Extended Abstracts of the 2017 International Conference on Solid State Devices and Materials, Sendai, 2017, pp189-190

D-3-02

An Energy Efficient and High Speed Architecture for Convolution Computing Based on Binary RRAMs

Chen Liu, Runze Han, Zheng Zhou, Peng Huang, Lifeng Liu, Xiaoyan Liu and Jinfeng Kang*

Institute of Microelectronics, Peking University, Beijing 100871, China
Phone: 0086-10- 62756745 Email: kangjf@pku.edu.cn;

Abstract

Based on RRAM arrays, a new hardware convolution computing architecture is proposed. In the architecture, the kernels are applied in form of voltages and the arrays store the images with binary RRAMs. The convolution results are parallel computed in the array. For a 28×28 image and 10 3×3 kernels, compared with the kernel storage, the architecture shows excellent performances including: 1) almost 100% accuracy within 20% LRS variation and 90% HRS variation; 2) more than 67 times speed boost; 3) 71.4% energy saving.

1.Introduction

Convolution is a fundamental operation in image processing and the convolution neural network (CNN). CNN and image processing has wide applications in computer vision area, and both include a large amount of convolution operations [1-2]. So it's essential to develop an efficient convolution computing architecture with high speed and low energy consumption. In previous works, the convolution computing system for CNN has been designed and experimentally demonstrated on the RRAM arrays [3-4]. However, the systems utilize the RRAM to store the kernels, resulting in the serial computing due to the slide of the kernels on the image, doubling the rows in the array to represent the negative kernels and introducing extra peripheral circuits for the subtraction of the currents [3]. In this work, we propose a convolution computing architecture in which the images are stored in the array. The architecture is parallel computing of images without redundant RRAMs and extra subtraction circuits. We further improve the architecture with a fixed series resistor connected with binary RRAM, reducing the impact of the device variation and the energy consumption.

2. Hardware Architecture

Unlike the kernel stored (KS) mode, in our architecture, the image is stored (IS) in the array represented by the conductance of RRAM and the voltages applied on the column represent the kernels. In this way convolution results of an image are the output currents of the rows, as shown in Fig.1. In IS mode, the results of an image are achieved in parallel, without the slide of the kernels on the image. Meanwhile, the negative kernels are represented by negative voltages on the column and the image pixels are all positive, making the negative RRAM rows and subtraction circuits needless. Because of that, though the KS mode can also speed computing up by expanding the array, the area-time product is still twice compared with the IS mode.

3.Results and Discussion

The fabrication process for the 8×16 Pt/Al_2O_3/HfO_2/TiN RRAM array is demonstrated in Fig.2(a)[5]. The typical consecutive I-V curves are shown in Fig.3. The device has multilevel resistance/conductance however the variation of the intermediate resistance is larger than the high resistance state (HRS) and the low resistance state (LRS), as shown in Fig.4. We especially focus on the variation of the device, which is the factor limiting the computing accuracy. The measured HRS/LRS resistance distribution of multiple and single RRAM is shown in Fig.5&6. The nonlinearity of the LRS/HRS is measured and fitted in Fig.7[6]. As the measuring voltage increases to 2.5V, due to the nonlinearity, the window is reduced to 18 and the induced variation can be nearly 50% for LRS and 99% for HRS, shown in Fig.8.

As for the device variation's influence on the computing accuracy, the basic variation requirement derived from single device is increasingly strict with the increasing of the number of voltage and conductance levels, as demonstrated in Fig.9. Also, considering the nonlinearity of RRAM, the amplitude range of the voltage should be as small as possible to avoid introducing extra variation. Noting that the value of an image often ranges from 0 to 256, significantly larger than the range of kernels, the IS mode is superior than the KS mode from the view of nonlinearity and variation.

As demonstrated in Fig.4&Fig.9, the variation of the intermediate resistances can hardly meet the requirement, and the resistance levels cannot cover the grayscale range. To solve the problem, we use multiple binary RRAMs to store a pixel and connect a fixed series resistor [7] with each RRAM, as shown in Fig.10. The simulated accuracy of single convolution operation with increasing LRS variation in Fig.11 indicates the improvement of the binary IS mode with series resistor. Besides reducing the relative variation, the series resistor can also reduce the power consumption on the device. Table I is an overall comparison between our developed IS mode and the KS mode in the condition of 28×28 binary image with 10 3×3 kernels. The energy consumption is reduced by 71.4% and the computing speed is raised by 67.6 times.

4.Conclusion

In this work we present a novel RRAM-based convolution computing architecture to process the image data stored in the RRAM arrays. By utilizing the binary storage and the series resistor, the accuracy is improved to almost 100%, the computing speed boost up to 67.6 times and the energy consumption is reduced by 71.4%.

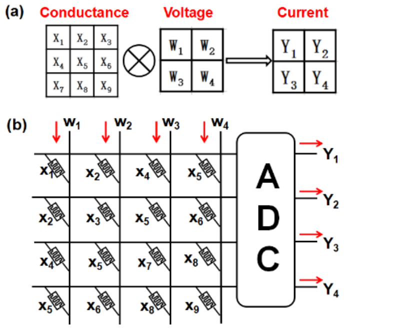

Fig.1. The architecture of the convolution computing based on multilevel RRAM. The image is stored in the RRAM array.

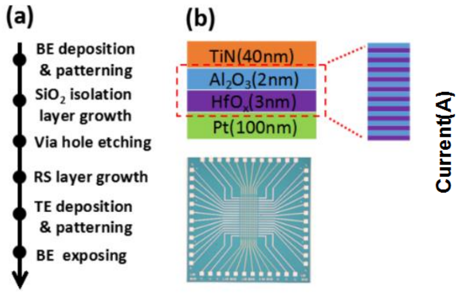

Fig.2. a) Process flow of the fabricated Al_2O_3/HfO_x crossbar RRAM array; b) Microscope micrograph of the RRAM array and the structure of RRAM cell.

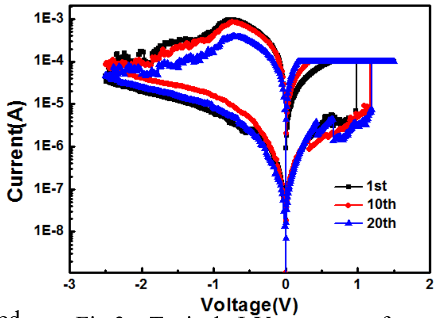

Fig.3. Typical I-V curves of a RRAM device in the array. The device shows abrupt set and gradual reset transitions.

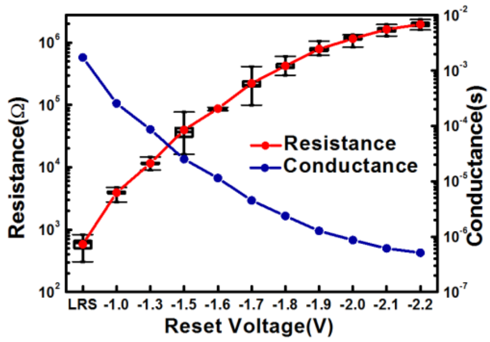

Fig.4. Measured multilevel resistance/ conductance of a device with different reset voltage in DC mode.

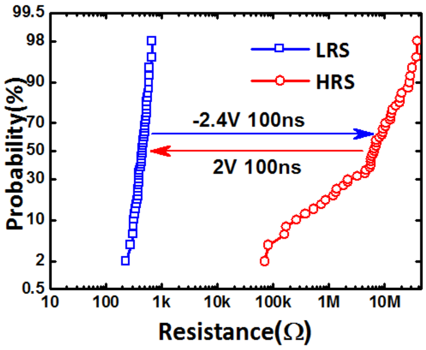

Fig.5. Measured HRS/LRS resistance distribution of 50 devices in the array. The devices show stable binary characteristic under positive/negative

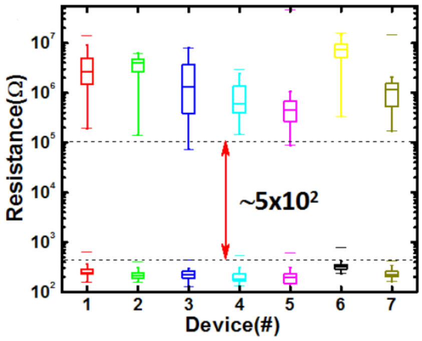

Fig.6. HRS/LRS resistance distribution of 7 typical devices. The HRS resistance variation is dramatically larger than LRS and the resistance window is over 500.

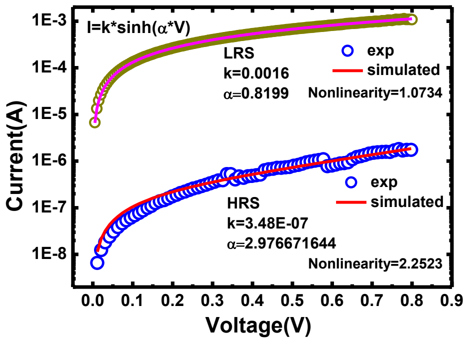

Fig.7. Experimental data and fitting of the nonlinear characteristic in the LRS and HRS of the RRAM device.

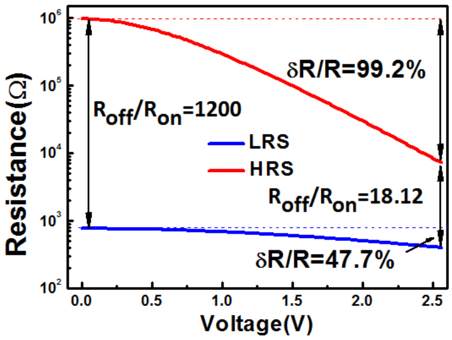

Fig.8. The simulated R-V curve based on the measured nonlinearity of the RRAM.

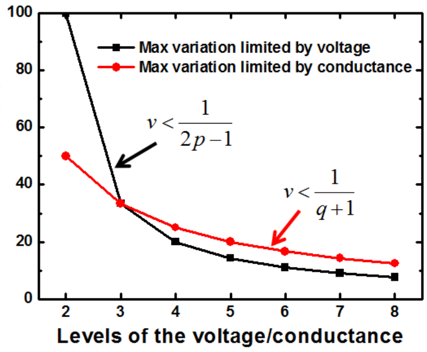

Fig.9. The max resistance variation(v) limited by the level of voltage(p) and the level of conductance(q).

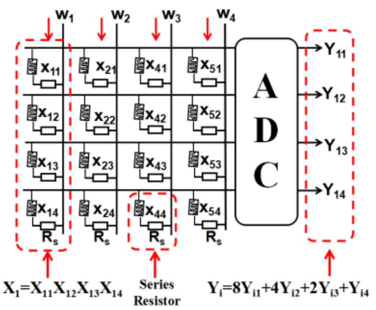

Fig.10. The improved architecture of the convolution computing based on binary RRAM. A fixed series resistor is connected with the RRAM.

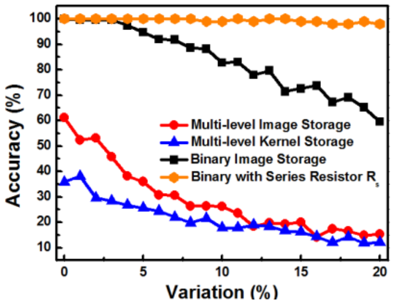

Fig.11. The simulation results of the computing accuracy with the increasing LRS variation. The HRS variation is set constantly 90%.

Table I. Comparison of the proposed architecture and the kernel storage architecture.

	RRAM Consumption(n)	Computing Pulse(n)	Energy (uJ/Image)	Redundant Device	Input Signals(n)
Kernel Storage	180	676	1.521	50%	784
Image Storage	6084	10	0.4346	0%	90

References:

[1] J.Qiu, et al, ACM Inter. Sym. on FPGA, 2016.
[2] E.Peterson, et al, Pattern Recognition, 35 (2002)
[3] L. Gao, et al, IEEE EDL 37 (2016)
[4] D.Garbin, et al, IEEE TED 62 (2015)
[5] Z.Chen, et al. Nanoscale Research Letters (2015)
[6] P. Huang, et al, IEEE TED, 60 (2013)
[7] P. Huang, et al, IEEE TED 64 (2017)