In-Memory Nearest Neighbor Search With Nanoelectromechanical Ternary Content-Addressable Memory

Jae Seong Lee, Jisoo Yoon, *Graduate Student Member, IEEE*, and Woo Young Choi, *Senior Member, IEEE*

***Abstract*— Nearest neighbor (NN) search is widely used in pattern classification and memory-augmented neural networks. To overcome the von Neumann bottleneck in conventional NN search architecture, in this study, nanoelectromechanical-switch-based ternary content-addressable memory (NEMTCAM) is introduced for the NN classifier. NEMTCAM can calculate the Hamming distance between the input vector and the stored vectors in a parallel search operation. The NEMTCAM operation was experimentally demonstrated. Furthermore, an analytical model for NN search accuracy, including cell-to-cell parasitic resistance, is presented. NEMTCAM can calculate up to 10 Hamming distances in 32-bit words owing to the high current ratio of the NEM memory switches.**

***Index Terms*— Ternary content-addressable memory (TCAM), nearest neighbor search, memory-augmented neural network (MANN), nanoelectromechanical (NEM) memory switch, CMOS-NEM hybrid circuit.**

I. Introduction

NEAREST neighbor (NN) search is essential in applications such as pattern classification [1], image processing [2], and memory-augmented neural networks (MANNs) [3] owing to its simplicity and versatility. For instance, MANNs have NN classifiers to compute the Hamming distance between the input feature vectors and stored ones in memory devices. However, conventional von Neumann architectures face a bandwidth challenge between the processing units and memory devices [4]. To address this problem, some solutions based on ternary content-addressable memory (TCAM) have been proposed [4]–[6]. As shown in Fig. 1, a locality-sensitive hashing (LSH) function converts a real value vector into a binary value vector; subsequently, the TCAM determines the NN vector through a parallel search operation within memory devices.

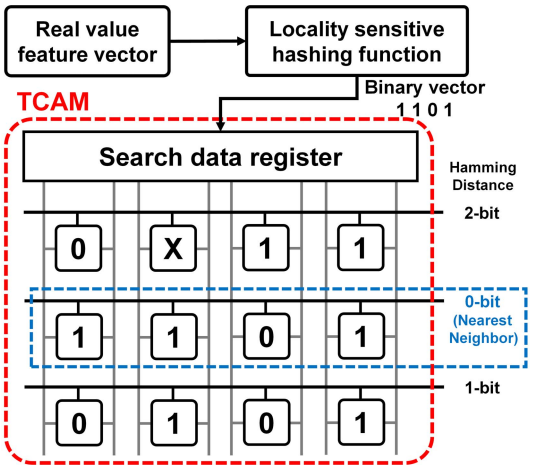

Fig. 1. Typical architecture of MANN-based on TCAM. LSH function converts a real value vector into a binary vector. TCAM calculates the hamming distance from the input binary vector.

Recently, nonvolatile memory devices such as resistive random-access memory (RRAM) cells [5] and ferroelectric field-effect transistors (FeFETs) [6] have been proposed for TCAM design rather than low-density and power-exhausting static RAM (SRAM). Fig. 2 shows various options of unit TCAM cells. Except for conventional ones, the rest of them show no meaningful footprint difference. On the contrary, the alternative TCAM technologies suffer from some challenges in addition to their own advantages. RRAM-based TCAM has a narrow sensing margin owing to its low resistance ratio between the match and mismatch states [7]. In the case of FeFET-based TCAM, the write disturbance must be considered, and the current ratio between the match and mismatch states changes depending on the memory window of FeFETs [8]. These resistance and current ratios are critical for computing the Hamming distance in the TCAM NN search operation.

Meanwhile, nanoelectromechanical-switch-based TCAM (NEMTCAM) has been considered as a promising option, as it features a near-infinite on-off current ratio of NEM memory switches during search operation [9]. Thus, in this study, NEMTCAM was introduced as a novel NN classifier. The manuscript is organized as follows. (i) The search operation

Manuscript received November 11, 2021; revised November 23, 2021; accepted November 24, 2021. Date of publication November 29, 2021; date of current version December 29, 2021. This work was supported in part by the National Research Foundation (NRF) of Korea funded by the Ministry of Science and ICT (MSIT) through the Intelligent Semiconductor Technology Development Program under Grant NRF-2019M3F3A1A02072089, Grant NRF-2021M3F3A2A01037927, and Grant NRF-2021R1A2C1007931 (Mid-Career Researcher Program); in part by the Institute for Information and Communications Technology Planning and Evaluation (IITP) funded by the MSIT through the Information Technology Research Center Program under Grant IITP-2020-2018-0-01421; and in part by the Ministry of Trade, Industry and Energy/Korea Semiconductor Research Consortium (MOTIE/KSRC) through the Technology Innovation Program under Grant 10080575. The review of this letter was arranged by Editor S. Yu. *(Corresponding author: Woo Young Choi.)*

The authors are with the Department of Electronic Engineering, Sogang University, Mapo-gu, Seoul 04107, Republic of Korea (e-mail: wchoi@sogang.ac.kr).

Color versions of one or more figures in this letter are available at https://doi.org/10.1109/LED.2021.3131184.

Digital Object Identifier 10.1109/LED.2021.3131184

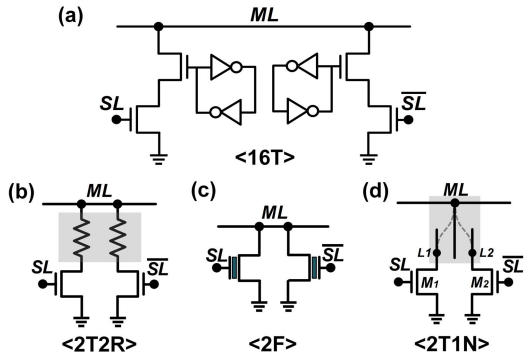

Fig. 2. (a) Conventional SRAM-based, (b) RRAM-based, (c) FeFET-based, and (d) NEMTCAM-based TCAM unit cells. The gray boxes indicate the vertically integrated devices.

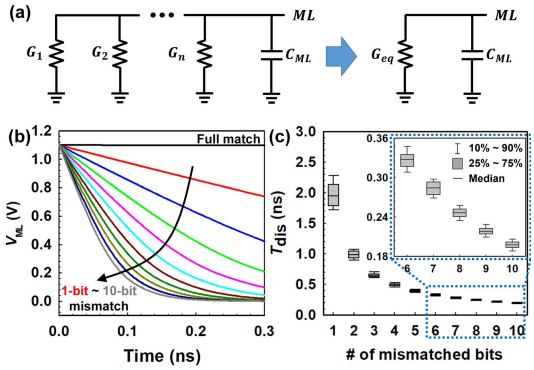

Fig. 3. (a) Equivalent circuit model of a TCAM single word. (b) Simulated V_{ML} increasing with the number of mismatched bits. (c) Calculated T_{dis} with the effect of device-to-device variations. The variations are considered through HSPICE Monte Carlo simulations with 100 trials.

of NEMTCAM will be discussed based on the analytic model and simulation results. (ii) The operation of NEMTCAM will be experimentally demonstrated. (iii) The parasitic effect on accuracy will be discussed.

II. Result and Discussion

Fig. 2d show the unit cell circuit diagram of NEMTCAM. The unit cell comprises a single NEM memory switch and two transistors (M_1 and M_2). NEM memory switches can be integrated in interconnect metal layers using a complementary metal–oxide–semiconductor (CMOS) baseline process [10]. The cantilever beam of an NEM memory switch can move mechanically between the selection line 1 (L1) and selection line 2 (L2) by selectively removing the surrounding inter-metal dielectric (IMD) layer. During the write operation, the write voltage is applied to either L1 or L2 through either M_1 or M_2. By inducing the electrostatic force between the beam and the selection lines, the beam position is toggled to L1 (represent bit 0) or L2 (bit 1) or the center of the gap (bit X). The detailed operation mechanism of NEM memory switches has already been presented and experimentally demonstrated in our previous work [10], [11]. For the search operation, a match line (ML) is first precharged to a high voltage. Subsequently, the search voltage is applied to the complementary search lines (SL and /SL). When search bits are matched with stored bits, the ML voltage (V_{ML}) is maintained at a high value. In contrast, when mismatched, the V_{ML} is discharged to the ground.

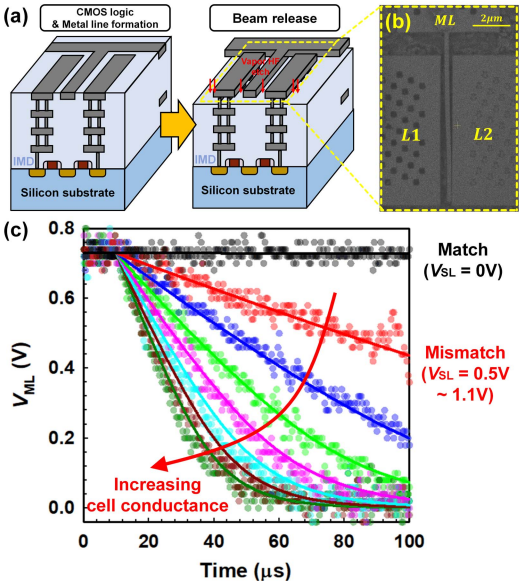

Fig. 4. (a) Key fabrication process of NEMTCAM. (b) Top view of fabricated NEMTCAM after writing bit '1'. (c) Transient response of V_{ML}. The circles represent the experimental results of a unit cell. The solid lines represent the fitted simulation results of 32-bit word.

To configure a word, multiple cells are connected to an ML in parallel. During the search operation, an n-bit word can be modeled using an equivalent circuit, as shown in Fig. 3a. The total equivalent conductance (G_{eq}) is calculated as the sum of all conductance of cells. Each cell has a low conductance when matched, and has a high conductance when mismatched. Thus, the total conductance shows a nearly linear distribution as a function of the number of mismatched bits. As the ML discharge action follows the exponential decay function with the RC constant, the number of mismatched bits can be distinguished by the discharge time (T_{dis}) difference. The search operation of 32-bit NEMTCAM was simulated by using HSPICE [12] with the NEMTCAM compact model [9]. As shown in Fig. 3b, V_{ML} decreases faster as more mismatched bits appear. In circuit level, clocked self-referenced sense amplifiers (CSRSAs) can be used to detect V_{ML} with reasonable accuracy.

The influence of device-to-device variation (or cycle-to-cycle) of transistors (M_1 and M_2) and NEM memory switches on NN search operation is also analyzed as shown in Fig. 3c. In NEM memory switches, the metal-to-metal contact resistance depends on the contact area. However, because the contact resistance is smaller than the on-resistance of a transistor, the contact resistance variation of NEM memory switches does not affect G_{eq}. Our results shows that T_{dis} can be discriminated referring to the number of mismatched cells even if the variations of transistors and NEM memory switches are considered.

To experimentally demonstrate the operation of the NEMTCAM, the unit cell circuits were fabricated. Fig. 4a shows the key process steps of the fabricated NEMTCAM. First, CMOS devices are integrated on a silicon substrate using the standard 65-nm CMOS frontend process. Subsequently, metal interconnect lines and vias are formed using a standard CMOS backend process. In this step, NEM memory switches

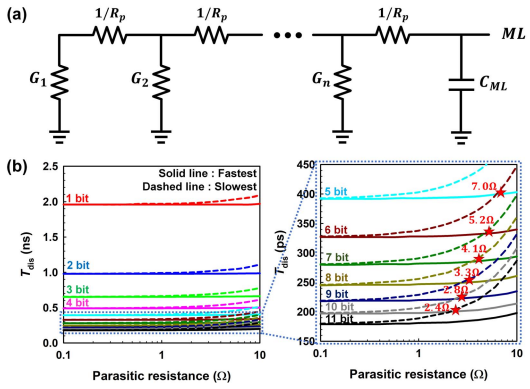

Fig. 5. (a) Modified equivalent circuit model including R_p. (b) Simulation results of T_{dis} of 32-bit word NEMTCAM. In our simulation, parasitic capacitances such as distributed line capacitance and parasitic capacitance of M_1/M_2 are also included, which has only little influence on accuracy.

are patterned using a dual damascene process. After patterning, the IMD layer surrounding the cantilever beam was selectively removed using vapor HF and a fluorine-based plasma etch process [10]. Fig. 4b shows the top view of the fabricated NEMTCAM when bit '1' is stored. Then, the NEMTCAM search operation is performed, as shown in Fig. 4c. After the MLs are precharged to a high voltage, V_{SL} is applied to the gate electrode of M_1. According to the V_{SL} value, the ML voltage is maintained or discharged to the ground. To observe T_{dis} difference according to G_{eq} difference in the unit cell, the magnitude of V_{SL} was adjusted as an alternative (increased from 0.5 V to 1.1 V in 0.1 V steps). It was observed that a higher conductance (applied higher V_{SL}) led to a faster V_{ML} drop, as shown by the circles in Fig. 4c. Similarly, according to the simulation results with 32-bit word, V_{ML} drops more rapidly as G_{eq} increases in proportion to the number of mismatched bits, as shown by the solid lines in Fig. 4c. The large delay of the measurement results are originated from the parasitic capacitance of large measurement pads and external capacitor for low-frequency measurements.

III. Parasitic Resistance Effect Analysis

The effect of parasitic elements on the NN search operation will be analyzed. Cell-to-cell parasitic resistance (R_p) has a considerable influence on accuracy depending on the location of the mismatched bits [5]. Fig. 5a shows the modified equivalent circuit model of an n-bit word, including R_p. Assuming that a word has k mismatched cells and the location of mismatched cells is the closest to ML, G_{eq} is the largest and T_{dis} is the smallest. Conversely, when the location is the farthest, G_{eq} is the smallest, and T_{dis} is the largest. Thus, T_{dis} ranges between the fastest and slowest cases. If the T_{dis} ranges of different numbers of mismatched cells overlap with each other, the accuracy of the NN search operation will decrease.

The overlap-starting point of R_p can be estimated based on the equivalent circuit analysis as follows. By solving the Kirchhoff matrix equation, G_{eq} can be calculated as [5]:

$$G_{eq} \approx \frac{\sum_{i=1}^{n} G_i}{1+\sum_{i=1}^{n} i G_i R_p} \left(R_p G_i \ll 1\right). \tag{1}$$

where G_i is the conductance of the i^{th} cell. Assuming that k bits are mismatched, G_i of the matched and mismatched cells are replaced with G_{match} and G_{miss}, respectively. G_{eq} can then be recalculated in the two extreme cases. G_{eq}'s when T_{dis} is the smallest and the largest are represented as $G_{eq,fastest}(k)$ and $G_{eq,slowest}(k)$, respectively. Both are derived as follows:

$$G_{eq,fastest}(k) = \frac{kG_{miss} + (n-k)G_{match}}{1 + R_p\left(\sum_{i=1}^{k} iG_{miss} + \sum_{j=k+1}^{n} jG_{match}\right)}, \tag{2}$$

$$G_{eq,slowest}(k) = \frac{kG_{miss} + (n-k)G_{match}}{1 + R_p\left(\sum_{i=1}^{n-k} iG_{match} + \sum_{j=n-k+1}^{n} jG_{miss}\right)}. \tag{3}$$

In the case of NEMTCAM, $G_{match}/G_{miss} \ll 1$ is valid owing to the high current ratio of the NEM memory switches. Subsequently, Eqs. (2) and (3) are simplified as

$$G_{eq,fastest}(k) = \frac{2kG_{miss}}{2 + k(k+1)R_pG_{miss}}, \tag{4}$$

$$G_{eq,slowest}(k) = \frac{2kG_{miss}}{2 + k(2n-k+1)R_pG_{miss}}. \tag{5}$$

To avoid the overlap between the k-bit and $(k+1)$-bit mismatched case, the following condition should be satisfied:

$$G_{eq,fastest}(k) < G_{eq,slowest}(k+1). \tag{6}$$

As a result, the condition of R_p is derived as

$$R_p < \frac{2R_{miss}}{k(k+1)(2n-2k-1)}). \tag{7}$$

The calculated R_p values at which the overlap occurs are consistent with the simulation results, as shown in Fig. 5b.

According to the International Technology Roadmap for Semiconductors, in the case of the 65-nm technology node, R_p is estimated to be 2.3 Ω [13]. This indicates that the NEMTCAM using the 65-nm node can discriminate up to 10 Hamming distances in a 32-bit word ($k = 10$ and $n = 32$), which is applicable to generic NN search [6]. For higher accuracy, when a larger bit width is needed, G_{miss} can be decreased. It will widen the interval between each T_{dis} range at the expense of the overall operating speed.

IV. Summary

An in-memory NN search operation using the NEMTCAM was successfully confirmed. In addition, the feasibility of NEMTCAM was experimentally demonstrated by using the fabricated cell using the 65-nm CMOS baseline process. NEMTCAM can calculate the Hamming distance by the discharge conductance distribution with good accuracy, even considering parasitic element effects. This will enable next-generation CAM architectures to transcend the existing neural network or pattern recognition systems.

Acknowledgment

The chip fabrication and EDA tool were supported by the IC Design Education Center and Inter-University Semiconductor Research Center.

REFERENCES

[1] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967, doi: 10.1109/TIT.1967.1053964.

[2] T. Liu, C. Rosenberg, and H. Rowley, "Clustering billions of images with large scale nearest neighbor search," in *Proc. IEEE Workshop Appl. Comput. Vis. (WACV)*, Feb. 2007, p. 28, doi: 10.1109/WACV.2007.18.

[3] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," *Mach. Learn. Res.*, vol. 48, pp. 1842–1850, Jun. 2016. [Online]. Available: http://proceedings.mlr.press/v48/santoro16.html

[4] M. Imani, Y. Kim, and T. Rosing, "NNgine: Ultra-efficient nearest neighbor accelerator based on in-memory computing," in *Proc. IEEE Int. Conf. Rebooting Comput. (ICRC)*, Nov. 2017, pp. 1–8, doi: 10.1109/ICRC.2017.8123666.

[5] Y. Liao, B. Gao, W. Zhang, P. Yao, X. Li, J. Tang, Z. Li, S. Cui, H. Wu, and H. Qian, "Parasitic resistance effect analysis in RRAM-based TCAM for memory augmented neural networks," in *Proc. IEEE Int. Memory Workshop (IMW)*, May 2020, pp. 1–4, doi: 10.1109/IMW48823.2020.9108137.

[6] K. Ni, X. Yin, A. F. Laguna, S. Joshi, S. Dünkel, M. Trentzsch, J. Müller, S. Beyer, M. Niemier, X. S. Hu, and S. Datta, "Ferroelectric ternary content-addressable memory for one-shot learning," *Nat. Electron.*, vol. 2, no. 11, pp. 521–529, 2019, doi: 10.1038/s41928-019-0321-3.

[7] D. R. B. Ly, B. Giraud, J.-P. Noel, A. Grossi, N. Castellani, G. Sassine, J.-F. Nodin, G. Molas, C. Fenouillet-Beranger, G. Indiveri, E. Nowak, and E. Vianello, "In-depth characterization of resistive memory-based ternary content addressable memories," in *IEDM Tech. Dig.*, Dec. 2018, pp. 20.3.1–20.3.4, doi: 10.1109/IEDM.2018.8614603.

[8] K. Ni, X. Li, J. A. Smith, M. Jerry, and S. Datta, "Write disturb in ferroelectric FETs and its implication for 1T-FeFET AND memory arrays," *IEEE Electron Device Lett.*, vol. 39, no. 11, pp. 1656–1659, Nov. 2018.

[9] J. S. Lee and W. Y. Choi, "Nanoelectromechanical-switch-based ternary content-addressable memory (NEMTCAM)," *IEEE Trans. Electron Devices*, vol. 68, no. 10, pp. 4903–4909, Oct. 2021, doi: 10.1109/TED.2021.3106886.

[10] H. S. Kwon, S. K. Kim, and W. Y. Choi, "Monolithic three-dimensional 65-nm CMOS-nanoelectromechanical reconfigurable logic for sub-1.2-V operation," *IEEE Electron Device Lett.*, vol. 38, no. 9, pp. 1317–1320, Sep. 2017, doi: 10.1109/LED.2017.2726685.

[11] G. Baek, J. Yoon, and W. Y. Choi, "Tri-state nanoelectromechanical memory switches for the implementation of a high-impedance state," *IEEE Access*, vol. 8, pp. 202006–202012, 2020, doi: 10.1109/ACCESS.2020.3036189.

[12] *HSPICE User Guide: Simulation and Analysis*, Version P-2019.06, Synopsys, Mountain view, CA, USA, Jun. 2008.

[13] Semiconductor Industry Association. (2005). *International Technology Roadmap for Semiconductors (ITRS)*. [Online]. Available: https://www.semiconductors.org/