Extended Abstracts of the 2019 International Conference on Solid State Devices and Materials, Nagoya, 2019, pp679-680

Weight Update / Inhibition Method for NOR Flash Based Polysilicon Synapse Array

Myung-Hyun Baek[1], Taejin Jang[1], Sungmin Hwang[1], Suhyeon Kim[1], and Byung-Gook Park[1]

[1] Seoul National University
1 Gwanak-ro, Gwanak-gu
Seoul 08826, Korea
Phone: +82-2-880-7282 E-mail: bgpark@snu.ac.kr

Abstract

A novel program inhibit scheme was demonstrated at NOR flash based asymmetric dual gate synapse array. Conventional NOR flash cannot inhibit unselected bit-line when PGM operation has done by FN tunneling method. However, by introducing additional gate, we proposed a weight update method through FN tunneling. The additional gate can boost the body potential of the inhibit cell to suppress FN tunneling even if a high PGM bias is applied. With this method, delicate weight control of the synapse array for off-line learning can be achieved.

1. Introduction

Recently, the accomplishment of artificial neural network (ANN), which is represented by deep-learning algorithms, has achieved brilliant results in various fields including pattern recognition, language detection, and automobile technology. However, software-based neural networks suffer from severe power consumption problem when calculate vector-matrix multiplication using conventional GPUs. To solve these issue, brain-inspired hardware neuromorphic systems have been widely studied. Memristor devices are one of the most popular candidate for synapse due to its non-volatile characteristics and high scalability, while also problematic in terms of device variation and reliability [1].

In this work, flash memory based 4-terminal synaptic transistor was investigated. Flash memory is CMOS-compatible technology, and has excellent reliability. Polysilicon was adopted as a body material because of simple fabrication process and low current drivability compared to single-crystalline silicon. Since the number of synapses in ANN is more than 100 times that of neurons, the current of each synapse device must be small in order to reduce the power consumption of the entire system [2].

Considering synapse array design, the NAND flash structure is not suitable for weighted sum operation because cells are connected in series to identical string. In contrast, NOR can automatically calculate current summation of independent cells. In this reason, our synapse array has similar structure to the NOR flash, where synaptic weight is updated by Fowler-Nordheim (FN) tunneling mechanism. Conventional NOR flash utilizes channel hot electron (CHE) for program (PGM) mechanism, which requires respectable current level. However, since our synapse array employs polysilicon as a body material for low current, FN tunneling was adopted for weight update instead of CHE. Basically, conventional NOR flash cannot use FN tunneling at PGM operation due to structural limitations. In this paper, we proposed novel weight update method utilizing FN tunneling by asymmetric double gate synapse array structure.

2. Device Structure and Synapse Array Design

Device structure and fabrication process

Fig. 1 shows a unit cell TEM image of fabricated dual gate polysilicon synapse array. The SiO_2 / Si_3N_4 / SiO_2 (ONO) layers were deposited above the bottom gate to store synaptic weight information. The opposite top gate exists for PGM inhibit with FN tunneling in our NOR-based synapse array. The body thickness was formed about 20 nm to allow the top gate to effectively control the channel potential during selective PGM operation.

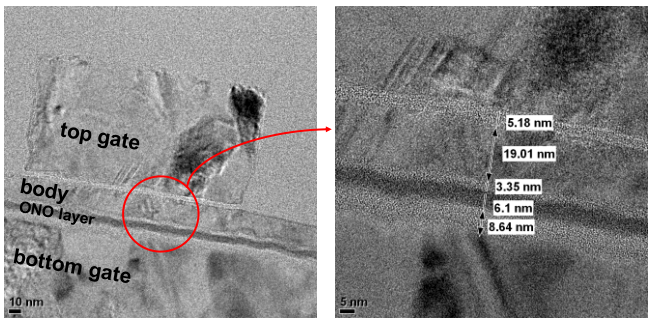

Fig. 1 TEM image of the fabricated dual gate synaptic transistor unit cell.

The brief fabrication flow is as follows. Firstly, 400-nm-thick buried oxide was grown by wet oxidation process and patterned for bottom gate. In-situ doped polysilicon CVD was followed by CMP process to shape the bottom gate. Then ONO layers and polysilicon body were sequentially deposited. After active region patterning, top gate oxide and also top gate polysilicon were formed. Finally, metallization and BEOL process was conducted.

Synapse array design and PGM scheme

Fig. 2 indicates the proposed synapse array architecture. For convenience, only 2×2 array was represented. When inference mode, synapse array is connected to the pre-/post-neurons. While in case of weight update mode, a high voltage is required to generate the FN tunneling. So the connections to the neurons are cut off and an additional peripheral circuits are connected to the synapse array.

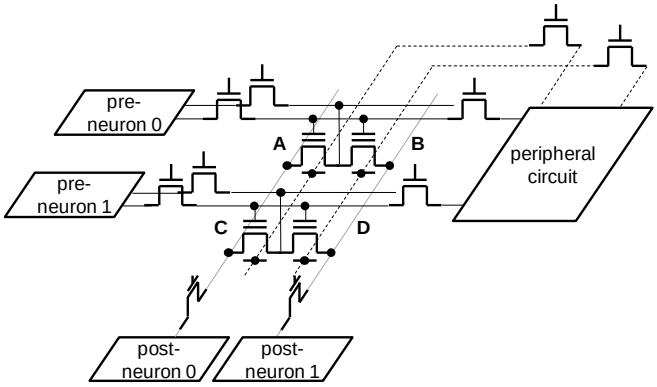

Fig. 2 Schematic diagram of the proposed synapse array and operating circuits.

Considering synapse A and B, because the two cells receive signals from the identical pre-neuron, the bottom gate and drain of both cells are tied respectively. Therefore, if a high PGM voltage was applied to the bottom gate line of a synapse A for weight update, the equivalent bias condition was produced at synapse B. In case of conventional NOR flash, PGM voltage applied to the word-line is low enough not to generate FN tunneling, and the selected cell is distinguished by the bit-line voltage. On the other hand, our synapse array utilizes FN tunneling mechanism, in which case tunneling occurs in both cells only by the gate voltage regardless of the bit-line voltage. To solve this dilemma, additional gate was built for PGM inhibit of unselected synapses. Top gates are tied (cells A and C) perpendicular to the bottom gate line. Table I shows bias schemes of proposed synapse array. Synapse A is a target cell and the others should be inhibited. In weight update mode, all source and drain lines are floating.

Table I Bias Condition for Weight Update Mode

Synapse	Bottom gate line	Top gate line
A	12 V	0 V
B	12 V	6 V
C	0 V	0 V
D	0 V	6 V

3. Measurement Results and Discussions

According to the Table I, each cells were programmed or inhibited as indicated at fig. 3. The gate length of the measured devices is 400 nm. 12 V was applied to the selected bottom gate line, while selected top gate line to which the target cell A is connected was grounded. Instead, the other unselected bottom gate lines were grounded when 6 V was applied to the other top gate lines to boost body potential. Compared to the NAND flash array, cell B has similar condition to the program disturb of which the body potential is boosted by bit-line voltage. Fig. 3 clearly documents that only the cell A was programmed, while synaptic weights of the other cells remained unchanged. Meanwhile, erase operation had also been demonstrated in fig. 4. Due to the proper scheme, only the target cell is erased and the other cells are inhibited.

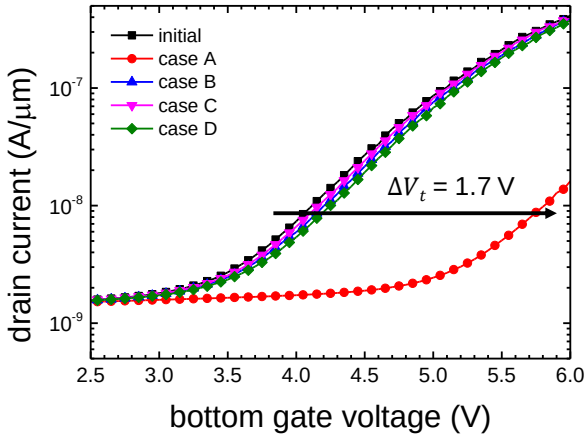

Fig. 3 Transfer characteristics of four independent cells after weight update operation. Applied pulse width was 50 μs.

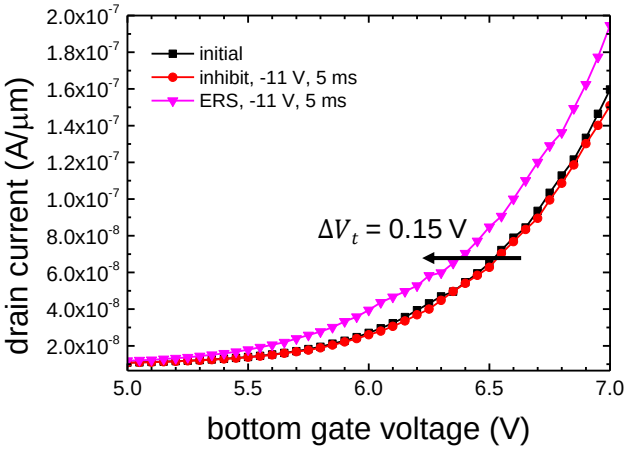

Fig. 4 Transfer characteristics of ERS / ERS inhibit cells.

4. Conclusions

We have demonstrated a novel synapse array structure and the corresponding weight update mechanism. Although proposed synapse array is similar to the NOR flash, independent weight update for each cell via FN tunneling was achieved. It is expected that low-power neuromorphic system can be implemented by applying this methodology.

Acknowledgements

This work was supported in part by the Brain Korea 21 Plus Project in 2019 and in part by Nano·Material Technology Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (2016M3A7B4910348).

References

[1] C. Sung *et al.*, J. Appl. Phys. **124** (2018) 151903.
[2] S. Hwang *et al.*, Electron Device Lett. **39** (2018) 1441.