A Reliable High-speed Compact In-memory Matching Circuit for CAM-Application Based on NV-RAM

Quang-Manh Duong, Quang-Kien Trinh, Hai-Duong Nguyen, Van Phuc-Hoang, Hoang-Gia Vu, Dinh-Ha Dao, Van-Toan Tran, Duy-Manh Luong
Le Quy Don Technical University, Hanoi, Vietnam

Correspondence: Quang-Kien Trinh (kien.trinh@lqdtu.edu.vn)
Communication: received 15 July 2022, revised 31 August 2022, accepted 26 September 2022
DOI: 10.32913/mic-ict-research.v2022.n2.1060

***Abstract*: This paper presents an effective approach for implementing content address memory (CAM) based on Non-volatile random-access memory (NV-RAM) technologies. We used the 2T-2R bitcell structure implemented on a 65nm CMOS process with a special in-memory matching circuit for realizing low-delay and energy-efficient lookup operations. The simulation results on Synopsys HSPICE indicate that the proposed CAM design can achieve a search error rate of 0.03-4.61%, search energy per bit of 4.36-6.47 fJ, and an extremely small search latency varying from 0.11-0.12 ns depending on the specific design configurations.**

Keywords: *In-memory computing, NV-RAM, NV-CAM, CAM matching circuit, high-speed.*

I. INTRODUCTION

Recently, Content Addressable Memory types (CAM/ TCAM, later collectively referred to as CAM) [1] have been widely used in applications that require the capability of high-speed access as well as real-time intense processing. Instead of receiving address and returning data as in traditional memories, CAM receives input patterns and returns the corresponding address of the data being stored, providing high-frequency lookup operations with the cost of integrated density and standby power consumption. The comparison with an input pattern is performed in parallel in a single clock cycle across all the rows of storage data, which are arranged by a lookup-table structure. In general, CAM is a type of associated memory, due to the outstanding characteristic of lookup speed, commonly used in massive pattern search applications such as network routers, searching engines or virus checkers, and so on.

The basic structure of a CAM is depicted in Fig. 1(a) and usually consists of two main components: a matrix of CAM cells that stores data patterns, where the number of columns equals the width of the search pattern, and a Priority

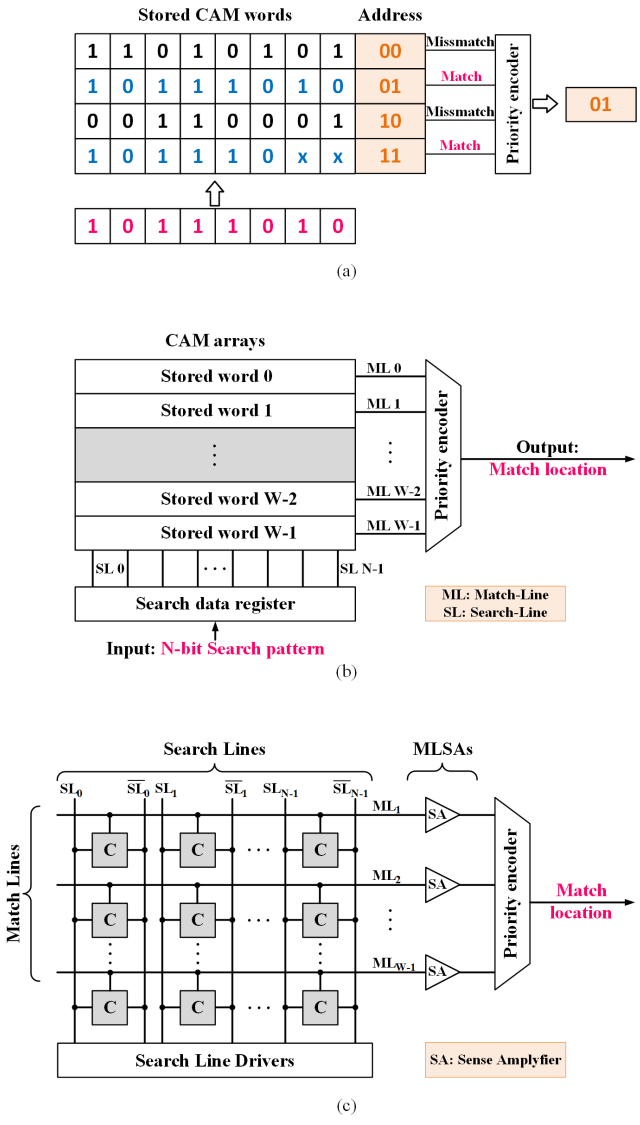


Figure 1. The basic structure and operations of a CAM.

Encoder (PE), which is used to select the address of the row with the highest priority among the similar rows that match the input pattern word. The CAM implementation diagram is shown in Fig. 1(b), where the stored words are arranged in rows, and each of those (a.k.a arrays) is connected to a single Match Line (ML), indicating that the search pattern and stored one at this location are identical (matched case) or different (mismatched case). The MLs are fed into the PE, which is normally responsible for selecting the most suitable position (with the highest priority) among them where the matched cases are found, normally the CAM row with the lower address will have a higher priority. Thus, simply, the CAM circuit receives at the input an N-bit search pattern and returns at the output the matching position of this pattern with a correspondingly selected stored word, either from a large input data space obtain a smaller data space at the output indicating the matching case in the CAM database. Finally, a practical CAM-application circuit model is depicted in Fig. 1(c) based on the diagram in Fig. 1(b) with additional peripheral circuits, including a sense amplifier at the output of the MLs (Match Line Sense Amplifiers - MLSAs) and a circuit that controls the operation of the search lines (Search Line Drivers). Each bit of the pattern word is stored in a CAM cell, which of the same column is connected to a pair of complementary Search Lines $(SL, \overline{SL})$. The lookup process in the CAM database starts from loading the search pattern into Search Line Drivers, followed by the high-level pre-charging for all the MLs, setting the matched state for all of them as well as the whole CAM arrays. In the next step, Search Line Drivers distribute search patterns through all the SL pairs, each CAM cell compares its stored bit with the value on the correspondingly connected SL pair. The comparison results of all CAM cells on the same row determine the final logical state of the MLs, which is then fed into the MLSAs. MLs whose all cells are matched retain high-prechanged, otherwise, they will be discharged to the ground if at least one cell is mismatched. The matching state of the MLs is determined by the sense amplifiers, they provide input signals to the PE to decide where the highest-priority matched location for the input search pattern.

CAM design challenges

The main design parameters that affect the performance of CAMs include the following: Energy, Area, Speed (latency), and Reliability, which are also challenges [2] when designing specific CAM applications. These parameters have different importance depending on the specific application, so it is necessary to determine which challenge as well as design parameter has the highest priority to optimize in an actual scenery. The energy consumed for the CAM operations is primarily composed of energy dissipated in ML, SL, and PE, as well as dissipated energy caused by leakage current, access transistors, decoders, search line drivers, and ML precharge circuits. Saving energy in CAM circuit design is a primarily critical research problem. Several innovations in CAM design improvement have recently been proposed such as a promising memory technology called Priority-Decision in Memory (PDM) [3] that eliminates the need for PEs to identify matched patterns, thereby significantly saving the search energy consumption; the precharge-free CAM schemes [4], which overcome the drawbacks of precharge-based designs and also improve performance during the search by reducing half of the ML evaluation time. The area (footprint) of the CAM array mainly depends on the area of the storage bit-cell. CAM based on a dynamic storage mechanism similar to the DRAM has been proposed to reduce the footprint but at the cost of refresh power and expensive process technology. Thus, several emerging memory technologies (discussed below in this section) have been proposed to address challenges for CAM both in terms of energy and area. The speed of the CAM is inversely proportional to the delay, which consists of two main components: the match delay and the delay of the PE. The match delay depends on several factors such as bitcell topology (NAND or NOR), ML capacitance, array (word) size, and the search pattern, which can be improved by a few techniques such as lowering the loading capacitance and ML voltage [5]; pipelined ML [6]. Meanwhile, multiple match detection (MMD) circuits and priority encoders (PEs) have been proposed and employed in TCAM chips [7], which show a great improvement not only in speed but also in the energy consumption of PE. Finally, the reliability of CAM is mainly governed by the characteristics of the memory employed for data storage, where RRAM or PCRAM typically suffer from poor endurance which limits the reliability of the CAM cells. In general, the reliability of CAM using CMOS is limited by the aging mechanisms of transistors, specifically, negative bias temperature instability (NBTI), positive bias temperature instability (PBTI), and time-dependent dielectric breakdown (TDDB), impacts of NBTI and PBTI effects on ternary CAM and solutions to reduce their aging effects are proposed in [8], [9].

Emerging CAM technologies

In previous decades, materials for CAM fabrication are mainly based on SRAM technology [10] (herein called S-CAM) with a cross-coupled inverter structure which offers fast read/write speeds and relatively high reliability but is associated with some major disadvantages such as large-power dissipation for both storage and search operations, non-zero standby energy consumption, low integration density in addition to the high cost. More specifically,

besides the large cell area caused by a large number of transistors (most of which employ 16 transistors per cell), S-CAM also suffers from high standby power induced by the increasing leakage current, especially when the process of technology shrinks below 45 nm, which has become a major barrier for the compact battery-operated devices. To solve these problems, the researchers proposed using emerging non-volatile memories (NV-CAM) as an alternative to S-CAM, although there are many challenges involved in realizing NV-CAM in mass production and some limitations compared to the previous technology (S-CAM) in terms of search speed and reliability. Emerging non-volatile technologies [11] such as Resistive RAM (RRAM), Magnetoresistive RAM (MRAM), Ferroelectric RAM (FeRAM), and Phase-change RAM (PCRAM) are capable of providing excellent solutions for bitcell area and energy efficiency. On the one side, NV-CAM reduces the number of transistors in a bitcell by replacing them with alternative storage elements, on the other side, being nonvolatile energy, the power consumption of NV-CAM is reduced by effectively blocking its leakage currents during the idle mode, making zero standby power. However, NV-CAM cells suffer from low reliability during search operations because the ML voltage difference (Delta) between the matched and mismatched cases is much smaller than that of the S-CAM cell. There exists a constraint when designing NV-CAM, in which to improve the reliability of NV-CAM it is necessary to increase the value of Delta, which depends on resistance ratio and process variation. To reduce process variation, the size of transistors should be increased, leading to cell area and power overheads. A recent study [12] revealed that PCRAM is the most energy-consuming memory with a medium bitcell area, whereas RRAM, which has inherent limitations in endurance compared to the virtually infinite endurance of other emerging non-volatile memories, has a wide spectrum of cell area values, and generally consumes less energy than PCRAM. Magnetic RAM, typically Spin-Torque Transfer MRAM (STT-MRAM), with low energy consumption, relatively small bitcell area also as high integration density, excellent CMOS compatibility, and extreme long-endurance ($\sim 10^{16}$), represents a promising candidate for CAM in the future compared to the other options [13], [14].

Notable CAM designs related to this work

The design of an NV-CAM circuit to perform data search and comparison operations with dependable reliability, high speed, low energy consumption, and small area occupancy is becoming increasingly essential for In-Memory Computing (IMC) applications. A variety of NV-CAM cell structures has been proposed to overcome the above-mentioned limitations of S-CAM including 2T-2R NV-CAM cell [15], 2.5T-1R NV-CAM cell [16], 3T-2R NV-CAM cell [17], 4T-2R NV-CAM cell [18]. These NV-CAM cell designs, so-called first class, have in common the use of a small number of transistors, ranging from 2T to 4T, which gives the advantage of cell area. In addition, most of them achieve very good results in terms of average energy consumption for search operations, but no metrics regarding their reliability (i.e. Search Error Rate - SER) have been revealed. The second class of CAM-cells that are considered in this paper has rather large bitcell structures, including 9T-2R NV-CAM cell [19], 10T-4R NV-CAM cell [20], and 15T-4R NV-CAM cell [21]. Despite their large size, compared to the cells in first class, these NV-CAM cells have surprising results in the speed of look-up operations, which is demonstrated through the sensing delay. The third class consists only of a 16T-CAM cell based on traditional SRAM memory [22], which is the best S-CAM design published with outstanding reliability (absolute zero SER) and competitive figures on search energy also as sensing delay.

From all of the presented CAM-cell classes, some impressive designs have achieved excellent results only on a single design parameter, but almost none of the proposals can balance all the design criteria, which is needed for applicability. For example, design [15] is the best in the cell area but also has the worst sensing delay, while design [21] and design [22], possessing excellent results on search energy and sensing delay respectively, are the worst designs in the cell area. The restrictions on some design parameters in the prior works [15]-[22] are the fundamental motivation for us to offer a balanced solution for the NV-CAM structure that can meet most of the aforementioned design criteria at an acceptable cost. In this paper, we propose a compact 2T-2R NV-CAM cell and a reliable in-memory matching circuit that employs the proposed cell for CAM applications which ensures very small SERs, the lowest search latencies with relatively low cost of search energies.

The main contributions of our work can be summarized as the followings:

- An in-memory matching circuit based on the 2T-2R NV-CAM cell structure which performs low-latency massive parallel lookup operations.
- A variety of evaluations on the reliability of the proposed design based on the SER value and sensing margin of the ML voltage against the effects of process variation using Monte Carlo simulations.
- An in-depth analysis of all design criteria and comparisons with other proposed CAM cells that employed traditional SRAM also as various emerging memory technologies including RRAM, PCRAM, and MRAM.

The remainder of the paper is organized as follows:

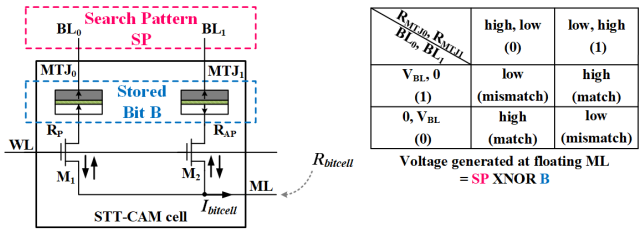


Figure 2. The structure of the proposed 2T-2R NV-CAM cell.

Section II proposes the 2T-2R bitcell configuration and the NV-CAM array architecture for the in-memory matching circuit. The impacts of process variations on the proposed design are analyzed by Monte Carlo simulation and presented in Section III. Design evaluation and comparison with several prior works are considered in Section IV. Section V concludes this work.

II. THE PROPOSED DESIGN FOR THE NV-CAM MATCHING CIRCUIT

A. 2T-2R NV-RAM Bitcell For In-Memory Matching circuit

Our proposed configuration in the bitcell level for STT-MRAM-based CAM is illustrated in Fig. 2, which can be extended to other NV-RAM devices (RRAM, PCRAM). Similar to the structure in [23], [24], a complementary bitcell structure is adopted here, which represents one stored bit in the CAM word, the complement BL to feed the search pattern bit, and the match line ML is left floating and will be used as the sensing node.

The proposed bitcell can perform a single XNOR function as described in the following. The two magnetic tunnel junctions (MTJ_0, MTJ_1) of STT-MRAM are programmed to be low (R_P) and high (R_{AP}) or revise, which is respected to '1' and '0' values and will be used as the stored bit B in the CAM word. Similarly, BL_0 and BL_1 are complementarily assigned to be (V_{BL}, 0) or (0, V_{BL}), respectively representing '1' and '0' of the input search pattern.

The gate of the access transistors M_1 and M_2 are connected to the word line WL, and the source is connected to the match line ML, two transistors are identical. It is obvious that this structure (1, 1) and (0, 0) combinations result in the same circuit, so as for (1, 0) and (0, 1). Therefore this circuit can perform the XNOR function, which is adopted as the matching operation (comparing the stored words with search pattern input) in the CAM applications.

The match line voltage V_{ML} generated by a single bitcell being activated within a matchline can be represented through the Norton equivalent circuit theorem [25].

For a single bitcell under a '1' ('0') input search pattern SP, $V_{BL,0} = V_{BL}$ and $V_{BL,1} = 0$ ($V_{BL,0} = 0$ and $V_{BL,1} = V_{BL}$) and the bitcell short-circuit current $I_{bitcell}$ is expressed as:

$$I_{bitcell}(SP=1) = \begin{cases} \frac{V_{BL}}{R_P+R_{access,0}} & if B = 1 \\ \frac{V_{BL}}{R_{AP}+R_{access,0}} & if B = 0 \end{cases} \quad (1)$$

$$I_{bitcell}(SP=0) = \begin{cases} \frac{V_{BL}}{R_{AP}+R_{access,1}} & if B = 1 \\ \frac{V_{BL}}{R_P+R_{access,1}} & if B = 0 \end{cases} \quad (2)$$

The overall resistance $R_{bitcell}$ seen from the ML terminal of the bitcell equals to $(R_{MTJ,0} + R_{access,0})||(R_{MTJ,1} + R_{access,1})$. By considering that $R_{access,0} = R_{access,1} = R_{access}$, from Eq. 1-2 the resulting voltage V_{ML} for an isolated active bitcell at position *(i, j)* within the memory array is:

$$V_{ML} = R_{bitcell} \cdot I_{bitcell,ij} = \frac{V_{BL}}{X_{ij}} \quad (3)$$

where X_{ij} is defined as:

$$X_{ij} = \begin{cases} \frac{R_P+R_{access}}{R_{bitcell}} & if \quad \overline{B_{ij} \oplus SP_J} = 1 \\ \frac{R_{AP}+R_{access}}{R_{bitcell}} & if \quad \overline{B_{ij} \oplus SP_J} = 0 \end{cases} \quad (4)$$

being B_{ij} the stored bit in CAM word, and SP_j is the input fed to column j. From Eq. 3-4 the match line voltage is proportional to the bit line voltage and directly depends on the XNOR of the search bit and the bit stored in the bitcell.

The resulting sensing margin $SM_{bitcell}$ defined as the ML voltage difference between the output XNOR is '1' and '0', which can be simplified as:

$$SM_{bitcell} = V_{BL} \cdot R_{bitcell} \left(\frac{1}{R_P + R_{access}} - \frac{1}{R_{AP} + R_{access}} \right) = V_{BL} \cdot TMR \cdot \frac{R_P}{R_P + R_{AP} + 2R_{access}} \quad (5)$$

where TMR is the MTJ technology-dependent tunneling magnetoresistance ratio $(R_{AP}-R_P)/R_P$ [26]. For extending to other NV-RAM devices (RRAM, PCRAM), we define the metric of resistance ratio $R_{ratio} = R_{AP}/R_P$. From Eq. 5, the sensing margin can be improved by adopting NV-RAM devices with a higher R_{ratio} or using the stronger access transistors, at the cost of a larger area per bitcell. The effects of the access transistors and resistance ratio (R_{ratio}) on the sensing margin, and hence on the reliability of our design will be discussed in the next section.

B. NV-CAM Array Architecture And In-Memory Matching Circuit

This section presents our proposed NV-CAM array architecture. The $M \cdot N$ array in Fig. 3 consists of M rows and N columns, with wordlines being used to enable computation in the respective rows, and the match lines as outputs. The search line drivers encode search patterns (SP) that are compared to stored bits in the selected word line by the XNOR functions, which are presented in detail in Fig. 2 and Section II. A. It could be proven that the XNOR operation in Eq. 1-3 could be generalized for the row level simply by merging all match lines of all bitcells in a row.

As all rows operate independently of each other, let us consider a single row as depicted in blue in Fig. 3. Let the number of XNOR outputs equal to '1' ('0') across the row be $N_1 (N_0 = N - N_1)$. Compared to the single bitcell resistance $R_{bitcell}$ in Section II. A, the resistance seen from the ML terminal in a row is reduced to $R_{bitcell}/N$. From Eq. 2-3, the select line voltage $V_{(ML,i)}$ in the considered row i-th is:

$$V_{ML,i} = \frac{R_{bitcell}}{N} \sum_{j=0}^{N-1} I_{bitcell,ij} = V_{BL} \frac{R_{bitcell}}{N} \sum_{j=0}^{N-1} \frac{1}{X_{ij}}$$
$$= V_{BL} \left(\frac{N_0}{N} \cdot \frac{R_{bitcell}}{R_{AP} + R_{access}} + \frac{N_1}{N} \cdot \frac{R_{bitcell}}{R_P + R_{access}} \right) \quad (6)$$

where X_{ij} is the resistance determining the bitcell current of the bitcell (i, j) in Eq. 1-3.

The matching output is determined by assigning the output OUT_i of the select line in the generic i-th row to '1' if $V_{(ML,i)} \geq V_{REF}$, and to '0' if $V_{(ML,i)} < V_{REF}$. (V_{REF} is the reference voltage between the V_{ML} in the FullMatch case and 1-bit MisMatch case). Hence, the logical output of the sense amp in the i-th row is equal to the logical conjunction of the XNOR computations across its bitcells:

$$OUT_i = \bigwedge_{j=0}^{N-1} \overline{B_{ij} \bigoplus SP_J} \quad (7)$$

In this work, a conventional current latch sense amplifier (CLSA) as in [14] is employed to generate the matching outputs, its schematic and timing diagram with Monte Carlo simulation are depicted in Fig. 4.

From Eq. 6, the row select voltage linearly depends on N_1, N and ranges from $V_{BL} \cdot \frac{R_{bitcell}}{(R_{AP}+R_{access})}$) to $V_{BL} \cdot \frac{R_{bitcell}}{(R_P+R_{access})}$, therefore, the sensing margin depends not only on the different widths of the access transistors and resistance ratio ($R_{ratio} = R_{AP}/R_P$) as aforementioned in sub-Section II.A but also the length of the stored word in CAM as illustrated in Fig. 5.

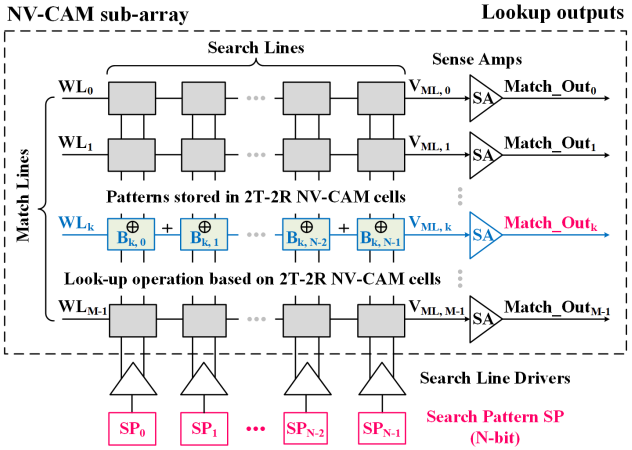


Figure 3. The proposed NV-CAM structure and operations.

As shown in Fig. 5(a), at the small CAM word length (N = 4 or 8 bits), an increase in access transistor size leads to a clear wider sensing margin. However, with a larger stored word length (i.e., N = 16, 32, or 64 bits), the sensing margin is almost independent of the access transistor.

Fig. 5(b) indicates the dependence of the sensing margin on different resistance ratios (R_{ratio}) when the width of the access transistor is 4X the minimum allowed by the technology (i.e., Width = 4W = 0.54 µm, Length = 0.06 µm). R_{ratio} is selected to evaluate in a relatively wide range of values from 2 to 100, where a small value is typical for STT-MRAM (R_{ratio} = 2), an average value commonly encountered in PCRAM (R_{ratio} = 10), and a large one correspondingly represents RRAM (R_{ratio} = 100). Similar above considerations, at a high level of the word length (N), R_{ratio} has a minor effect on the sensing margin. For example, when N = 64 bits, the SM value is 2.7 mV (17.4) when R_{ratio} = 2 (R_{ratio} = 100), meanwhile when N = 4 bits, the SM value is 37.3 mV (116.9) when R_{ratio} = 2 (R_{ratio} = 100).

Our simulation was performed by using the Hspice model with a commercial 65 nm CMOS library. Besides, the analysis of process variations was performed through the Monte Carlo simulations using statistical models for both CMOS and resistance values (the standard deviation of the resistances is 7%). For the CMOS transistors, local and global variations were accounted for by using the statistical models provided by the foundry. The impacts of process variations on our design will be discussed in the next section.

III. IMPACTS OF PROCESS VARIATIONS

Process variations degrade the sensing margin, therefore, errors in the logical value at the output of each row for the

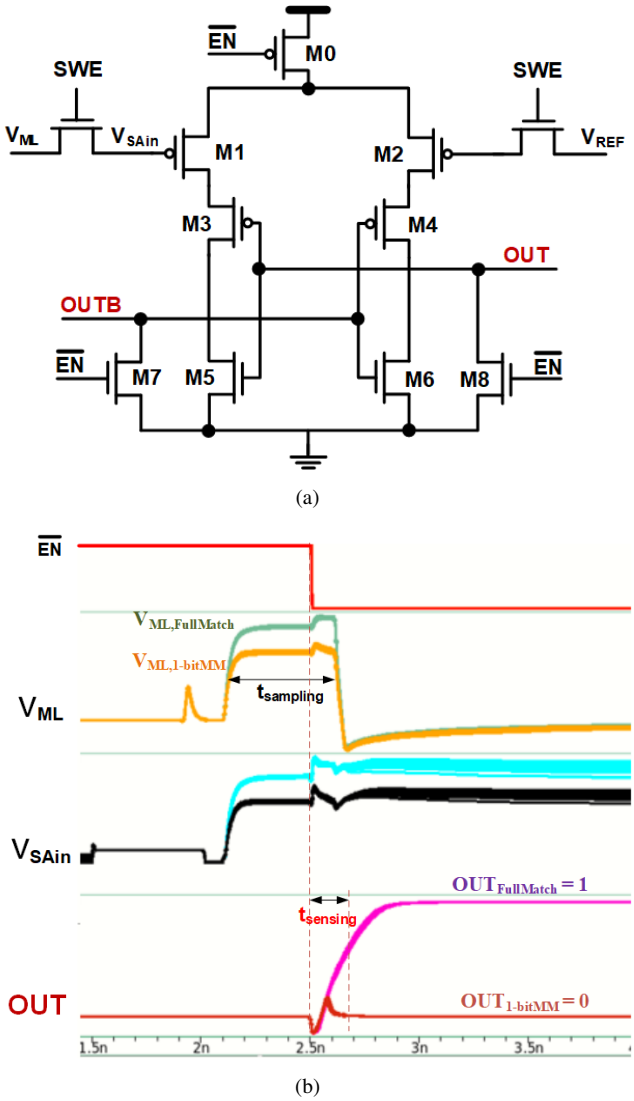


Figure 4. The schematic of CLSA (a), and corresponding timing diagram with Monte Carlo simulation (b).

matching operation are induced. To quantify the impact of variations, Monte Carlo simulations were run under global and local variations in both transistors and resistances. Typical arrays with M = 64 rows and N = 4, 8, 16, 32, and 64 columns will be investigated below. All building blocks from decoding to read-out are included to make the array fully functional and self-consistent.

To quantify the impact of process variations on robustness, we introduce, the SER is the probability of an erroneous output caused by variations in an elementary matching operation [27]:

$$SER = Pr[SM < 0] = \frac{1}{2}\left[1 + erf\left(-\frac{1}{\sqrt{2}} \cdot \frac{1}{\sigma_{SM}/\mu_{SM}}\right)\right] \quad (8)$$

In this work, we investigate SER as the worst-case

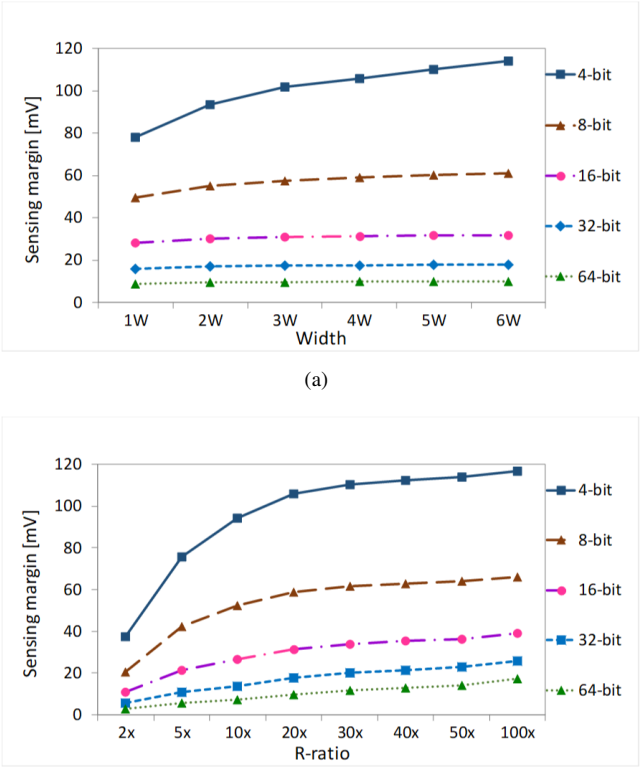


Figure 5. Sensing margin at different word lengths versus width of the access transistor with fixed R_{ratio} = 20 (a), and R_{ratio} with fixed Width = 4W (b).

search error rate (smallest sensing margin) (i.e., when distinguishing fully matched and 1-bit mismatched output). Therefore, $SM = V_{ML_{FM}} - V_{ML_{1MM}}$ is the sensing margin or the difference between ML voltages in cases of FullMatch and 1-bit MisMatch, μ_{SM} and σ_{SM} denote the mean and the standard variation of the sensing margin SM:

$$\sigma_{SM} = \sqrt{\sigma_{ML}^2 + \sigma_{REF}^2 + \sigma_{offset}^2} \quad (9)$$

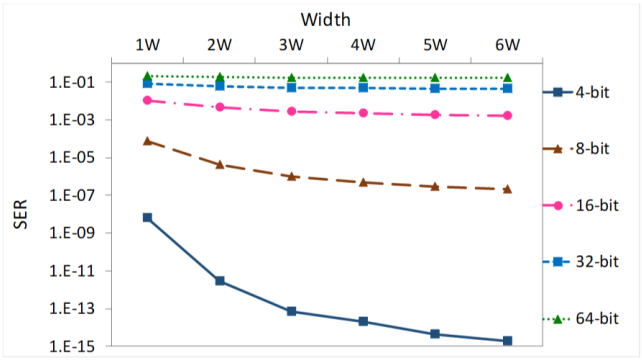
$$\mu_{SM} = |\mu_{ML} - \mu_{REF}| \quad (10)$$

where μ_{ML}, μ_{REF} (σ_{ML}, σ_{REF}) are respectively the mean (standard variation) of the match line voltage and the reference voltage. As usual, the offset of the sense amplifier has zero mean and standard deviation σ_{offset} [28].

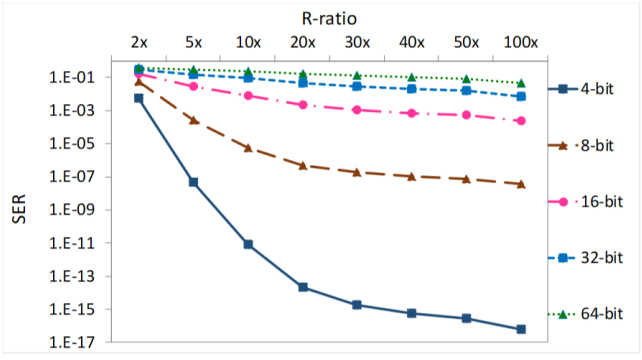
In this work, we consider σ_{offset} is 5mV for a reasonable area of circuit design (sense amplifier size of 16x transistor width and 2x length) [14].

The dependence of the SER on the width of the access transistor, the word length, and R_{ratio} are demonstrated in Fig. 6. Similar to the dependence of the sensing margin on those parameters (N, width of access transistor, R_{ratio}) shown in Fig. 5, the clear differences between the SER

value at low-level of the word-length (N = 4, or 8 bits) versus the SER value at higher N (i.e., N = 16, 32, or 64 bits) is once again depicted. At 4-bit word length, SER can reach extremely small values, specifically, it can be lowered to 6.04e-17 in the case of an R_{ratio} equal to 100, corresponding to RRAM technology. In contrast, SER values become quite large and pose some concerns when increasing the word length to 64 bits, especially, when using STT-RAM technology (R_{ratio} = 2), SER is 3.98 e-1 (i.e 39.8%). As shown in Fig. 6, SER values decrease with increasing the size of the access transistor (Fig. 6(a)) or increasing R_{ratio} (Fig. 6(b)), but the decrease rate in Fig. 6(b) is slightly faster, which presents that the width of the transistor has little effect on SER. For example, with the configuration (N = 64 bits, R_{ratio} = 20), SER only changes from 2.12e-01 (Width = 1W) to 1.66e-01 (Width = 6W). The effects of temperature on the SER are also conducted, which are revealed in Fig. 7. Ignoring the effect of word length in SER, it is clear that ambient temperature causes negligible changes of SER, which decrease from 1.7e-1 to 1.53e-1 when the temperature increases from 25 degrees to 85 degrees respectively.



(a)



(b)

Figure 6. SER at different word lengths vs. width of the access transistor with fixed R_{ratio} = 20 (a), and R_{ratio} with fixed Width = 4W (b).

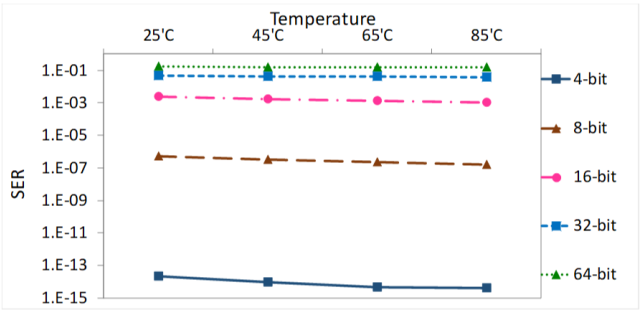


Figure 7. SER of the proposed design with different temperatures with fixed configuration (R_{ratio} = 20, Width = 4W) (b).

IV. PERFORMANCE EVALUATION AND DISCUSSIONS

We conducted the performance evaluations of the proposed NV-CAM design on energy consumption and pattern search latency, which are shown in Fig. 8 and Fig. 9 correspondingly. For the energy consumption measurement, we consider a wide range of match cases including FullMatch (FM), 1-bit MisMatch (1MM), 2-bit MisMatch (2MM), etc., up to the case of Full MisMatch (FMM). At 64-bit word length, the energy consumption of NV-CAM per bit per one access for searching based on the low-R_{ratio} device (R_{ratio} = 2) as STT-MRAM is the highest, gradually decreasing in the case of higher R_{ratio} technologies (PCRAM, RRAM). When R_{ratio} is 100, energy per bit on one search (Esearch/bit) varies from 4.36 fJ to 6.47 fJ depending on the match case.

For the measurement of search latency, we only consider typical word lengths of 16-bit, 32-bit, and 64-bit so that the comparison of search latencies (actually sensing time as illustrated in Fig. 4(b)) with other proposed designs ensures objectivity. In contrast to the decrease in power consumption, the search latency steadily increases for STT-MRAM, PCRAM, and RRAM respectively. The smallest search latencies could be achieved with NV-CAM based on STT-MRAM in the range from 0.11 ns to 0.12 ns depending on the considered word lengths.

The evaluation of the reliability and optimization of a CAM-cell design must be based on a combination of design parameters mentioned in Section I. In addition, the number of proposed designs for comparison should be large enough to ensure objectivity. For those reasons, a series of CAM designs [15]-[22] has been chosen for evaluation. The comparisons in detail over all the design aspects including Cell area, SER, Energy per bit on one search (Esearch/bit), and Sensing latency (Tsensing) are shown in Table I.

Cell area: The smallest cell area belongs to the 2T-2R cell proposed in [15], which was fabricated in the 90 nm technology process. Most of the compact CAM-

TABLE I
COMPARISON WITH PRIOR WORKS

	ISCAS'18 [22]	**TCASI'18** [21]	**TCASII'16** [20]	**JJAP'12** [19]	**JSSC'16** [18]	**E-Let'17** [17]	**ISSCC'16** [16]	**VLSI'13** [15]	**This work**
Memory type	CMOS NOR-type	MRAM	MRAM	MRAM	RRAM	RRAM/MRAM	RRAM	PCRAM	**RRAM/PCRAM/MRAM**
Technology process [nm]	65 nm	**40 nm**	45 nm	90 nm	180 nm	180 nm	65 nm	90 nm	65 nm
Cell structure	16T-SRAM	15T-4R	10T-4R	9T-2R	4T-2R	3T-2R	2.5T-1R	**2T-2R**	**2T-2R**
Cell area [μm^2]	11.43	10.76	2.78	6.84	9.7	8.6	0.59	**0.41**	0.51
Cell area [F^2]	3175	6725	1372.8	844.4	1197.5	265.4	163.8	**50.6**	142.4
Search Error Rate (SER)	**0%**	2.7%	18.5%	30%	N/A	N/A	N/A	N/A	0.03 - 4.61%[1]
Esearch/bit [fJ]	2.07	**0.17**	40.5	0.73	1.3 – 4.36[2]	0.96 – 3.67[2]	0.28 – 0.71[2]	0.75	4.36 - 6.47[2]
Tsensing [ns]	0.52	0.17	1.0	0.22	1.2	0.68	1.0	1.9	**0.11 - 0.12**[3]

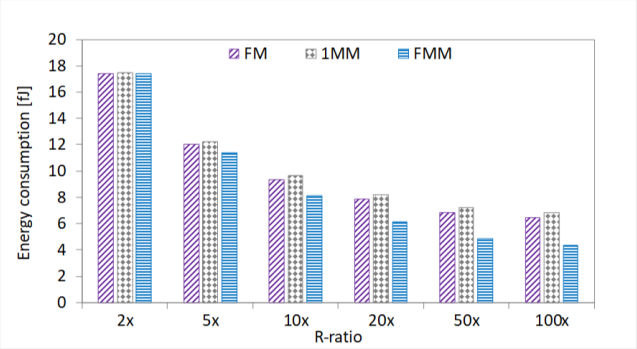


Figure 8. Energy consumption per bit in the proposed NV-CAM matching circuit on one search with fixed Width = 4W at 64-bit word length.

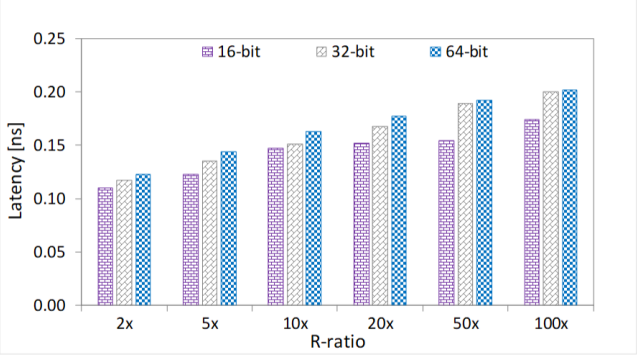


Figure 9. Sensing latency versus different R_{ratio} and word lengths.

cell structures (2.5T-1R, 2T-2R) come with the advantage of cell area. For technology independence, the cell area in minimum feature size (F^2) should be considered for the whole proposed CAM cells.

Search Error Rate: The traditional S-CAM design with 16T-SRAM based on CMOS NOR-type [22], has absolute reliability with an SER value of 0%, which outperforms all the rest.

Search energy: Usually, the search energy in CAM matching circuits is evaluated per bit on one search and this parameter is dependent on the presence of leakage current, ML discharging scheme, the applied bitline voltages, etc. The best search energy consumption is surprisingly belonging to a seemingly very bulky and power-hungry NV-CAM structure (15T-4MJT) [21] with the energy consumption per bit on one search only about 0.17 fJ.

Sensing latency: Our proposed 2T-2T cell structure has shown superiority in sensing latency compared to the remaining designs on this criterion. Moreover, this design achieved a much lower sensing latency compared to the second 2T-2R cell proposed. Besides techniques in circuit design, the distinction in the sensing latency of this work comes from the following applied measures: *First*, we used the current latch sense amplifier, which should have advantages in latency [29]. *Second*, we accept a small compromise in search energy caused by a high level of bitline voltage in return for lower sensing latency.

We summarized the highlights of all evaluated designs as follows: design [22] is the best in the classification of reliability, achieving absolute zero of SER as announced by the authors; design [21] is the best in the classification of energy-efficiency, presenting excellent results in terms of energy consumption; design [15] is the best in the classification of area occupancy according to both standard size (μm^2) and minimum feature size (F^2). Meanwhile, compared to the design [15], our proposed 2T-2R NV-CAM cell is about 2.8x larger in the normalized feature size but offers an extremely low sensing latency, approximately 15.8x-17.3x smaller. Among three CAM-cell structures based on the 65 nm technology process, design [22] despite getting good SER parameter but has a huge disadvantage related to the cell area, in contrary to design [16] and this work (take 3-nd and 2-nd place on this criterion). Compared to this work, design [16] requires two steps to perform the search operation, causing a larger sensing delay and increasing the complexity of the sensing circuit. However, our design requires higher search energy per bit than [16], which is explained by the large applied bitline voltage to achieve low SER values, while SER are absolutely not mentioned in [16].

[1]achieved with configuration (R_{ratio} = 100, the width of the access transistor = 4W, the word length is 64-bit)

[2]measured in cases of 1-bit MisMatch and Full MisMatch

[3]achieved with configuration (R_{ratio} = 2, the width of the access transistor = 4W, the word length varies from 16-bit to 64-bit)

V. CONCLUSION

Emerging non-volatile memory-based CAM cell structures (NV-CAM) have demonstrated potential ability as an alternative to traditional SRAM-based CAM cell structures integrated into frequent-search seldom-write IMC applications. In this paper, a compact 2T-2R NV-CAM cell structure is proposed with an efficient in-memory matching circuit for realizing reliable low-delay lookup operations. To verify the reliability and applicability of the proposed design, a variety of different NV-CAM designs have been chosen for comparative evaluation, which revealed that our 2T-2R CAM-cell structure outperforms the rest in search latency with an extremely small value in the range of 0.11-0.12 ns when employed STT-MRAM. In addition, our proposed design is one of the most compact bitcell structures with a cell area of only 0.51 μm^2, which can guarantee dependable lookup operations with search error rates in the range of 0.03-4.61%. Compared to the other CAM-cell structures evaluated in this paper, our proposed design is the best in search latency and at the same time is one of the most area-efficiency solutions. Moreover, this design can well meet almost the remaining design criteria with a small cost of average search energy. Due to all the presented advantages, our design could be the appropriate candidate for the NV-CAM cell under the 65 nm technology node.

ACKNOWLEDGMENT

This research is funded by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.01-2018.310.

REFERENCES

[1] T. Kohonen, *Content-Addressable Memories*, ser. Springer Series in Information Sciences. Springer Berlin Heidelberg, 2012. [Online]. Available: https://books.google.com.vn/books?id=A-KpCAAAQBAJ

[2] R. Karam, R. Puri, S. Ghosh, and S. Bhunia, "Emerging trends in design and applications of memory-based computing and content-addressable memories," *Proceedings of the IEEE*, vol. 103, no. 8, pp. 1311–1330, 2015.

[3] H.-J. Tsai, K.-H. Yang, Y.-C. Peng, C.-C. Lin, Y.-H. Tsao, M.-F. Chang, and T.-F. Chen, "Energy-efficient tcam search engine design using priority-decision in memory technology," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 3, pp. 962–973, 2017.

[4] T. Venkata Mahendra, S. Wasmir Hussain, S. Mishra, and A. Dandapat, "Energy-efficient precharge-free ternary content addressable memory (tcam) for high search rate applications," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 67, no. 7, pp. 2345–2357, 2020.

[5] T. V. Mahendra, S. W. Hussain, S. Mishra, and A. Dandapat, "Low match-line voltage swing technique for content addressable memory," in *2019 7th International Conference on Smart Computing Communications (ICSCC)*, 2019, pp. 1–5.

[6] S. Jiang, P. Yan, and R. Sridhar, "A high speed and low power content-addressable memory(cam) using pipelined scheme," in *2015 28th IEEE International System-on-Chip Conference (SOCC)*, 2015, pp. 345–349.

[7] N. Mohan, W. Fung, and M. Sachdev, "Low-power priority encoder and multiple match detection circuit for ternary content addressable memory," in *2006 IEEE International SOC Conference*, 2006, pp. 253–256.

[8] Y.-H. Lee, I.-C. Lin, and S.-W. Wang, "Impacts of nbti and pbti effects on ternary cam," in *International Symposium on Quality Electronic Design (ISQED)*, 2013, pp. 38–45.

[9] I.-C. LIN, Y.-H. LEE, and S.-W. WANG, "Reducing aging effects on ternary cam," *IEICE Transactions on Electronics*, vol. E99.C, pp. 878–891, 07 2016.

[10] K. Pagiamtzis and A. Sheikholeslami, "Content-addressable memory (cam) circuits and architectures: a tutorial and survey," *IEEE Journal of Solid-State Circuits*, vol. 41, no. 3, pp. 712–727, 2006.

[11] R. Bez and A. Pirovano, "Non-volatile memory technologies: emerging concepts and new materials," *Materials Science in Semiconductor Processing*, vol. 7, no. 4, pp. 349–355, 2004, papers presented at the E-MRS 2004 Spring Meeting Symposium C: New Materials in Future Silicon Technology. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1369800104001003

[12] H.-S. P. Wong and S. Salahuddin, "Memory leads the way to better computing," *Nature nanotechnology*, vol. 10, no. 3, pp. 191–194, 2015.

[13] K. T. Quang, S. Ruocco, and M. Alioto, "Modeling the impact of dynamic voltage scaling on 1t-1j stt-ram write energy and performance," in *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2015, pp. 2313–2316.

[14] Q. K. Trinh, S. Ruocco, and M. Alioto, "Novel boosted-voltage sensing scheme for variation-resilient stt-mram read," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 63, no. 10, pp. 1652–1660, 2016.

[15] J. Li, R. Montoye, M. Ishii, K. Stawiasz, T. Nishida, K. Maloney, G. Ditlow, S. Lewis, T. Maffitt, R. Jordan, L. Chang, and P. Song, "1mb 0.41 μm^2 2t-2r cell nonvolatile tcam with two-bit encoding and clocked self-referenced sensing," in *2013 Symposium on VLSI Technology*, 2013, pp. C104–C105.

[16] C.-C. Lin, J.-Y. Hung, W.-Z. Lin, C.-P. Lo, Y.-N. Chiang, H.-J. Tsai, G.-H. Yang, Y.-C. King, C. J. Lin, T.-F. Chen, and M.-F. Chang, "7.4 a 256b-wordlength reram-based tcam with 1ns search-time and 14× improvement in wordlength-energyefficiency-density product using 2.5t1r cell," in *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, 2016, pp. 136–137.

[17] C. Kim and K.-W. Kwon, "3t-2r non-volatile tcam with voltage limiter and self-controlled bias circuit," *Electronics Letters*, vol. 53, no. 13, pp. 837–839, 2017. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/el.2017.1027

[18] M.-F. Chang, L.-Y. Huang, W.-Z. Lin, Y.-N. Chiang, C.-C. Kuo, C.-H. Chuang, K.-H. Yang, H.-J. Tsai, T.-F. Chen, and S.-S. Sheu, "A reram-based 4t2r nonvolatile tcam using rc-filtered stress-decoupled scheme for frequent-off instant-on search engines used in iot and big-data processing," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 11, pp. 2786–2798, 2016.

[19] S. Matsunaga, A. Katsumata, M. Natsui, T. Endoh, H. Ohno, and T. Hanyu, "Design of a nine-transistor/two-magnetic-tunnel-junction-cell-based low-energy nonvolatile ternary content-addressable memory," *Japanese Journal of Applied*

Physics, vol. 51, no. 2, p. 02BM06, feb 2012. [Online]. Available: https://doi.org/10.1143/jjap.51.02bm06

[20] B. Song, T. Na, J. P. Kim, S. H. Kang, and S.-O. Jung, "A 10t-4mtj nonvolatile ternary cam cell for reliable search operation and a compact area," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 64, no. 6, pp. 700–704, 2017.

[21] C. Wang, D. Zhang, L. Zeng, E. Deng, J. Chen, and W. Zhao, "A novel mtj-based non-volatile ternary content-addressable memory for high-speed, low-power, and high-reliable search operation," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 66, no. 4, pp. 1454–1464, 2019.

[22] W. Choi, K. Lee, and J. Park, "Low cost ternary content addressable memory using adaptive matchline discharging scheme," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2018, pp. 1–4.

[23] T.-N. Pham, Q.-K. Trinh, I.-J. Chang, and M. Alioto, "Stt-bnn: A novel stt-mram in-memory computing macro for binary neural networks," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 12, no. 2, pp. 569–579, 2022.

[24] V.-T. Nguyen, Q.-K. Trinh, R. Zhang, and Y. Nakashima, "Stt-bsnn: An in-memory deep binary spiking neural network based on stt-mram," *IEEE Access*, vol. 9, pp. 151 373–151 385, 2021.

[25] R. E. Thomas, A. J. Rosa, and G. J. Toussaint, *The analysis and design of linear circuits*. John Wiley & Sons, 2016.

[26] C. Augustine, N. N. Mojumder, X. Fong, S. H. Choday, S. P. Park, and K. Roy, "Spin-transfer torque mrams for low power memories: Perspective and prospective," *IEEE Sensors Journal*, vol. 12, no. 4, pp. 756–766, 2012.

[27] B. S. Kim, S. G. Park, Y. K. You, and S. I. Jung, "Probability & statistics for engineers & scientists," 2011.

[28] M. H. Abu-Rahma, Y. Chen, W. Sy, W. L. Ong, L. Y. Ting, S. S. Yoon, M. Han, and E. Terzioglu, "Characterization of sram sense amplifier input offset for yield prediction in 28nm cmos," in *2011 IEEE Custom Integrated Circuits Conference (CICC)*, 2011, pp. 1–4.

[29] A. K. Mishra, U. Chopra, and V. Dhandapani, "Comparative analysis in terms of power and delay of the different sense amplifier topologies," *Journal of Engg. Research EMSME Special Issue pp*, vol. 47, p. 57, 2021.

Quang-Manh Duong received BE Degree (2006) from the Moscow Aviation Institute in Moscow, Russia, and MS Degree (2012) from the Le Quy Don Technical University in electronic engineering in Hanoi, Vietnam. He was a lecturer and researcher in the Department of Microelectronics and Microprocessor Engineering, Faculty of Radio Electronics Engineering, Le Quy Don Technical University. His research interest includes low-power integrated circuit design, emerging memory technologies, and big data processing. Currently, he is working as a doctoral student at Le Quy Don Technical University.
Email: manhdq@lqdtu.edu.vn.

Quang-Kien Trinh (Member, IEEE) received a Ph.D. degree in computer engineering from the National University of Singapore, Singapore in 2018. He is working as the Deputy Head of The Department of Microprocessor Engineering, Le Quy Don Technical University, Hanoi, Vietnam. His research interests include low-power integrated circuit design, emerging memory technologies, and hardware security. He is the author and co-author of more than 40 publications (mostly IEEE) and 02 patents, and was TPC of many international conferences such as McSoC, ATC, NICS, INISCOM.
Email: kien.trinh@lqdtu.edu.vn.

Hai-Duong Nguyen received Ph.D. (2007) from The Bauman Moscow State Technical University in Russia. He specializes in teaching computer engineering since 1995 at Le Quy Don Technical University, Viet Nam. Currently, he is working as a senior researcher and the Head of the Department of Microelectronics and Microprocessor Engineering, Faculty of Radio Electronics Engineering, Le Quy Don Technical University. His research interest includes computing architecture, embedded systems, VLSI architecture for digital signal processing, and hardware security.
Email: mta.haiduongnguyen@gmail.com.

Van-Phuc Hoang received a Ph.D. degree from The University of Electro-Communications, Tokyo, Japan in 2012. He is working as an Associate Professor, and Director with the Institute of System Integration, Le Quy Don Technical University, Hanoi, Vietnam. His research interests include hardware security, digital circuits and systems, embedded systems for the Internet of Things, and VLSI architecture for digital signal processing. He was the Technical Program Chair of several IEEE international conferences such as ICDV 2017, MCSoC 2018, SigTelCom 2019, APCCAS 2020, ATC 2020, and ICICDT 2022. He is a member of IEEE.
Email: phuchv@lqdtu.edu.vn.

Hoang-Gia Vu received BS (2007) and MS (2010) Degrees in Electronic Engineering respectively from Le Quy Don Technical University, Vietnam; he received Ph.D. (2018) in Computer Engineering from Nara Institute of Science and Technology, Japan. Currently, he is a lecturer and researcher in the Department of Microelectronics and Microprocessor Engineering, Faculty of Radio Electronics Engineering, Le Quy Don Technical University. His research interest includes reconfigurable computing, computing architecture, and embedded systems.
Email: giavh@lqdtu.edu.vn.

Dinh-Ha Dao received BS (2011), MS (2015), and Ph.D. (2018) Degrees in Micro and Nanoelectronics respectively from the Faculty of Radioengineering and Electronics (Belarusian State University of Informatics and Radioelectronics, BSUIR) in Minsk, Belarus. Currently, he is a researcher in the Department of Microelectronics and Microprocessor Engineering, Faculty of Radio Electronics Engineering, Le Quy Don Technical University. He has authored or co-authored more than 30 publications. His research interest includes Sensor, Microelectronics and Semiconductor Engineering, Semiconductor Fabrication, and GaN Heterostructures.
Email: daodinhha@lqdtu.edu.vn.

Van-Toan Tran received his bachelor's and master's degrees in cybernetics and automation engineering from Le Quy Don Technical University, Hanoi, Vietnam in 2004 and 2014, respectively. He is currently pursuing a Ph.D. degree in electronic engineering with Le Quy Don Technical University. His research interest includes integrated circuit design and hardware security.
Email: trantoantbb@gmail.com.

Duy-Manh Luong received the B.S (2005) and M.S (2007) degrees from Vietnam National University, and the D.E (2016) degree in electronics engineering from the University of Electro-Communications, Japan. He worked as a postdoctoral researcher at Osaka University, Japan. He is currently deputy head of the department at Le Quy Don Technical University, Hanoi, Vietnam. His research interests include microwave semiconductor devices and circuits and THz integrated systems for wireless communication applications based on resonant tunneling diodes and photonic crystals.
Email: manhld@lqdtu.edu.vn.