IEEE Journal on Exploratory Solid-State Computational Devices and Circuits

Received 4 May 2023; revised 24 August 2023; accepted 29 August 2023.
Date of publication 4 September 2023; date of current version 5 December 2023.

Digital Object Identifier 10.1109/JXCDC.2023.3311899

Modeling of Bilayer Modulated RRAM and Its Array Performance for Compute-in-Memory Applications

JIA-WEI LEE, TZU-CHIN CHOU, PO-AN CHEN, and MENG-HSUEH CHIANG (Senior Member, IEEE)

Department of Electrical Engineering, National Cheng Kung University, Tainan 701, Taiwan
CORRESPONDING AUTHOR: J.-W. LEE (Q78051011@gs.ncku.edu.tw).

This work was supported in part by the National Science and Technology Council, Taiwan, under Grant 111- 2221-E-006-225, in part by the Higher Education Sprout Project from National Cheng Kung University (NCKU), in part by the National Center for High-Performance Computing, Taiwan, in part by the Semiconductor Research Institute, and in part by NCKU QFort Center.

ABSTRACT This article presents a modified compact model of resistive random access memory (RRAM) with a tunneling barrier. The bilayer modulated RRAM can be integrated into a higher density array, reducing leakage current in standby mode. The model demonstrates current transition behavior from low- to high-bias regions by considering both bulk-limited and electrode-limited transport mechanisms. This model can evaluate RRAM array performance under various pulsing conditions and device parameter variations with calibrated model cards. The compute-in-memory application requires precise current sum results hindered by the wire resistance loading effect. This study also evaluates various sizes of arrays suitable for performance improvement.

INDEX TERMS Bilayer resistive random access memory (RRAM), compact model, computing-in-memory, interconnect resistance, RRAM.

I. INTRODUCTION

As the semiconductor industry continues to push the scaling of MOS transistors, information technology has proliferated, mainly improving the application of artificial intelligence (AI) through artificial neural networks (ANNs). However, the limitations of traditional von Neumann machines have reached a bottleneck in terms of data transfer between memory and arithmetic units. Additionally, power consumption per chip area has become increasingly challenging to scale down. Moreover, due to their high computational requirements, AI applications are often associated with high energy consumption and heat generation, which can limit their performance. Emerging memories have shown their potential to solve these problems by mimicking the neuron operation in the human brain. Therefore, studying memory cell characteristics and array performance, also known as device-circuit interaction (DCI), is crucial for technological development.

In recent years, there has been extensive research on the bilayer transition metal oxide (TMO) structure of resistive random access memory (RRAM), including HfOx/AlOx, HfOx/TiOx, and TiOx/TaOx. These bilayer RRAMs, with different material combinations, show promising results as they increase the nonlinearity of current, making them advantageous for large-array applications due to their low standby leakage current. Several studies have reported that the inserted dielectric layer significantly increases the nonlinear factor at read voltage compared to devices with only switching oxide [1], [2], [3].

This article introduces a novel 2-D material, hexagonal boron nitride (h-BN), which acts as an ultrathin tunneling barrier (TB) for bilayer modulated RRAM and is an excellent insulating layer for RRAM applications [4]. By combining transport mechanisms, the characteristics of bilayer RRAM can be captured and applied to DCI studies, further improving AI applications.

II. EXPERIMENT

To maintain the original function of the hBN layer, a 150-nm-thick titanium nitride (TiN) layer was sputtered onto a SiO_2/Si substrate as the bottom electrode (BE). TiN, with a proper work function (WF) of 4.45 eV, formed an appropriate barrier height with the conduction band minimum (CBM) of the hBN layer, achieving better nonlinearity. A thin multilayer hBN was transferred onto TiN via bubbling transfer, with a thermal release tape as the supporting layer [5]. A 5-nm-thick HfO_2 layer was deposited via thermal atomic layer deposition to serve as the switching layer. The device was fabricated by depositing the oxygen-exchange layer and top electrode (TE)-Ti and Al via e-gun evaporation. The device structure was Al(TE)/Ti/HfO_2/hBN/TiN(BE). DC measurements of the bilayer RRAM devices were performed using the Cascade/Agilent B1500A.

III. MODELING METHODOLOGY

This proposed model is built by the SPICE-compatible language, Verilog-AMS, widely used in the semiconductor industry. To simulate the characteristics of RRAM cells, the model mainly comprises a current calculation module, a gap (filament) formation module, and a temperature calculation module. The transport mechanisms in the insulator have been extensively discussed throughout the development of the semiconductor industry. The main categories of transport mechanisms are bulk-limited current and electrode-limited current. Many RRAM studies have focused on a single transport mechanism.

Therefore, this article proposes a compact model that includes both bulk-limited and electrode-limited mechanisms with a smooth transition that can be applied to various fabricated RRAM devices. To achieve this, we introduce Fowler–Nordheim tunneling (FNT), space charge limited current (SCLC), and ohmic conduction, consistent with the experimental extraction results. Many studies have shown that the bilayer structure that creates high-nonlinearity RRAM demonstrates FNT behavior [6], [7], [8]. The transport mechanism in the insulator shows multiple possibilities. This is caused by the fabrication method. By observing the extracted data and some other similarly fabricated bilayer RRAM studies, the coexistence of SCLC and FNT mechanisms is suitable for this experimental result. Hopping is indeed one of the leading transport mechanisms in reported studies of RRAM, but in this study, we found it not as significant as SCLC and FNT. In the high-resistance state (HRS), one of the benefits of the double layer is that the second dielectric layer creates filament seeds that improve cycle-to-cycle uniformity [9].

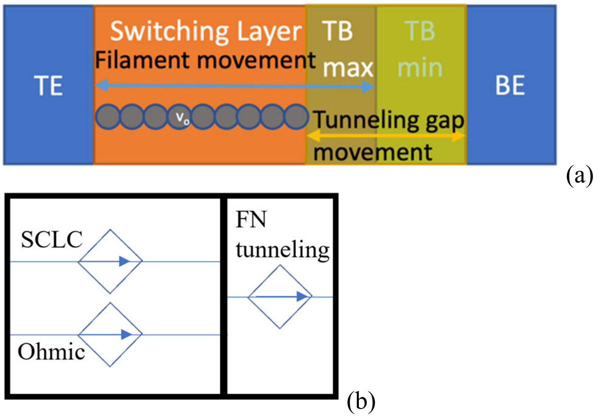


FIGURE 1. Schematic of tunneling gap and filament formation. (a) RRAM cell is composed of TE, switching layer, TB, and BE, and the conductive filament is formed or ruptured during the set/reset process under forward-/reverse-bias condition. In the actual measurement, the RRAM conductance shows the distribution in a group of devices. In terms of model, only some devices grow to full switching layer. This is an assumption made by the authors that the filament formed in switching layer until blocked by TB. (b) Description of current flow is mainly calculated by SCLC, ohmic, and FNT. If the RRAM device is fabricated with only a single layer of switching oxide, it is simulated with ohmic and SCLC. If the device is fabricated with a bilayer structure, the simulation includes both SCLC and FNT.

As shown in Fig. 1, the gap formation model calculates the movement of filaments, which determines the tunneling gap. The electric field across the dielectric layers can be defined as the tunneling gap changes. Since the tunneling current is highly dependent on the electric field, the thickness of the tunneling gap determines the current level. A conducting path responsible for SCLC and ohmic current is formed as the filament grows. The time and thermal energy-related positive feedback behavior are shown in Fig. 1.

A. CURRENT CALCULATION

Regarding high field, the FNT becomes the predominant output current. Thus, the TB creates a high selectivity of RRAM devices. The tunneling current can be expressed as follows [10]:

$$J_{\mathrm{FN}} = \left(q^2/8\pi h\varphi\right) E^2 \exp\left(-8\pi\sqrt{2m^* q\varphi^3}/3hE\right) \tag{1}$$

where h is the Plank constant, φ is the band offset between electrode and tunneling oxide, d is the direct tunneling distance, m^* is the effective mass, and E is the electric field across the tunneling gap. The tunneling current could be lumped into

$$J_{\mathrm{FN}} = \left(AE^2/\varphi\right) \exp\left(-B\varphi^{1.5}/E\right) \tag{2}$$

where A and B represent $(q^2/8\pi h)$ and $(8\pi(2m^* q)^{1/2}/3h)$, respectively, which can be extracted from the experiment data.

The typical SCLC and ohmic conduction behavior is described as follows [10]:

$$J_{\mathrm{sclc}} = (9/8)\,\mu\varepsilon\left(V^2/d^3\right) \tag{3}$$

and

$$J_{\mathrm{ohm}} = q\mu n\,(V/d) \tag{4}$$

where μ is the mobility in the dielectric, ε is the dielectric constant, and d is the dielectric thickness. Finally, the total current is calculated by multiplying J and the filament contact area. The tunneling current at a high electric field is determined by FNT.

The above mechanisms are smoothed with the window function, which can be constructed from tanh functions [11], such as

$$\begin{aligned} F(x) = {} & f_0(x)\left[1 - \tanh(x - x_0)\right]/2 \\ & + f_1(x)\left[\tanh(x - x_0) + 1\right)/2 \end{aligned} \tag{5}$$

where $f_0(x)$ and $f_1(x)$ represent different equations used here.

B. GAP FORMATION

The tunneling gap changes with the filament movement, whose range is defined between the maximum gap (Gapmax) and minimum gap (Gapmin), as indicated in Fig. 1(a). The process occurs when the oxygen vacancies start to generate. It could quickly form a conductive path that leads to a low-resistance state (LRS) ohmic current and reduces the tunneling gap distance. The time and the thermal-related positive feedback behavior are expressed as [12]

$$\frac{\mathrm{dg}}{\mathrm{dt}} = v_0 \exp\left(\frac{-U_{\mathrm{a}}}{kT}\right)\sinh\left(\frac{qa\gamma E}{kT}\right) \tag{6}$$

where g is the gap between electrode and filament, U_{a} is the activation energy, and a is the oxide lattice constant. v_0 and γ represent the escape attempt frequency and local enhancement factor that considers high-k dielectrics' polarizability and the device structure's nonuniform potential distribution, respectively [13]. E is the electric field across the oxide.

C. TEMPERATURE CALCULATION

The gap evolution would increase the current and the temperature due to Joule heating. Then, the increased temperature would accelerate the gap formation. A compact temperature calculation model helps the evaluation of this thermal run-away process. The temperature in the active region (T_{RRAM}) is evaluated with Joule heating (W_{j}) and thermal dissipation (W_{d}) as

$$T_{\mathrm{RRAM}} = \int_{t0}^{t1} \left(W_{\mathrm{j}} - W_{\mathrm{d}}\right)/CV\,\mathrm{d}t. \tag{7}$$

The above equation is further derived into a closed-form 1-D case as [8]

$$T_{\mathrm{RRAM}} = \frac{d_{\mathrm{hd}} W_{\mathrm{j}}}{kA}\left(1 - \exp\left(\frac{-kA}{d_{\mathrm{hd}} CV} t\right)\right) + T_{\mathrm{room}}. \tag{8}$$

C is the material's heat capacity, V is the device's volume, and k is the heat transfer coefficient. A is the filament contact area. d_{hd} is the heat dissipation distance assumed to be the same as total device thickness. W_{j} and W_{d} are defined as follows:

$$W_{\mathrm{j}} = I_{\mathrm{pn}} V_{\mathrm{pn}} \tag{9}$$

and

$$W_{\mathrm{d}} = \partial Q/\partial t = -k\nabla T. \tag{10}$$

The temperature factors accelerate the time, and the thermal-related positive feedback formation of conductive filament is described in (7).

TABLE 1. **Model card of the simulated RRAM.**

Parameter description	Symbol	Value (nm)
The initial temperature	T_0	300 K
Activation energy	U_a	0.5 eV
Lattice constant	A	0.25 x 10^{-9} m
Escape attempt velocity	v_0	0.1 m/s
Local enhancement factor	γ	10
The heated volume of RRAM	*Volume*	10^{-24} m^3
Filament contact area	*F_cr*	8 x 10^{-15} m^2
Heat dissipation distance of filament	*dr_T*	7 x 10^{-9} m
Minimum thickness of tunneling gap	*Gapmin*	4 x 10^{-9} m
Maximum thickness of tunneling gap	*Gapmax*	7 x 10^{-9} m
Free electron density	N	1x10^{25}-1x10^{26} m^{-3}

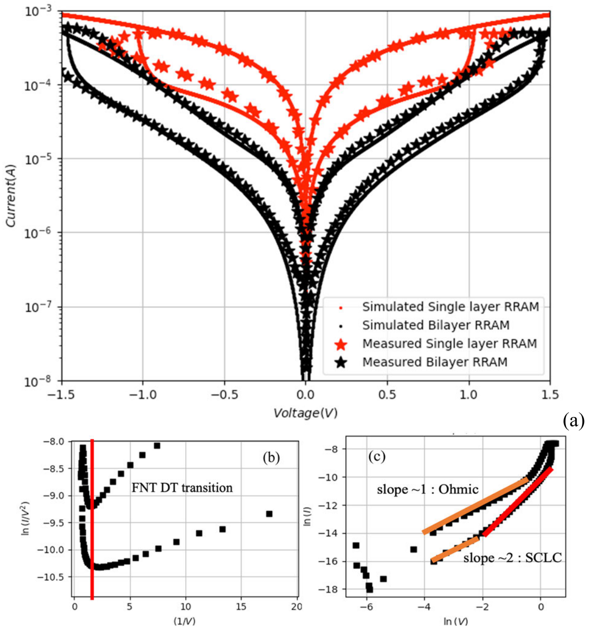


FIGURE 2. **Measured device and simulated result. (a) Result of the device with TB and the result of RRAM without TB. The simulated quasi-dc is under the ramp rate of 0.5 V/s. (b) FNT extraction from (a). (c) Extraction of SCLC and ohmic conduction from (a).**

Using the model discussed above, the simulated I–V characteristics of the RRAM cell with an embedded TB are shown in Fig. 2, where the case without the embedded barrier is included for comparison. The simulated parameters are listed in Table 1. The not well-fit LRS at 1.1–1.5 V is due to an internal numerical calculation. To capture the current suppression effect at low-voltage region in LRS, which increases the nonlinearity of RRAM, two kinds of transport mechanisms are smoothed together to make it continuous. At present, the nonlinearity and low-current value at LRS are well-fitted. More fitting parameters could be introduced into the model to improve the accuracy at higher bias regions.

To predict the reliability of RRAM, the Arrhenius equation A$\exp(E_{\mathrm{a}}/KT)$ may be utilized [14]. This equation corresponds to the retention degradation and the redox reaction involved in the process. Typically, the temperature dependence of the reaction rate adheres to the Arrhenius law. The prediction of temperature could be introduced here. This equation signifies the self-diffusion of the mobile ion within the device. Since the generation of oxygen ion or oxygen vacancy pairs dictates the conductivity of RRAM, the associated resistance drift can be described using this self-diffusion term. Such a drift is linked to the reliability of RRAM, enabling the prediction of temperature to evaluate the retention of RRAM.

Although this study focuses on the interconnect issue of computer integrated manufacturing (CIM) applications and does not consider the interface defect, the impact of the trap must be considered in further studies. The trap's influence on memory variation and failure is a critical aspect that must not be overlooked. The defect/trap would cause variation issues like random telegram noise (RTN) that would cause read errors [15], [16]. Due to the device being under a smaller read current, the variational effect would be even higher. The trap impacts the HRS/LRS failure because the resistance window shifts toward HRS or LRS after the Set/Reset cycles. This would affect the device application for online learning and offline learning. For online learning, the device is under continuous weight update, which means hundreds or thousands of Set/Reset (Program/Erase in other memories) processes depending on the choice of algorithm (NN model and hyperparameters). As for offline learning, the model is trained on CPU/GPU and once programmed to RRAM array. To avoid the HRS/LRS failure caused by the trap and the device Set/Reset asymmetry in weight update (asymmetry Set/Reset IV behavior), RRAM in this study assumes an offline learning scenario, mainly focusing on the benefits of IV nonlinearity that bilayer RRAM brings.

IV. APPLICATION FOR COMPUTING-IN-MEMORY

A. SINGLE-CELL COMPARISON

This section discusses two types of RRAM with the same switching oxide, HfO_2, which is commonly used in RRAMs. The nonlinearity difference between the two types is increased by the TB, which is commonly determined by [6], [17]

$$\eta = (I@V_{\text{ref}})\,/\,(I@1/2V_{\text{ref}}) \text{ or } (I@1/3V_{\text{ref}})\,.$$

In this study, V_{ref} is defined as 1.5 V. The higher the nonlinearity is, the higher the selectivity is for RRAM devices. Resistive switching with and without a capping layer exhibits significant differences in characteristics, especially in the mid- and low-biasing regions. Different transport mechanisms can be extracted in different bias regions for bilayer devices.

As shown in Fig. 2(a), the device without a TB can be well-fit with a simple ohmic conduction mechanism. In contrast, the device with a TB shows a dramatically decreased read current in the low-bias region. This observation is consistent with the findings of [10], which demonstrated that in the bulk-limited transport mechanism of RRAM operation, the LRS primarily exhibits ohmic conduction, while the HRS exhibits both space-charge-limited conduction and ohmic conduction.

In Fig. 2(b) and (c), the extracted characteristics from measurement data in the highly biased region transport mechanism show FNT behavior. However, looking at the right-hand side of the FNT extraction plot, the slope difference suggests that compared to a purely tunneling mechanism, the current could not cover the entire transition. Thus, the extraction in the mid- and low-bias regions is shown in Fig. 2(c).

B. NEUROMORPHIC APPLICATION

The operation of deep NNs can be accelerated by using memory arrays. In an ANN, the weight connecting neurons with different weights can be represented, as shown in Fig. 3. The input signal sent from input neurons goes through a synapse with a certain weight. For a fully connected ANN, the operation between two layers of neurons can be represented as a vector multiplication. The input voltage multiplied by the conductance generates current. The sum of currents in each column represents the final input to the next layer of hidden neurons.

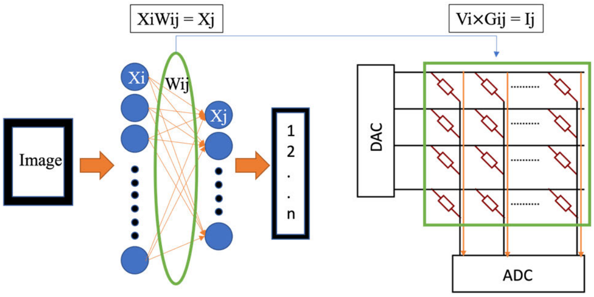


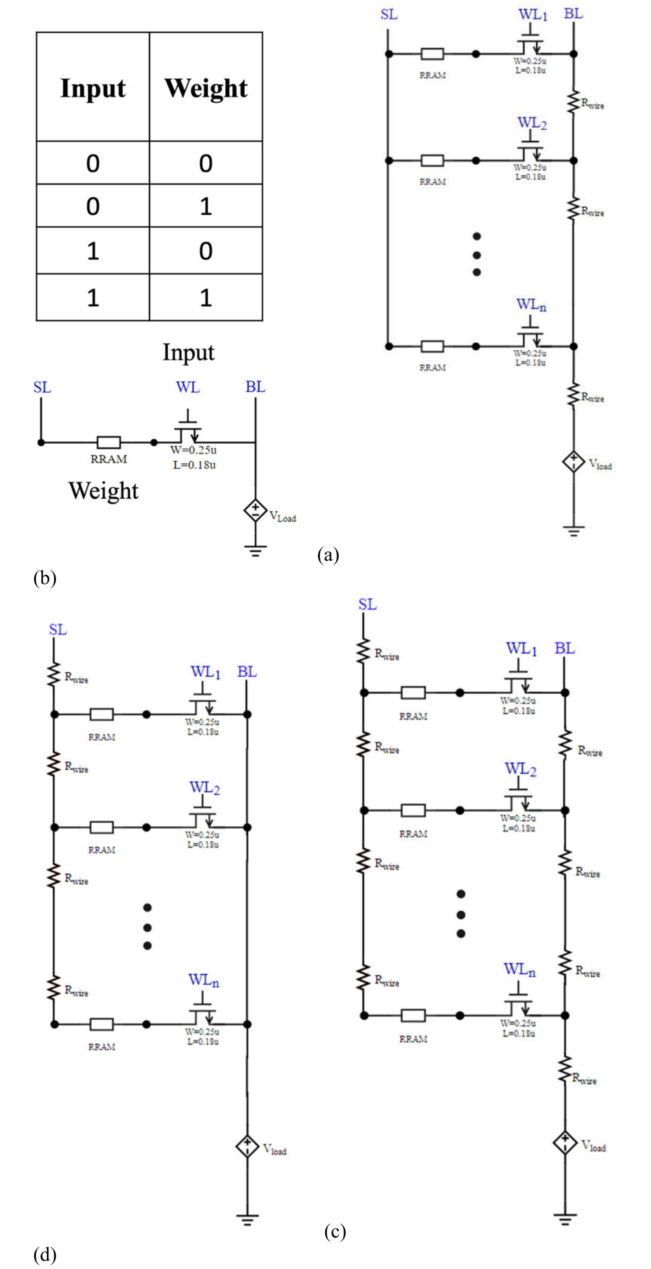
FIGURE 3. Schematic of NN and memory array. The signal transfer between the layers of NN could be emulated by vector matrix multiplication (VMM) of memory array by Ohm's law.

TABLE 2. Tested RRAM characteristics with technology assumption [18], [19] for transient simulation.

Parameter	Value
Set/Reset voltage	2.5V/-2.5V
Read voltage	0.5V
Read pulse duration	10ns
Write pulse duration	50ns
Rise/fall time	10ns
Wire resistance per branch	3Ω
Single layer RRAM LRS/HRS	41.325k/216.295k
Single layer RRAM η	3.15
Bilayer RRAM LRS/HRS	162.410k/845.87k
Bilayer RRAM η	12.88
Filament contact area for 1T1R array	4 x 10^{-17}
Local enhancement factor γ	10 / 20

To simulate the circuit-level behavior in the NN application, Table 2 lists the assumptions and parameter values used in the array simulation. The fabricated device was assumed to be scaled to the current level that fits within the operation range of a 0.18-μm transistor. The nonlinearity ratio remained the same for both the single-layer and bilayer devices. The bilayer RRAM, which has a higher nonlinearity, was set at the same voltage amplitude as that for the single-layer device. Consequently, the bilayer RRAM shows reduced current at V_{read}, relieving the loading effect issue. The comparison results are presented in Section IV-C. When the application is used for offline learning, the whole ANN is trained in CPU/GPU, and the result is then programmed into the RRAM array. In this case, the array mainly performs the inference task.

The inference accuracy of the NN is highly depending on the result of the output neuron, which collects the output signal from the previous layer. Even if the weight/synapse is ideal, the output signal inevitably shifts due to the loading effect on the interconnect resistance. Although the actual connection of R_{wire} might be parallel or serial depending on the weight-input combination, which would be described in Section IV-C.1), one can qualitatively describe the effect as R_{synapse} in series with R_{wire}. The impact from R_{wire} is higher, and the value of R_{synapse} is less recognizable. The loading



Input	Weight
0	0
0	1
1	0
1	1

FIGURE 4. Schematic of memory array structure considering the parasitic R_{wire} loading effect on: (a) cell without parasitic R_{wire} would be compared with: (b) BL with R_{wire}; (c) SL with R_{wire}; and (d) both BL and SL with R_{wire}. The WL would determine whether to read or write the memory cell. The total current-sum error rate is compared with that of the memory cell array without considering R_{wire}. The circuit simulation calculates the current sum of a single column from 1 to 256 cells.

effect is caused by the IR drop on the wire resistance; the actual current sum is reduced.

C. LOADING EFFECT ON BACKEND INTERCONNECT

1) INPUT-WEIGHT COMBINATION SCENARIOS

Performing this type of operation requires a significant amount of current summing, but the loading effect on the wordline (WL), bitline (BL), and source line (SL) takes place.

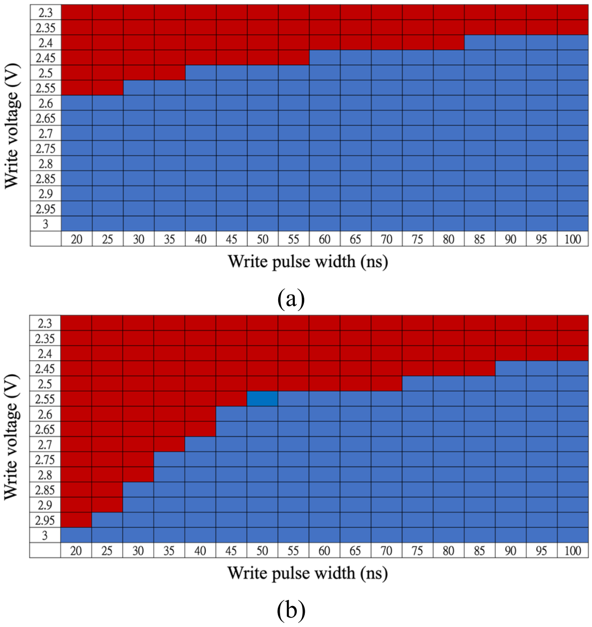


FIGURE 5. (a) Single RRAM cell and (b) RRAM cell selected from one of the columns in a 256 x 256 array. The block in red means the corresponding pulse condition cannot let the device reach MRS. The R_{wire} loading leads to the increase in minimum write voltage.

When the SL is charged to V_{dd}, the state of WL determines whether the input would affect the synapses. The result of the current sum is then observed by examining the node V_{load}. Hence, the WL voltage represents the input of 1 or 0. On the other hand, the resistance state of the selected synapse itself is either HRS or LRS, which means 0 or 1 of weight. As shown in Fig. 4, there will be four kinds of conditions for each neuron cell.

Note that the WL resistance is not in the current path, so this study does not consider the resistance impact from WL. Also, considering the whole circuit, the driver/sink resistance would share the loading effect. Depending on the array size and input-weight scenarios, the impact of wire resistance would vastly vary at this stage. This work focuses on the R_{wire} impact, and the driver/sink effect could be included in the overall power, performance, area, and cost (PPAC) evaluation in further studies.

Without considering the wire resistance, the current sum is multiplied by the number of each case in (a) and summed together. As for cases considering R_{wire} on BL and SL, there are several combinations that the cell with input-weight condition 1-1 (1-1 cell) would be connected to R_{wire} both in series and parallel in the whole circuit. Note that the 1-1 cell contributes most of the current while the 0-0, 0-1, and 1-0 cells contribute much less current than the 1-1 case. The 1-0 and 0-1 cells represent the same digital output, but the currents differ. By utilizing bilayer RRAM, the difference could be suppressed. In the following discussion, all the cells are first programmed to the LRS and controlled by WL, which means the cell state is either 1-1 or 0-1.

Here, we discuss the possible scenarios that would result in different current sum values with the loading effect. For

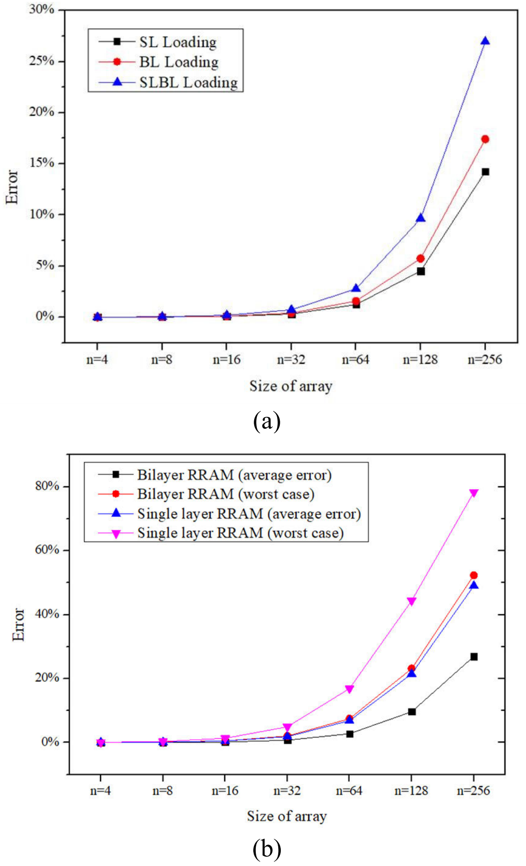


FIGURE 6. Simulated errors versus the size of the array. (a) Error rate is evaluated either with SL loading effect, BL loading effect, or source line and bit line (SLBL) loading effect. Regardless of the scenario, the error rate increases quickly as the size of the array increases. (b) With lowered output current by bilayer RRAM (with barrier), the error rate would be reduced on both worst-case scenario and average errors.

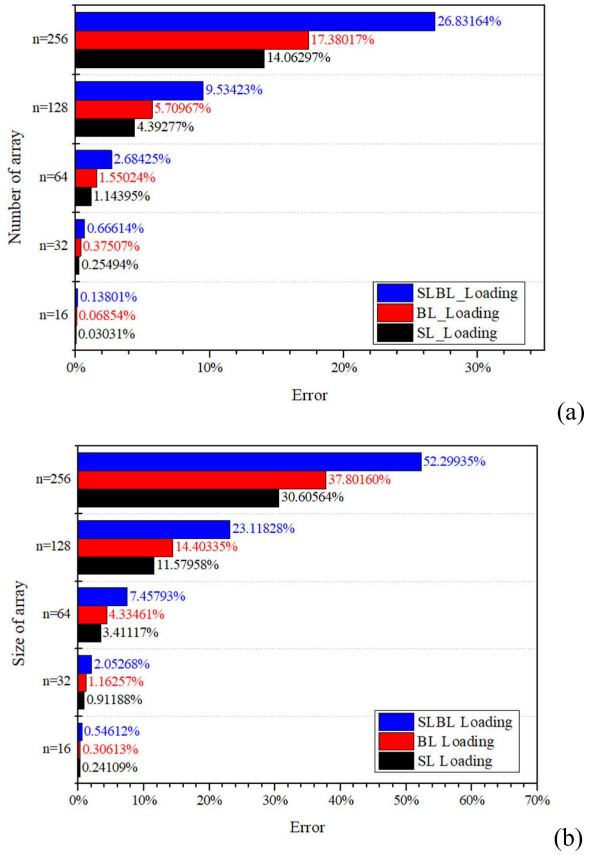


FIGURE 7. Average error rate improved by applying bilayer RRAM for the NN application. (a) Current sum error rate of bilayer RRAM with SLBL, SL-only, and BL-only loading effects. (b) Current sum error rate of single-layer RRAM with SLBL, SL-only, and BL-only loading effects. The IR drop effect increases with the size of the array, which contains more R_{wire} in the whole circuit. Thus, the overall average loading effect is significantly reduced by applying bilayer RRAM, which reduces the LRS current leading to less IR drop in the circuit.

instance, half of the WLs could be pulled up at the upper part of the column, and another half remains off at the lower part.

In contrast, half of the WLs could also be pulled up at the lower part of the column. In the following discussion, both cases account for 50% of the total column number.

The scenario described from the top column to the bottom column would be (50%, 1-1)–(50%, 0-1) (case A). Another case for the same level of the current sum would be (25%, 1-1)–(25%, 0-1)–(25%, 1-1)–(25%, 0-1) (case B). Although these two cases represent the same level of input-weight multiplication, the actual output would be affected by the R_{wire}. In case A, the half side of SL/BL is connected with R_{wire} in series since WL turns off the adjacent cells. As for case B, 25% of R_{wire} on SL/BL is connected in series, and the mid column is connected in parallel with WL being pulled up. The combination could be further expanded into like (25%, 0-1)–(25%, 1-1)–(25%, 1-1)–(25%, 0-1) and (25%, 1-1)–(25%, 0-1)–(25%, 0-1)–(25%, 1-1).

In the application of ANN, the TB reduces the error rate and the total current sum value at the output end. Each RRAM cell's conductance in the NN application represents the synapse's weight. However, not all neurons receive signals during actual operation, and not all weights are in high conducting states. As a result, the loading effect can vary depending on the distribution of weight and input signals.

Connecting the devices can significantly impact the actual current sum due to the nature of utilizing Kirchhoff's law. The loading effect is evaluated by the ratio of the array's current sum without considering the wire resistance over the array's current sum considering the wire resistance. The current sum is calculated at the output of BL by the nature of Ohm's law. The loading effect on BL and the impact on the SL are the first things to be considered. If this effect is considered for SL, the phenomenon mentioned earlier would become the opposite: the top RRAM would be series connected with the least R_{wire} component. The combination of case A- and case B-like scenarios is then averaged together and evaluated in different array sizes in the following section. The worst-case scenario would happen at, for example, case B-like (25%, 1-1)–(25%, 0-1)–(25%, 0-1)–(25%, 0-1). Only the top 25% side of the column cell is enabled. The BL side is series connected with R_{wire} on BL reset of the 75% column cell. If the case is 5%–95%, the result would be more extreme. However, to cover up most of the input-weight distribution, the column combination discussed in this article is divided into four parts as case B.

2) CURRENT SUM ERROR EVALUATION

To write RRAM, WL is enabled and SET voltage pulse with various amplitudes and widths is applied to BL. As shown in Fig. 5, the write shmoo plot shows the switching condition under various pulsewidths and amplitude. Due to the more connected R_{wire} in a larger array, there will be a higher voltage drop along the SL and BL. Thus, the voltage across the RRAM would be less than expected, leading to the write error in the device. The write error case is defined as the device not reaching the mid-resistive state (MRS), which means the current would be less than half of the LRS current. The minimum pulse voltage or width needs to be increased to switch RRAM.

To study the impact of the loading effect with wire resistance and avoid the write error condition, the following cases are simulated by assuming a higher local enhancement factor γ for the filament change in which the state 1 of weight is ensured. Only the enabled WL determines the combination of the R_{wire} connection. The value of the current sum impacted by the wire resistance is compared with the cases without wire resistance. As the size of the array increases, so does the ratio.

Hence, the current sum error would be inevitably high at a larger array scale. However, the bilayer RRAM that operates at the same voltage range with a smaller current would decrease its error rate, as shown in Fig. 6. In Fig. 7, more R_{wire} values are considered by including both SL and BL. In most severe error cases, the bilayer RRAM could decrease its error rate by almost 50%.

As described earlier, the loading effect can significantly impact the analog-to-digital converter (ADC) output for the input of the next neuron layer. The current sum error would affect the ADC that converts the current sum into a digital value due to the wire resistance loading effect. When it is sent to the next layer of NN, the current sum error rate is directly converted into the error rate of the output bit.

The resolution of ADC is defined as the smallest incremental voltage that can be recognized and thus causes a change in the digital output. It is expressed as the number of bits output by the ADC. Therefore, an ADC converts the analog signal to a digitized value. For example, if the tested cases have 256 rows, choose an ADC with a resolution of 8 bits. Assume that a voltage ADC with current-to-voltage (I/V) conversion is used. The current sum is converted into voltage for digitalization steps with a reference voltage V_{ref}. The output would be $V_{\text{ref}}/256$ for each digital step.

Also, the simulation data indicate that the current sum error rate increases dramatically with the array size after $n = 32$. An error rate below 10% is more acceptable, considering the array size and ADC resolution. Take a 5 -V 8-bit ADC as an example, and the resolution would be about 0.0195 V, which is about 0.4% for each case. Thus, the current sum error is significant for this application. If the design chooses a 6-bit ADC, the resolution is about 1.5% for each case, and the error rate of the bilayer structure is close to this range, which is more acceptable than that in the 8-bit case. For a more robust design, a size $n = 256$ current sum array could be replaced by four size $n = 64$ current sum arrays.

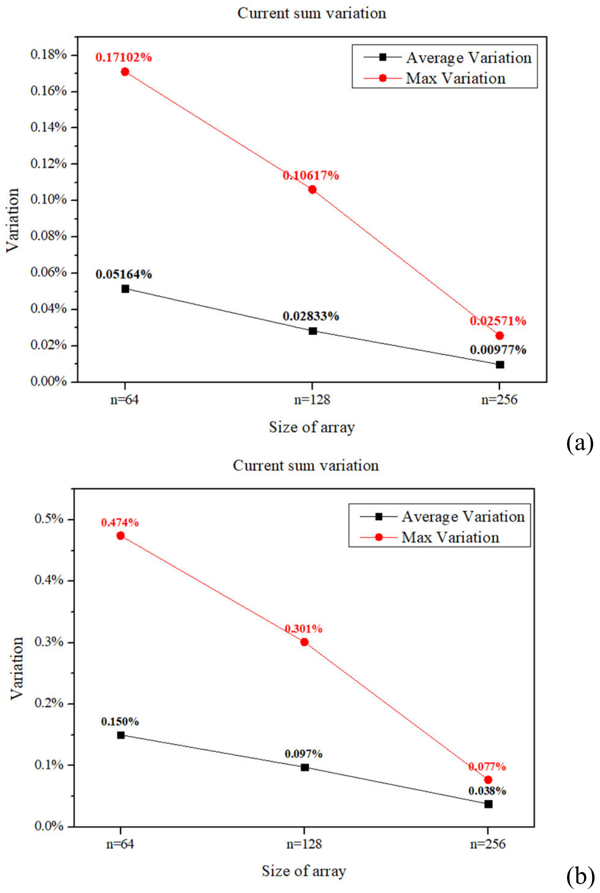


FIGURE 8. Comparison of different current sum with thickness variation. The total variation drops slightly as the size of arrays increases due to the total current sum increases. The variation of each of the switching layer and barrier layer is set up with: (a) 0.5- and (b) 1-nm variation, respectively. The variation would arise during the deposition process. The Monte Carlo simulation is done by using Gaussian distribution with three standard deviations.

The resistance variation of RRAM also needs to be taken into consideration. A significant drawback hinders RRAM from applications and mass production in the industry. Because the resistance switching phenomenon is dominated by the change of conduction filaments, its physical dimension would directly affect the resistance distribution of every RRAM resistance state. Since the current calculation is mainly determined by the electric field within the device, it is sensitive to the variation of dielectric layer thickness during the thin-film deposition process. The presented model with Gapmin and Gapmax could be helpful in the study of process variation. The variation of Gapmax and Gapmin might vary the set voltage V_{set} from HRS to LRS and LRS current, respectively. As predicted in Fig. 8, the overall variation of the current sum is more minor as the array size becomes larger. Because the value of the current sum is higher as the array size gets larger, the variation per array is more minor due to the nature of the NN algorithm.

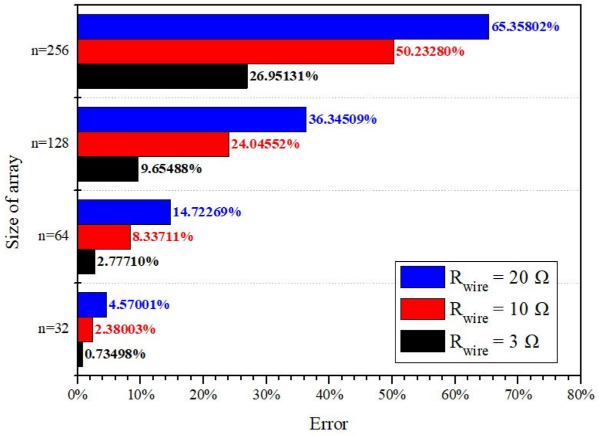


FIGURE 9. Comparison of different R_{wire} according to different technology nodes. R_{wire} values of 20, 10, and 3 Ω correspond to the 10-, 14-, and 22-nm technology nodes [20], respectively. As R_{wire} increases with the technology node, the error rate for the current sum becomes even more severe.

As technology nodes advance, critical dimensions (CDs) are pushed to smaller sizes, increasing wire resistance. This, in turn, can negatively impact the current sum, as shown in Fig. 9. Enlarging the array size can also raise the error rate. Developing advanced technology nodes further exacerbates this problem by contributing to the rising wire resistance. Although scaling CD can enhance circuit density and performance, it also results in larger interconnect parasites, which must be carefully evaluated for further CIM performance in following-generation applications.

V. CONCLUSION

This study presents a novel model that combines the bulk-limited and electrode-limited transport mechanisms for bilayer RRAM. This model enables predictive evaluation of the device current in NN applications, where each RRAM cell's conductance represents the synapse's weight. However, the loading effect can impact the evaluation of device current, as neuron/weight combination would vary the final current sum during operation. To overcome this issue, we propose using bilayer modulated RRAM, which significantly reduces the impact of the loading effect. By leveraging the unique properties of bilayer RRAM, we can achieve a more reliable and accurate evaluation of device current, improving the overall performance of NN applications. Our findings pave the way for developing more advanced and efficient NN models with a wide range of potential applications in areas such as image and speech recognition, natural language processing, and autonomous driving.

REFERENCES

[1] U. Chand, K.-C. Huang, C.-Y. Huang, and T.-Y. Tseng, "Mechanism of nonlinear switching in HfO$_2$-based crossbar RRAM with inserting large bandgap tunneling barrier layer," *IEEE Trans. Electron Devices*, vol. 62, no. 11, pp. 3665–3670, Nov. 2015, doi: 10.1109/TED.2015.2471835.

[2] K.-C. Chuang et al., "Impact of the stacking order of HfO$_x$ and AlO$_x$ dielectric films on RRAM switching mechanisms to behave digital resistive switching and synaptic characteristics," *IEEE J. Electron Devices Soc.*, vol. 7, pp. 589–595, 2019, doi: 10.1109/JEDS.2019.2915975.

[3] K.-S. Li et al., "Study of sub-5 nm RRAM, tunneling selector and selector less device," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2015, pp. 385–388, doi: 10.1109/ISCAS.2015.7168651.

[4] L. Shan et al., "Non-linear resistive switching characteristics in HFO$_2$-based RRAM with low-dimensional material engineered interface," in *Proc. China Semiconductor Technol. Int. Conf. (CSTIC)*, Mar. 2021, pp. 1–3, doi: 10.1109/CSTIC52283.2021.9461454.

[5] P.-A. Chen, "Application of 2D materials to advanced resistive random-access memory and planar interconnect," Ph.D. dissertation, Nat. Cheng Kung Univ., Tainan City, Taiwan, 2022. [Online]. Available: https://thesis.lib.ncku.edu.tw/thesis/detail/2f6550405f02e790ddbc74bf9099c725/

[6] Y.-C. Chen, C.-C. Lin, S.-T. Hu, C.-Y. Lin, B. Fowler, and J. Lee, "A novel resistive switching identification method through relaxation characteristics for sneak-path-constrained selectorless RRAM application," *Sci. Rep.*, vol. 9, no. 1, p. 12420, Aug. 2019, doi: 10.1038/s41598-019-48932-5.

[7] G. Kim et al., "Retention secured nonlinear and self-rectifying analog charge trap memristor for energy-efficient neuromorphic hardware," *Adv. Sci.*, vol. 10, no. 3, Jan. 2023, Art. no. 2205654, doi: 10.1002/advs.202205654.

[8] B. Kim, I.-S. Kim, J.-U. Woo, S.-J. Chae, S.-H. Go, and S. Nahm, "Self-rectifying and artificial synaptic characteristics of amorphous Ta$_2$O$_5$ thin film grown on two-dimensional metal-oxide nanosheet," *Appl. Surf. Sci.*, vol. 609, Jan. 2023, Art. no. 155353, doi: 10.1016/j.apsusc.2022.155353.

[9] P.-A. Chen, W.-C. Hsu, and M.-H. Chiang, "Bilayer modulation with dual vacancy filaments by intentionally oxidized titanium oxide for multilayer-hBN RRAM," *IEEE Trans. Nanotechnol.*, vol. 20, pp. 687–694, 2021, doi: 10.1109/TNANO.2021.3110899.

[10] S. M. Sze, Y. Li, and K. K. Ng, *Physics of Semiconductor Devices*. Hoboken, NJ, USA: Wiley, 2021.

[11] K. Xia, "Smoothing globally continuous piecewise functions based on limiting functions for device compact modeling," *J. Comput. Electron.*, vol. 18, no. 3, pp. 1025–1036, Sep. 2019, doi: 10.1007/s10825-019-01356-w.

[12] J.-W. Lee and M.-H. Chiang, "Modeling of RRAM with embedded tunneling barrier and its application in logic in memory," *IEEE J. Electron Devices Soc.*, vol. 8, pp. 1390–1396, 2020, doi: 10.1109/JEDS.2020.3008172.

[13] J. McPherson, J.-Y. Kim, A. Shanware, and H. Mogul, "Thermochemical description of dielectric breakdown in high dielectric constant materials," *Appl. Phys. Lett.*, vol. 82, no. 13, pp. 2121–2123, Mar. 2003, doi: 10.1063/1.1565180.

[14] Z. Wei et al., "Highly reliable TaO$_x$ ReRAM and direct evidence of redox reaction mechanism," in *IEDM Tech. Dig.*, Dec. 2008, pp. 1–4, doi: 10.1109/IEDM.2008.4796676.

[15] A. Calderoni, S. Sills, and N. Ramaswamy, "Performance comparison of O-based and Cu-based ReRAM for high-density applications," in *Proc. IEEE 6th Int. Memory Workshop (IMW)*, May 2014, pp. 1–4, doi: 10.1109/IMW.2014.6849351.

[16] A. Calderoni, S. Sills, C. Cardon, E. Faraoni, and N. Ramaswamy, "Engineering ReRAM for high-density applications," *Microelectron. Eng.*, vol. 147, pp. 145–150, Nov. 2015, doi: 10.1016/j.mee.2015.04.044.

[17] X. Peng, R. Madler, P. Y. Chen, and S. Yu, "Cross-point memory design challenges and survey of selector device characteristics," *J. Comput. Electron.*, vol. 16, no. 4, pp. 1167–1174, 2017, doi: 10.1007/s10825-017-1062-z.

[18] A. A. Gismatulin, G. N. Kamaev, V. N. Kruchinin, V. A. Gritsenko, O. M. Orlov, and A. Chin, "Charge transport mechanism in the forming-free memristor based on silicon nitride," *Sci. Rep.*, vol. 11, no. 1, p. 2417, Jan. 2021, doi: 10.1038/s41598-021-82159-7.

[19] Y. Xie, *Emerging Memory Technologies: Design, Architecture, and Applications*. New York, NY, USA: Springer, 2013.

[20] R. Han, P. Huang, Y. Zhao, X. Cui, X. Liu, and J. Kang, "Efficient evaluation model including interconnect resistance effect for large scale RRAM crossbar array matrix computing," *Sci. China Inf. Sci.*, vol. 62, no. 2, p. 22401, Feb. 2019, doi: 10.1007/s11432-018-9555-8.