Received 28 September 2023; revised 21 November 2023; accepted 26 November 2023. Date of publication 29 November 2023; date of current version 29 January 2024. The review of this article was arranged by Editor S. Sadana

Digital Object Identifier 10.1109/JEDS.2023.3337399

Self-Organizing Mapping Neural Network Implementation Based on 3-D NAND Flash for Competitive Learning

ANYI ZHU[1,2], LEI JIN[1,2], WEN ZHOU[1,2], TIANCHUN YE[1,2] (Senior Member, IEEE), AND ZONGLIANG HUO[1,2]

1 Institute of Microelectronics of the Chinese Academy of Sciences, Beijing 100029, China
2 University of Chinese Academy of Sciences, Beijing 100049, China

CORRESPONDING AUTHORS: Z. HUO AND L. JIN (e-mail: huozongliang@ime.ac.cn; jinlei@ime.ac.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1107700.

ABSTRACT Self-organizing Map (SOM) neural network is a prominent algorithm in unsupervised machine learning, which is widely used for data clustering, high-dimensional visualization, and feature extraction. However, the hardware implementation of SOM is limited by the von Neumann bottleneck. Herein, a SOM neural network is implemented by the combination of 3D NAND flash memory arrays and in-memory Euclidean distance (ED) calculation. The weights in the SOM network are mapped to the conductance of the 3D NAND differential pair. It is experimentally demonstrated that the differential pair in 3D NAND flash array possesses superior characteristics for neuromorphic computing during increasing and decreasing synaptic weight. Using the 3D NAND-based SOM, a competitive learning neural network is established and used for the unsupervised classification of a set of Gaussian distribution data points. The experimental results illustrate the excellent performance and efficiency of the proposed architecture, highlighting the potential of 3D NAND-based in-memory computing for artificial intelligence applications.

INDEX TERMS Self-organizing map (SOM), 3D NAND Flash, competitive learning, in-memory computing.

I. INTRODUCTION

As the demand for state-of-the-art computational tasks increases, traditional computing architectures based on the von Neumann architecture face limitations in terms of memory bandwidth and energy efficiency. To tackle this problem through, it has led to the exploration of alternative computing paradigms such as in-memory computing, where computations are performed within the memory itself, thus eliminating the need for data transfer between the processor and memory [1].

In-memory computing has been widely applied to neural networks, where the data-intensive computations are typically performed by the memory itself. Self-organizing mapping (SOM) neural network, based on the Euclidean distance (ED) calculation, is a type of unsupervised learning that consists of a set of interconnected neurons arranged in a two-dimensional grid [2]. Each neuron in the network represents a weight vector in a high-dimensional input space.

Implementing SOM with conventional CMOS-based hardware is restricted by the complexity in calculating the ED and winner determination, which imposes an enormous increase in time cost and energy consumption when the number of neurons increases. Until now, several emerging nonvolatile memories, containing resistive random access memory (RRAM) and ferroelectric field-effect transistor (FeFET), were proposed as synaptic devices to implement SOM neuromorphic systems [3], [4], [5], [6], [7]. However, compared with these devices, 3D NAND Flash has obvious advantages in terms of cost, retention, density, and technology maturity [8], [9], [10]. And compared to RRAM, 3D NAND also has lower write energy per bit. Thus, using 3D NAND as weight mapping hardware for the SOM can be

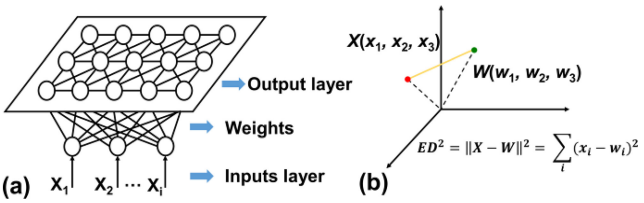

FIGURE 1. **(a) Basic Structure of SOM Networks. (b) Illustration of ED calculation in Euclidean space.**

chosen as a promising candidate, and as of now there is a lack of related research about this topic.

In this paper, a SOM neural network implementation based on ED calculation in 3D NAND Flash array is presented for competitive learning. The feasibility and efficiency of our approach is demonstrated through experimental results of two-dimensional Gaussian distribution data point clustering.

II. DEVICE AND EXPERIMENTS

The SOM neural network is experimentally demonstrated on charge-trapping 3D NAND flash mini-array test structure. More device details are described in our previous works [9], [12], [13]. Customized waveform and algorithms for SOM are developed on Keysight B1500A semiconductor device analyzer.

III. RESULTS AND DISCUSSION

A. SELF-ORGANIZING MAPPING NEURAL NETWORK BASED ON EUCLIDEAN DISTANCE CALCULATION

The SOM neural network consists of one input layer, one output layer, and the weights connections as depicted in Fig. 1(a). The neurons of the input layer are arranged in a single layer with the same number as the sample dimension. The output layer is also a competitive layer, consisting of a low-dimensional (usually one or two) grid of neurons. The SOM network works by iteratively adjusting the weight vectors to map the input data onto the grid. This process involves a competition between neurons to determine which neuron best represents the input data, based on the Euclidean distance (ED) between the input vector and the neuron's weight vector.

As exemplified in Fig. 1(b), the equation for calculating ED between an input vector X and a weight vector W of a neuron can be expressed as:

$$\mathrm{ED}^2 = \|\mathrm{X} - \mathrm{W}\|^2 = \sum_i \mathrm{x_i}^2 - 2\sum_i \mathrm{x_i w_i} + \sum_i w_i^2 \quad (1)$$

where ED is the Euclidean distance between X and W, which are the n-dimensional vectors [14]. After calculating the ED values, the neuron with the smallest ED is the winner, or the Best Matching Unit (BMU). Additionally, the weight vectors of the winner and its neighboring neurons also updated to form a smooth transition from the BMU to the surrounding neurons. The equation of updating the activated neuron connection weights can be expressed as follows:

$$\mathrm{w_i}(t+1) = \mathrm{w_i}(t) + \mathrm{h_i} \cdot (\mathrm{x}(t) - \mathrm{w_i}(t)) \quad (2)$$

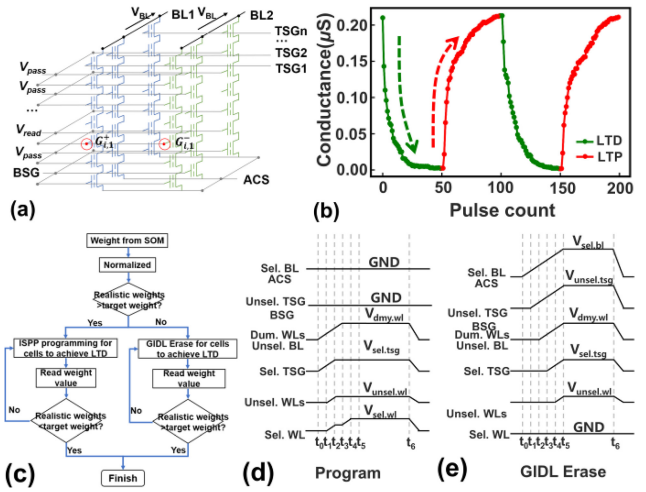

FIGURE 2. **(a) The differential pair in 3D NAND array to achieve different positive and negative conductivity. (b) LTD and LTP properties through cell erase and program operation as a function of pulse number. (c) Flow chart of the LTD and LTP processes for cell weight modulation. Corresponding schematic waveforms is shown for (d) the programming operation and (e) the soft Erase operation.**

where t is the training epochs, $\mathrm{h_i}$ is the neighborhood function, which defines the neighborhood update mechanism for preserving topological relationships in SOM, typically Choosing the Gaussian function:

$$\mathrm{h_i} = \eta(t) \cdot \exp\left(\frac{-\mathrm{ED}^2(c, i)}{2\sigma(t)^2}\right) \quad (3)$$

where $\eta(t)$ is the learning rate, ED(c, i) represents the Euclidean distance between neuron c and neuron i, and $\sigma(t)$ is the kernel radius, which decreases with training epochs.

B. IMPLEMENTING ED CALCULATION WITH 3D NAND DEVICE ARRAY

To experimentally verify the feasibility of the SOM network, we first validated the ED calculation based on 3D NAND array. In our proposed approach, the calculation of Euclidean distance in SOM network is implemented by using the differential conductance of 3D NAND Flash memory cells. Specifically, the conductance difference between the two corresponding cells with adjacent bit lines (BL) represents the weight vector of a neuron as shown in Fig. 2(a). The corresponding weight value is obtained by calculating the differential conductance of cells between two neighboring BLs.

In order to implement ED calculation with 3D NAND array, we need to normalize both the input vector and weight vector. Since the input terms x^2 of (1) are the same for all weights and do not affect the comparison when determining the winner index, we can ignore the first term and obtain the BMU with the smallest ED by comparing the last two terms:

$$\mathrm{ED}' = -2\sum_i \mathrm{x_i w_i} + \sum_i \mathrm{w_i^2} \quad (4)$$

where ED′ is relative ED value that ignores the x^2 term. The multiplication accumulation calculation (MAC) operations of the $\sum_i x_i w_i$ term in (4) can be directly accelerated by Ohm's Law and Kirchhoff's Current Law with 3D NAND array that using normalized x_i as the input voltage and w_i as differential pairs conductance. However, we cannot obtain the $\sum_i w_i^2$ term concurrently because the weight term is different for each mapping grid node, so w_i cannot be normalized together with x_i as the input voltage. To address this problem, we choose a word line (WL) row as the square row to mapping the normalized $-\sum_i w_i^2$ term with differential pairs and fix the input of WL in this row to 0.5 as shown in Fig. 4(a). As a result, the output at the n^{th} column can be expressed as:

$$Y_n = \sum_i x_i w_i - \frac{1}{2}\sum_i w_i^2 = -\frac{1}{2}ED'. \tag{5}$$

Thus, the column with the largest output has the smallest ED value with the same input, and the corresponding neuron will be the BMU. Fig. 2(b) shows that one single 3D NAND cell exhibits prominent gradual conductance responses over a few hundred cycles in both of the long term depression (LTD) regime and the long term potentiation (LTP) regimes through program and 1-bit erase operation. The modulated line of cell can prove that differential pairs possess continuously tunable conductance meanwhile, which demonstrates the in-memory computing capability. The flow chart of the synaptic weight reduction and increase processes resemble the long term depression (LTD) regime and the long term potentiation (LTP) regime achieved by WL programming and soft Erase operation respectively is illustrated in Fig. 2(c). Fig. 2(d) and (e) shows the detailed corresponding schematic waveforms concurrently. Previous research has demonstrated that using soft erase operations with gradient bias to the dummy WL can inhibit unintended erase and reduce disturbance [13]. And the entire read operation of WL in the same row is performed within a single time step, so the latency of SOM is the read latency of 3D NAND and does not scale with the size of the array.

In order to verify the accuracy of ED calculation with 3D NAND array, we have set a target ED value and modulated the conductance and input voltage based on the numerical calculation to bring the result closer to that value. The experimental results with our mapping method for 100 temporal cycles is indicated in Fig. 3(a). The test value maintains a stable variation in the vicinity of the set target value, and the average value is very close to the target value. After analyzing the error between the original data and the target value, we found that the error statistics show a Gaussian distribution with the average value $\mu = 0.0269$ and the variance $\sigma = 0.00728$. The results suggest that the accuracy of ED calculation with 3D NAND array can satisfy the requirements of SOM neural network.

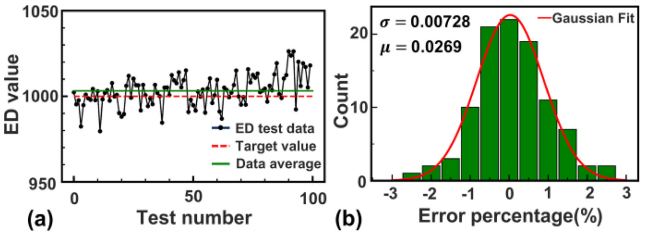

FIGURE 3. (a) The experimental results of the ED calculation with the 3D NAND array. (b) Error distribution between experimental test results and target values for 100 time ED calculations, which can be fitted by a Gaussian distribution curve with $\mu = 0.0269$, $\sigma = 0.00728$.

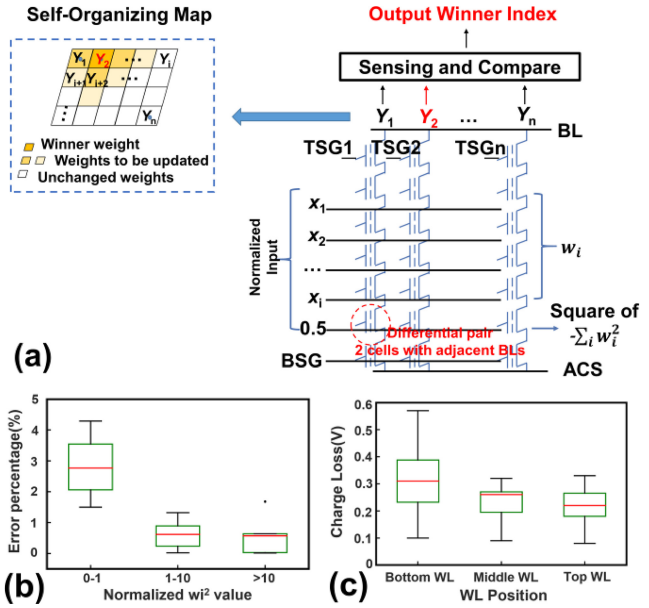

FIGURE 4. (a) The implementation of the 2D SOM with the 3D NAND array. The inputs in our SOM network are sent by layer to get the output. (b) Error percentage of w_i^2 with 3D NAND differential pair. (c) The charge loss of 3D NAND device after 3000 conductance modulations.

C. IMPLEMENTING SOM WITH 3D NAND DEVICE ARRAY FOR COMPETITIVE LEARNING

Based on the $2 \times 3 \times 15$ 3D NAND mini-array, we have constructed a 2D-SOM neural network for competitive learning as depicted in Fig. 4(a), consisting of a one-dimensional input layer and a two-dimensional output layer. Each node in the input layer is connected to all nodes in the output layer by synapses and forms a 2D planar topological map. As shown in Fig. 4(b), the error percentage of small w_i^2 term mapping with 3D NAND differential pair conductance is around 2.7%, which is within an acceptable range and will not affect the accuracy of the SOM network. In addition, we conducted 3000 conductance modulations on the 3D NAND synaptic device and measured the charge loss. The experimental results in Fig. 4(c) demonstrate the robustness and reliability of the device.

To implement SOM, first of all, the threshold voltages of all synaptic devices are initialized to approximately the same value. After that, the **i**-dimensional synaptic weights are stored as the differential pairs conductance in the $\mathbf{i}^{th}$ data

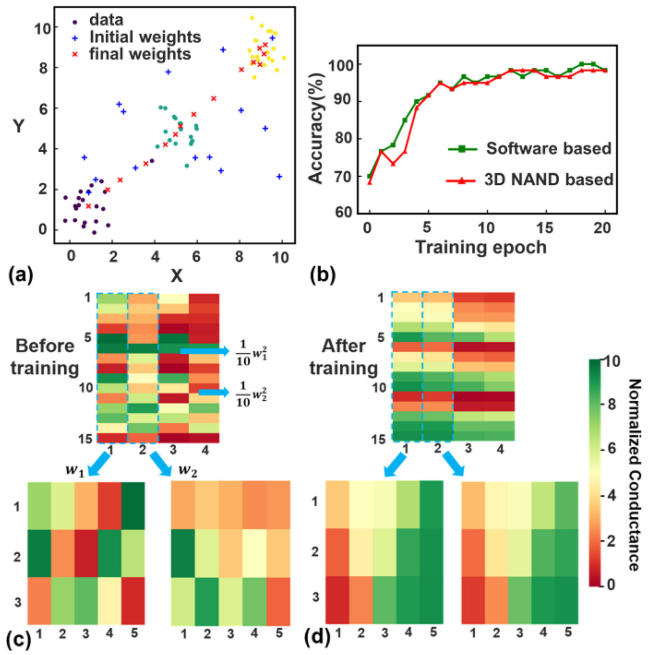

FIGURE 5. **(a) The Experimental results of SOM classification on two-dimensional Gaussian dataset. (b) The traces of accuracy during the training epochs with software and 3D NAND separately. (c)&(d) The heat map of SOM weights before and after training.**

rows, whereas the squared values of weights are stored in the square row. The input data is normalized into the read voltage in the linear region of each device, where input of the square row is fixed at 0.5. Afterwards, perform read operations on the array as described in [9] and use the winner-takes-all (WTA) rule to determine the winner neuron index with the maximum output and the smallest Euclidean distance. Subsequently, the synaptic weights of the winner neuron and the neighboring neurons will be updated according to the scheme in Section III-B, and the negative weights of the square rows are updated concurrently.

We use the 3D NAND-based SOM for data clustering and clustered 60 data points into a 3×5 SOM map. The input of the designed SOM neural network is a two-dimensional dataset containing three sets of 20 coordinate points with Gaussian distribution. In the visualization results of Fig. 5(a), the data points are successfully clustered into three different regions with different colors after 20 training epochs. With a gradually decreasing neighborhood function in the training process, the originally discrete weights are gradually concentrated and distributed to the dense areas of the data distribution. As portrayed in Fig. 5(b), the success rate of the 3D NAND based SOM network approaches 98.3% within 12 training epochs, which means that only one or two points were misclassified. The normalized conductance heat map of weight of neurons before and after training are shown in Fig. 5(c) and Fig. 5(d). The conductance distribution has changed from a random and unordered distribution to an ordered distribution after training.

TABLE 1. **Comparison of mainstream non-volatile memories for SOM.**

Items	3D NAND (This work)	FeFET [3]	RRAM [6]
On/off ratio	$4*10^5$	10^2-10^4	10^4
Cell conductance	~0.1μS	~2μS	~0.1mS
Energy consumption per synapse	~100nW	70nW	~4μW
Write energy/bit	10 fJ	~1 fJ	0.1-1pJ
Accuracy for small SOM	98.3% (3*5)	0.22 error (10*10)	94.6% (8*8)
Classification task	Data point set	RGB dataset	IRIS data set
Retention	Long	Long	Medium
Device Area	$4F^2/N$*	$6F^2$	$4F^2$
Maturity	High	Low	Low
Cost	Low	High	High

*Here N denotes the number of word line layers in 3D NAND flash memory

To demonstrate the energy efficiency of our 3D NAND-based SOM in data clustering, we performed an energy estimation. With a trained SOM, we only need an inference process to implement data clustering. In the inference or reading process, the read voltage is about 1 V, the voltage width is 10μs, and the average conductance is around 0.1μS. The Joule heat dissipated by the 3D NAND array is around:

$$E_{test} = t \times V^2 G \times size_{array} = 90pJ. \quad (6)$$

The result showed that the low energy consumption of the array is only ~90pJ for each clustering task. In order to support the merits of the proposed 3D NAND-based SOM, we quantitatively compare this work with other mainstream memory technologies for SOM solving application. Despite suffering from high operation voltage and long write latency, 3D NAND shows larger on/off ratio and lower cell conductance than other eNVM technologies, indicating that 3D NAND has good potential to realize not high energy consumption per synapse.

Table 1 shows the comparison of mainstream NVM for SOM application. Device integration and density per bit are crucial factors for expanding CIM applications including SOM networks. Table 1 reveals that 3D NAND devices exhibit high technology maturity in terms of 3D integration, resulting in a higher density per bit compared to other non-volatile memories (NVMs). Furthermore, the retention characteristics of the device cell are also of utmost importance for SOM, as it necessitates the ability to maintain weight over a certain period of time. And 3D NAND has better retention characteristics and device-to-device variation than new memory devices. On this basis, 3D NAND SOM has achieved excellent performance in predicting small datasets by multibit storage, which is beneficial for enhancing the learning accuracy of the neural network.

IV. CONCLUSION

This paper presents a novel approach for implementing a SOM network based on 3D NAND Flash for competitive

learning. The ED calculation of SOM is executed with 3D NAND array and modifications of synaptic weights are achieved through conductance tuning on 3D NAND differential pairs. The results of data clustering with Gaussian distributions demonstrate the feasibility and efficiency of the neuromorphic computing based on 3D NAND flash memory.

REFERENCES

[1] G. Singh et al., "A review of near-memory computing architectures: Opportunities and challenges," in *Proc. 21st Euromicro Conf. Digit. Syst. Design (DSD)*, Prague, Czech Republic, 2018, pp. 608–617, doi: 10.1109/DSD.2018.00106.

[2] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biol. Cybern.*, vol. 43, no. 1, pp. 59–69, 1982, doi: 10.1007/BF00337288.

[3] S. Barve et al., "NeuroSOFM: A neuromorphic self-organizing feature map heterogeneously integrating RRAM and FeFET," *IEEE J. Explor. Solid-State Computat. Devices Circuits*, vol. 7, no. 2, pp. 97–105, Dec. 2021, doi: 10.1109/JXCDC.2021.3119489.

[4] H. Zhou et al., "Energy-efficient memristive Euclidean distance engine for brain-inspired competitive learning," *Adv. Intell. Syst.*, vol. 3, no. 11, Nov. 2021, Art. no. 2100114.

[5] A. Mouallem et al., "1T1R in-memory compute for winner takes all application in Kohonen neural networks," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Austin, TX, USA, 2022, pp. 1561–1565, doi: 10.1109/ISCAS48785.2022.9937656.

[6] R. Wang et al., "Implementing in-situ self-organizing maps with memristor crossbar arrays for data mining and optimization," *Nat. Commun.*, vol. 13, no. 1, p. 2289, 2022.

[7] G. Milano et al., "In materia reservoir computing with a fully memristive architecture based on self-organizing nanowire networks," *Nat. Mater.*, vol. 21, no. 2, pp. 195–202, 2022.

[8] W. Shim and S. Yu, "Technological design of 3D NAND-based compute-in-memory architecture for GB-scale deep neural network," *IEEE Electron Device Lett.*, vol. 42, no. 2, pp. 160–163, Feb. 2021, doi: 10.1109/LED.2020.3048101.

[9] W. Zhou et al., "Unsupervised learning in winner-takes-all neural network based on 3D NAND flash," *IEEE Electron Device Lett.*, vol. 43, no. 3, pp. 374–377, Mar. 2022, doi: 10.1109/LED.2022.3144584.

[10] S.-T. Lee and J.-H. Lee, "Neuromorphic computing using random synaptic feedback weights for error backpropagation in NAND flash memory-based synaptic devices," *IEEE Trans. Electron Devices*, vol. 70, no. 3, pp. 1019–1024, Mar. 2023, doi: 10.1109/TED.2023.3237670.

[11] C. Mu et al., "A 200M-query-vector/s computing-in-RRAM ADC-less k-nearest-neighbor accelerator with time-domain winner-takes-all circuits," in *Proc. IEEE 4th Int. Conf. Artif. Intell. Circuits Syst. (AICAS)*, Incheon, South Korea, 2022, pp. 222–225, doi: 10.1109/AICAS54282.2022.9869962.

[12] Y. Ouyang, Z. Xia, T. Yang, D. Shi, W. Zhou, and Z. Huo, "Optimization of performance and reliability in 3D NAND flash memory," *IEEE Electron Device Lett.*, vol. 41, no. 6, pp. 840–843, Jun. 2020, doi: 10.1109/LED.2020.2987087.

[13] W. Zhou et al., "Winner-takes-all neural network based on 3D NAND flash with 1-bit erase scheme," *IEEE Electron Device Lett.*, vol. 44, no. 5, pp. 761–764, May 2023, doi: 10.1109/LED.2023.3262962.

[14] A. Rauber, D. Merkl, and M. Dittenbach, "The growing hierarchical self-organizing map: Exploratory analysis of high-dimensional data," *IEEE Trans. Neural Netw.*, vol. 13, no. 6, pp. 1331–1341, Nov. 2002, doi: 10.1109/TNN.2002.804221.