Extended Abstracts of the 2020 International Conference on Solid State Devices and Materials, VIRTUAL conference, 2020, pp103-104

Computation-In-Memory (CIM) with Logic-Process-Compatible Embedded Non-Volatile Memory for Neural Network

Hsueh-Wei Chen, Ying-Je Chen, Woan-Yun Hsiao, Wei-Ren Chen, Hung-Yi Liao, Wein-Town Sun and Ching-Yuan Lin

Business Group II, eMemory Technology Inc.,
8F, No.5, Tai-Yuan 1st St., Jhubei City, Hsinchu County, 302082, Taiwan

Abstract

This paper introduces a cost-effective and highly accurate in-memory computing macro using logic process compatible non-volatile memory (NVM). With high-precision current control in each individual cell, the memory macro is suitable for multi-level storage of synaptic weights in AI application. A proof-of-concept 3Mb NVM array with multi-level state is implemented using 55-nm CMOS technology. From MNIST benchmark, it shows 94.87% classification fidelity for MLP and 97.69% classification fidelity for CNN. The neural network achieves the power efficiency of 15 TOPS/W.

1. Introduction

Non-volatile floating gate memory devices, because of the tunable programming states, are very attractive for analog computing applications [1, 2]. In this work, we demonstrate in-memory computing macro to serve as the multi-level weight storage based on the proprietary embedded non-volatile memory [3], as shown in Fig.1. Each level of weights could be precisely controlled by a delicate program-verify (PV) sequence. The proposed macro could compute multiply-accumulate (MAC) in MLP at low power and also support mainstream CNNs with high accuracy. The results demonstrate a stable and robust neuromorphic network performance, which can perform high-fidelity classification of images of the standard MNIST benchmark dataset.

2. Array Characterization and Experimental Approach

The in-memory computing macro can map to convolutional and fully-connected layers of CNNs and Multi-Layer Perceptron (MLPs), which consists of 3Mb NeoMTP memory array, a row decoder, wordline (WL) driver and highly precise current-controlled programming circuit. In order to achieve weights with high accuracy, the cell current is written and verified by program-verify (PV) algorithm (Fig. 2). For writing mode, all selected cells will go through PV flow from Lv1 to Lv16. In each writing level, the PV flow includes 8 preset conditions with a total of 200 pulse shots to cover die-to-die or wafer-to-wafer variation. When the cells reach the writing level or the target level, the inhibiting conditions will be applied and wait for the rest cells to complete PV. The cell current as weights provides 16 individual current state (Fig. 3a) with standard deviation less than 0.005uA (Fig. 3b) in the range of 0.2uA to 3.2uA. Each individual current shows highly precise current-controlled capability (Fig. 3c). Memory cells as weights of all individual current also show good retention property at 125°C (Fig. 4a) for more than 168 hours and 250°C (Fig. 4b) for more than 24 hours.

The in-memory computing characteristics mentioned above are then introduced to MLP with three fully-connected layers and CNN with 2 convolution layers and 2 fully-connected layers for accuracy and performance evaluation based on MNIST benchmark dataset. Furthermore, the impact of non-ideal effects of memory cell on recognition accuracy can be evaluated by error models.

3. Results and Discussions

Figure 5 shows all exported cell current value of MLP and CNN for MNIST test. Each individual level shows well-controlled analog cell current and the maximum of 0.005uA variation of 1 sigma that passes the criteria of 0.05uA. It results in high recognition accuracy. Fig. 6 shows the accuracy measured by our macro achieving 94.87% for MLP and 97.69% for CNN, which is close to the software (floating point) 95.19% for MLP and 99.02% for CNN. Even after the accelerated retention test at 125°C and 250°C, the accuracy remains almost the same as that before retention test as shown in Fig. 6. Considering memory cell behavior variation in actual process manufacturing, Figure 7 reveals there is no significant degradation with the cell fluctuation of +/-3 sigma and 5% failure bit rate in the whole memory array. Figure 8 shows the sum of cell current (weights) for each neuron of each layer in one test image of MLP and CNN. Lower summation current is due to lower cell current for each individual current level as weights. In the characterization, proposed in-memory computing macro achieves a peak energy efficiency of 15 TOPS/W in recognition mode, consisting of the power consumption of major circuits including WL driver, sensing amplifier, and activation circuit. The brief summary of the macro performance is shown in Table I.

4. Conclusion

We propose a low-cost and high-accuracy multi-level in-memory computing macro with a 3Mb embedded non-volatile memory cell array. It is successfully demonstrated in a small neural network classifier on MNIST benchmark dataset and achieves a peak power efficiency of 15 TOPS/W. The macro obtains good recognition accuracy even under the non-ideal cases that include data retention, cell fluctuation and failure bits.

References

[1] C. Mead, Analog VLSI and Neural Systems, Addison-Wesley, 1989.

[2] CG. Indiveri et al., “Neuromorphic silicon neuron circuits”, Frontiers in Neuroscience, vol. 5, art. 73, 2011.

[3] C. C. H. Hsu, Y. T. Lin, E. C. S. Yang, R. S. J. Shen, “Logic Nonvolatile Memory,” World Scientific, ISBN: 978-981-4460-90-3, May2014, pp. 17-46.

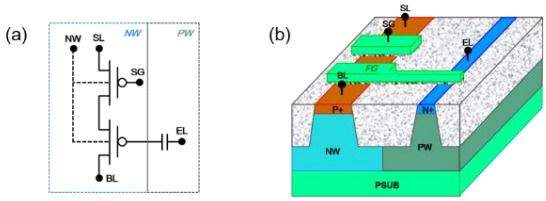
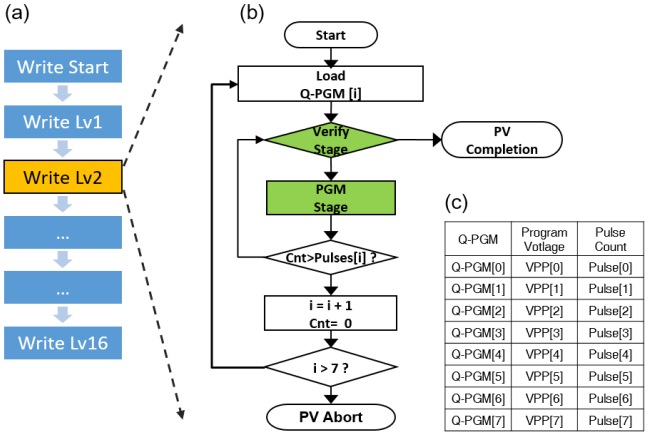


Fig. 1 (a) Schematic and (b) Layout of proprietary memory cell.



(c)

Q-PGM	Program Votlage	Pulse Count
Q-PGM[0]	VPP[0]	Pulse[0]
Q-PGM[1]	VPP[1]	Pulse[1]
Q-PGM[2]	VPP[2]	Pulse[2]
Q-PGM[3]	VPP[3]	Pulse[3]
Q-PGM[4]	VPP[4]	Pulse[4]
Q-PGM[5]	VPP[5]	Pulse[5]
Q-PGM[6]	VPP[6]	Pulse[6]
Q-PGM[7]	VPP[7]	Pulse[7]

Fig.2 Proposed PV algorithm (a) all selected cells are processed by PV flow from Lv1 to Lv16. (b) The detailed PV flow within single writing level flow (c) Example of PV bias conditions including program voltage and pulse count.

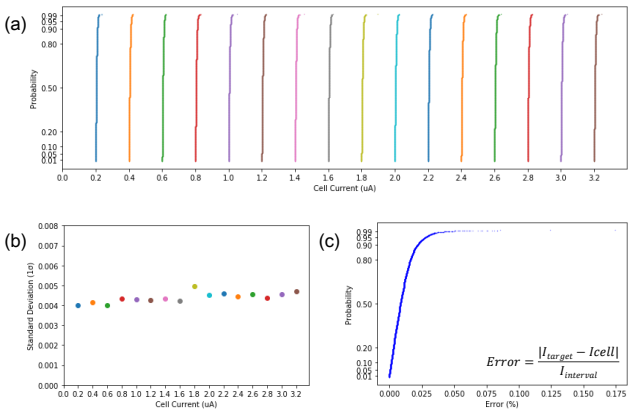


Fig. 3 Current distribution of multi-level nonvolatile memory cell. Individual (a) cell current and (b) standard deviation. (c)The error rate of whole current state.

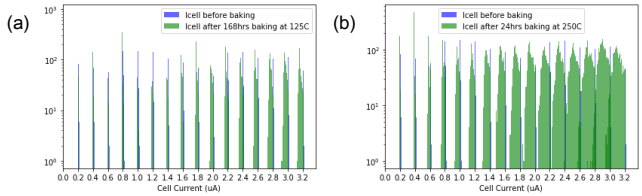


Fig. 4 Retention characteristics at (a) 125°C and (b) 250°C.

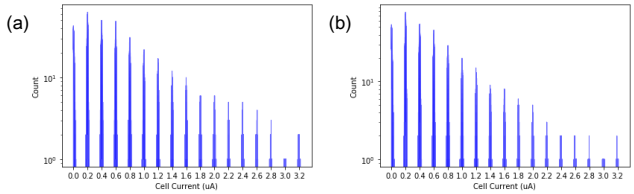


Fig. 5 Weight export statistics. Histogram shows the exported cell current values (weights) of (a) whole MLP layers and (b) whole CNN layers in this work.

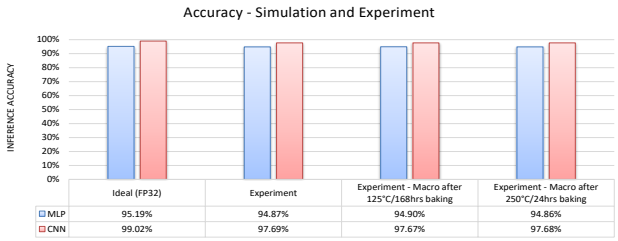


Fig. 6 Simulation and actual results of inference accuracy.

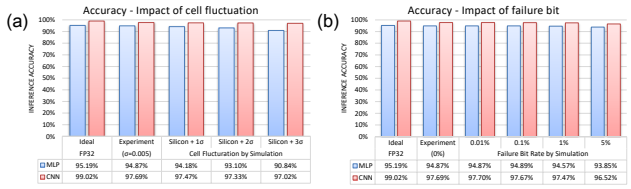


Fig. 7 Experimental results and error model by simulation. Impact of (a) cell fluctuation and (b) failure bit on inference accuracy.

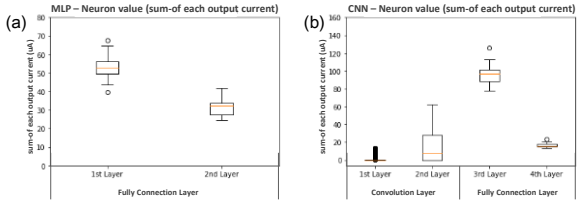


Fig. 8 The measured value of each neuron (sum of cell current (weights) for 1 MNIST images by each layer of MLP and CNN.

Technology			55nm CMOS process
Weight Storage			NeoMTP
Synapses			3 M
Cell Bit-precision			4 bits
Supply Voltage			1.2V / 2.4V
Power			670 mW
Peak Perf. [TOPS]			10.04
TOPS/W			15.0
Inference accuracy on MNIST benchmark (MLP/CNN)	Ideal (FP32)		95.19% / 99.02%
	Experiment		94.87% / 97.69%
	Experiment	Macro after 125°C/168hrs baking	94.90% / 97.67%
		Macro after 250°C/24hrs baking	94.86% / 97.68%
	Error Model	Cell fluctuation with 4 sigma	90.84% / 97.02%
		5% failure bits rate of whole array	93.85% / 96.52%

Table I Summary of macro characteristics