A 4M Synapses integrated Analog ReRAM based 66.5 TOPS/W Neural-Network Processor with Cell Current Controlled Writing and Flexible Network Architecture

Reiji Mochida, Kazuyuki Kouno, Yuriko Hayata, Masayoshi Nakayama, Takashi Ono, Hitoshi Suwa,
Ryutaro Yasuhara, Koji Katayama, Takumi Mikawa, Yasushi Gohou
Panasonic Semiconductor Solutions Co.,Ltd., 1 Kotari-yakemachi, Nagaokakyo City, Kyoto 617-8520 Japan
E-mail:mochida.reiji@jp.panasonic.com

Abstract

This paper presents low-power neural-network (NN) processor using ReRAM to store weights as analog resistance for future AI computing. We propose ReRAM perceptron circuit for realizing large scale integration, highly accurate cell current controlled writing scheme, and flexible network architecture (FNA) in which any NNs can be configured. Fabricated 180nm test chip shows well-controlled analog cell current with linear 30μA dynamic range and 0.59μA variation of 1 sigma, results in 90.8% MNIST numerical recognition rate. Furthermore, 4M synapses integrated 40nm test chip achieves lower analog cell current and 66.5 TOPS/W power efficiency.

Introduction

The NN carries out enormous calculations of multiply accumulate (MAC) operation between weights and input data, and thus it needs high-performance hardware such as graphics processing unit (GPU), in which it consumes a great amount of power. To overcome this issue, MAC operation circuits equipped with memristor using ReRAM is proposed [1]. However, writing analog cell current accurately to ReRAM is a challenge. In addition, analog based data transfer between perceptrons is difficult, which requires lots of A/D and D/A converters. Thus executing large-scale MAC operations using analog cell current has not been reported yet. This paper presents a newly-developed low-power NN processor using ReRAM to store weights as analog cell current for realizing large scale MAC operations. We call it Resistive Analog Neuro Device (RAND) chip.

ReRAM Perceptron Circuit

The NN is made up of a combination of computation nodes called "perceptron", as shown in Fig. 1. At each perceptron, multiple inputs and weights are multiplied and accumulated and the result of MAC operation is delivered to an activation function to obtain its output. Proposed ReRAM perceptron circuit has a group of word lines (WLs), bit lines (BLs), and source lines (SLs). A memory cell (MC) is composed of one select transistor and one resistive switching element (1T-1R). Because weights in the NN have positive and negative values, two MCs connected to the same WL are used to express one weight. For example, MC connected to BL0 holds positive weight, while MC connected to BL1 holds negative weight. WLs are assigned to corresponding inputs. With these configurations, BL0 transfers results of the MAC operation of positive weights, as analog cell current, to a sense amplifier (SA), and BL1 similarly transfers that of negative weights. SA compares these current and outputs digital value, which is equivalent to a step function output. It allows MAC operation of multiple inputs in a single reading (inference-READ), thus enables MAC operation to be carried out at high speed with lower power consumption. In addition, both input and output are digitized, therefore A/D and D/A converters are unnecessary, which is suitable for large scale MAC operations.

Highly Accurate Cell Current-Controlled Writing

Fig. 2 shows highly accurate cell current-controlled writing circuit, RAND device structure, and evaluation results of analog cell current. Generally, NN operation is improved by writing analog cell current linearly with a wide dynamic range. Analog cell current of ReRAM depends on writing current. The writing current can be set in the following sequence. First, a constant current generated by current supply circuit is sent to weight control circuit, which amplifies the constant current to a desired writing current and then generates the corresponding gate clamp voltage (VCLP). Second, a write driver copies the writing current in current-mirror structure by applying VCLP, and finally it supplies writing current to MC. This analog writing circuit demonstrates the ability of writing analog cell current with linear 30uA dynamic range and 0.59μA variation of 1 sigma.

Flexible Network Architecture

Fig. 3 shows RAND architecture. This architecture can be used as both analog NN processor and digital storage memory by changing cell select control. In NN processor mode, XDRV simultaneously selects multiple WLs, and it includes a pair of latch units LAT1 and LAT2. YMUX selects BLs to be connected to positive and negative input terminals of SA. SA input selector selects BL to be connected as input line to SA.

The inference operation procedure with FNA in the manner of pipeline operation is summarized in Fig. 4. Based on NN information, NN-controller manages the cell addresses, to which weights are assigned respectively, and carries out the following sequence using a single RAND array. At the XDRV, input data A are latched by LAT1 and then copied to LAT2. Next, BLs are selected, and SA obtains data B. Thus, data B are latched by LAT1. The BL selection is changed and SA obtains data C. Data C is latched by LAT1, in addition to Data B. Finally, the data latched by LAT1 are copied to LAT2, which is followed by BL selection and SA obtains output data D, the FNA completes the NN computation. In this manner, the FNA enables a single chip to be applied to various deep NNs.

Measurement Results

Fig. 5 shows results of MNIST handwritten digit datasets recognition. NNs have input layer with 196 nodes and output layer with 10 nodes. Input data are in the form of 14×14-bit images created by compressing the MNIST datasets. The network also includes middle layers with 64 nodes 1, 2, or 3. The processing results gave a maximum accuracy of 87.3% for the NN with 3-middle layer. However, these results are affected by SA's offset and the variation of cell current. To solve this problem, we have developed a circuit for max value search of MAC operation (MSMA). With this MSMA architecture, RAND chip can add current through a fixed resistance to the positive side or to the negative side. MSMA architecture shown in Fig. 5 depicts a case of adding to the negative side. Both SA0 and SA1 output "1" in a normal inference-READ. The difference between positive current

978-1-5386-4218-4/18/$31.00 ©2018 IEEE

(Ipos) and negative current (Ineg) is larger at SA0 than at SA1, so we use the difference between each current in SA0 (50uA) and SA1 (5uA). In the case of MSMA inference-READ, adding fixed resistance current (Imsma) to negative side eliminates minute difference between Ipos and Ineg. Applying MSMA inference-READ has achieved 90.8% accuracy.

Fig. 6 shows 180nm test chip micrograph, in which we performed NN processing, and 40nm test chip micrograph. Fig. 7 shows results of writing analog cell current demonstrated by both test chips. Since it is possible to scale the filament with fine process technology [5], the cell current of 40nm is lower than that of 180nm, which leads to better power efficiency.

Conclusion

We proposed low-power and high-accuracy NN processor using ReRAM. Table I compares this paper's work with various technologies. RAND chip fabricated by 180nm process consumes power of 15.8mW on a 1024 input inference-READ, achieving power efficiency of 20.7 TOPS/W. In addition, 40nm ReRAM reduces power consumption during an inference-READ to 9.9mW, thus achieving power efficiency of 66.5 TOPS/W.

Acknowledgements

This work is supported by NEDO program. The authors also thank Professor T.Asai of Hokkaido University for valuable discussions.

References

[1] M. Prezioso, et al., Nature, vol. 521, no. 14441, pp.61-64, 2015.
[2] D. Miyashita, et al., ASSCC, pp.25-28, 2016.
[3] B. Moons, et al., ISSCC, pp.246-247, 2017.
[4] K. Ando, et al., VLSI Circuit, pp.24-25, 2017.
[5] Y. Hayakawa, et al., VLSI Technology, pp.14-15, 2015.

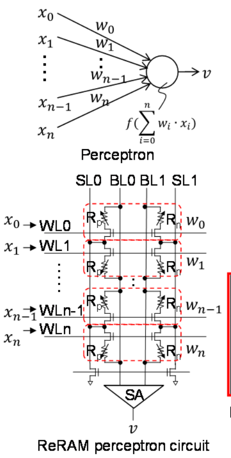


Fig 1. Proposed ReRAM perceptron circuit

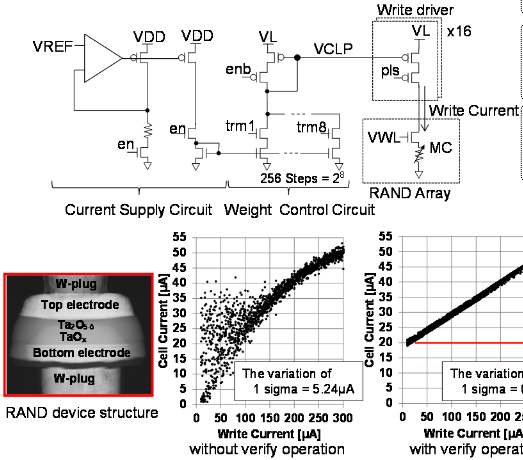


Fig 2. Analog writing circuit and results of analog cell current

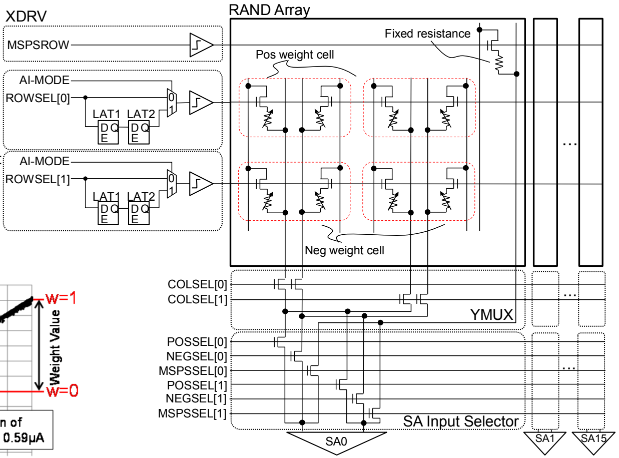
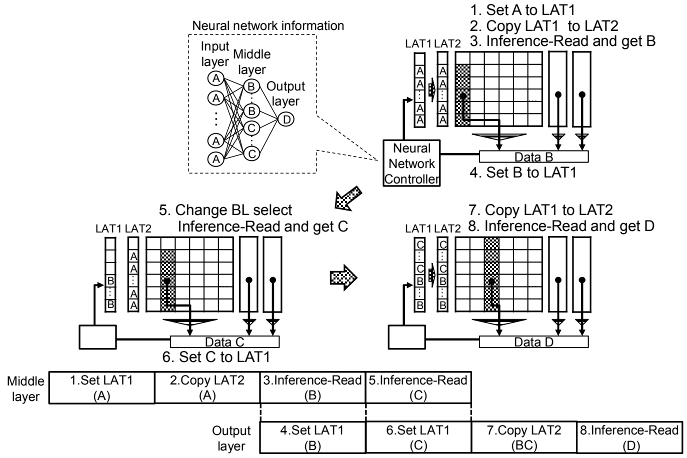


Fig 3. RAND architecture.
It can be used as both analog NN processor and digital storage memory



Middle layer	1.Set LAT1 (A)	2.Copy LAT2 (A)	3.Inference-Read (B)	5.Inference-Read (C)		
Output layer			4.Set LAT1 (B)	6.Set LAT1 (C)	7.Copy LAT2 (BC)	8.Inference-Read (D)

Fig 4. Inference operation procedure with FNA

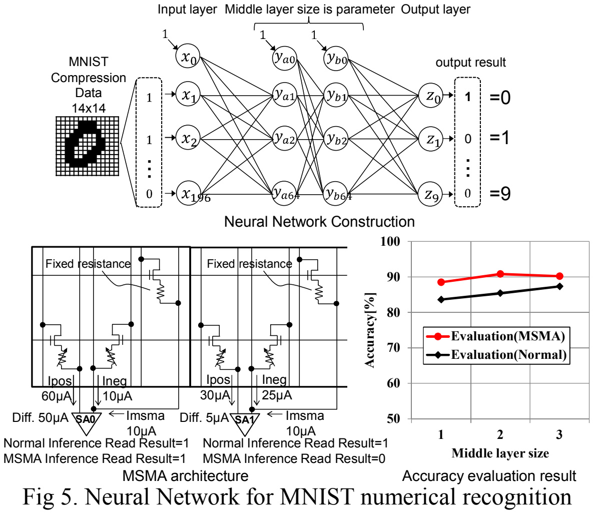


Fig 5. Neural Network for MNIST numerical recognition

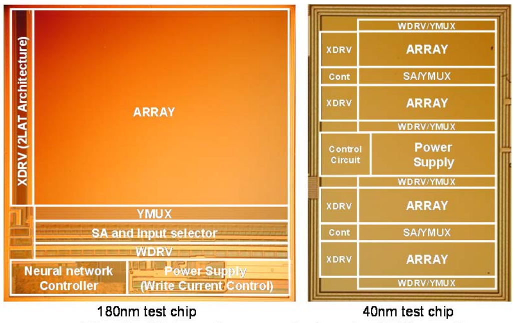


Fig 6. Chip micrograph (scale different)

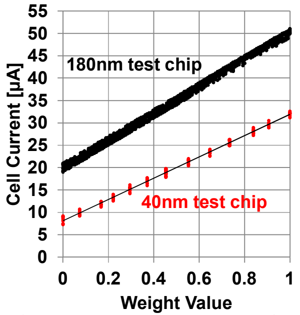


Fig 7.Comparison of analog cell current between 180nm/40nm

TABLE I. Comparison Table

	ASSCC 2016 [2]	ISSCC 2017 [3]	VLSI 2017 [4]	This Work	
Technology	65nm	28nm	65nm	180nm	40nm
Weight Storage	SRAM	SRAM	SRAM	ReRAM	ReRAM
Synapses	32K	128K	0.8M	2M	4M
Area(mm^2)	3.61	1.87	3.9	12.6	2.71
Synapses/mm^2	0.01M	0.07M	0.21M	0.16M	1.48M
Voltage(V)	N/A	0.65-1.1	0.55-1.0	1.8	1.1
Power(mW)	N/A	7.6	50-600	15.8	9.9
TOPS	N/A	0.076	1.38	0.33	0.66
TOPS/W	48.2	10	6.0	20.7	66.5