Xia L, Gu P, Li B *et al.* Technological exploration of RRAM crossbar array for matrix-vector multiplication. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 31(1): 3–19 Jan. 2016. DOI 10.1007/s11390-016-1608-8

Technological Exploration of RRAM Crossbar Array for Matrix-Vector Multiplication

Lixue Xia [1,2], *Student Member, IEEE*, Peng Gu [1,2], *Student Member, IEEE*
Boxun Li [1,2], *Student Member, IEEE*, Tianqi Tang [1,2], *Student Member, IEEE*, Xiling Yin [1,2]
Wenqin Huangfu [1,2], Shimeng Yu[3], *Member, IEEE*, Yu Cao[3], *Senior Member, IEEE*
Yu Wang[1,2,*], *Senior Member, IEEE*, and Huazhong Yang [1,2], *Senior Member, IEEE*

[1] *Department of Electronic Engineering, Tsinghua University, Beijing 100084, China*

[2] *Tsinghua National Laboratory for Information Science and Technology (TNList), Tsinghua University Beijing 100084, China*

[3] *School of Electrical, Computer and Energy Engineering, Arizona State University, Arizona 85281, U.S.A.*

E-mail: xialx13@mails.tsinghua.edu.cn; gupeng9259@163.com; {dpxlbx, ttq1008, yxlsarah, wenqinhf}@gmail.com {Shimeng.Yu, Yu.Cao}@asu.edu; yu-wang@mail.tsinghua.edu.cn; yanghz@tsinghua.edu.cn

Received September 1, 2015; revised December 8, 2015.

Abstract Matrix-vector multiplication is the key operation for many computationally intensive algorithms. The emerging metal oxide resistive switching random access memory (RRAM) device and RRAM crossbar array have demonstrated a promising hardware realization of the analog matrix-vector multiplication with ultra-high energy efficiency. In this paper, we analyze the impact of both device level and circuit level non-ideal factors, including the nonlinear current-voltage relationship of RRAM devices, the variation of device fabrication and write operation, and the interconnect resistance as well as other crossbar array parameters. On top of that, we propose a technological exploration flow for device parameter configuration to overcome the impact of non-ideal factors and achieve a better trade-off among performance, energy, and reliability for each specific application. Our simulation results of a support vector machine (SVM) and Mixed National Institute of Standards and Technology (MNIST) pattern recognition dataset show that RRAM crossbar array based SVM is robust to input signal fluctuation but sensitive to tunneling gap deviation. A further resistance resolution test presents that a 6-bit RRAM device is able to realize a recognition accuracy around 90%, indicating the physical feasibility of RRAM crossbar array based SVM. In addition, the proposed technological exploration flow is able to achieve 10.98% improvement of recognition accuracy on the MNIST dataset and 26.4% energy savings compared with previous work. Experimental results also show that more than 84.4% power saving can be achieved at the cost of little accuracy reduction.

Keywords resistive switching random access memory (RRAM), machine learning, electronic design automation, matrix-vector multiplication, non-ideal factor

1 Introduction

Machine learning is becoming popular in a wide range of domains. Many emerging applications, ranging from image and speech recognition to natural language processing and information retrieval, rely heavily on machine learning techniques[1]. Matrix-vector multiplication is of significant importance in many applications[2-3], such as support vector machine (SVM)[4] and deep learning algorithms[5]. Therefore, the performance of matrix-vector multiplication has become one of the most crucial considerations in accelera-

Regular Paper

Special Section on Computer Architecture and Systems with Emerging Technologies

This work was supported by the National Basic Research 973 Program of China under Grant No. 2013CB329000, the National Natural Science Foundation of China under Grant Nos. 61373026, 61261160501, the Brain Inspired Computing Research of Tsinghua University under Grant No. 20141080934, Tsinghua University Initiative Scientific Research Program, and the Importation and Development of High-Caliber Talents Project of Beijing Municipal Institutions.

*Corresponding Author

tor designs for machine learning applications[3].

Recently, the emerging metal oxide resistive switching random access memory (RRAM) device and RRAM crossbar array have demonstrated an efficient hardware implementation of matrix-vector multiplication[6-8]. Based on the multilevel resistance characteristic of RRAM device and the cross-point structure, RRAM crossbar array can use the input voltage signal as the vector data and save the matrix data into the RRAM cells, which realizes matrix-vector multiplication efficiently with $O(1)$ time complexity by the nature of merging all cells' current in each row. Furthermore, as a nonvolatile memory device, RRAM is an emerging approach to merging the memory and computation, which has potential to break the "memory wall" bottleneck of traditional von Neumann architecture. Many studies have explored the potential of computing with RRAM crossbar array. For example, a low power approximate computing system, which is based on the RRAM crossbar implementation of matrix multiplication and neural network, has demonstrated power efficiency of more than 400 GFLOPS/W[9].

However, although many researchers have adequately demonstrated the benefit of RRAM crossbar based computing systems, many important non-ideal factors are neglected. Most of the previous work is based on a simplified circuit model[8,10-11] and uses a linear resistor to represent an RRAM device, which may lead to inaccurate conclusions[12]. Moreover, some non-ideal factors, such as the nonlinear voltage-current relationship of RRAM devices, the interconnect resistance, and the resistance state deviation, may significantly influence the performance of RRAM crossbar array based computing systems. Therefore, a detailed and comprehensive analysis of the impact of these non-ideal factors is still lacking.

The contributions of this paper include:

1) We analyze the impact of various non-ideal factors on the performance of RRAM crossbar array. We demonstrate that the RC delay of the array could be ignored (about 10 ps for a 100×100 crossbar according to our simulation). We also propose that the nonlinearity of RRAM devices, the variation of device processing and write operation, and interconnect resistance will have a major influence on the computation accuracy of output voltage. Moreover, we present that the minimum resistance state of RRAM devices has little direct impact on computation accuracy while increasing load resistance will significantly improve computation accuracy.

2) We propose a technological exploration flow of RRAM crossbar array to mitigate the impact of non-ideal factors and realize a better trade-off among performance, energy, and reliability for each specific application. The proposed flow includes: the configuration of technology node, RRAM resistance range, and load resistance; the algorithm of mapping matrix parameters to RRAM resistance states; and an iterative solution to optimize the power and performance.

3) Finally, we use the Mixed National Institute of Standards and Technology (MNIST) dataset and a linear SVM classifier as a case study to test the performance of the proposed technology exploration flow. Our simulation results demonstrate that the exploration flow can achieve 10.98% improvement of recognition accuracy and 26.4% power reduction compared with previous work[10], and can further receive a 84.4% power saving at the cost of little accuracy reduction.

2 Preliminaries

2.1 RRAM Characteristics and Device Model

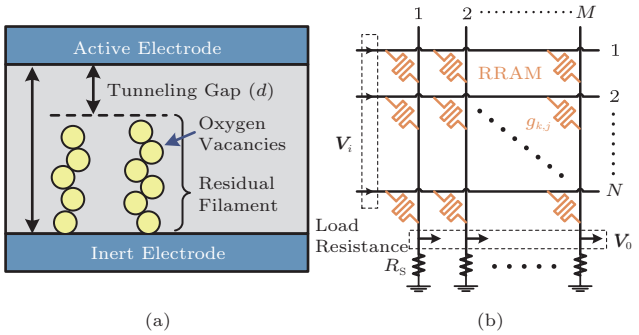
RRAM device is a passive two-port element based on metal oxide materials like TiO_x[13], WO_x[14], and HfO_x[15] with variable resistance. In this paper, we use HfO_x-based RRAM for study because it is one of the most mature RRAM materials explored[16].

Fig.1(a) demonstrates a 2D filament model of the HfO_x-based RRAM[17]. Its conductance is exponentially dependent on the tunneling gap distance (d). When a large voltage is applied on the electrodes, the tunneling gap distance d will change due to the electric filed and temperature-enhanced oxygen ion migration, and the resistivity of RRAM device will switch between the highest resistance state R_{OFF} and the lowest resistance state R_{ON}. Theoretically, an RRAM device can achieve any resistance in the range between R_{ON} and R_{OFF}. This work focuses on the choice of the resistivity of RRAM devices and other device parameters. How to tune the RRAM device to the specific resistance state will not be discussed in the paper.

For the HfO_x-based RRAM device, the nonlinear I-V relationship can be empirically expressed as follows[17]:

$$I = I_0 \times \exp\left(-\frac{d}{d_0}\right) \times \sinh\left(\frac{V}{V_0}\right), \tag{1}$$

where d is the average tunneling gap distance, V is the voltage across the RRAM device, and I is the current. I_0 (around 1 mA), d_0 (around 0.25 nm) and V_0 (around 0.25 V) are fitting parameters through experiments.



(a) (b)

Fig.1. (a) Physical model of the HfO_x-based RRAM. (b) Structure of the RRAM crossbar array.

In order to analyze the device and circuit interaction issues for the RRAM crossbar array based computation, we use HSPICE to simulate the circuit performance based on a recent Verilog-A model described in [17].

2.2 RRAM Crossbar Array

RRAM crossbar array is able to perform the analog matrix-vector multiplication efficiently. Fig.1(b) illustrates the structure of the RRAM crossbar array. The relationship between the input voltage vector ($\boldsymbol{V}_i$) and the output voltage vector ($\boldsymbol{V}_o$) can be expressed as follows[8]:

$$\begin{pmatrix} V_{o,1} \\ \vdots \\ V_{o,M} \end{pmatrix} = \begin{pmatrix} c_{1,1} & \cdots & c_{1,N} \\ \vdots & \ddots & \vdots \\ c_{M,1} & \cdots & c_{M,N} \end{pmatrix} \begin{pmatrix} V_{i,1} \\ \vdots \\ V_{i,N} \end{pmatrix}. \quad (2)$$

The index numbers of input and output voltages are denoted by k ($k = 1, 2, ..., N$) and j ($j = 1, 2, ..., M$) respectively, and the matrix parameter $c_{k,j}$ can be represented by the conductivity of the RRAM device ($g_{k,j}$) and the load resistor (g_s) as:

$$c_{k,j} = \frac{g_{k,j}}{g_s + \sum_{l=1}^{N} g_{k,l}}. \quad (3)$$

Since both g_s and $g_{k,j}$ can only be positive, two RRAM crossbar arrays are required to represent a matrix with both positive and negative parameters. The input voltage vectors of the positive RRAM crossbar array and the negative RRAM crossbar array should be $\boldsymbol{V}_i$ and $-\boldsymbol{V}_i$, respectively. The relationship between the input and the output voltage vectors can be expressed as:

$$\begin{aligned} \boldsymbol{V}_o &= \boldsymbol{C}^+ \cdot \boldsymbol{V}_i + \boldsymbol{C}^- \cdot (-\boldsymbol{V}_i) \\ &= (\boldsymbol{C}^+ - \boldsymbol{C}^-) \cdot \boldsymbol{V}_i = \boldsymbol{C} \cdot \boldsymbol{V}_i, \end{aligned} \quad (4)$$

where $\boldsymbol{C}^+$ and $\boldsymbol{C}^-$ are the matrixes represented by the positive and the negative RRAM crossbar arrays as described in (2) and (3) respectively.

2.3 Related Work

Recently, lots of researchers are devoted to fabricating RRAM devices with different kinds of materials and technologies. Some researchers use the technology at the device level to improve the performance of RRAM, such as using emerging oxide material with special characteristic[15,18-19] and optimizing the thickness of oxide layer[20-21]. Moreover, some researchers further analyze the causes and consequences of some non-ideal factors of RRAM[22] and propose some models to describe the performance of RRAM[23-25]. These device level results provide sufficient preliminary knowledge about the non-ideal factors and then support the analysis and optimization of circuit level design, which provides the basis of our work.

On the other hand, some researchers focus on the RRAM-based computation architecture and propose some RRAM-based designs for applications such as approximate computing[9,26] and neuromorphic applications[27]. However, these results do not consider the non-ideal factors of device and circuit. Actually, for a certain resistance level (i.e., 0/1 for memory device), the maximum difference between two fabricated RRAM cells may be larger than one order of magnitude[28]. For example, the measured resistance value of high resistance state of a 2-value RRAM varies from 1 MΩ to 10 MΩ[29]. Meanwhile, the write operation (i.e., SET/RESET for memory device) cannot precisely adjust the resistance of RRAM cell, which also results in a stochastic resistance deviation ranging from 10% to more than 60%[22]. These resistance variations introduce noise into the computation circuit, which will obviously influence the computation result of RRAM crossbar array.

There are a few researchers introducing some variations into weight matrixes in algorithms to reflect the RRAM device variations for simulation, and trying to resist them by improving the reliability of behavior-level algorithms[30-32]. But these results do not consider other device level phenomena of the non-ideal factors, such as the nonlinear *I-V* characteristics of RRAM, which limits the improvement space of the RRAM circuit.

Moreover, some circuit level non-ideal factors may cause considerable impacts on the computation if we ignore them when configuring the circuit design. For

instance, the interconnect resistance between two adjacent RRAM cells is 2.97 Ω for 22 nm technology node[33]. The resistance of a wire in a 100×100 crossbar would be as large as 300 Ω. Since the lowest resistance state of an RRAM cell is only around 500 Ω, such a large interconnection resistance may have a significant impact on the voltage distribution[17]. If we take these impacts into consideration, there is a large design space in the detailed circuit design of RRAM crossbar array for computing, and thus a technological exploration of RRAM crossbar array is necessary to provide a guidance about how to choose the technology node, the resistance levels of RRAM, the load resistance and other parameters to reduce the influence of the non-ideal factors from a basic circuit level and improve the circuit design.

3 Design Challenge Discussion

In this section, the non-ideal factors of RRAM crossbar based computing circuit are studied. Generally, the non-ideal factors can be classified into two levels: RRAM device level and circuit level. The device's non-ideal factors include the nonlinear I-V relationship of RRAM devices, the process variation[28,34], and the stochastic behavior of write operation[22,35-36]. These device factors not only have impact on the computing accuracy of RRAM-based system, but also interact with other circuit level factors and further influence some design decisions of the crossbar circuit. The structure factors contain the IR-drop phenomenon[37] and the RC delay caused by interconnect resistance. These structure factors are directly related to the behavior level performance and restrict the limit of some design parameters. Therefore, to get an optimized design considering the trade-off relationship among accuracy, power and other performance, the impacts of the non-ideal factors of both device level and circuit level need to be analyzed first.

Especially, the sneak path problem[38] will not be a major problem when RRAM crossbar array is used for computation. To further explain, the sneak path problem occurs only in memory applications when one word line and one bit line are selected for each write or read operation and the unselected lines will have negative impact on the accuracy of output signals. In matrix-vector multiplication applications, all the lines will be selected and the sneak path problem will be eliminated.

As the goal of this paper is to explore design methodologies for efficient computing systems based on RRAM crossbar array, the computation error rate in different cases should be one of the major metrics to evaluate the impact of different factors. The computation error rate of output voltage can be defined as:

$$\epsilon = \max \left| \frac{V_{\text{actual}} - V_{\text{theoretical}}}{V_{\text{theoretical}}} \right| \times 100\%,$$

where $V_{\text{theoretical}}$ is the ideal output voltage calculated by (2). Other performance, such as the operating speed of the crossbar array, is also analyzed in this section.

3.1 Non-Ideal Factors of Devices

As an emerging kind of devices, the existing practical RRAM devices cannot be directly seen as an ideal rheostat, and the non-ideal device factors can cause impacts on the behavior level computation. In this paper, the major two non-ideal device factors are analyzed: the nonlinear I-V characteristic and the variations caused by device processing and write operation.

3.1.1 Impact of Nonlinear Characteristics of RRAM Devices

As shown in (1), the I-V relationship of RRAM devices is nonlinear. However, the resistance states of RRAM devices should be constant to represent a specific matrix stably when the RRAM devices are used to perform the matrix-vector multiplication. Therefore, to confine the resistance deviations of RRAM devices, the range of the voltage applied on the RRAM devices should be limited. According to (1), the linearity of RRAM devices is mainly determined by the term $\sinh(\frac{V}{V_0})$. The RRAM device comes into an ideal linear resistance state when $V \approx 0$:

$$\sinh\left(\frac{V}{V_0}\right) \sim \frac{V}{V_0}.$$

Fig.2 illustrates the resistance states of an RRAM device under different tunneling gap distances (d) and different applied voltages (V). The tilted dotted line tracks the maximum voltage that could be applied on an RRAM device under a specific maximum deviation from the approximate linear resistance state at $V \approx 0$. For example, a voltage of 0.5 V will cause a 5% resistance deviation for $d = 0.2$ nm. Considering the same (5%) resistance deviation, the voltage is limited to the range of 0.15 V for $d = 1.9$ nm. These results demonstrate that the RRAM resistance states vary with the applied voltage and both d and V have influence on the stability of the RRAM resistance states. Since Ohmic current dominates in the low resistance state while tunneling current dominates in the high resistance state, a

smaller RRAM resistance state with a smaller tunneling gap distance d will result in a more linear I-V relationship under different voltages. Therefore, in order to achieve a more linear I-V relationship of RRAM devices, both the RRAM resistance state (the tunneling gap distance d) and the applied voltage (V) should be confined.

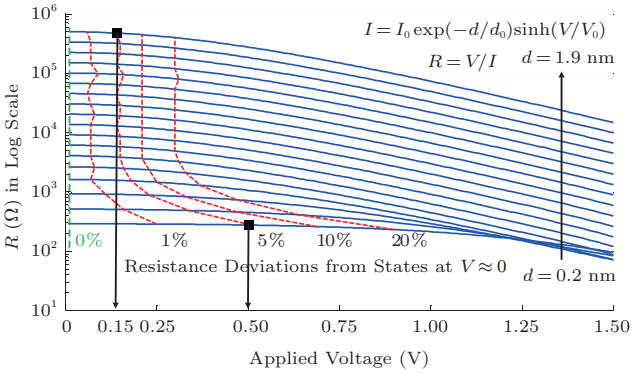


Fig.2. RRAM resistance states under different tunneling gap distances (d) and different applied voltages (V). The two vertical lines intersect the tilted dotted line with two points, representing the same voltage deviation (5%) from approximate linear resistance state at different distances (d) with distinct applied voltage. Both the tunneling gap distance d and the applied voltage (V) should be limited to achieve an approximate linear resistance state at $V \approx 0$ for a better computation result.

3.1.2 Variation of Device and Operation

The variations of RRAM device when processing computation can be caused by the variations in device fabrication and the stochastic write behavior during write operation[22]. The fabrication variation includes geometric device-to-device variations, such as length and width variations, oxide thickness variations, and the surface roughness[39]. Just as mentioned in (1), the RRAM resistance state has an exponential dependence on the tunneling gap distance (d). Therefore, the device variations may have obvious impact on the RRAM-based computing system accuracy. The write variations are mainly determined by the fluctuation of the number of vacancies and the changes in filament geometry during set and reset transient of RRAM[22,36], which leads to a stochastic write result even for the same RRAM cell.

This work focuses on the influence of non-ideal factors on computing operation instead of the factors themselves. From a more general view, both these two kinds of variations can be regarded as a stochastic fluctuation on RRAM resistance of each cell. Although the RRAM device can be theoretically tuned to any resistance value, the unpredictable deviation makes two resistance values indistinguishable if they are in the resistance's deviation range of each other. Therefore, given the variation degree, the maximum amount of resistance levels is limited by the maximum and minimum value range of an RRAM cell, which further determines the quantization precision of the numerical value saved in RRAM device.

Specifically, if the maximum deviation ratio of the device is δ, the neighboring two resistance levels R_{lower} and R_{higher} should satisfy the following inequality to distinguish them:

$$R_{\text{lower}} + R_{\text{lower}} \times \delta < R_{\text{higher}} - R_{\text{higher}} \times \delta.$$

Thus the constraint of neighboring layers is:

$$\frac{R_{\text{higher}}}{R_{\text{lower}}} > \frac{1+\delta}{1-\delta}.$$

This constraint can be further extended to the whole resistance range of RRAM. Given that the maximum resistance of an RRAM cell is R_{OFF}, while the minimum resistance is R_{ON}, if we want to put k resistance levels into the range, the constraint inequality can be expressed as:

$$\left(\frac{1+\delta}{1-\delta}\right)^k < \frac{R_{\text{OFF}}}{R_{\text{ON}}}.$$

Therefore, the maximum value of k is limited by both the variation degree δ and the resistance range, as (5) shows:

$$k < \log_{\left(\frac{1+\delta}{1-\delta}\right)} \frac{R_{\text{OFF}}}{R_{\text{ON}}}. \tag{5}$$

According to the fabrication result[28,40], the on/off ratio of HfO_2-based RRAM cell is about 10^5, and the variation ratio is about 5%∼20%. By substituting these values back to (5), we can find that the whole resistance range can only provide about 105 resistance levels at the 5% variation, and the amount will reduce to 28 when the variation is 20%. Therefore, when we consider the influence of variation of current RRAM devices, the maximum precision of the numerical value saved in RRAM can only be 4∼7 bits in practical, which is determined by the on/off ratio and the variation degree of the device. Further analysis about the detailed impact of this constraint is shown in Subsection 4.3.

3.2 Non-Ideal Factors of Crossbar Structure

As the technology node continues to scale down, the parasitic parameters induced by interconnects in crossbar structure can exert negative influence on the performance of the circuit. In this paper, two major impacts are studied: the RC delay and the interconnect resistance.

3.2.1 RC Delay

RC delay may have a negative impact on the operating speed of RRAM crossbar array based computation[41]. However, the RC delay for RRAM crossbar array is trivial (around 10 ps according to our simulation results) when the wire length between two adjacent junctions is around tens of nanometers for a 100×100 RRAM crossbar array. Therefore, the RC delay is not a major consideration of the RRAM crossbar array based computing system design. The design should focus on the performance of peripheral circuits which may significantly impact the operating speed.

3.2.2 Impact of Interconnect

In order to analyze the impact of interconnect resistance on output voltage computation accuracy, a SPICE simulation of the worst-case scenario is conducted as a corner case to guarantee the computation accuracy in normal cases. A worst-case scenario is defined that all the input voltages of the RRAM crossbar array are of the same amplitude and the worst result can be reflected by the output port which is the farthest away from the input ports, while all the RRAM cells are in the lowest resistance states R_{ON}. The load resistance (R_{S}) is set to 5 kΩ and the lowest resistance state of RRAM cells (R_{ON}) is set to 1 kΩ. The amplitude of input voltages is set to 0.9 V. The crossbar size is varied from 5×5 to 100×100 and the computation error rate is tested as defined in (2) under different technology nodes. The interconnect resistance between two adjacent junctions is 4.53 Ω, 2.97 Ω, and 1.55 Ω, respectively, for a $4F^2$ RRAM crossbar structure, where F is the feature size of RRAM device, under 16 nm, 22 nm, and 32 nm technology node according to the International Technology Roadmap for Semiconductors 2013[33]. An ideal case without any interconnect resistance is also simulated as a comparison.

The results are demonstrated in Fig.3. When the interconnect resistance ($R_{\mathrm{Interconnect}}$) is neglected, the computation error rate decreases with the rise of crossbar size $N \times N$. To be specific, the equivalent resistance of the N shunt RRAM cells in a column will drop while the load resistance in that column remains the same. The decreased voltage applied on the RRAM cells will result in better linearity, making the crossbar array represent the matrix more accurately as described in (3). Therefore, the computation accuracy increases with the crossbar size. However, when the interconnect resistance is taken into consideration, the computation error rate will decrease at the beginning and finally increase due to the voltage drop on the interconnect resistance. Therefore, under the interaction of the nonlinearity of RRAM cells and interconnect resistance, there will be an optimal crossbar size $N \times N$ for each technology node in the worst-case scenario, and the optimal crossbar size will shift slightly as the technology node scales down. On the other hand, if the crossbar size is restricted by the application, the smaller technology node leads to higher error rate, as shown in Fig.3. These results imply that the nonlinearity of RRAM cells and interconnect resistance should be considered together to realize a better implementation of the matrix-vector multiplication operations.

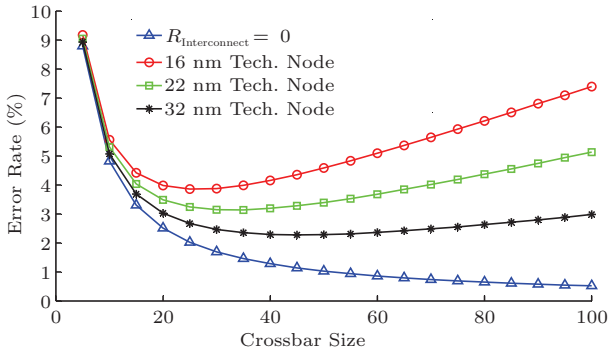


Fig.3. Worst-case computation error rates (ϵ) of RRAM crossbar arrays with different crossbar array sizes ($N \times N$) and different technology nodes. The RRAM resistance states are calculated at $V \approx 0$.

4 Technological Exploration Flow of RRAM Crossbar Array

According to the analysis in Section 3, the non-ideal factors can influence the performance of RRAM crossbar array through three design parameters: the load resistance R_{S} interacted with nonlinear I-V relationship, the resistance levels limited by variation, and the technology node of interconnect lines that influences the IR-drop phenomenon. These three parameters form the design space of an RRAM crossbar array for a matrix-vector multiplication application like SVM. In order to

overcome the impact of these non-ideal factors, we describe the proposed technological exploration flow of RRAM crossbar array and achieve a better trade-off among accuracy, energy and reliability.

Among the above three non-ideal factors, the interconnect line's influence is independent with the other two device factors, and is always chosen considering the design of other peripheral CMOS circuits. Thereby we first determine the technology node of interconnect lines. For the other two factors caused by the device itself, the nonlinear I-V characteristic mainly comes from the physical mechanism of RRAM (residual filament shown in Fig.1(a)) while the variation is essentially caused by the stochastic process of moving atoms. These two factors are also independent with each other and can be separately optimized by choosing the load resistance R_{S} and the resistance levels of RRAM. From the design view, R_{S} is more important because it captures a part of input voltage from every RRAM cell, which introduces a computation bias into the whole crossbar array. To deal with this problem, we propose a numerical iteration algorithm to map the data onto the crossbar considering the influence of R_{S} and embed this mapping algorithm into the design flow to improve the computation accuracy. The resistance range of RRAM can also influence the computation accuracy considering the variations of RRAM cells, but increasing R_{ON} can reduce the power consumption. In order to optimize the trade-off among power, accuracy and other parameters of the circuit, we propose an iterative flow to explore the design space and to find the optimal design configuration.

Fig.4 demonstrates the overview of the proposed flow. The flow consists of five stages: 1) determine the technology node according to characteristics of the application; 2) choose a proper initial R_{S} to reduce the impact of interconnect resistance; 3) reset the resistance range to the maximum range for iteratively optimizing

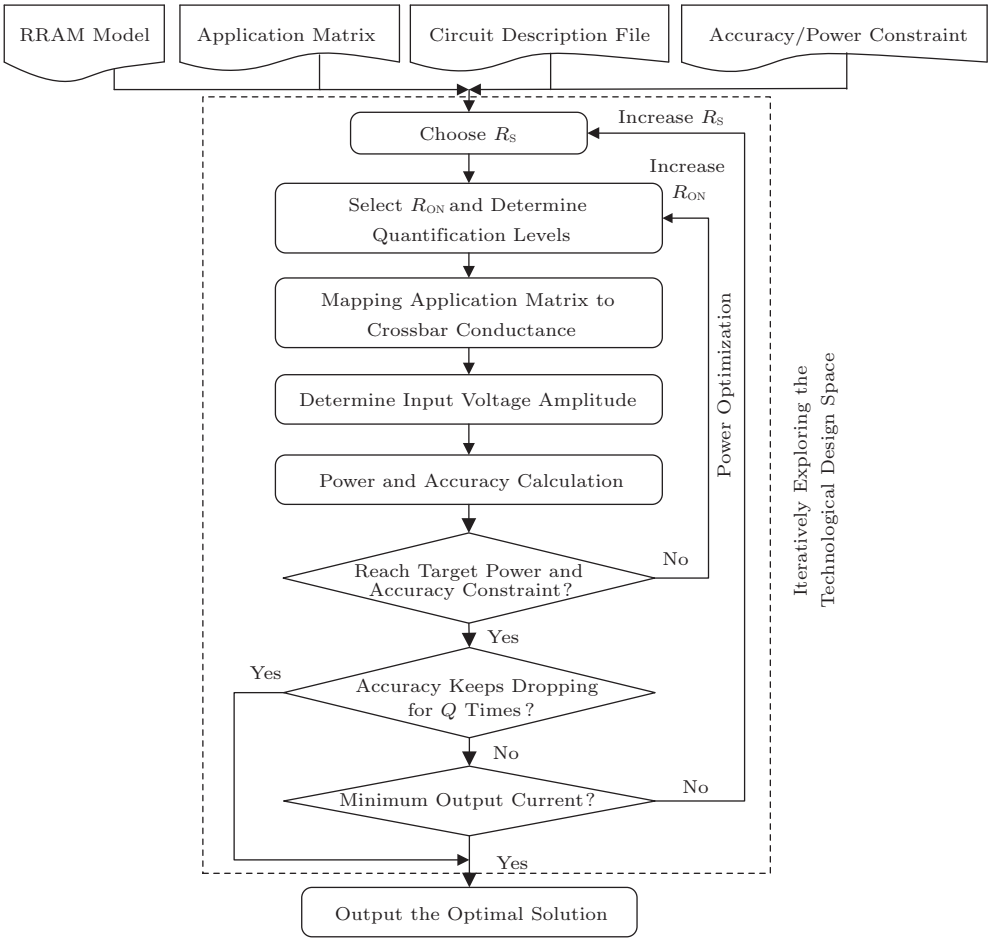


Fig. 4. Proposed technological exploration flow of RRAM crossbar array. The flow includes the configuration of technology node, resistance levels, load resistance, the algorithm of matrix mapping to crossbar array, and a consideration on the trade-off between power and performance.

the power and the accuracy; 4) map the application matrix $\boldsymbol{C}$ to the RRAM conductance matrixes $\boldsymbol{G}$ robustly; and 5) iteratively explore the technological design space and optimize the performance, energy and reliability of the system. Q is the maximum number of continuously accuracy reduction, which is used to stop iteration. In addition, although the crossbar size can also influence the effect of IR-drop as shown in Fig.3, the practical crossbar size is constrained by the characteristics of the specific application. Consequently, the proposed design flow does not consider the configuration of RRAM crossbar array size.

4.1 First Stage: Determining the Technology Node

As the interconnect resistance has negative impact on the computation accuracy of RRAM crossbar array, the technology node should be scaled up to support applications that require a large crossbar array or high computation accuracy. Meanwhile, the scaling down of technology node will shrink the area of RRAM crossbar array. Therefore, there may exist a trade-off between the area and the computation accuracy. After the setup of crossbar size and technology node, device level parameters can be further configured as discussed in the next stage.

4.2 Second Stage: Choice of R_{S}

Besides the value of interconnect resistance, many other parameters, such as the value of R_{S} and the resistance states of RRAM cells, also influence the computation accuracy of RRAM crossbar based computation. Since the practical computation accuracy is heavily dependent on the pattern of input signals and the resistance distribution of RRAM cells, large quantities of variables form a complex design space. In order to extract the key parameters and simplify the design options, the worst-case scenario is studied so that the negative influence of interconnect resistance can be fully exposed.

The value of R_{S} needs to be determined considering R_{ON} since R_{S} and R_{ON} influence the linearity of RRAM cells together. Theoretically, when R_{S} increases or R_{ON} decreases, the voltage applied on the RRAM cells will decline. As discussed in Subsection 3.1.1, a smaller applied voltage will result in better linearity of RRAM devices and better computation accuracy. However, a smaller R_{ON} can also lead to more serious impact of the interconnect resistance. The impact of R_{ON} on the computation accuracy is hard to predict. In order to better study the impact of R_{S} and R_{ON} in the worst-case scenario as defined in Subsection 3.2, where all the RRAM cells are set to R_{ON}, a simulation is conducted. The crossbar size is set to 50×50 and the amplitude of input voltages (which are the same) are set to 0.9 V (about 0.1 V will be applied on the RRAM cells). The technology node is set to 22 nm. We vary R_{ON} from 500 Ω to 5 kΩ and vary R_{S} from 1 kΩ to 11 kΩ.

The simulation results are illustrated in Fig.5. It demonstrates that the computation error rate decreases exponentially with the rise of R_{S}. Compared with R_{S}, the computation accuracy improves less than 1% when R_{ON} varies from 500 Ω to 5 kΩ under the same R_{S}. This result indicates that R_{ON} has little direct impact on the computation accuracy when not considering the limited resistance levels caused by variation. Therefore, the choice of R_{ON} can be neglected for convenience, and the technological exploration flow should focus on the choice of R_{S}. To be specific, the simulation results illustrated in Fig.5 can serve as a look-up table and the technological exploration flow will first choose a proper initial R_{S} to satisfy the worst case and reduce the impact of interconnect resistance. In addition, since the application performance is also influenced by the practical resistance distribution of RRAM cells, a larger R_{S} cannot guarantee a better computation accuracy. A smaller initial R_{S} can be used and the optimal choice of R_{S} can be achieved by iteratively exploring the technological design space in the next stages of the technological exploration flow.

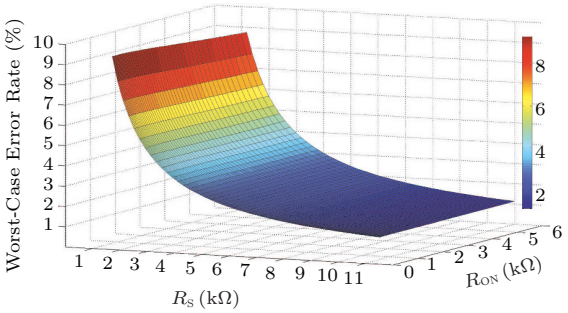


Fig.5. Computation error rates of RRAM crossbar array with different R_{S}s and R_{ON}s. The simulation results demonstrate that the computation error rate decreases exponentially with R_{S}, while R_{ON} has little impact on the computation accuracy. Therefore, the technological exploration flow of RRAM crossbar array for matrix-vector multiplication should focus on the choice of R_{S}. The size of crossbar array is 50×50.

4.3 Third Stage: Choice of Resistance Range

Low power consumption is one of the main advantages of RRAM-based circuit[42]. Therefore many researchers concern power consumption more than accuracy especially for the low-power applications like approximate computing[9,43]. Obviously, if we increase the resistance of R_{ON}, the resistance value of each level after mapping will increase, leading to a lower power consumption of the whole crossbar shown in Fig.1(b). However, the increasing of R_{ON} results in a smaller resistance range of RRAM. As discussed in Subsection 3.1, given the device variation, the range of RRAM resistance restricts the maximum amount of resistance levels for error-free separation. For a practical computation, the precision of the number saved in RRAM cells is determined by the application requirement. As a result, when R_{ON} rises, the distance between two neighboring resistance levels gets smaller, and finally breaks the error-free constraint. On the other hand, if the precision given by application is already larger than that the RRAM's characteristic can support, the decreasing resistance range will further reduce the computation accuracy of RRAM crossbar array. Fig.6 shows the SPICE simulation result of different R_{ON}s and variations. Considering that the Verilog-A RRAM model contains the resistance range from 300 Ω to 500 kΩ[17], we select 200 kΩ as the value of R_{OFF} to support the variation range and change the value of R_{ON} from 500 Ω to 50 kΩ, and the data precision is set to be 6-bit (64 resistance levels). The result shows the relationship between accuracy and R_{ON} influenced by variation.

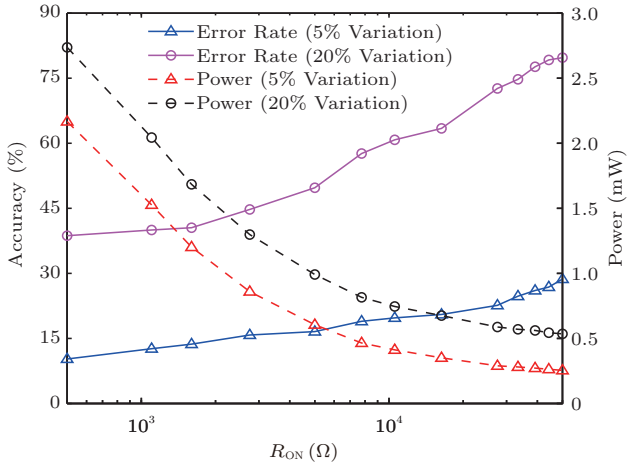


Fig.6. Practical relationship between accuracy, power and R_{ON} at 5% and 20% variation when using 6-bit precision (64 resistance levels). The crossbar size is 50×50. The accuracy and the power data are the average results of 150 matrix samples and 150 input samples.

To further analyze the trade-off relationship between accuracy and power, we simulate the power consumption of a 50×50 crossbar when processing matrix-vector multiplication with 5% variation. The relationship between power and error rate is shown in Fig.7. The result shows that the power reduces rapidly at first, which means we can obtain considerable power saving at the cost of a little accuracy. This is because when R_{ON} is small enough, the low resistance RRAM cells cost most of the power in the whole crossbar circuit, and increasing their resistance can significantly reduce the power according to the inversely proportional relationship between resistance and power. However, when R_{ON} has already been large enough, further increasing the RRAM resistance only has a little effect on power saving, but can cause the rapid drop of accuracy as shown in Fig.6. As a result, there is an inflection point in the trade-off line. Designers can choose this point as the optimized result, or use one parameter as a constraint to optimize the other one.

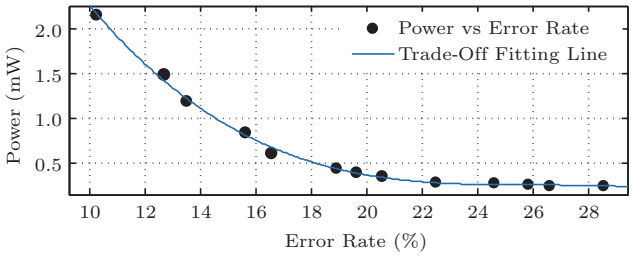


Fig.7. Trade-off relationship between power and accuracy for a 50×50 crossbar array among different mapping resistance ranges. The amount of resistance levels is 64 (6-bit precision) and the variation is 5%.

As shown in Fig.4, the proposed flow can optimize the above trade-off, which is a small nested loop. After choosing R_{S}, we need to select a resistance range (or actually an R_{ON} for most cases) to determine the final quantization levels of RRAM resistance according to the precision of application. During the mapping phase, the mapping results of resistances (or conductances) need to be quantified into the determined levels. Finally, the estimated power and accuracy are tested if they can satisfy the restriction provided by designers according to the monotonic relationship between power and accuracy. For example, if we give minimum accuracy as a constraint, we can gradually increase R_{ON} and reduce the power consumption until the accuracy is lower than the threshold, which reaches a minimum power. Oppositely, given a maximum power cost as a constraint, we can also gradually increase R_{ON} and find

for the first time that the power consumption is lower than the threshold, which reaches maximum accuracy.

In addition, when the resistance range gets too small, the difference between different computation results will also reduce according to (2), resulting in a new challenge to the precision of read circuit. This phenomenon can be regarded as another restriction like the power or accuracy restriction in the above flow and we can also introduce this condition into the judgment phase to further limit the design space.

4.4 Fourth Stage: Mapping Matrix Parameters to RRAM Device Conductivities Robustly

The conductance states of RRAM cells in the crossbar array must be configured properly to realize the multiplied matrix $\boldsymbol{C}$. However, as shown in (3), $c_{k,j}$ not only relies on the conductivity of the corresponding RRAM cell $g_{k,j}$, but also depends on all the RRAM cells' conductance states in the same j-th column in the crossbar array. In order to realize a one-to-one mapping between matrix $\boldsymbol{C}$ and the conductance matrix of the RRAM crossbar array, some previous work proposed a few simple and fast approximations to the mapping problem like [10]:

$$g_{k,j} = c'_{k,j} \times (g_{\mathrm{ON}} - g_{\mathrm{OFF}}) + g_{\mathrm{OFF}}.$$

When

$$g_s \gg (g_{\mathrm{ON}} - g_{\mathrm{OFF}}) \times \sum_{l=1}^{N} c'_{k,l}, \tag{6}$$

(3) can be approximated to:

$$c_{k,j} \approx c'_{k,j} \times \frac{g_{\mathrm{ON}}}{g_s} = c'_{k,j} \times g_{\mathrm{ON}} \times R_{\mathrm{S}},$$

where $c_{k,j}$ is the matrix parameter of a specific application and $c'_{k,j}$ is a decayed version of $c_{k,j}$. g_{ON} and g_{OFF} are the maximum and the minimum conductance states of the RRAM cells in the crossbar array.

The above equation demonstrates a linear one-to-one mapping between matrix $\boldsymbol{C}$ and the RRAM conductance matrixes $\boldsymbol{G}$ when g_s is determined. However, the precondition of the approximation may be difficult to be satisfied and may decrease computation accuracy. For example, $R_{\mathrm{ON}} \approx 1$ kΩ for a physical RRAM device[17,44] and $\sum_{l=1}^{N} c'_{k,l} \approx 5$ for a 256×256 RRAM crossbar array[10]. And according to (6), R_{S} should be about 100 Ω. However, as described in Subsection 4.2, such a small R_{S} will lead to a large computation error because of the interconnect resistance and the nonlinearity of RRAM devices.

In this work, we propose a numerical iteration algorithm to map matrix $\boldsymbol{C}$ to the conductance matrixes $\boldsymbol{G}$ without any approximation, which improves the computation accuracy of RRAM crossbar array.

(3) can be expressed as:

$$g_{k,j} - c_{k,j} \times \sum_{l=1}^{N} g_{k,l} = g_s \times c_{k,j}. \tag{7}$$

If we combine the N equations (for $j = 1, 2, ..., N$ in (7)) together, all of the RRAM cells in the k-th column in the crossbar array can form a system of linear equations of N variables together:

$$\begin{pmatrix} 1-c_{k,1} & -c_{k,1} & \cdots & -c_{k,1} \\ -c_{k,2} & 1-c_{k,2} & \cdots & -c_{k,2} \\ \vdots & \vdots & \ddots & \vdots \\ -c_{k,N} & -c_{k,N} & \cdots & 1-c_{k,N} \end{pmatrix} \begin{pmatrix} g_{k,1} \\ g_{k,2} \\ \vdots \\ g_{k,N} \end{pmatrix} = \begin{pmatrix} g_s c_{k,1} \\ g_s c_{k,2} \\ \vdots \\ g_s c_{k,N} \end{pmatrix}. \tag{8}$$

The accurate conductance states of RRAM cells ($g_{k,j}$) can be achieved by solving the above equations when the matrix parameters $c_{k,j}$ are provided. However, several constraints must be considered to guarantee the solved conductance states can be realized by physical RRAM cells. The first constraint results from the range of conductance states that can be realized by physical RRAM devices. Suppose the minimum and the maximum conductance states of RRAM cells in a crossbar array are g_{OFF} and g_{ON}, respectively. The parameters $c_{k,j}$ must be of the following range to enable all the solved $g_{k,j}$ to be within the range between g_{OFF} and g_{ON}.

$$\chi_{\min} \leqslant c_{k,j} \leqslant \chi_{\max}, \tag{9}$$

where $\chi_{\max}$ and $\chi_{\min}$ are the maximum and the minimum matrix parameters that can be represented by a physical RRAM crossbar array:

$$\chi_{\min} = \frac{g_{\mathrm{OFF}}}{g_s + g_{\mathrm{OFF}} + (N-1)g_{\mathrm{ON}}}, \tag{10}$$

$$\chi_{\max} = \frac{g_{\mathrm{ON}}}{g_s + g_{\mathrm{ON}} + (N-1)g_{\mathrm{OFF}}}. \tag{11}$$

Moreover, as described in (4), two crossbars are required to represent a matrix with both positive and

negative parameters. In order to satisfy the condition described in (9), (4) should be revised to:

$$\boldsymbol{C}_0 = \boldsymbol{C}_0^+ - \boldsymbol{C}_0^- = \alpha((\boldsymbol{C}^+ + \Delta) - (\boldsymbol{C}^- + \Delta)), \quad (12)$$

where:

$$c_{k,j}^+ = \begin{cases} c_{k,j}, & \text{if } c_{k,j} > 0, \\ 0, & \text{if } c_{k,j} \leqslant 0, \end{cases} \quad (13)$$

$$c_{k,j}^- = \begin{cases} -c_{k,j}, & \text{if } c_{k,j} < 0, \\ 0, & \text{if } c_{k,j} \geqslant 0. \end{cases} \quad (14)$$

α and Δ are parameters to map $\boldsymbol{C}_0^+$ and $\boldsymbol{C}_0^-$ to the range described in (9). The choice of α and Δ can be achieved by exhausted search. In order to reduce the search space, a restriction of α and Δ is required. We set $c_{\max} = \max(|c_{k,j}|)$. According to (9)$\sim$(12), the constraints of α and Δ can be expressed as:

$$\frac{\chi_{\min}}{\alpha} \leqslant \Delta \leqslant \frac{\chi_{\max}}{\alpha} - c_{\max}, \quad (15)$$

$$\alpha \leqslant \frac{\chi_{\max} - \chi_{\min}}{c_{\max}}. \quad (16)$$

Finally, Algorithm 1 demonstrates the steps of mapping matrix $\boldsymbol{C}$ to the conductance matrixes $\boldsymbol{G}^+$ and $\boldsymbol{G}^-$. Lines 1$\sim$4 in the algorithm are used to set up parameter constraints. Lines 7, 8, 10, and 11 calculate candidate $\boldsymbol{G}^+$ and $\boldsymbol{G}^-$. Lines 9 and 12 check the feasibility of candidate solutions.

Algorithm 1. Robust Parameter Mapping Algorithm

Input: $\boldsymbol{C}$, $g_{\rm ON}$, $g_{\rm OFF}$, g_s, *SearchStep*
Output: $\boldsymbol{G}^+$, $\boldsymbol{G}^-$

1 Calculate $\boldsymbol{C}^+$ and $\boldsymbol{C}^-$ according to (13) and (14);
2 Calculate $\chi_{\min}$ and $\chi_{\max}$ according to (10) and (11);
3 Calculate $\alpha_{\max}$ according to (16);
4 Calculate $\Delta_{\max}$ and $\Delta_{\min}$ according to (15);
5 **for** $\alpha = \alpha_{\max} : -SearchStep : 0$ **do**
6 **for** $\Delta = \Delta_{\min} : SearchStep : \Delta_{\max}$ **do**
7 $\boldsymbol{C}_0^+ \leftarrow \alpha(\boldsymbol{C}^+ + \Delta)$;
8 Calculate $\boldsymbol{G}^+$ by solving the equation set in (8)
9 **if** $g_{\rm OFF} \leqslant \boldsymbol{G}^+ \leqslant g_{\rm ON}$ **then**
10 $\boldsymbol{C}_0^- \leftarrow \alpha(\boldsymbol{C}^- + \Delta)$;
11 Calculate $\boldsymbol{G}^-$ by solving the equation set in (8)
12 **if** $g_{\rm OFF} \leqslant \boldsymbol{G}^- \leqslant g_{\rm ON}$ **then**
13 **return** $\boldsymbol{G}^+$ *and* $\boldsymbol{G}^-$
14 **end**
15 **end**
16 **end**
17 **end**
18 **return** *'Bad Parameters'*

4.5 Fifth Stage: Trade-Off Among Performance, Energy and Reliability

The proposed technological exploration flow iteratively tests the performance of different parameters and tracks the optimal point. To be specific, the technological exploration flow will first choose a proper initial $R_{\rm S}$ as discussed in Subsection 4.2. The selected $R_{\rm S}$ should be a small one to guarantee the amplitude of output current. Afterwards, the technological exploration flow will calculate the corresponding $\boldsymbol{G}^+$ and $\boldsymbol{G}^-$ according to the selected $R_{\rm S}$ and the proposed robust mapping algorithm. The calculated parameters will be used for simulating the detailed performance of the RRAM crossbar array based computing systems. As a larger $R_{\rm S}$ may lead to better computation accuracy, and less energy consumption but smaller amplitude of output current, the technological exploration flow will keep increasing $R_{\rm S}$ gradually to track the change of the system performance, energy and reliability. The exploration of the design space will stop once the output current becomes too small. In addition, the rise of $R_{\rm S}$ can only guarantee that the computation accuracy increases in the worst case. The input pattern and the RRAM resistance state distribution may lead to worse computation accuracy for a larger $R_{\rm S}$. Therefore, the exploration of the design space will also stop when the computation accuracy begins to decrease continuously for a period of time. Finally, by comparing all the tracked solutions, the technological exploration flow is able to provide a solution with better trade-off among performance, energy, and reliability.

5 Experimental Results

In this section, we use the support vector machine (SVM) as a case study to demonstrate the performance of the proposed technological exploration flow.

Support vector machine (SVM) is one of the most crucial machine learning algorithms[45] with considerable matrix-vector multiplication workload. Supposing the data can be represented as $\boldsymbol{x}$, SVM focuses on learning the hyperplane $\boldsymbol{w}$ with max-margin to distinguish $\boldsymbol{x}$ and other data. The decision of the class of $\boldsymbol{x}$ is determined by the sign of calculating $\boldsymbol{w}^{\rm T}\boldsymbol{x} + b = \boldsymbol{w'}^{\rm T}\boldsymbol{x'}$, where $\boldsymbol{x'} = (1; \boldsymbol{x})$ and $\boldsymbol{w'} = (b; \boldsymbol{w})$. Since many hyperplanes $\boldsymbol{w}$ can form a matrix $\boldsymbol{W}$ together, the major operation of an SVM is the matrix-vector multiplication. Therefore, we use the RRAM crossbar array and the proposed technological exploration flow to implement an SVM and test its performance.

5.1 Experimental Setup

In our experiment, the MNIST dataset is used to test the performance of RRAM-based SVM. MNIST

is a widely used dataset with more than 60 000 handwritten digits for optical character recognition. In our experiment, we choose 20 000 examples of handwritten digits of "0"∼"9" to train the SVM. We extract a 49-dimension feature through principal component analysis (PCA)[46] from the original 28×28 images. In other words, the dimension of input data $\hat{x}$ is 50 when one dimension for the offset b is considered. As there are 10 classes of handwritten digits in the MNIST dataset, we train 10 different SVMs to distinguish only one digit from the others. The recognition accuracy of SVM trained on CPU is 94%. And the size of the combined matrix $\boldsymbol{W}$ of 10 SVMs is 50×10. We realize this matrix with a 50×50 RRAM crossbar array. All the other 40 output ports are regarded as virtual nodes whose states will not be considered. The unused RRAM cells in the crossbar array are set to the highest resistance states to reduce the extra energy consumption and negative impact. Other 5 000 examples in the MNIST dataset are used to test the performance of RRAM-based SVM. The maximum amplitude of input voltage is set to 1 V to achieve better linearity of RRAM devices. Most of the input voltages applied on the RRAM cells are around tens to hundreds of millivolt. A current comparator is used to select the port with the highest output current and provide the recognition results. We use SPICE to simulate the circuit performance of RRAM crossbar, such as the power consumption and the output voltage. A recent Verilog-A model described in [17] is chosen as the RRAM device model. The simulation results are provided in Table 1. Some comparisons are made between the proposed technological exploration flow and the method based on [10] under different technology nodes.

5.2 Performance of Matrix Mapping Algorithm

We first compare the proposed matrix mapping algorithm with the one proposed in [10] under the same technology node. The experimental results demonstrate that both algorithms work well when R_{S} is very small ($R_{\mathrm{S}} = 100\ \Omega$). However, as discussed in Subsection 4.2, such a small R_{S} will lead to bad computation accuracy because of interconnect resistance. Only around 80% recognition accuracy is achieved in this situation. As for the cases with a larger R_{S} of 3 kΩ, the recognition accuracy of the proposed technological exploration flow significantly increases to more than 90%, while a dramatic decrease from 90% to 9% is observed for the previous method. These results demonstrate that the approximation used in the previous work does not work well for a larger R_{S}. And the proposed method is robust since there is no approximation used in the mapping algorithm.

5.3 Impact of R_{S} and Interconnects

We also increase R_{S} to 10 kΩ to test the impact of R_{S} on the performance of RRAM-based SVM. We first

Table 1. Experimental Results of RRAM-Based SVM with Differnet Parameters ($R_{\mathrm{ON}} = 500\ \Omega$)

Map Algorithm	Technology Node (nm)	R_{S} (Ω)	Signal Fluctuation (%)	Device Variation (%)	Accuracy (%)	Accuracy Improvement (%)	Power (mW)	Power Savings (%)
[10]	22	100	0	0	82	–	1.96	–
[10]	22	3k	0	0	9	–89.02	1.93	2.02
[10]	Ideal	1	0	0	90	9.76	3.00	–52.73
Proposed	22	100	0	0	83	1.22	4.07	–106.94
Proposed	22	3k	0	0	93	13.41	2.02	–3.04
Proposed	16	3k	0	0	90	9.76	1.97	–0.40
Proposed	16	10k	0	0	83	1.22	1.40	28.99
Proposed	22	10k	0	0	86	4.88	1.42	27.64
Proposed	32	10k	0	0	91	**10.98**	1.45	**26.40**
Proposed	22	3k	0	5	90	9.76	2.16	–7.18
Proposed	22	3k	0	10	74	–9.76	2.13	–8.36
Proposed	22	3k	0	20	53	–35.37	2.62	–33.26
Proposed	22	3k	5	0	92	12.20	2.03	–3.33
Proposed	22	3k	10	0	90	9.76	2.11	–7.59
Proposed	22	3k	20	0	87	6.10	2.07	–5.51

fix the technology node to test the impact of R_S. Compared with the cases when $R_S = 3$ kΩ, the recognition accuracy does not increase but drops from 93% to 86%. The reason lies in that a different R_S will lead to different RRAM conductance matrixes. The RRAM conductance matrix at R_S may be affected more seriously by the variation of RRAM resistance states and the interconnect resistance. Such results verify the discussion in Subsection 4.5 that a larger R_S is not necessary to lead to better computation accuracy in practical machine learning tasks instead of the worst case. Then, we vary the technology node of interconnection from 16 nm to 32 nm fixing R_S. The results demonstrate that a lower interconnect resistance is beneficial to the recognition accuracy for RRAM-based SVM.

5.4 Power Saving by Resistance Range Optimization

As mentioned in Subsection 4.3, the proposed flow can further optimize the power consumption at the cost of a little accuracy reduction. To verify the power optimization effect of the proposed method, we use an accuracy threshold and find the minimum power consumption with different interconnect technology nodes from 18 nm to 36 nm, and the device variation is still 5%. Considering that the related RRAM-based work's result is about 82%[10], we use 80% as the classification accuracy constraint. The results are shown in Table 2. The result shows that by utilizing the trade-off relationship between power and accuracy, about 80% power consumption can be saved in various interconnect technologies. Another experiment shows that if we use 85% as the classification accuracy constraint, the power consumption saving is about 70%, which is also a considerable gain. However, according to the trade-off relationship shown in Fig.7, further reducing the accuracy threshold only has little effect, and thus 80%∼85% is a relatively reasonable range for power optimization.

Table 2. Power Saving with a Restricted Accuracy Threshold (Initial $R_{ON} = 500$ Ω)

Technology Node (nm)	Accuracy Threshold (%)	Optimal R_{ON} (kΩ)	Initial Power (mW)	Optimal Power (mW)	Power Savings (%)
16	80	17.0	2.10	0.340	83.7
22	80	16.3	2.16	0.347	83.9
28	80	16.0	2.19	0.351	84.0
36	80	16.0	2.26	0.353	84.4
22	85	5.0	2.16	0.611	71.7

5.5 Robustness of RRAM Crossbar Array

The above results demonstrate that the RRAM-based SVM works well under ideal conditions. However, several non-ideal factors may influence the RRAM-based SVM performance. In this subsection, we discuss the impact of device variation, resistance resolution, and signal fluctuation to test the robustness of the RRAM-based SVM.

5.5.1 Impact of Device Variation and Resistance Resolution

Given a certain 8-bit data precision (256 resistance levels), we test the performance of RRAM-based SVM with different maximum deviations of 5%, 10%, and 20% respectively, as shown in Table 1. The simulation results verify the above hypothesis and the recognition accuracy significantly drops from 90% to only 53%. The RRAM-based SVM is very sensitive to the variation of RRAM resistance.

The above experiments are based on the condition that the precision of data saved in RRAM crossbar arrays is determined by the application. As a result, when the variation degree is large enough, the neighboring resistance levels of RRAM cannot be clearly distinguished, as (5) shows. In other words, the data precision has already lost the effect, which means we can reduce the data precision (namely the RRAM resistance resolution) to match the variation level in order to obtain an error-free result of data mapping. We reverse (5) to find the maximum variation degree that a specific data precision can support:

$$\delta < \frac{\sqrt[k]{\frac{R_{OFF}}{R_{ON}}} - 1}{\sqrt[k]{\frac{R_{OFF}}{R_{ON}}} + 1},$$

where δ is the variation degree and k is the amount of RRAM resistance levels. For instance, when we reduce the data precision down to 4-bit, namely, 16 resistance levels, the maximum δ is 18.51%. We test the performance of SVM based on different data precisions with corresponding maximum variation degree, where the practical data is 8-bit in this experiment. The simulation results are illustrated in Fig.8. The result shows that 6-bit is an inflection point for accuracy (90% for SVM), which means the variation degree needs to be less than 4.68%. When we use less data precision to realize error-free recognition of resistance levels, the accuracy still drops rapidly, which is similar to the results

of directly using the original data precision as Table 1 shows.

5.5.2 Impact of Signal Fluctuations

The electrical noise from the input ports will lead to input signal fluctuation. Here we simulate the performance of RRAM-based SVM under different fluctuations of input signals. The results show that the proposed RRAM-based SVM is robust to the signal fluctuations. For example, a 10% variation of the input signal only reduces the recognition accuracy from 92% to 90%. These results demonstrate that the RRAM-based SVM is able to work in the environments with large signal fluctuations.

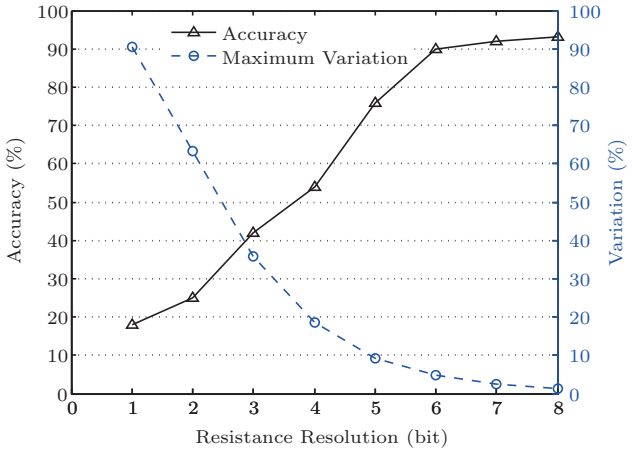


Fig.8. Recognition accuracy with different RRAM resistance resolutions under maximum variation. Each accuracy result is the average value of four variation matrixes.

6 Conclusions

In this paper, we studied the impact of a wide range of parameters and proposed a technology exploration flow to configure these parameters to achieve a better trade-off among performance, energy and reliability for RRAM crossbar array based computing system design. We first analyzed the impact of both device level and circuit level non-ideal factors, including the nonlinear I-V relationship of RRAM devices, the variation of device processing and write operation, the interconnects, and other device parameters. In order to overcome the impact of these non-ideal factors and achieve a better trade-off among performance, energy and reliability, we proposed a technological exploration flow for device parameter configuration of RRAM crossbar array based computation, including the technology node and load resistance configuration, and the algorithm of matrix mapping to crossbar array with considerations on the trade-off between power and performance. We used the MNIST dataset and a linear SVM classifier as a case study to test the performance of the proposed framework. The simulation results show 10.98% improvement of recognition accuracy and 26.4% power reduction compared with previous work[10], and can further receive an 84.4% power saving at the cost of little accuracy reduction. In addition, although this work focuses on the load resistance based read scheme for RRAM crossbar, other kinds of read peripheral circuits such as sense amplifiers and analog to digital converters can also be regarded as equivalent resistances or impedances connected to the crossbar array. Therefore, the proposed design flow can be extended to other read schemes with little modification, and the experimental results are still valuable for these designs.

In the future, we will further explore how to compensate the impact of IR-drop problem in mapping, which can improve the computation accuracy of RRAM-based matrix-vector multiplication especially when using large crossbars, we will also develop a friendly design automation tool.

References

[1] Franklin J. The elements of statistical learning: Data mining, inference and prediction. *The Mathematical Intelligencer*, 2005, 27(2): 83-85.

[2] Jang J W, Choi S B, Prasanna V K. Energy- and time-efficient matrix multiplication on FPGAs. *IEEE Transactions on Very Large Scale Integration* (*VLSI*) *Systems*, 2005, 13(11): 1305-1319.

[3] Williams S, Oliker L, Vuduc R *et al.* Optimization of sparse matrix-vector multiplication on emerging multicore platforms. *Parallel Computing*, 2009, 35(3): 178-194.

[4] Catanzaro B, Sundaram N, Keutzer K. Fast support vector machine training and classification on graphics processors. In *Proc. the 25th International Conference on Machine Learning*, July 2008, pp.104-111.

[5] Dean J, Corrado G, Monga R *et al.* Large scale distributed deep networks. In *Advances in Neural Information Processing Systems 25*, Pereira F, Burges C, Bottou L, Weinberger K (eds.), Curran Associates, Inc., 2012, pp.1232-1241.

[6] Xu C, Dong X, Jouppi N P *et al.* Design implications of memristor-based RRAM cross-point structures. In *Proc. Design, Automation and Test in Europe Conference and Exhibition* (*DATE*), March 2011.

[7] Wang Y, Li B, Luo R *et al.* Energy efficient neural networks for big data analytics. In *Proc. Design, Automation and Test in Europe Conference and Exhibition* (*DATE*), March 2014.

[8] Hu M, Li H, Wu Q *et al.* Hardware realization of BSB recall function using memristor crossbar arrays. In *Proc. the*

49th Annual Design Automation Conference, June 2012, pp.498-503.

[9] Li B, Shan Y, Hu M *et al*. Memristor-based approximated computation. In *Proc. the International Symposium on Low Power Electronics and Design*, September 2013, pp.242-247.

[10] Hu M, Li H, Chen Y *et al*. Memristor crossbar-based neuromorphic computing system: A case study. *IEEE Transactions on Neural Networks and Learning Systems*, 2014, 25(10): 1864-1878.

[11] Li B, Wang Y, Wang Y *et al*. Training itself: Mixed-signal training acceleration for memristor-based neural network. In *Proc. the 19th Asia and South Pacific Design Automation Conference* (*ASP-DAC*), January 2014, pp.361-366.

[12] Deng Y, Huang P, Chen B *et al*. RRAM crossbar array with cell selection device: A device and circuit interaction study. *IEEE Transactions on Electron Devices*, 2013, 60(2): 719-726.

[13] Seo K, Kim I, Jung S *et al*. Analog memory and spike-timing-dependent plasticity characteristics of a nanoscale titanium oxide bilayer resistive switching device. *Nanotechnology*, 2011, 22(25): 254023.

[14] Chang T, Jo S H, Lu W. Short-term memory to long-term memory transition in a nanoscale memristor. *ACS Nano*, 2011, 5(9): 7669-7676.

[15] Fang Z, Yu H, Li X *et al*. Multilayer-based forming-free RRAM devices with excellent uniformity. *IEEE Electron Device Letters*, 2011, 32(4): 566-568.

[16] Wong H S P, Lee H Y, Yu S *et al*. Metal-oxide RRAM. *Proceedings of the IEEE*, 2012, 100(6): 1951-1970.

[17] Yu S, Gao B, Fang Z *et al*. A low energy oxide-based electronic synaptic device for neuromorphic visual systems with tolerance to device variation. *Advanced Materials*, 2013, 25(12): 1774-1779.

[18] Jiao B, Deng N, Yu J *et al*. Resisitive switching variability study on 1T1R ALOX/WOx-based RRAM array. In *Proc. International Conference of Electron Devices and Solid-State Circuits* (*EDSSC*), June 2013.

[19] Goux L, Fantini A, Kar G *et al*. Ultralow sub-500nA operating current high-performance TINAL 2O3 HfO2 HFTiN bipolar RRAM achieved through understanding-based stack-engineering. In *Proc. Symposium on VLSI Technology* (*VLSIT*), June 2012, pp.159-160.

[20] Young-Fisher K G, Bersuker G, Butcher B *et al*. Leakage current-forming voltage relation and oxygen gettering in HfO_x RRAM devices. *IEEE Electron Device Letters*, 2013, 34(6): 750-752.

[21] Yu S, Guan X, Wong H S P. On the stochastic nature of resistive switching in metal oxide RRAM: Physical modeling, monte carlo simulation, and experimental characterization. In *Proc. International Electron Devices Meeting* (*IEDM*), December 2011, pp.17.3.1-17.3.4.

[22] Degraeve R, Fantini A, Raghavan N *et al*. Causes and consequences of the stochastic aspect of filamentary RRAM. *Microelectronic Engineering*, 2015, 147: 171-175.

[23] Long S, Lian X, Cagli C *et al*. A model for the set statistics of RRAM inspired in the percolation model of oxide breakdown. *IEEE Electron Device Letters*, 2013, 34(8): 999-1001.

[24] Guan X, Yu S, Wong H S. A SPICE compact model of metal oxide resistive switching memory with variations. *IEEE Electron Device Letters*, 2012, 33(10): 1405-1407.

[25] Guan X, Yu S, Wong H S P. On the switching parameter variation of metal-oxide RRAM—Part I: Physical modeling and simulation methodology. *IEEE Transactions on Electron Devices*, 2012, 59(4): 1172-1182.

[26] Li B, Gu P, Shan Y *et al*. RRAM-based analog approximate computing. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2015, 34(12): 1905-1917.

[27] Wang Y, Tang T, Xia L *et al*. Energy efficient RRAM spiking neural network for real time classification. In *Proc. the 25th Edition on Great Lakes Symposium on VLSI*, May 2015, pp.189-194.

[28] Chen A, Lin M R. Variability of resistive switching memories and its impact on crossbar array performance. In *Proc. International Reliability Physics Symposium* (*IRPS*), April 2011, pp.MY.7.1-MY.7.4.

[29] Lee H, Che P, Wu T *et al*. Low power and high speed bipolar switching with a thin reactive Ti buffer layer in robust HfO2 based RRAM. In *Proc. International Electron Devices Meeting*, December 2008.

[30] Serrano-Gotarredona T, Masquelier T, Prodromakis T *et al*. STDP and STDP variations with memristors for spiking neuromorphic learning systems. *Frontiers in Neuroscience*, 2013, 7(7): 2.

[31] Tang T, Luo R, Li B *et al*. Energy efficient spiking neural network design with RRAM devices. In *Proc. the 14th International Symposium on Integrated Circuits* (*ISIC*), December 2014, pp.268-271.

[32] Querlioz D, Bichler O, Gamrat C. Simulation of a memristor-based spiking neural network immune to device variations. In *Proc. International Joint Conference on Neural Networks* (*IJCNN*), July 31-Aug.5, 2011, pp.1775-1781.

[33] ITRS teams. International technology roadmap for semiconductors: 2013 edition executive summary. http://public.itrs.net/ITRS%201999-2014%20Mtgs,%20Presentations%20&%20Links/2013ITRS/2013Chapters/2013ExecutiveSummary.pdf, Dec. 2015.

[34] Dongale T, Patil K, Mullani S *et al*. Investigation of process parameter variation in the memristor based resistive random access memory (RRAM): Effect of device size variations. *Materials Science in Semiconductor Processing*, 2015, 35: 174-180.

[35] Walczyk D, Walczyk C, Schroeder T *et al*. Resistive switching characteristics of CMOS embedded HfO2-based 1T1R cells. *Microelectronic Engineering*, 2011, 88(7): 1133-1135.

[36] Lee S R, Kim Y B, Chang M *et al*. Multi-level switching of triple-layered TaOx RRAM with excellent reliability for storage class memory. In *Proc. Symposium on VLSI Technology* (*VLSIT*), June 2012, pp.71-72.

[37] Liu B, Li H, Chen Y *et al*. Reduction and IR-drop compensations techniques for reliable neuromorphic computing systems. In *Proc. International Conference on Computer-Aided Design* (*ICCAD*), November 2014, pp.63-70.

[38] Kannan S, Rajendran J, Karri R *et al*. Sneak-path testing of crossbar-based nonvolatile random access memories. *IEEE Transactions on Nanotechnology*, 2013, 12(3): 413-426.

[39] Prakash A, Jana D, Samanta S *et al.* Self-compliance-improved resistive switching using Ir/TaO_x/W cross-point memory. *Nanoscale Research Letters*, 2013, 8(1): 527.

[40] Sheu S S, Chiang P C, Lin W P *et al.* A 5ns fast write multi-level non-volatile 1 K bits RRAM memory with advance write scheme. In *Proc. Symposium on VLSI Circuits*, June 2009, pp.82-83.

[41] Wong S C, Lee G Y, Ma D J. Modeling of interconnect capacitance, delay, and crosstalk in VLSI. *IEEE Transactions on Semiconductor Manufacturing*, 2000, 13(1): 108-111.

[42] Govoreanu B, Kar G, Chen Y *et al.* 10×10 nm^2 Hf/HfO_x crossbar resistive RAM with excellent performance, reliability and low-energy operation. In *Proc. International Electron Devices Meeting* (*IEDM*), December 2011, pp.31.6.1-31.6.4.

[43] Yu S, Gao B, Fang Z *et al.* A neuromorphic visual system using RRAM synaptic devices with sub-pJ energy and tolerance to variability: Experimental characterization and large-scale modeling. In *Proc. International Electron Devices Meeting* (*IEDM*), December 2012, pp.10.4.1-10.4.4.

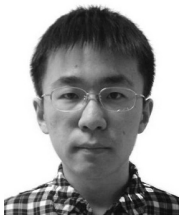
[44] Kawahara A, Kawai K, Ikeda Y *et al.* Filament scaling forming technique and level-verify-write scheme with endurance over 107 cycles in ReRAM. In *Proc. International Solid-State Circuits Conference Digest of Technical Papers* (*ISSCC*), February 2013, pp.220-221.

[45] Wang L (ed.). Support Vector Machines: Theory and Applications, Volume 177. Springer-Verlag Berlin Heidelberg, 2005.

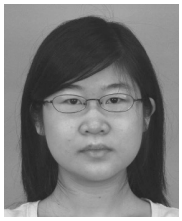
[46] Bishop C M. Pattern Recognition and Machine Learning. Springer, 2006.

Lixue Xia received his B.S. degree in electronic engineering from Tsinghua University, Beijing, in 2013. He is currently pursuing his Ph.D. degree in the Department of Electronic Engineering, Tsinghua University, Beijing. His research mainly focuses on energy efficient hardware computing system design and neuromorphic computing system based on emerging non-volatile device.

Peng Gu received his B.S. degree in electronic engineering from Tsinghua University, Beijing, in 2015. He is currently pursuing his Ph.D. degree with the SEALab, the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA, USA. His current research interests include low power system design and hardware acceleration and computing with emerging devices. He has authored and co-authored several papers in DAC, DATE and ASP-DAC.



Boxun Li received his B.S. degree in electronic engineering from Tsinghua University, Beijing, in 2013. He is currently pursuing his M.S. degree in the Department of Electronic Engineering, Tsinghua University, Beijing. His research mainly focuses on energy efficient hardware computing system design and parallel computing based on GPU.



Tianqi Tang received her B.S. degree in electronic engineering from Tsinghua University, Beijing, in 2014. She is currently pursuing her M.S. degree in the Department of Electronic Engineering, Tsinghua University, Beijing. Her research mainly focuses on on-chip neural network system and emerging non-volatile-memory technology.

Xiling Yin is currently pursuing her B.S. degree in electronic engineering from Tsinghua University, Beijing. Her research mainly involves RRAM (resistive random access memory) circuit design and test.

Wenqin Huangfu is currently pursuing his B.S. degree in electronic engineering in Tsinghua University, Beijing. His research interests include non-volatile memory, power efficient system design, hardware acceleration and computing with emerging devices.

Shimeng Yu received his B.S. degree in microelectronics from Peking University, Beijing, and his M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 2009, 2011, and 2013, respectively. He is currently an assistant professor of computer engineering and electrical engineering with Arizona State University, Tempe, AZ, USA.

Yu Cao received his Ph.D. degree in electrical engineering from the University of California, Berkeley, CA, USA, in 2002. He is currently an associate professor of electrical engineering with Arizona State University, Tempe, AZ, USA.

Yu Wang received his B.S. degree in 2002 and Ph.D. degree (with honor) in 2007 from Tsinghua University, Beijing. He is currently an associate professor with the Department of Electronic Engineering, Tsinghua University, Beijing. His research interests include parallel circuit analysis, application specific hardware computing (especially on the brain related problems), and power/reliability aware system design methodology. Dr. Wang has authored and coauthored over 130 papers in refereed journals and conferences. He is the recipient of IBM X10 Faculty Award in 2010, the Best Paper Award in ISVLSI 2012, and six Best Paper Nominations in ASPDAC/CODES/ISLPED. He serves as the associate editor for IEEE Trans. CAD, Journal of Circuits, Systems, and Computers. He is the TPC CoChair of ICFPT 2011, finance chair of ISLPED 2012∼2015, and serves as TPC member in many important conferences (DAC, FPGA, DATE, ASPDAC, ISLPED, ISQED, ICFPT, ISVLSI, etc.).

Huazhong Yang received his B.S. degree in microelectronics and his M.S. and Ph.D. degrees in electronic engineering from Tsinghua University, Beijing, in 1989, 1993, and 1998, respectively. In 1993, he joined the Department of Electronic Engineering, Tsinghua University, Beijing, where he is currently a specially appointed professor of the Cheung Kong Scholars Program. He has authored and co-authored over 300 technical papers and holds 70 granted patents. His current research interests include wireless sensor networks, data converters, parallel circuit simulation algorithms, nonvolatile processors, and energy-harvesting circuits.