Extended Abstracts of the 2019 International Conference on Solid State Devices and Materials, Nagoya, 2019, pp231-232

Neuromorphic computing based on Highly reliable Analog ReRAM by filament control

Takumi Mikawa, Ryutaro Yasuhara, Koji Katayama, Kazuyuki Kouno, Takashi Ono, Reiji Mochida, Masayoshi Nakayama, Hitoshi Suwa, Yasushi Gohou and Toru Kakiage

Panasonic Semiconductor Solutions Co., Ltd.,
1 Kotari-yakemachi, Nagaokakyo, Kyoto 617-8520, Japan
Phone: +81-80-9940-4734 E-mail: mikawa.takumi@jp.panasonic.com

Abstract

We have developed neuromorphic computing based on Analog ReRAM, Resistive Analog Neuromorphic Device (RAND), as low power solution. We have proposed perceptron circuit which has resistive elements to store weights as analog resistance. We have demonstrated 180nm test chip and verified the concept of low power potential. We present reliability issues on analog nonvolatile memories and approach to improve it by filament control in resistive switching element.

1. Introduction

In the big data era in which data is collected from IoT devices everywhere, there are several issues, such as explosive increase of data processing on the Cloud, protection of personal information, real-time feedback between the Cloud and edge application. In order to solve these issues, it is highly expected that edge AI will make user interface more intelligent. For edge AI, it is necessary to realize overwhelmingly low power consumption, 10 mW class, or less. So as to realize super low power solution, ReRAM has a good potential as a neuromorphic device that schematically reproduces the behavior of neurons in the brain [1]. We can form the structure of a neural network by utilizing the advantages of low power consumption and high speed operation of ReRAM. Simultaneous calculation of a large amount of data achieves superior power efficiency.

2. Resistive Analog Neuromorphic Device (RAND)

Fig.1 shows RAND concept which structure is compatible with deep neural network which is composed of input layer, hidden layers, output layer and weights. RAND has analog resistive switching elements (RSEs) and diodes at the intersection of the wiring input layer and the wiring output layer. The weights are stored to RSEs whose continuous resistance value can be recorded as its reciprocal number.

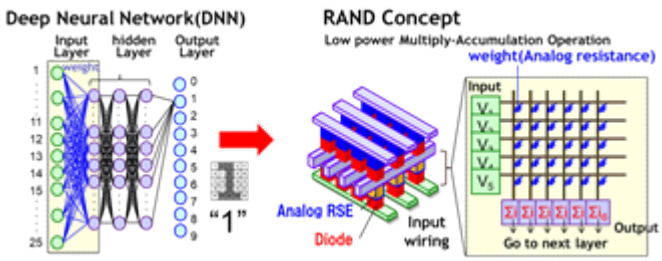
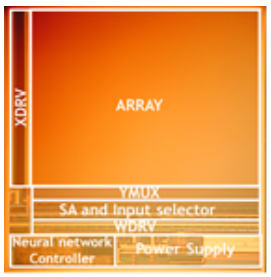


Fig. 1 RAND concept for low power solution

When a voltage is applied to the input layer as data, current flows through RSEs based on Ohm's law. This current is added along the output layer and MAC operation can be executed. By stacking layers or increasing the capacity of RSE, it is possible to calculate a large scale neural network.

Fig. 2 shows 180nm test chip micrograph, in which we performed neural network processing and performance of the fabricated RAND chip [2]. This RAND chip has only one memory layer, not multi layers, in order to verify the principle of our concept. This RAND chip has shown very low power consumption of 15.8mW on a 1024 input inference-READ, and high power efficiency of more than 20 TOPS/W as the principle



Technology	180nm
Weight Storage	ReRAM
Synapses	2M
Area(mm^2)	12.6
Synapses/mm^2	0.16M
Voltage(V)	1.8
Power(mW)	15.8
TOPS	0.33
TOPS/W	>20

Fig. 2 Micrograph of 180nm RAND chip

In RAND array, as shown in Fig.3 (a), at each perceptron, multiple inputs and weights are multiplied and accumulated. So MAC operation obtains output thorough an activation function. Proposed ReRAM perceptron circuit in Fig.3 (b) has a group of word lines (WLs), bit lines (BLs), and source lines (SLs). A cell is composed of one select transistor and one resistive switching element (1T-1R). Generally weights in the neural network have positive and negative values, so two cells connected to the same WL are used to express one weight.

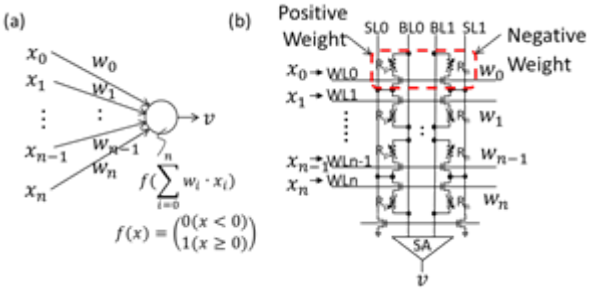


Fig. 3 (a) Schematic diagram of a perceptron and (b) Proposed ReRAM-based perceptron circuit.

3. Reliability issue ReRAM in Analog-type RAND

We have reported to increase variation in read current inversely with the magnitude of the write current, in other words, inversely with the magnitude of the read current [3]. This relationship also holds for an analog resistive switching element to store analog weight. We have evaluated the relationship between cell current distribution and writing and reading current in Fig.4. From the figures, it can be seen that the variation at writing to cell is more influential.

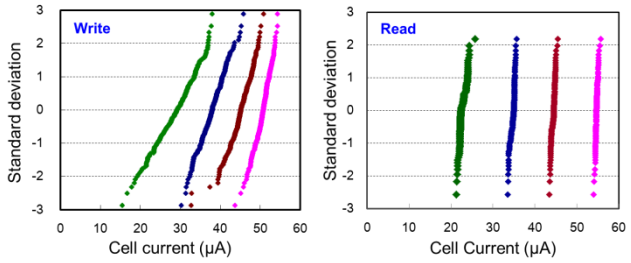


Fig. 4 Distribution of cell current at writing and reading to cell

Fig.5 shows the room temperature retention characteristics after 24 hours. Current distribution for 24 hours at room temperature retention does not include so-called retention degradation (e.g.10 years at 85C), but includes only write and read variability. Even after current control and verify operation, current fluctuation is observed, especially in the low-current region [4].

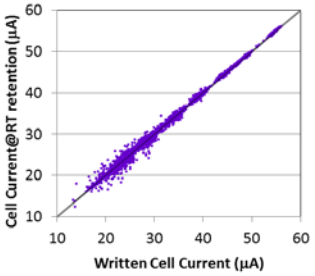


Fig. 5 Room temperature retention characteristics of RAND

These characteristics shown can be explained by conductive filament (CF) characterization method shown in Fig.6. In our model, CF is formed in the $Ta_2O_{5-\sigma}$layer and its properties can be characterized as a function of filament size S, filament length L, density of oxygen vacancies N (Vo) and residual oxygen density N (Ox). These parameters can be extracted from the distribution of the read current, shown by resistance switching under certain operating conditions.

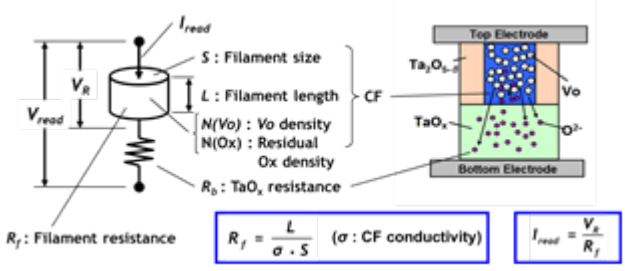


Fig. 6 Components of conductive filament and formula with schematic cross section of ReRAM cell with Ta_2O_5/TaOx structure

As we mentioned above, large variations are observed at lower cell current at writing and after room temperature retention. When cell current is around 50uA, filament has high N (Vo). That means percolation network paths in the filament for hopping conduction are dense. So a stable current flows. When cell current is around 20uA, filament has low N (Vo). Percolation network paths are sparse and variable depend on the placement of oxygen vacancies. So a current variation is larger. Same phenomena are observed in the retention degradation in the digital memory.Fig.7 shows the correlation between oxygen vacancies N (Vo) and filament size S with contour plot of data retention ratio [5]. Good retention is characterized by high N (Vo), and percolation network paths are preserved even if some of paths are cut by the diffusion of oxygen. On the other hand, the number of percolation network paths is lower in bad retention cases: the paths can be easily cut, so current decreases. These results suggest that good retention characteristics can be realized by selecting operation conditions that create filaments with higher Vo density and lower filament size. This conductive filament characterization method makes it possible to get highly reliable filament for each analog resistivity by optimized operating conditions.

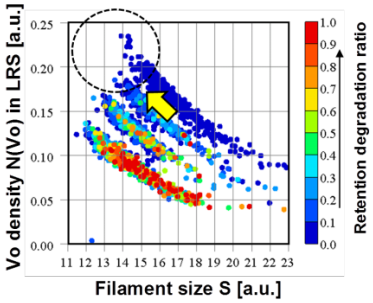


Fig. 7 Relationship between retention degradation in low resistive state and characteristics of conductive filament

4. Conclusion

We have proposed Resistive Analog Neuromorphic Device (RAND) as low power solution. RAND chip fabricated by 180nm process shows good potential for edge AI. In addition, we have mentioned the issue of reliability and have suggested solution to get highly reliable filament for each analog resistivity based on the developed reliability model.

Acknowledgements

We would like to thank Dr. H. Akinaga of National Institute of Advanced Industrial Science and Technology (AIST) and Prof. T.Asai of Hokkaido University for valuable discussions. This study "RAND" work in the paper was supported by NEDO program.

References

[1] S.R. Nandakumar, *et. al., Journal of Applied Physics, Vol.124, No.15, art. no.152135, 2018.*

[2] R. Mochida, *et.al., 2018 Symposium on VLSI Technology, Tech. Dig., pp.175–176, 2018.*

[3] S. Muraoka *et al., VLSI Tech. Dig., p. T62-T63.*

[4] R.Yasuhara, *et.al, IPRS2019 Tech. Dig.* 3C.4

[5] T. Mikawa, *et.al, IMW 2019, Tech. Dig., pp.56–59.*