. *Multiple-Bits-per-Cell RRAM Controller Verification*

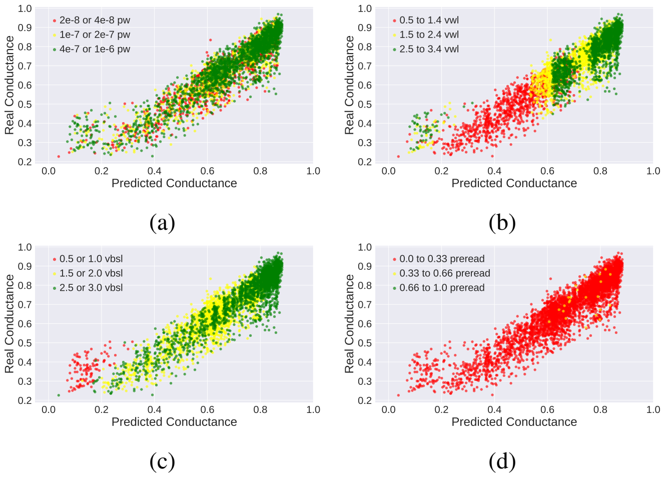
The DNN models presented here can be instrumented as fast verification components for circuits involving RRAM arrays, as depicted in Fig. 8. During functional verification of a digital multiple-bits-per-cell RRAM controller, for example, the DNN models can be queried to get the predicted final conductance distribution based on a pulse. By sampling the RRAM conductance from these distributions after each pulse, an on-chip write-verify process can be accurately modeled and



(a) (b)

(c) (d)

Fig. 4: The PDN's predicted final conductance values are plotted against the actual final conductance data for the SET dataset. The results are color stratified by input variable: (a) pulse width, (b) wordline voltage, (c) bitline voltage, (d) pre-read conductance. (b) shows that low wordline voltages are difficult to predict accurately, and (c) indicates that low and medium bitline voltages are easier to predict than high voltages. (d) indicates that the final conductance states of cells with a high pre-read conductance are more difficult to predict.



(a) (b)

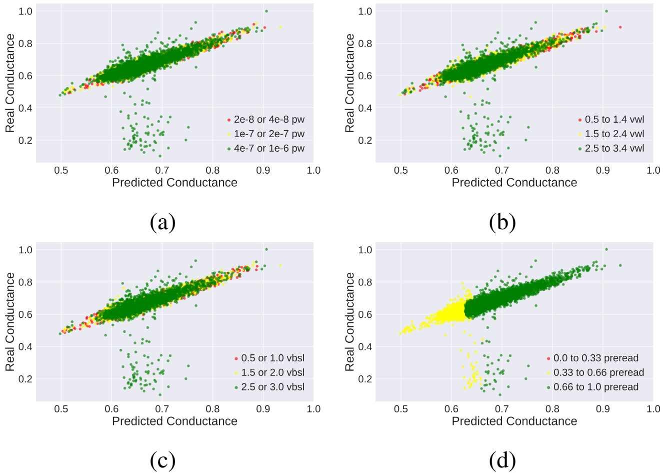
(c) (d)

Fig. 5: PDN results for predictions on SET data with >20% change in scaled conductance (where the pulse has a significant impact on the cell conductance). The data are depicted in plots corresponding to input variables: (a) pulse width, (b) wordline voltage, (c) bitline voltage, and (d) pre-read conductance. Despite the significant change in conductance, the model can accurately predict final conductance, with the low bitline voltage data in (c) being slightly harder to predict.

validated with realistic behaviors [13]. Furthermore, the write-verify process can be optimized for the RRAM technology that is characterized.

V. Conclusion

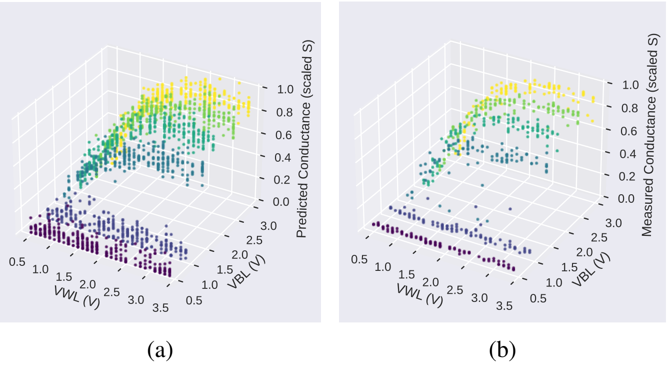
Our neural network models prove effective in predicting the final conductance value of an RRAM cell following a SET/RESET pulse as well as its probability distribution. The models' ability to make predictions over a broad range of pulsing conditions enables functional verification of circuit blocks containing RRAM arrays. These DNN models can be fitted quickly to experimental data without prior knowledge of



(a) (b)

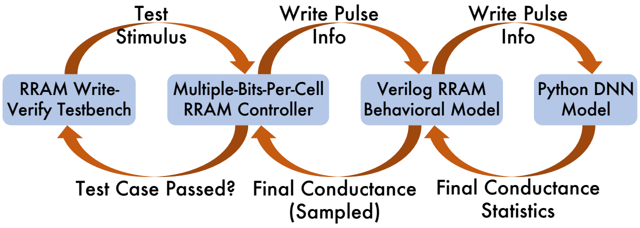
(c) (d)

Fig. 6: The PDN is trained on the RESET dataset and tested to predict final conductance. The predictions are plotted against the actual conductance values, colored by (a) pulse width, (b) wordline voltage, (c) source-line voltage, and (d) pre-read conductance. Aside for some of the data with high wordline voltage and high source- line voltage, the model successfully predicts most of the conductance values with low standard deviation.



(a) (b)

Fig. 7: Example plots showing the predicted vs. measured conductance for SET sweep, filtered for t_{pw} = 100 ns and lowest 25% of initial conductance. Different V_{BL} voltages are stratified by color as a visual aid. The V_{WL} and V_{BL} are graphed on the x- and y-axes, with the z-axis in (a) representing scaled predicted conductance and in (b) representing scaled measured conductance.



Test Stimulus
Write Pulse Info
Write Pulse Info
RRAM Write-Verify Testbench
Multiple-Bits-Per-Cell RRAM Controller
Verilog RRAM Behavioral Model
Python DNN Model
Test Case Passed?
Final Conductance (Sampled)
Final Conductance Statistics

Fig. 8: Integration of DNN-based RRAM model into digital verification flow for multiple-bits-per-cell RRAM controller.

the RRAM's geometry or electrical behaviors. They serve as a powerful empirical tool for enabling the further development of RRAM technology.

References

[1] M. M. Sabry Aly, M. Gao, G. Hills, C.-S. Lee, G. Pitner, M. M. Shulaker, T. F. Wu, M. Asheghi, J. Bokor, F. Franchetti, K. E. Goodson, C. Kozyrakis, I. Markov, K. Olukotun, L. Pileggi, E. Pop, J. Rabaey, C. Ré, H.-S. P. Wong, and S. Mitra, "Energy-Efficient Abundant-Data Computing: The N3XT 1,000x," *Computer*, vol. 48, no. 12, pp. 24–33, 2015.

[2] H.-S. P. Wong and S. Salahuddin, "Memory leads the way to better computing," *Nature nanotechnology*, vol. 10, no. 3, pp. 191–194, 2015.

[3] H.-S. P. Wong, H.-Y. Lee, S. Yu, Y.-S. Chen, Y. Wu, P.-S. Chen, B. Lee, F. T. Chen, and M.-J. Tsai, "Metal–Oxide RRAM," *Proceedings of the IEEE*, vol. 100, no. 6, pp. 1951–1970, 2012.

[4] E. Hsieh, M. Giordano, B. Hodson, A. Levy, S. Osekowsky, R. Radway, Y. Shih, W. Wan, T. F. Wu, X. Zheng, M. Nelson, B. Le, H.-S. Wong, S. Mitra, and S. Wong, "High-Density Multiple Bits-per-Cell 1T4R RRAM Array with Gradual SET/RESET and its Effectiveness for Deep Learning," in *2019 IEEE International Electron Devices Meeting (IEDM)*, 2019, pp. 35.6.1–35.6.4.

[5] H. Li, P. Huang, B. Gao, X. Liu, J. Kang, and H.-S. Philip Wong, "Device and Circuit Interaction Analysis of Stochastic Behaviors in Cross-Point RRAM Arrays," *IEEE Transactions on Electron Devices*, vol. 64, no. 12, pp. 4928–4936, 2017.

[6] S. Ambrogio, S. Balatti, V. McCaffrey, D. Wang, and D. Ielmini, "Impact of low-frequency noise on read distributions of resistive switching memory (RRAM)," in *2014 IEEE International Electron Devices Meeting*, 2014, pp. 14.4.1–14.4.4.

[7] H. Li, W.-C. Chen, A. Levy, C.-H. Wang, H. Wang, P.-H. Chen, W. Wan, W.-S. Khwa, H. Chuang, Y.-D. Chih *et al.*, "SAPIENS: A 64-kb RRAM-based non-volatile associative memory for one-shot learning and inference at the edge," *IEEE Transactions on Electron Devices*, vol. 68, no. 12, pp. 6637–6643, 2021.

[8] R. Liu, D. Mahalanabis, H. J. Barnaby, and S. Yu, "Investigation of single-bit and multiple-bit upsets in oxide RRAM-based 1T1R and crossbar memory arrays," *IEEE Transactions on Nuclear Science*, vol. 62, no. 5, pp. 2294–2301, 2015.

[9] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain." *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[10] A. Likas, "Probability density estimation using artificial neural networks," *Computer physics communications*, vol. 135, no. 2, pp. 167–175, 2001.

[11] Y. Zhao, J. Hu, P. Huang, F. Yuan, Y. Chai, X. Liu, and J. Kang, "A physics-based compact model for material-and operation-oriented switching behaviors of CBRAM," in *2016 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2016, pp. 7–6.

[12] A. Grossi, D. Walczyk, C. Zambelli, E. Miranda, P. Olivo, V. Stikanov, A. Feriani, J. Sune, G. Schoof, R. Kraemer *et al.*, "Impact of intercell and intracell variability on forming and switching parameters in RRAM arrays," *IEEE Transactions on Electron Devices*, vol. 62, no. 8, pp. 2502–2509, 2015.

[13] B. Q. Le, A. Levy, T. F. Wu, R. M. Radway, E. R. Hsieh, X. Zheng, M. Nelson, P. Raina, H.-S. P. Wong, S. Wong *et al.*, "RADAR: A fast and energy-efficient programming technique for multiple bits-per-cell RRAM arrays," *IEEE Transactions on Electron Devices*, vol. 68, no. 9, pp. 4397–4403, 2021.