This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TNANO.2019.2922838

A modeling methodology for Resistive RAM based on Stanford-PKU model with extended multilevel capability

John Reuben, Dietmar Fey, and Christian Wenger

***Abstract*—Modeling of Resistive RAMs (RRAMs) is a herculean task due to its non-linearity. While the exigent need for a model has motivated research groups to formulate realistic models, the diversity in RRAMs' characteristics has created a gap between model developers and model users. This paper bridges the gap by proposing an algorithm by which the parameters of a model are tuned to specific RRAMs. To this end, a physics-based compact model was chosen due to its flexibility, and the proposed algorithm was used to exactly fit the model to different RRAMs, which differed greatly in their material composition and switching behavior. Further, the model was extended to simulate multiple Low Resistance States (LRS), which is a vital focus of research to increase memory density in RRAMs. The ability of the model to simulate the switching from a high resistance state to multiple LRS was verified by measurements on 1T-1R cells.**

***Index Terms*—RRAM, physics-based models, multilevel modeling, SET/RESET process, Stanford model, memristor, resistive switching, 1T-1R**

I. Introduction

RESISTIVE RAMs (RRAMs) are two terminal devices capable of changing their resistance in response to an applied voltage. Initially RRAM was researched as an emerging Non-Volatile Memory (NVM) and a possible replacement to FLASH memory. In the recent past, RRAM has also extended its influence beyond memory to logic circuits/computing. The emergence of the RRAM as a NVM device which can compute, at a time when computer architects are facing the memory wall problem, has set the stage for RRAM to be efficiently deployed for in-memory computing. A new field called 'memristive logic' [1] has emerged, which is the methodology of designing logic circuits using RRAM as the primary computing device. Consequently, research in RRAM-based memories [2] and RRAM-based computing circuits [3] are very active and ever increasing. Such efforts need reliable models for RRAM to be used in SPICE simulation for predictive analysis, feasibility analysis and design space exploration. Considering the fact that RRAM is fabricated in a few labs, the majority of the RRAM community heavily depends on models for research.

Being a non-linear device, the RRAM is difficult to be modeled and formulating a realistic model which reproduces its behavior involves significant effort. According to [4], the fundamental reason why RRAM modeling is non-trivial is that one is attempting to solve an inverse problem in a complicated non-linear system. In spite of this, the exigent need for a model has motivated research groups and at the time of this writing, there are more than 15 distinct models for RRAM. A detailed survey of these models, elaborating their features and capabilities was carried out recently in [5]. While more models are being developed by device researchers, circuit designers and system architects who work at higher levels of abstraction, need to 'plug-in' a model for RRAM (with specific characteristics) without understanding the device physics $e.g.$ the properties of the switching oxide used. There exists a 'knowledge gap' in the RRAM research community with circuit/system designers perceiving the RRAM as a simple switch (between two resistive states), while device engineers perceiving the same as a field-directed movement of oxygen vacancies with certain stochasticity involved. RRAM model developers bridge this gap, to a certain extent, by describing the complex switching process in a compact model to be used by circuit/system designers. In Section II-A, we classify RRAM models according to the modeling philosophy and give a broad overview on RRAM modeling.

Although models bridge the gap between device and circuit/system designers, there still exists a barrier between model developers and model users. In Section II-B, we sample few recently fabricated RRAMs and highlight the diversity in their characteristics (like threshold voltage at which the RRAM switches its state, low resistance state (LRS), high resistance state (HRS), resistance window, $i.e.$ HRS/LRS ratio etc). We identify the need for a model to be able to simulate such diverse RRAM characteristics (Section II-C). In Section III-A, we justify why we chose the Stanford-PKU model to be that model which can simulate the heterogeneity in RRAMs. We briefly describe the Stanford-PKU RRAM model in Section III-B to familiarize the reader with the modeling approach used in this model and the associated parameters. Having described the key model features, we present our algorithm which takes the target RRAM's specification as

John Reuben and Dietmar Fey are with Chair of Computer Science 3 - Computer Architecture, Friedrich-Alexander-University (FAU) Erlangen-Nürnberg, Erlangen, Germany. (email:johnreubenp@gmail.com, dietmar.fey@fau.de)

Christian Wenger is with IHP – Leibniz-Institut für innovative Mikroelektronik, 15236 Frankfurt (Oder), Germany (e-mail: wenger@ihp-microelectronics.com).

Manuscript received December 21, 2018; revised 11 March, 2019 and 24 April, 2019. Copyright (c) 2019 IEEE. Personal use of this material is permitted. However, permission to use this material for any other other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org

Copyright (c) 2019 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TNANO.2019.2922838

input and tunes the parameters of the model to exactly fit the RRAM's characteristics (Section III-C).

Since data is stored as resistance of the switching oxide in RRAM (which, in turn depends on the geometry of the conductive filament formed), it is possible to store more than two states. Memory designers have rightly exploited this capability to increase memory density because, in principle, the storage of n bits per cell will result in n times increase in the storage density, achieving scaling of the silicon die area by 1/n [6]. Moreover, for novel computing paradigms like ternary computing, RRAMs are being explored for storing 'trits' and implementing ternary adders [7], [8]. In Section IV-A, we review three ways by which multiple states have been implemented in RRAM. Since implementing multiple LRS by varying the compliance current (also called multilevel SET process) in 1T-1R is a viable option, we investigate it further. We extend the Stanford-PKU model to model the multilevel SET process (Section IV-B). In Section IV-C, we contend the case for a composite 1T-1R model, in which the transistor and RRAM are modeled together as a single unit. We then propose a 1T-1R fitting algorithm, in which we enumerate the steps to fit the composite 1T-1R model to the behaviour of a fabricated 1T-1R structure (Section IV-D). Finally, we corroborate our modeling methodology on 1T-1R cells fabricated at IHP[1] in Section IV-E and confirm that our model indeed conforms to the 1T-1R cell's behavior. Section V concludes our work with some outlook for the future.

II. Progress in RRAM modeling and motivation for this work

A. RRAM models: Classification

Based on the modeling philosophy, RRAM models can be broadly classified as either 'physics-based' models or 'black-box' models [4]. In the former modeling philosophy, the physical properties of the device and the switching physics are understood and modeled by appropriate equations. In the 'black-box' approach (also called 'measurement' approach), the RRAM is modeled based on how it responds to different stimulus and the approach is agnostic to the device structure. The measured experimental data are formulated as mathematical equations, resulting in a model for RRAM. Physics-based models tend to be more accurate since they consider temperature related phenomenon like joule heating, which influence the resistance to which the RRAM is programmed during SET/RESET operation. Moreover, certain physics-based models not only capture switching physics ($i.e.$ SET/RESET operation), but also certain temporal characteristics like reliability and technology related characteristic like scaling [9].

Among physics-based models, the models can be further classified on the basis of their resolution/scale as Atomistic, Kinetic Monte Carlo (KMC), Finite Element Method (FEM) and Compact models [9]. While 'atomistic' models try to capture intricate details like ion/atom diffusion and migration mechanisms at atomistic scale (few nm^3), the 'compact' models capture macroscopic details like geometry of the conductive filament and temperature. Detailed classification of physics-based models is presented in [9]. The authors in [9] point out an important trade-off for circuit designers: the more detailed a physical model is, the higher the computational cost/simulation time. The simulation time of physics-based model increases drastically from 'compact' models to 'atomistic' models and hence accuracy and simulation time must be judiciously balanced.

[1]Institute for High Performance Microelectronics – Leibniz-Institut für innovative Mikroelektronik, Germany

TABLE I: A sample of recently fabricated RRAMs with their median characteristics – SET/RESET voltage in volts, LRS/HRS in Ω

Device	SET	RESET	LRS	HRS	Ref
$Pt/HfO_2(5nm)/Ti/TiN$	0.88	-0.5	3.65K	5.1M	[10]
$TiN/Hf_{1-x}Al_xO_y/Ti/TiN$	0.9	-1.07	6.66K	66.66K	IHP [11]
$Al/Ge/TaO_x(10nm)/Pt$	2	-0.96	826	37M	[12]
$Ti/SiO_2(5nm)/C$	2.4	-1.25	20K	100M	[13]

B. Diversity in RRAM characteristics

Different materials for the switching layer (metal oxide) and top/bottom electrodes have been explored by researchers to fabricate RRAMs with different characteristics. In Table I, we list four different RRAMs and their median characteristics (switching voltages for SET/RESET and the LRS/HRS). The table is not comprehensive and represents a sample of recently fabricated RRAMs. We deliberately chose different switching materials/electrodes to highlight the diversity in materials and their characteristics. One can easily observe that the RESET voltage varies from as low as -0.5 V to -1.25 V while the high resistance state (HRS) varies from 66.66 KΩ to as high as 100 MΩ, among a random sample of RRAMs. Each of these RRAMs have been optimized in some manner: the RRAM device in [11] have been optimized for less cell-to-cell variability (in the array) and post-programming instabilities; [12] optimizes cycle-to-cycle and cell-to-cell variability; [10], [12] have the advantage of a high resistance window and [13] reports less HRS variability in addition to a high resistance window. How can one model RRAMs with such varied attributes?

C. The need for a fitting algorithm

Developing a physics-based model and extracting the model parameters from the experimental data of the RRAM is a complex research problem. Generally, research groups develop a model and verify their model on a device fabricated in their lab. After calibrating their model on measurements, the model is released/published with a default set of parameters which are fitted to that particular device (for example, the model proposed in [14] was verified on Al-doped HfOx devices and released with corresponding parameters which match their device's characteristics. Similarly, the 'cone' model in [15] was verified on $Ti/ZrO_2/Pt$ devices and released with fitting parameters). Often, there arises a need to be able to use a

Copyright (c) 2019 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TNANO.2019.2922838

model (or a modeling philosophy) for a device other than the device on which the model was verified. A memory designer might want to simulate a memory array with two different RRAMs (one with high resistance window [12] and another with low SET/RESET voltages [10]) and compare the energy consumption during write/read operation. In the realm of memristive logic, certain logic operations need RRAM with specific characteristics. The implementation of boolean NOR gate in an array needs a RRAM with a SET voltage twice it's RESET voltage [16]. We argue that every RRAM manufacturer cannot formulate a model from scratch and one must be able to use an already developed model to mimic a specific RRAM's behavior. All this necessitates an algorithm/methodology by which a model can be fitted to a particular RRAM which suits the application.

III. Proposed algorithm to fit the Stanford-PKU RRAM model to a specific RRAM

A. Why Stanford-PKU RRAM model?

The Stanford-PKU is a physics-based, compact model developed for metal oxide bipolar RRAMs [14], [17]–[19]. The model was well characterized on HfO_2 and HfO_x/TiO_x bi-layer devices [14]. The resistive switching behavior is modeled by the growth (during SET) and rupture (during RESET) of the conductive filament. Since filamentary switching is believed to the switching phenomenon in both oxide-based RRAM and conductive bridge RRAM (also called Electrochemical Metallization (ECM)) [9], [20], the model is generic enough to model a wide variety of RRAMs. For a model to be flexible enough to simulate the switching characteristics of a wide variety of RRAMs, it needs to have 'structural stability' (i.e. the qualitative properties of the model do not change in response to small changes in model parameters [21]). In other words, the model parameters can be varied (rather tuned) without affecting the key qualitative property of the model i.e. resistive switching behavior. This observation forms the basis for choosing the Stanford-PKU model to formulate our generic fitting methodology. In [14], the developers of the model depict graphically, how a small change in each parameter influences the overall behavior of the model marginally. This was the starting point for our analysis and the proposed methodology. In addition to structural stability, the model can simulate other phenomena like non-linear switching kinetics and 'fading memory' observed in fabricated devices [22].

The Stanford-PKU model has a few limitations. Adapting the model to different devices involves significant effort and this holds true of any physics-based model for RRAM. The Stanford-PKU model does not incorporate Random Telegraph Noise (RTN) which is believed to affect 'read margin'. Furthermore, being a compact model, it cannot capture endurance and retention behaviour of RRAMs.

B. Stanford-PKU model in a nutshell

This model simplifies the resistive switching into the growth and rupture of a single dominant filament. The gap distance, g (between the tip of the filament and the counter electrode) is the crucial parameter which determines the resistive state. The parameter g is programmable between gap_{min} and gap_{max} with the device being in HRS at gap_{max} and in LRS at gap_{min} (default values of gap_{min} and gap_{max} are 0.2 nm and 1.8 nm). Fig. 1 lists the equations governing the resistive switching process. The key equation (shaded yellow) describes the current (I) through the RRAM as a function of voltage across (V) across it and the gap in the conductive filament (g). The current has an exponential dependence on g, which together with hyperbolic dependence on the V, implements the sudden increase (or decrease) of the current resulting in a transition to LRS (or HRS). The reader is referred to [14] for an elaborate description of the model.

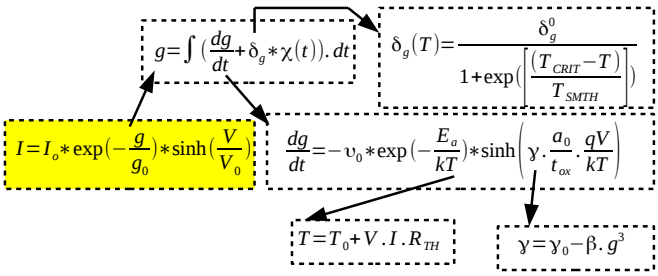


Fig. 1: Equations of Stanford-PKU RRAM model: Key equation (yellow) describes current through RRAM as a function of the voltage across it and filament gap. The other equations describe the evolution of the gap and the associated temperature change [14]

The parameters E_a (activation energy), a_0 (atomic spacing of the switching oxide), t_{ox} (thickness of the switching oxide), T_0 (environment temperature) and R_{TH} (thermal resistance) are determined by device structure, material properties and test environment. They can be easily obtained from the RRAM manufacturer since they are process parameters. δ_g^0 is the fitting parameter for variations in the gap. T_{CRIT} denotes the threshold temperature, above which significant variations in the gap occurs and T_{SMTH} is the variations smoothing parameter [14]. These three parameters (δ_g^0, T_{CRIT}, T_{SMTH}) are related to variations in switching phenomenon and are usually not relevant to model deterministic switching. The parameters – I_0, g_0, V_0, υ_0, γ_0 and β are called 'switching parameters' by the model developers, and they determine the median switching characteristics. The process parameters are dictated by the fabrication aspects of the device and are not tunable. Therefore, the six switching parameters are the key knobs to tune the model to a particular RRAM.

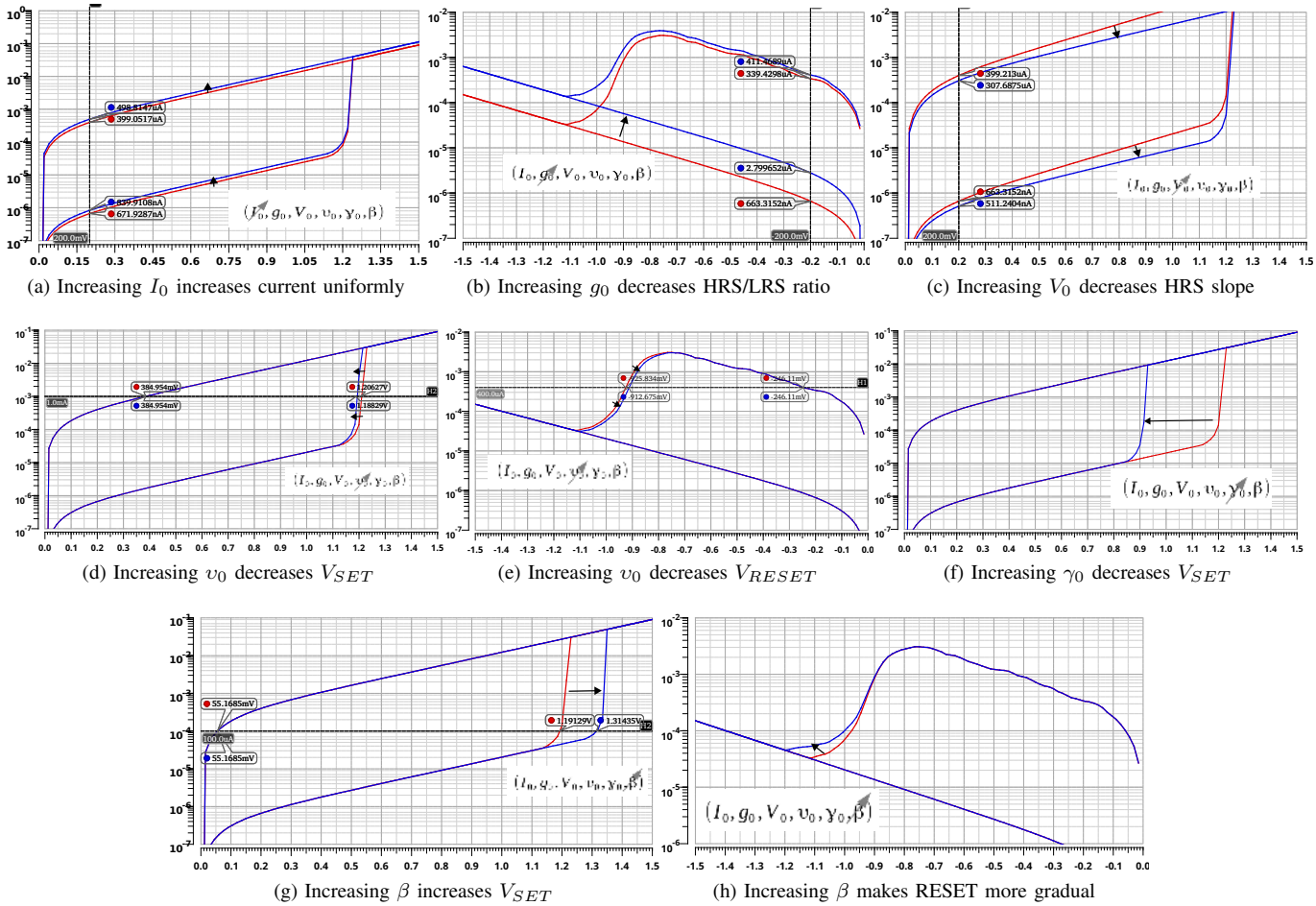
C. Fitting algorithm

By observing Fig. 1, one can decipher that tuning the six switching parameters to fit the model to a RRAM's specifications is a multi-faceted optimization problem. To do so with minimal effort, one must be aware of the following:

1) which switching parameter affects which aspect of the switching behavior? *i.e.* the role of each parameter in the resistive switching behavior.
2) to what extent? *i.e.* the degree to which a change in the switching parameter affects the overall switching behavior.

Copyright (c) 2019 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TNANO.2019.2922838



(a) Increasing I_0 increases current uniformly (b) Increasing g_0 decreases HRS/LRS ratio (c) Increasing V_0 decreases HRS slope

(d) Increasing υ_0 decreases V_{SET} (e) Increasing υ_0 decreases V_{RESET} (f) Increasing γ_0 decreases V_{SET}

(g) Increasing β increases V_{SET} (h) Increasing β makes RESET more gradual

Fig. 2: The effect of 25 % perturbation in the switching parameters on the switching behavior: V along x-axis and I along y-axis. The red curve corresponds to the default parameters and the blue curve corresponds to the particular parameter increased by 25%, while all other parameters remain the same. If the perturbation of a particular parameter affects both SET and RESET process equally, either of them is plotted. If no effect, it is not plotted (e.g. varying γ_0 has no effect on the RESET process)

TABLE II: The effect of 25% perturbation in the switching parameters of the model

	Quantitative	Qualitative (Predominant role)
I_0	I_{HRS} and I_{LRS} increase by 25%	scales current uniformly (Fig. 2-a)
g_0	I_{HRS} and I_{LRS} increase by 420% and 20%, respectively	determines resistance window i.e. HRS/LRS ratio (Fig. 2-b)
V_0	I_{HRS} and I_{LRS} both decrease by 23%	determines slope of HRS or LRS (Fig. 2-c)
υ_0	No change in current levels; V_{SET} and V_{RESET} decrease by 1.5 %	Since γ_0 has a stronger influence on V_{SET} , this parameter can be used to tune the voltage at which the device RESETs (Fig. 2-d, e)
γ_0	No change in current levels; V_{SET} decreases by 25%	determines the voltage at which the device SETs (Fig. 2-f)
β	No change in current levels; V_{SET} increases by 10% (Fig. 2-g); RESET process becomes more gradual	determines RESET curvature* (Fig. 2-h)

* the model implements gradual RESET while SET is abrupt and this is typical of RRAM devices

To investigate the correlation between the parameters and resistive switching behavior, we applied perturbations (of different degrees) to each of the six parameters and studied its effect on the I-V curves. When the parameters were perturbed simultaneously, the results were inconclusive and not useful. Therefore, we perturbed a single parameter (keeping other five constant) to find the predominant role of each parameter. The default values of each of these switching parameters (I_0 = $1e^{-3}$, g_0 = $2.5e^{-10}$, V_0 = 0.25, υ_0 = 10, γ_0 = 16, β = 0.8) were increased by 25%, one parameter at a time, and the resulting change in the I-V curves are plotted in Fig. 2 (t_{ox} was set to 8 nm to represent a typical RRAM). In Table II, we summarize the vital effects of the perturbation in each of the six switching parameters. Except υ_0, which influences both V_{SET} and V_{RESET} equally, all the parameters have a specific role. Understanding this role gives insights into the parameter needed to be perturbed/tuned to fit the model to a particular switching behavior. Next, we need to tune the switching parameters in a particular order. For example, tuning γ_0 (to fix V_{SET}) followed by β (to fix RESET curvature) will alter V_{SET} because β also influences V_{SET}. Moreover, tuning the parameters in a random order will delay the convergence process due to the large solution space. The amount to which a parameter is tuned is important since a larger step in tuning

Copyright (c) 2019 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TNANO.2019.2922838

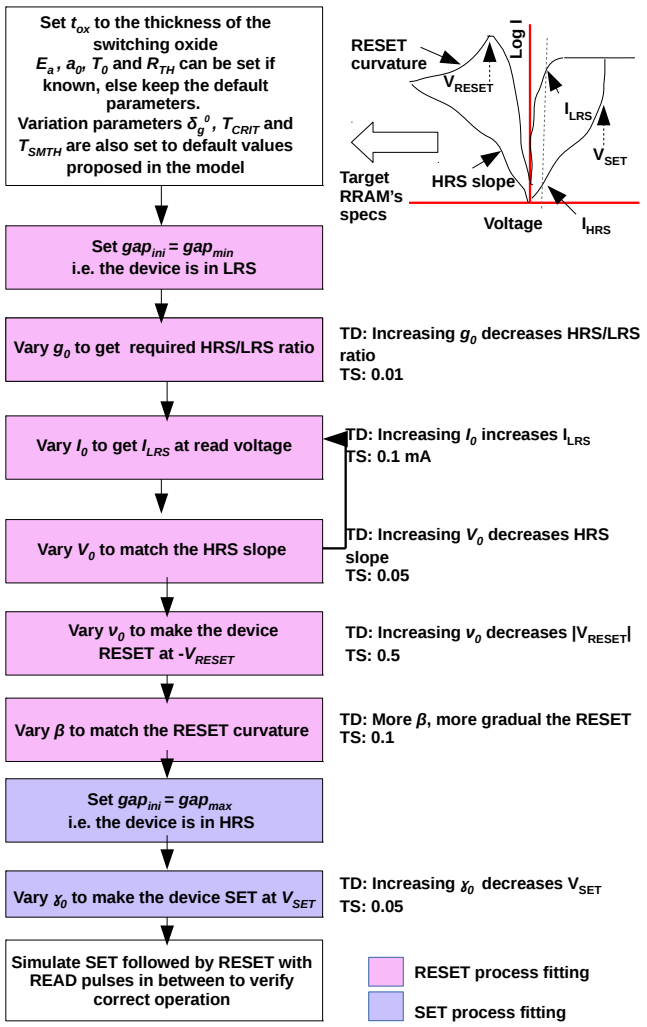
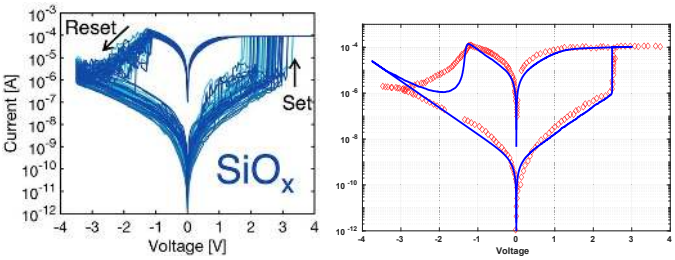


Fig. 3: Fitting Algorithm: In each step of the algorithm, 'TD' denotes 'Tuning Direction' to give direction to the tuning process and 'TS' denotes the 'Tuning Scale' to avoid unfeasible results

will result in unfeasible results like current in kA (due to exponential and hyperbolic dependence). Based on these insights, we propose an algorithm which efficiently fits the model to a specific RRAM with minimum effort in terms of simulation iterations and time to fit.

Starting with the default parameters released with the model, the algorithm described in Fig. 3 tunes the different parameters of the model to fit to a target RRAM. In simulation, the device must be set to an initial state, which is stable. This is implemented by the 'gap_{ini}' parameter which must be set to gap_{min} or gap_{max}. It is assumed that the RRAM to be simulated is already 'formed' and hence the initial state is known. We fit the RESET process first, due to the lack of a parameter which strongly influences V_{RESET}. A negative pulse greater than the target device's V_{RESET} is applied with 'READ' pulses before and after the RESET process. The order of tuning is $g_0 \rightarrow I_0 \rightarrow V_0$ for the RESET process. This is because g_0 has the largest influence on the current levels, followed by I_0 and V_0. When tuning I_0, if the device reads out I_{LRS} at LRS, it will read out I_{HRS} at HRS since g_0 is already tuned. Since varying V_0 disturbs the I_{LRS} fixed in the previous step, I_0 is tuned again after V_0. Once the base current levels are tuned to match the measured IV curves, we proceed to fit the voltage at which the device RESETs. It must be noted that v_0 must be tuned to match the voltage at which the device **starts** to RESET and β must be tuned to match the gradual nature of the RESET process. The fitting of the SET process is simpler and requires the tuning of only γ_0 to fix V_{SET} (the current levels during SET process are already tuned since I_0, g_0, V_0 affect both SET and RESET equally). Since γ_0 has no effect on the RESET process, the tuning of γ_0 towards the end does not 'disturb' the RESET curve already fitted. Since the RESET curvature and the slope of HRS/LRS are usually not quantified, V_0 and β are tuned till the simulated curves and measurement curves match sufficiently. This algorithm was used to fit the model to different RRAMs we considered in Table I. The simulated curves could not be overlaid on the measured IV curves due to multiple curves published by authors, showing cycle-to-cycle switching on the same device. As a sample, we reproduce the multiple curves published for $Ti/SiOx/C$ device in Fig. 4-(a). We extracted the mean values and overlaid them on the model curve to highlight the fitting obtained in simulation (Fig. 4-(b)). The simulated waveforms of all the RRAMs of Table I, during positive and negative voltages are graphed in Fig. 5 and Fig. 6 (the fitting of IHP's RRAM is performed in Section IV-E). The Stanford-PKU RRAM model is flexible enough to model RRAMs of different materials (HfO_x, TaO_x, SiO_x) and different oxide thickness (5 nm $< t_{ox} <$ 10 nm). The tuned parameters are tabulated in Table III to make the curves reproducible. However, this fitting algorithm couldn't be used to fit RRAMs with t_{ox} as high as 60 nm [23] and as low as 3 nm [24]. To fit those RRAMs, the algorithm or the model itself may need to be significantly enhanced.



(a) Measured IV curves for 50 cycles: Reproduced from Ref. [13] with permission from the Royal Society of Chemistry

(b) Mean curve of (a) extracted (red) and overlaid on simulated curve (blue)

Fig. 4: The Stanford-PKU model fitted to $Ti/SiOx/C$ device shows good correspondence

IV. Extending Stanford-PKU RRAM model to model multiple low resistance states

A. Multi-Level Cell storage (MLC) in RRAMs

In RRAMs, MLC can be achieved in the following ways [25]:

Copyright (c) 2019 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TNANO.2019.2922838

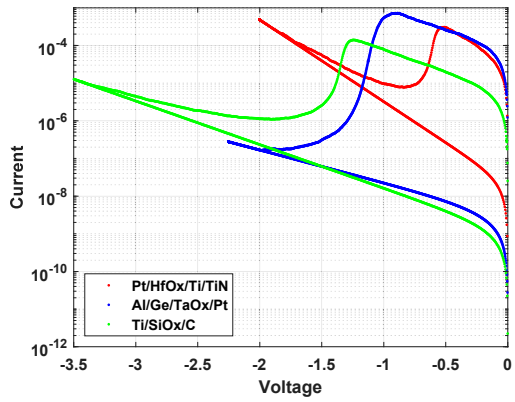


Fig. 5: Using the proposed algorithm, the model was fitted to RRAMs with different RESET voltages, RESET curvatures, HRS/LRS ratio, switching oxides and thickness

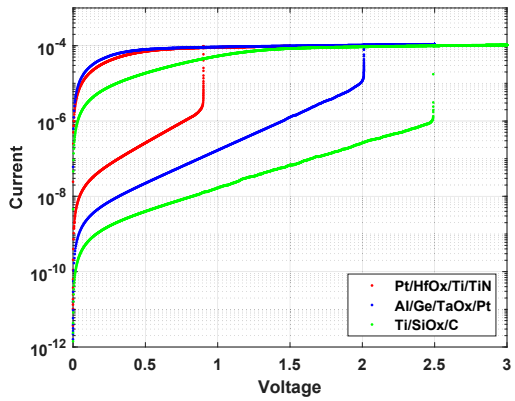


Fig. 6: Using the proposed algorithm, the model was fitted to RRAMs with different SET voltages. A transistor was used to enforce a Current Compliance (CC) of 100 μA

1) By varying compliance current (also called multilevel SET): from the initial HRS, the RRAM is programmed to different LRS during the SET process.
2) By varying reset voltage (also called multilevel RESET): from the initial LRS, the RRAM is programmed to different HRS during the RESET process.
3) By varying programming pulse widths (not popularly pursued due to being energy inefficient).

In arrays, multilevel SET is achieved in 1T-1R by varying the gate voltage, which in turns varies the compliance (drain) current. The physical phenomenon behind this process is believed to be the formation and subsequent widening of the conductive filament with increasing drain current [26], [27]. Multilevel RESET, which is implemented by varying the maximum voltage applied during RESET can also be implemented in passive arrays and 1S-1R (RRAM fabricated in series with a 'Selector' having bi-directional diode-like characteristics) configuration. The physical phenomenon is believed to be the larger gap (between the tip of the conductive filament and bottom electrode) with increasing reset voltage, $i.e.$ the device goes to a deeper RESET with higher reset voltage [25]. Consequently, V_{SET} also increases for the subsequent SET operation because more energy is needed to form the conductive filament again. Therefore, implementing multilevel RESET in an array will necessitate a peripheral circuitry capable of applying different voltages (V_{SET1}, V_{SET2}). Moreover, HRS variability is more compared to LRS variability in RRAM arrays [24], [28]. Since implementing multiple states narrows the separation between neighboring states, multilevel SET may be preferred due to lower variability in LRS, and consequently an error-free sensing circuitry to distinguish between the different LRS states.

TABLE III: The parameters of the model corresponding to different RRAMs. Only t_{ox} and the parameters tuned in the fitting algorithm are tabulated, the remaining parameters are default values from the model release. δ_g^0=0.005 to minimize variations

Device	g_0	I_0	v_0	β	γ_0	V_0
$Pt/HfO_2(5nm)/Ti/TiN$	$2.176e^{-10}$	$0.17e^{-3}$	10.5	2.1	20.75	0.2
$Al/Ge/TaO_x(10nm)/Pt$	$1.495e^{-10}$	$1.04e^{-3}$	15	1.5	12.15	0.25
$Ti/SiO_2(5nm)/C$	$1.8525e^{-10}$	$0.374e^{-4}$	$1e^{-9}$	1.8	18	0.375

B. Proposed modification to Stanford-PKU RRAM model

The Stanford-PKU model, as presented in [14], has the capability to simulate multilevel RESET process. However, it does not have the capability to simulate multilevel SET process [29]. An enhanced version of the Stanford-PKU model (beta version [29]) reported to model multilevel SET process by incorporating the width of the conductive filament. But the number of 'switching parameters' in this model doubled (due to the inclusion of the radius of the filament), making it difficult to tune the parameters to fit different RRAMs, in a deterministic manner. Therefore, to retain its flexibility and augment it with capability to simulate multilevel SET process, a modification to the model proposed in [14] was necessary.

As already stated, the gap, g, is the state variable which decides the resistance to which the RRAM is programmed ($gap_{min} < g < gap_{max}$). During SET process, g changes from gap_{max} (HRS) to gap_{min} (LRS). In the original model [14], gap_{min} is a constant and hence the device goes from HRS to a single LRS, $i.e.$ binary switching. The key modification to the model was to make gap_{min} a variable. The LRS to which the 1T-1R cell is programmed is a strong function of the gate voltage, $i.e.$ higher the gate voltage, higher the drain current and consequently a thicker filament or a lower LRS. Furthermore, the inverse relation between the LRS and gate voltage has a technology dependence, $i.e.$ drive strength. The W/L ratio of the transistor accounts for this drive strength.

$$LRS \propto \frac{1}{V_{gate}} \tag{1}$$

$$\text{Therefore,} \quad gap_{min} = K_{th} \cdot \frac{(\frac{W}{L})}{V_{gate}} + C \tag{2}$$

where K_{th} and C are fitting constants for a particular 1T-1R cell. Intuitively, K_{th}, being the slope, is a measure of how

Copyright (c) 2019 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TNANO.2019.2922838

fast the filament grows laterally. Thickness co-efficient, K_{th} and C can be calculated from measurements, as demonstrated in Section IV-E. Since gap_{min} is no more a constant, the modified model could simulate the transition from a HRS to multiple LRS.

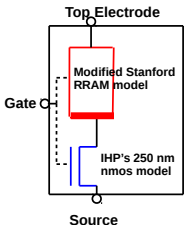


Fig. 7: Composite 1T-1R model

C. Composite 1T-1R model

The multilevel SET process is implemented by varying the compliance current, $i.e.$ the drain current of the transistor. Since this can be implemented only in 1T-1R structure, it is justifiable to create a composite 1T-1R model. In 1T-1R structure, during the SET process, the drain current (which is decided by the voltage at the gate terminal of the NMOS transistor) influences the LRS to which the RRAM is programmed. Although the transistor and RRAM are modeled separately, they must be 'synchronized' together to mimic the 1T-1R behavior. This is because, in practice, the transistor is integrated with the RRAM (the bottom electrode is fused with the drain terminal) and neither the transistor nor RRAM can be characterized separately after fabrication. Moreover, enforcing an external compliance current is not recommended because of the latency involved (the current abruptly rises during SET, destroying the RRAM) and therefore, current compliance must be implemented inherently in the RRAM by integrating it with a transistor. The composite 1T-1R modeling approach (Fig. 7) is based on the premise that if RRAM and transistor are fabricated together and characterized together as one unit, they need to be modeled together.

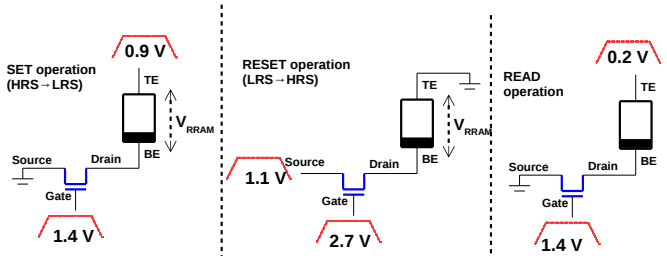


Fig. 8: Two pulses are applied simultaneously (one at the Gate and other at Top Electrode (TE)/Source to program and read from IHP's 1T-1R devices

D. 1T-1R fitting algorithm

Given 1T-1R measurements from a lab, we propose the following algorithm to fit the composite 1T-1R model to it.

TABLE IV: Multilevel SET process in IHP's 1T-1R devices

State transition	Voltages	Read-out current at 0.2 V
(LRS1,2,3) → HRS	2.7 V at gate, 1.1 V at source, TE grounded	3μA i.e. 66.66 KΩ
HRS → LRS1	1.2 V at gate, 0.9 V at TE, source grounded	20μA i.e. 10 KΩ
HRS → LRS2*	1.4 V at gate, 0.9 V at TE, source grounded	30μA i.e. 6.66 KΩ
HRS → LRS3	1.6 V at gate, 0.9 V at TE, source grounded	40μA i.e. 5 KΩ

* denotes the default LRS during binary switching

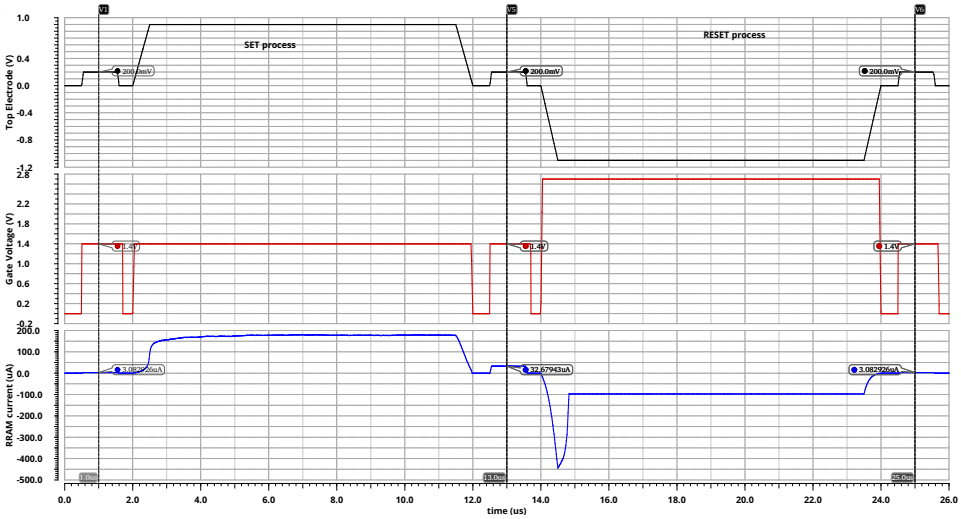
1) From 1T-1R measurements, extract V_{RRAM} (Fig. 8) during SET and RESET process by DC analysis (this is performed by replacing RRAM with a fixed resistor).
2) The extracted V_{RRAM} values, LRS, HRS values, I_{LRS}, I_{HRS} are fed to the algorithm presented in Section III-C to fit the Stanford-PKU model to the RRAM.
3) Scale gap_{min} to a higher value to have enough space around it for neighboring LRS.
4) Calculate gap_{min} corresponding to different LRS with other parameters fixed as in step 2.
5) Calculate K_{th} and C by plotting gap_{min} as a function of transistor's gate voltage.
6) Create a composite model with modified Stanford-PKU RRAM model and NMOS model file as depicted in Fig. 7.
7) Simulate SET and RESET process and tune I_0 and g_0 if needed, to fit measurement results.

E. Model corroboration

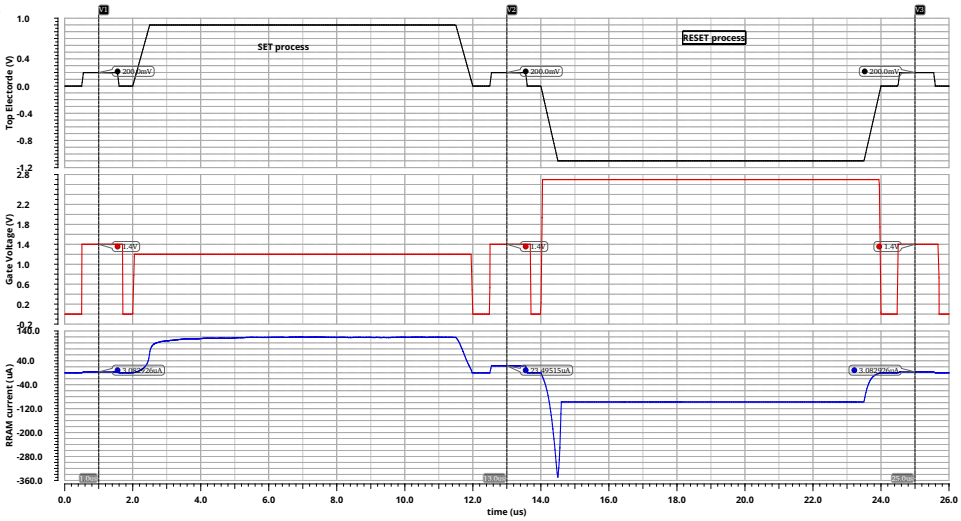
To corroborate our modeling approach, we used the 4k bit 1T-1R array fabricated at IHP [11]. The 1T-1R is constituted by a NMOS transistor manufactured in IHP's 0.25 μm CMOS technology, whose drain is connected in series to the RRAM. The RRAM is a Metal-Insulator-Metal ($TiN/Hf_{I-x}Al_xO_y/Ti/TiN$) stack integrated on the metal line 2 of the CMOS process. The pulses used for SET, RESET and READ process are illustrated in Fig. 8 and the duration of all pulses were 10 μs to maximize switching yield. Reader is referred to [11] for detailed electrical characterization. To achieve multiple LRS, the gate voltage was varied as shown in Table IV. Gate voltage of 1.4 V was considered the default LRS for binary switching and it resulted in a HRS/LRS of 10. When gate voltage was increased to 1.6 V, the corresponding increase in the drain current programmed the RRAM to an even lower LRS of 5 KΩ. Similarly, when the gate voltage was reduced from 1.4 V to 1.2 V, the decrease in drain current programmed the RRAM to higher LRS of 10 KΩ. In the former case, the conductive filament formed was thicker due to the increased drain current and in the latter case, it was thinner because of the decreased drain current. With the aforementioned measurements (Table IV, read-out currents are average values from cell-to-cell measurements in 4K bit array), the generalized 1T-1R fitting algorithm presented in Section IV-D was used to fit the composite 1T-1R model to IHP's 1T-1R cells.

Copyright (c) 2019 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

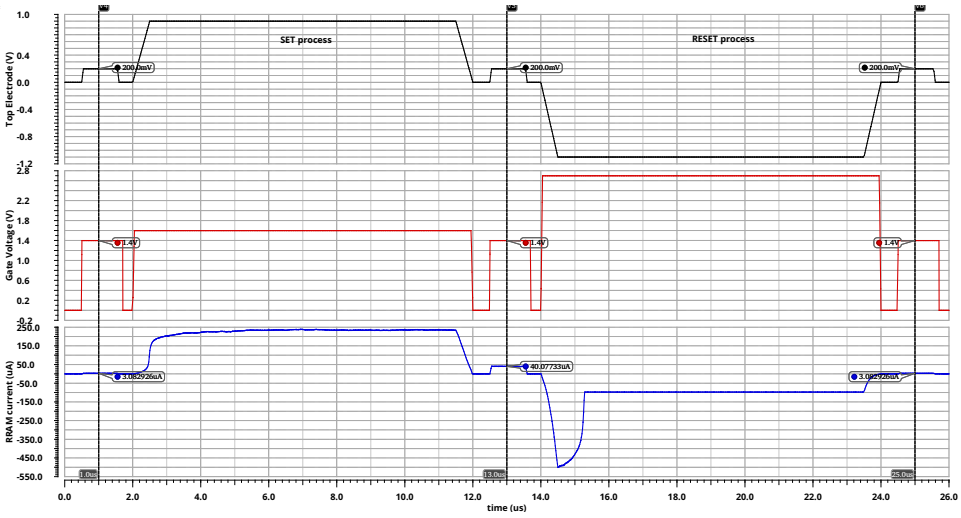
This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TNANO.2019.2922838



(a) Gate Voltage of 1.4 V during SET process programs the RRAM to LRS2 (Read-out current of 32 μA)



(b) Gate Voltage of 1.2 V during SET process programs the RRAM to LRS1 (Read-out current of 23 μA)



(c) Gate Voltage of 1.6 V during SET process programs the RRAM to LRS3 (Read-out current of 40 μA)

Fig. 9: Composite 1T-1R model simulates IHP's 1T-1R characteristic with multilevel capability. Irrespective of the LRS it is programmed to during the SET process, the RRAM resets to the same HRS of 3 μA

Copyright (c) 2019 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TNANO.2019.2922838

As already stated, the measurements were obtained from 1T-1R structure, and it was not possible to obtain the voltage drop across the individual RRAM (denoted 'V_{RRAM}' in Fig. 8). By replacing the RRAM with 66.66 KΩ resistor (at the start of SET process, the RRAM is in HRS), a DC analysis of the 1T-1R circuit was performed to confirm that the transistor was operating in the triode region. In triode region, the transistor behaves as a linear resistor of resistance,

$$r_{DS} = \frac{1}{k'_n.\frac{W}{L}.V_{GS} - V_t} \tag{3}$$

For IHP's 250 nm process parameters, this drain–to–source resistance (r_{DS}) was calculated to be 544 Ω. Consequently, the voltage drop across the transistor was calculated to be negligible (6.8 mV) and most of the voltage applied at TE appeared across the RRAM at the start of the SET process i.e. $V_{RRAM} = 0.89$ V. A similar analysis was carried out to calculate the voltage drop across RRAM at the start of the RESET process (by replacing RRAM with 6.66 KΩ resistor followed by DC analysis) and V_{RRAM} was found to be 1.03 V. Thus, V_{SET} of 0.89 V and V_{RESET} of -1.03 V was fed to the fitting algorithm presented in Fig. 3 along with other values (I_{LRS} of 30 μA and I_{HRS} of 3 μA) to arrive at an initial fitting for IHP's standalone RRAM. It must be noted that this initial fitting can only simulate binary switching, $i.e.$ from a single HRS to a single LRS and vice versa. To accommodate the fitting of multiple LRS, the binary switching is fitted to the central LRS of 6.66 KΩ. Although the end result is a composite 1T-1R model, such an initial fitting (to RRAM alone) is necessary to start with a good initial solution (just as in simulated annealing optimization, a good initial solution is necessary to guarantee faster convergence).

In the initial fitting to IHP's individual RRAM, gap_{min} was 0.2 nm. This value is too small to accommodate other LRS around it. So, gap_{min} was scaled to 1 nm and the model parameters were retuned to match HRS/LRS currents. Next, the individual RRAM is simulated with different gap_{min} and the exact gap_{min} corresponding to LRS currents of 20 μA, 30 μA and 40 μA were calculated as 1.15 nm, 1 nm and 0.89 nm, respectively. The W/L of transistors was 4.75 (1140 nm/240 nm). Substituting these gap_{min} and the corresponding gate voltages in Equation 2, a line was fitted and K_{th} and C were found to be $2.589e^{-10}$ and 1.21 e^{-10}. Therefore,

$$gap_{min} = 2.589e^{-10}.\frac{(\frac{W}{L})}{V_{gate}} + 1.21e^{-10} \tag{4}$$

The Verilog-A code of Stanford-PKU model was modified to reflect this change and connected with 250 nm NMOS model file of IHP to create a composite 1T-1R model capable of simulating IHP's multilevel SET process. Fig. 9 shows three wave-forms from simulation, where the SET process was followed by a RESET process in a single simulation, with READ process in between them. The gate voltage alone was varied during the SET process to implement multiple LRS. The model parameters used in simulation are provided in Table V.

TABLE V: Fitted model parameters for IHP's 1T-1R cells

E_a = 0.6	a_0 = 2.5 e^{-10}	t_{ox} = 6 e^{-9}	T_0 = 298 K
R_{TH} = 1500	I_0 = 8.54 e^{-4}	g_0 = 0.346 e^{-9}	V_0 = 0.26
υ_0 = 0.05	γ_0 = 19.5	β = 0.4	δ_g^0 = 0.005
T_{CRIT} = 450	T_{SMTH} = 500	gap_{max} = $18.8e^{-10}$	gap_{min} = Eq. 4

NMOS model parameters: t_{ox} = 10 nm, W = 1.14 μm, L = 0.24 μm, V_t = 0.6 V

V. Conclusion and Future Work

In this work, we have proposed a modeling methodology in which we demonstrate how to take a model with its default parameters (usually published with the model) and tune the parameters to a specific device. We chose the Stanford-PKU RRAM model due to its versatility and demonstrated its capability to model filamentary switching RRAMs of different materials and characteristics. The presented methodology can be automated, which is an important direction of future work. The fitting algorithm can be used by circuit designers and computer architects to fit the model to a device with no a priori knowledge of RRAM switching mechanisms and associated device physics. Although we restricted ourselves to one model in this paper, our modeling methodology is generic and can be summarized in three steps: perturb the parameters of a model; understand its effect on the resistive switching behavior; tune the parameters accordingly. Therefore, in principle, the presented modeling methodology can be adopted for any RRAM model with good structural stability. Consequently, our approach will relieve RRAM based system designers of modeling effort, in future. Since multilevel storage is an emerging focus of RRAM research, we extended the Stanford-PKU RRAM model to model multilevel SET process. The composite 1T-1R model is able to simulate the switching from a single HRS to multiple LRS states. The modeling approach was verified on IHP's 1T-1R cells and could model four distinct states.

Acknowledgment

This research was funded by Deutsche Forschungsgemeinschaft (DFG) -Integrierte Memristor-Basierte Rechner-Architekturen (IMBRA) (Project number 389549790). The authors would like to thank Zizhen Jiang, Stanford University Nanoelectronics lab for useful discussions.

References

[1] J. Reuben, R. Ben-Hur, N. Wald, N. Talati, A. Ali, P.-E. Gaillardon, and S. Kvatinsky, "Memristive logic: A framework for evaluation and comparison," in *Power And Timing Modeling, Optimization and Simulation (PATMOS)*, September 2017, pp. 1–8.

[2] T.-C. Chang, K.-C. Chang, T.-M. Tsai, T.-J. Chu, and S. M. Sze, "Resistance random access memory," *Materials Today*, vol. 19, no. 5, pp. 254 – 264, 2016.

[3] D. Ielmini and H.-S. P. Wong, "In-memory computing with resistive switching devices," *Nature Electronics*, vol. 1, pp. 333 – 343, 2018.

[4] R. S. Williams and M. D. Pickett, *The Art and Science of Constructing a Memristor Model*. New York, NY: Springer New York, 2014, pp. 93–104.

Copyright (c) 2019 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TNANO.2019.2922838

[5] D. Panda, P. P. Sahu, and T. Y. Tseng, "A collective study on modeling and simulation of resistive random access memory," *Nanoscale Research Letters*, vol. 13, no. 1, p. 8, Jan 2018.
[6] X. Yang, A. B. K. Chen, B. Joon Choi, and I.-W. Chen, "Demonstration and modeling of multi-bit resistance random access memory," *Applied Physics Letters*, vol. 102, no. 4, p. 043502, 2013. [Online]. Available: https://doi.org/10.1063/1.4790158
[7] D. Wust, D. Fey, and J. Knödtel, "A programmable ternary CPU using hybrid cmos/memristor circuits," *IJPEDS*, vol. 33, no. 4, pp. 387–407, 2018.
[8] D. Fey, M. Reichenbach, C. Söll, M. Biglari, J. Röber, and R. Weigel, "Using memristor technology for multi-value registers in signed-digit arithmetic circuits," in *Proceedings of the Second International Symposium on Memory Systems, MEMSYS 2016, Alexandria, VA, USA, October 3-6, 2016*, 2016, pp. 442–454.
[9] D. Ielmini and V. Milo, "Physics-based modeling approaches of resistive switching devices for memory and in-memory computing applications," *J. Comput. Electron.*, vol. 16, no. 4, pp. 1121–1143, Dec. 2017.
[10] J. Sandrini, "Fabrication, characterization and integration of resistive random access memories," p. 241, 2017.
[11] E. Prez, A. Grossi, C. Zambelli, P. Olivo, R. Roelofs, and C. Wenger, "Reduction of the cell-to-cell variability in hf1-xalxoybased rram arrays by using program algorithms," *IEEE Electron Device Letters*, vol. 38, no. 2, pp. 175–178, Feb 2017.
[12] V. Y. Zhuo, Y. Jiang, R. Zhao, L. P. Shi, Y. Yang, T. C. Chong, and J. Robertson, "Improved switching uniformity and low-voltage operation intao_x-based rram using ge reactive layer," *IEEE Electron Device Letters*, vol. 34, no. 9, pp. 1130–1132, Sept 2013.
[13] E. Ambrosi, A. Bricalli, M. Laudato, and D. Ielmini, "Impact of oxide and electrode materials on the switching characteristics of oxide reram devices," *Faraday Discuss.*, pp. –, 2018.
[14] Z. Jiang, Y. Wu, S. Yu, L. Yang, K. Song, Z. Karim, and H. . P. Wong, "A compact model for metaloxide resistive random access memory with experiment verification," *IEEE Transactions on Electron Devices*, vol. 63, no. 5, pp. 1884–1892, May 2016.
[15] G. Gonzlez-Cordero, J. B. Roldan, F. Jimnez-Molinos, J. Su, S. Long, and M. Liu, "A new compact model for bipolar rrams based on truncated-cone conductive filamentsa verilog-a approach," *Semiconductor Science and Technology*, vol. 31, no. 11, p. 115013, 2016. [Online]. Available: http://stacks.iop.org/0268-1242/31/i=11/a=115013
[16] J. Reuben, R. Ben-Hur, N. Wald, N. Talati, A. H. Ali, P.-E. Gaillardon, and S. Kvatinsky, "A taxonomy and evaluation framework for memristive logic," in *Handbook of Memristor Networks*, A. Adamatzky, L. Chua, and G. Sirakoulis, Eds. Springer International Publishing, 2019.
[17] X. Guan, S. Yu, and H. . P. Wong, "A spice compact model of metal oxide resistive switching memory with variations," *IEEE Electron Device Letters*, vol. 33, no. 10, pp. 1405–1407, Oct 2012.
[18] S. Yu, B. Gao, Z. Fang, H. Yu, J. Kang, and H. . P. Wong, "A neuromorphic visual system using rram synaptic devices with sub-pj energy and tolerance to variability: Experimental characterization and large-scale modeling," in *2012 International Electron Devices Meeting*, Dec 2012, pp. 10.4.1–10.4.4.
[19] Z. Jiang, S. Yu, Y. Wu, J. H. Engel, X. Guan, and H. . P. Wong, "Verilog-a compact model for oxide-based resistive random access memory (rram)," in *2014 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, Sept 2014, pp. 41–44.
[20] S. Menzel, "Comprehensive modeling of electrochemical metallization memory cells," *Journal of Computational Electronics*, vol. 16, no. 4, pp. 1017–1037, Dec 2017.
[21] L. Chua, "Device modeling via nonlinear circuit elements," *IEEE Transactions on Circuits and Systems*, vol. 27, no. 11, pp. 1014–1044, November 1980.
[22] S. Menzel, A. Siemon, A. Ascoli, and R. Tetzlaff, "Requirements and challenges for modelling redox-based memristive devices," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2018, pp. 1–5.
[23] Y. Huang, Z. Shen, Y. Wu, M. Xie, Y. Hu, S. Zhang, X. Shi, and H. Zeng, "Cuo/zno memristors via oxygen or metal migration controlled by electrodes," *AIP Advances*, vol. 6, no. 2, p. 025018, 2016. [Online]. Available: https://doi.org/10.1063/1.4942477
[24] A. Bricalli, E. Ambrosi, M. Laudato, M. Maestro, R. Rodriguez, and D. Ielmini, "Resistive switching device technology based on silicon oxide for improved onoff ratiopart i: Memory devices," *IEEE Transactions on Electron Devices*, vol. 65, no. 1, pp. 115–121, Jan 2018.
[25] A. Prakash and H. Hwang, "Multilevel cell storage and resistance variability in resistive random access memory," *Physical Sciences Reviews*, vol. 1, no. 6, pp. –, 2016.
[26] A. Prakash, J. Park, J. Song, J. Woo, E. Cha, and H. Hwang, "Demonstration of low power 3-bit multilevel cell characteristics in a taox-based rram by stack engineering," *IEEE Electron Device Letters*, vol. 36, no. 1, pp. 32–34, Jan 2015.
[27] U. Russo, D. Kamalanathan, D. Ielmini, A. L. Lacaita, and M. N. Kozicki, "Study of multilevel programming in programmable metallization cell (pmc) memory," *IEEE Transactions on Electron Devices*, vol. 56, no. 5, pp. 1040–1047, May 2009.
[28] A. Grossi, E. Perez, C. Zambelli, P. Olivo, and C. Wenger, "Performance and reliability comparison of 1t-1r rram arrays with amorphous and polycrystalline hfo2," in *2016 Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS)*, Jan 2016, pp. 80–83.
[29] H. Li, Z. Jiang, P. Huang, Y. Wu, H. . Chen, B. Gao, X. Y. Liu, J. F. Kang, and H. . P. Wong, "Variation-aware, reliability-emphasized design and optimization of rram using spice model," in *2015 Design, Automation Test in Europe Conference Exhibition (DATE)*, March 2015, pp. 1425–1430.

John Reuben received B.E (Hons) degree from BITS, Pilani in 2004 and Masters and PhD from VIT University, India in 2008 and 2015, respectively. He was a post-doctoral researcher in Technion, Israel from January 2017-January 2018 where he was a recipient of Viterbi fellowship. He is currently working as a post-doctoral researcher in Friedrich Alexander University, Erlangen, Germany. His research interests are Resistive RAMs, memristive logic and beyond-CMOS computing.

Dietmar Fey is a full Professor of Computer Science at Friedrich-Alexander-University Erlangen-Nürnberg (FAU) where he leads the Chair for Computer Architecture since 2009. Before he was associate professor from 2001-2009 for Computer Engineering at University Jena. His research interests are in parallel computer architectures, memristive computing, parallel programming environments, and embedded systems. He was and is involved in several national and international research projects and initiatives on parallel and embedded computing. He authored or co-authored over 140 articles including 3 books, and about 20 papers in journals. He is a member of German Computer Society and and of HiPEAC (European Network of Excellence on High Performance and Embedded Architecture and Compilation) and a contributor for the HiPEAC roadmap.

Christian Wenger received the Diploma degree in physics from the University of Konstanz in 1995 and the Ph.D. degrees from the Technical University of Dresden in 2000 and 2009. Since 2002, he has been with Innovations for High Performance Microelectronics, where he is currently involved in the field of functional devices for medical and space applications. He has authored or co-authored more than 170 papers and holds eight patents. He received the Microelectronics for Medical Engineering Professorship from the Brandenburg Medical School Theodor Fontane in 2018.

Copyright (c) 2019 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.