SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

April 2023, Vol. 66 142402:1–142402:11
https://doi.org/10.1007/s11432-021-3502-4

An ultra-high-density and energy-efficient content addressable memory design based on 3D-NAND flash

Haozhang YANG, Peng HUANG*, Runze HAN, Xiaoyan LIU & Jinfeng KANG*

School of Integrated Circuits, Peking University, Beijing 100871, China

Received 25 November 2021/Revised 12 February 2022/Accepted 27 May 2022/Published online 9 March 2023

Abstract In this article, the design of a novel 3D-NAND-based content addressable memory (CAM) consisting of two flash transistors with ultra-high density and low power consumption is proposed for data-intensive computing. Hewlett simulation program with integrated circuit emphasis (HSPICE) of a 3D-NAND-based CAM array is performed to study the functionality and properties of the presented CAM design, and the results indicate that the energy consumption is 0.196 fJ/bit/search. The cell density of a 16-layer 3D-NAND flash is 157 times higher than CAMs based on conventional static random access memory. Furthermore, to exploit the multibit storage property of 3D-NAND flash, we also propose a multilevel CAM design, which significantly boosts the cell density and expands the functionality. As a proof-of-concept illustration, we take a 4-level CAM to successfully implement the search operation. Furthermore, the impacts of 3D-NAND layers and parasitic effects on the performance of the proposed CAM design are also discussed.

Keywords content addressable memory (CAM), 3D-NAND flash, data-intensive computing, in-memory computing, multilevel CAM

Citation Yang H Z, Huang P, Han R Z, et al. An ultra-high-density and energy-efficient content addressable memory design based on 3D-NAND flash. Sci China Inf Sci, 2023, 66(4): 142402, https://doi.org/10.1007/s11432-021-3502-4

1 Introduction

Data-intensive computing applications such as pattern recognition, video processing, database engine and network router have drastically increased due to the rapid development of big data and artificial intelligence (AI), which pose stringent requirements for storing and processing massive data resources [1–5]. Content addressable memory (CAM) has different working schemes from random access memory (RAM) and powerful functionality of high parallelism and fast speed, making it an excellent solution for data-intensive computing systems [6, 7]. A simplified structure of the general CAM system is shown in Figure 1. The CAM compares input data entered from search lines (SLs) with stored contents and outputs the match address through match lines (MLs) [8]. However, the existing conventional CMOS-based CAM designs prevent further development in data-intensive computing systems because of their excessively large circuit area and nontrivial standby power consumption as the scaling down of technology [9, 10].

To address the aforementioned challenges, considerable research has been conducted, which exploits the emerging nonvolatile memories (NVMs) to constitute the CAM cell for storage and search operations, conducing to higher storage capacity, zero standby power and lower search energy consumption [11]. Thus far, various types of NVMs have been proposed to implement large-scale integration CAMs, including phase change memory (PCM) [12], resistive RAM (RRAM) [13–16], ferroelectric field effect transistor (Fe-FET) [17–19] and magnetic tunnel junction (MTJ) [20–22]. However, most emerging NVMs mass production techniques are still not mature enough for further large-scale integrations. There have also been some research efforts that employ flash devices to realize CAM structures. A 1 MB binary CAM (BCAM) based on flash memory technologies was first demonstrated for area reduction and nonvolatility [23]. An analog CAM was implemented using 5b analog flash based on an integrator circuit [24]. Moreover,

* Corresponding author (email: phwang@pku.edu.cn, kangjf@pku.edu.cn)

© Science China Press 2023

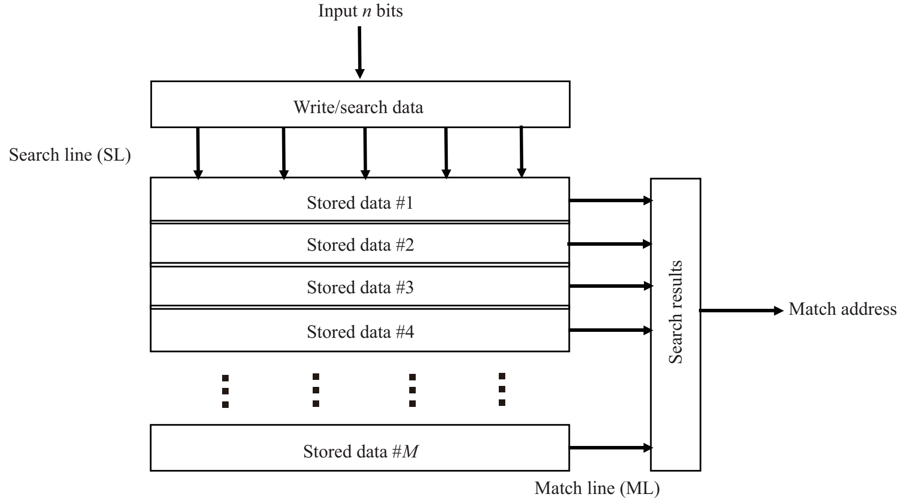


Figure 1 Structure of a general CAM system.

the first flash-based ternary CAM (TCAM) was demonstrated, utilizing two flash transistors and thereby reducing the circuit area [25]. However, all the above approaches utilize the NOR architecture and planar technology, which immensely limit the scalability of the CAM array. Recently, a sort of search scheme based on NAND flash has also been proposed [26]. Currently, 3D-NAND flash rapidly develops and has a vertical stacking architecture and industrial mass production technology, that showcases an exceptional advantage of ultra-high cell density. Therefore, it possesses great potential for implementing large-scale integration CAMs and may be a highly competitive candidate for data-intensive computing systems.

In this work, a novel 3D-NAND-based CAM design with ultra-high density and low power is proposed for data-intensive computing, with one CAM cell constitutive of two flash transistors. The rest of this paper is organized as follows. In Section 2, the proposed CAM design based on 3D-NAND flash is introduced in detail. In Section 3, we perform Hewlett simulation program with integrated circuit emphasis (HSPICE) simulations of a 3D-NAND-based CAM array to analyze the functionality and properties of the proposed design after building an experimentally calibrated NAND transistor model. In Section 4, considering 3D-NAND flash has a good and reliable feature of multibit storage, we further propose a multilevel CAM design based on 3D-NAND, which significantly boosts the density of CAM cells and expands the functionality of the CAM systems oriented to data-intensive computing. As a proof-of-concept illustration, we successfully implement the search operation using a 4-level CAM. Furthermore, the impacts of 3D-NAND layers and parasitic effects on the performance of the proposed CAM design are also discussed based on HSPICE simulations. Finally, Section 5 concludes this paper.

2 Proposed CAM design

The proposed CAM design employs two adjacent NAND transistors in the word line (WL) direction to constitute one CAM cell as shown in Figure 2(a). The data stored in the CAM cell are determined by the threshold voltage ($V_{\rm th}$) of two transistors together. Most of the existing CAM systems can be divided into two categories: BCAM and TCAM, depending on the number of logic states pre-defined. Compared with BCAM which stores “0” and “1” in one storage element, TCAM can store an additional state, “X” in TCAM cell, which means we do not care its specific value and matches any input. The logic of storage and search operation in our proposed CAM design is workable for both of the two categories and as a proof-of-concept demonstration, we take the TCAM logic to illustrate.

First, we will illustrate the operating principle of the NAND flash transistors. The $V_{\rm th}$ is determined by the number of electrons in the floating gate or charge trapping layer. By applying different bias conditions, we can implement erasing and programming operations to regulate the $V_{\rm th}$ of NAND flash, as shown in Figure 2(b) [27]. Erasing operation is performed in the block unit, which corresponds to a row of NAND strings. All the WLs are connected to the ground and all the drain select lines (DSLs), bit lines (BLs), and source select line (SSL) are floating. A high voltage pulse (20 V) is applied to the bulk. In this condition, the electrons are swept out from the floating gate to the bulk by F-N tunneling, resulting

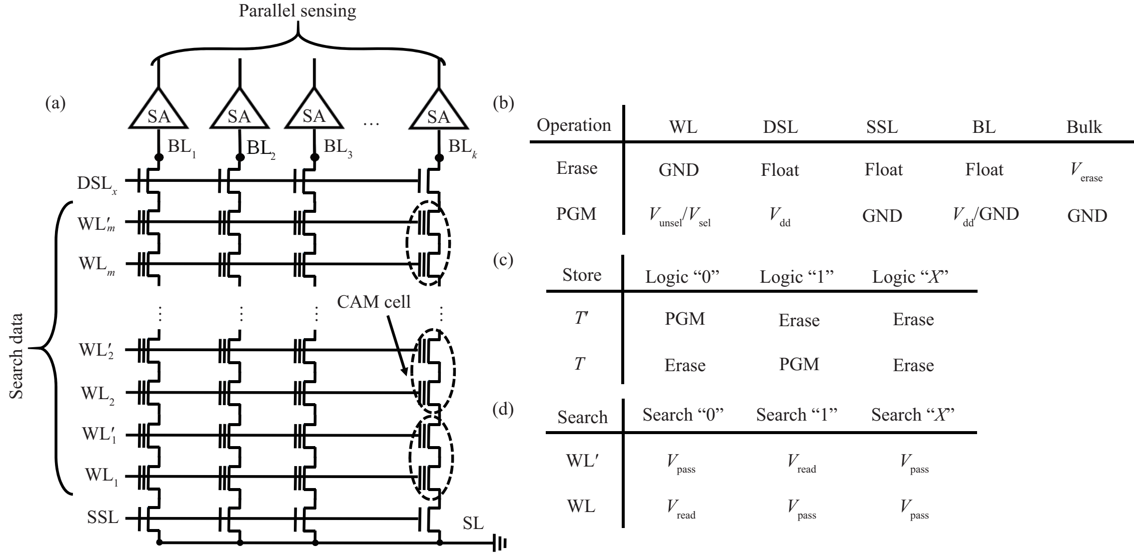


Figure 2 (a) Schematic of the implementation of the CAM cell based on 3D-NAND flash; (b) bias conditions of the erasing and programming operation; (c) definition of data stored by the NAND transistor according to TCAM logic; (d) search scheme with "0", "1", and "X" in TCAM logic, respectively.

in a decrease of the V_{th}. In the programming operation, the basic unit is in the page, which consists of a row of NAND transistors and corresponds to a certain WL. The corresponding DSL is connected to the supply voltage (V_{dd}), and the SSL and bulk are connected to the ground. The BLs that need programming are connected to the ground and BLs which inhibit programming are connected to V_{dd}. The selected WL is connected to a programming voltage (18 V) and the unselected WLs are connected to a pass voltage (10 V). In this condition, there is a large potential difference between the control gate and channel, resulting in the F-N tunneling of electrons to the floating gate and therefore increasing the V_{th} for the programming cells. For those programming inhibited cells, the channel potential is self-boosted to a high voltage (8 V) which prevents the F-N tunneling.

We mark the upper transistor in a CAM cell as T′ and the lower one as T. The cell stores "0" ("1") when T is erased (programmed) and T′ is programmed (erased). The wildcard "X" is represented by both of the transistors erased. The combination of V_{th} is summarized in Figure 2(b). All of the data should be loaded to the CAM array in advance according to the above rules. After that, search data will be transformed into voltage pulses and applied on WLs. For example, to search logic "0", a read voltage V_{read} is applied on the lower WL and a pass voltage V_{pass} is applied on the upper WL. The search scheme is concluded in Figure 2(c), which is different from the read operation in the normal storage mode of 3D-NAND flash. Different cells of one CAM word are serially connected between BLs and ground through drain selectors and source selectors, constituting a NAND string. Before each search operation, all of the BLs will be pre-charged to V_{dd} via the pre-charge transistors. The search results are read out from BL voltages in parallel by a sense amplifier (SA). To illustrate the match principle of the proposed CAM cell, we assume that it stores logic "0", which means T is of low V_{th} and T′ is of high V_{th}. If the input data is "1", with a high voltage applied on WL, T is in ON state and, on the contrary, T′ is still in OFF state under V_{read}, which leads to a cut-off path. If the input data is "0" or "X", both of the transistors will be in ON state and form a discharge path. We can see the NAND string will be in ON state and discharge BL quickly only when all of the cells are matched. So long as there is one bit mismatched, the path will be cut off and the BL will discharge slowly. In a word, we can distinguish the match and mismatch conditions by discharge rate of BLs.

Figure 3 shows the structure of the proposed CAM design based on the vertical channel 3D-NAND array. To search a word of m-bit, we need a 3D-NAND block of $2 \times m$ layers. Besides, assuming the page size is k and the number of strings is n, there will be $k \times n$ words stored in the block altogether. According to the general scale of 3D-NAND array, the arrangement of our proposed CAM design can significantly increase search parallelism because the page size is of several KBs [28]. During search operation, all of the BLs will be pre-charged firstly and after that, search data will be converted into voltage pulses and applied on WLs. If the voltage level of some BL is low, each bit of the stored word in this corresponding

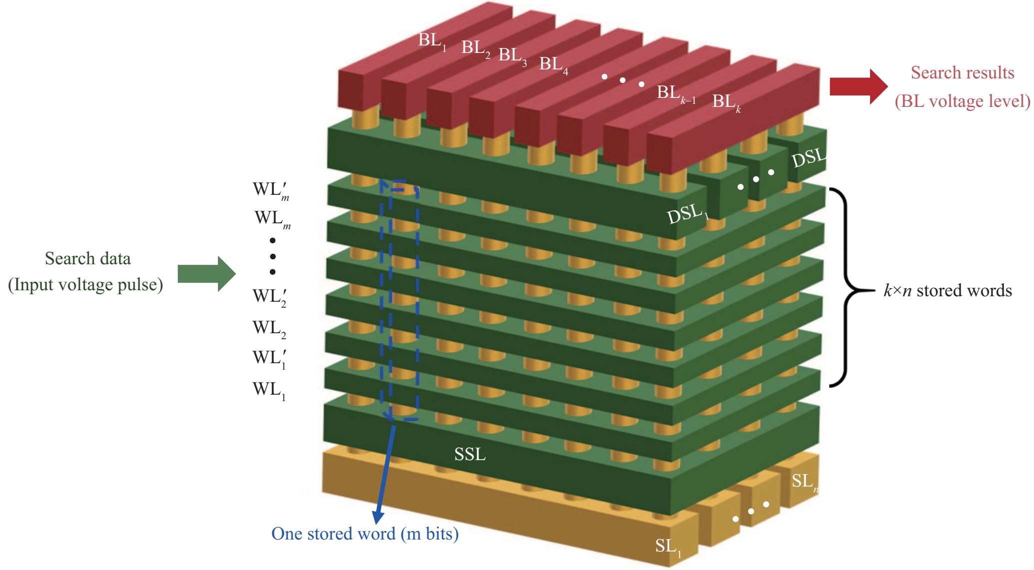


Figure 3 Structure of the proposed CAM design based on a 3D-NAND flash array with $k \times n$ stored words of m bits.

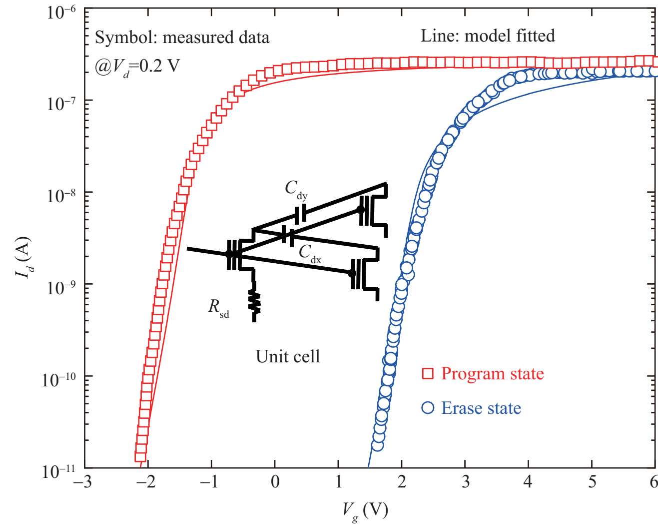


Figure 4 I_d-V_g curves based on the BSIM model calibrated with the experimental data in [30]. Inset: the unit cell used in HSPICE simulations of 3D-NAND-based CAM arrays. The parasitic effects are considered.

NAND string is totally matched with the input data. If the voltage level is high, there must be at least one bit mismatched of stored contents with the input data.

3 Functionality and property analysis

3.1 Simulation method

As a proof-of-concept demonstration of the proposed CAM design, we build a compact model of NAND transistor based on Berkeley short-channel IGFET model (BSIM) 3v3 calibrated with experimental data firstly, of which the device size is $W = 70$ nm, $L = 70$ nm [29, 30]. Figure 4 shows the I_d-V_g curves with good fitting. Besides, we also consider parasitic effects within the 3D-NAND array, including channel series resistance ($R_{\rm sd} = 4\ \Omega$), parasitic resistance in BLs and SLs ($R_{\rm BL} = 0.5\ \Omega$, $R_{\rm SL} = 10\ \Omega$) and parasitic capacitor between transistors in x direction and y direction ($C_{\rm dx} = 0.03$ fF, $C_{\rm dy} = 0.02$ fF), which are calculated according to [30, 31].

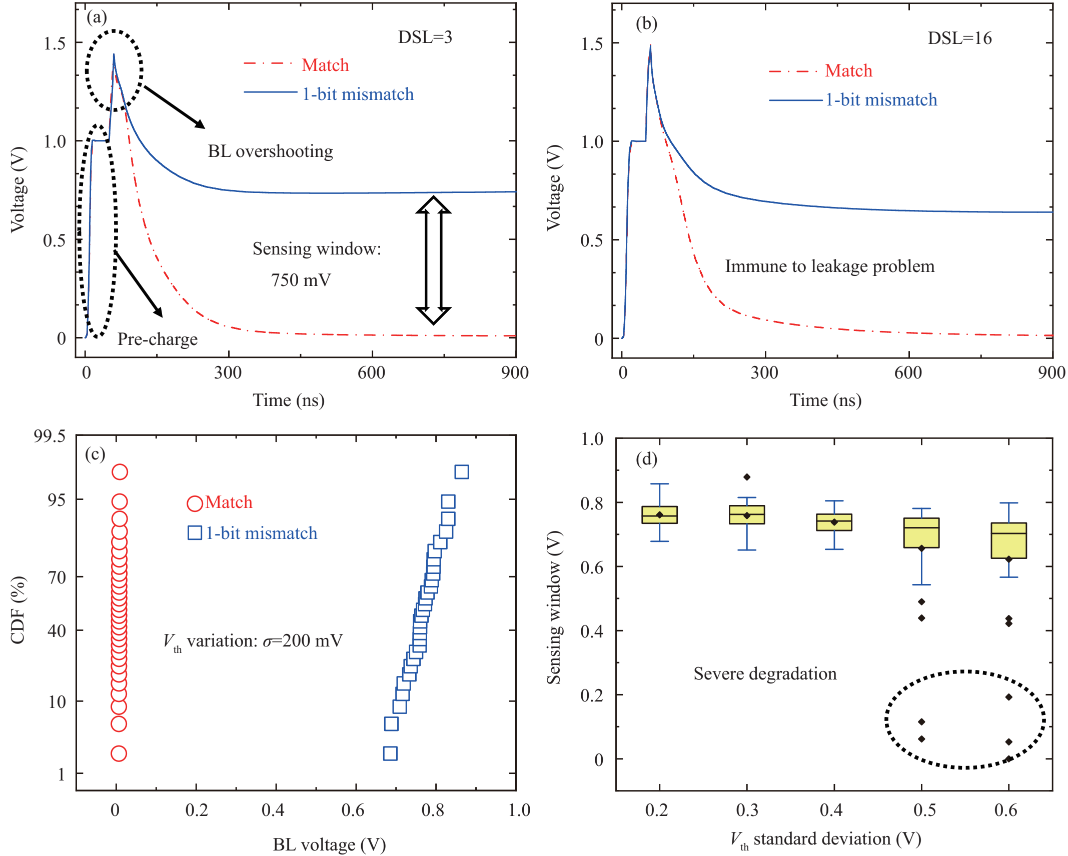


Figure 5 (Color online) (a) The simulated search results of 16-bit words for full match and only 1-bit mismatch conditions ($V_{\rm read} = 0$ V, $V_{\rm pass} = 5$ V, and $V_{\rm dd} = 1$ V); (b) comparison worst case with the number of DSL is 16; (c) cumulative distributions of BL voltages with $V_{\rm th}$ of 200 mV standard deviation for fully match and only 1-bit mismatch conditions, respectively; (d) the sensing window with different $V_{\rm th}$ standard deviations. The stack layer number is set as 32 for 16-bit words, and BL number as 2 for the two cases.

3.2 Functionality verification

Based on the compact NAND transistor model, we perform HSPICE simulations of the array level to verify the functionality of the proposed CAM design. We set $V_{\rm read} = 0$ V, $V_{\rm pass} = 5$ V, and $V_{\rm dd} = 1$ V. Figure 5(a) shows the corresponding BL voltage waveform of two 16-bit words. We can see that the BL voltage is pre-charged to $V_{\rm dd}$ firstly and after the input data is entered, the voltage level will tend to overshoot briefly and then descend to a relatively stable level. Because 1-bit mismatch is the most difficult case to distinguish from fully match conditions in practice, we are mainly focused on the sensing window of these two conditions. The margin is around 750 mV, which is big enough for discriminative sense and therefore proves the functionality of our proposed CAM design. For the BL voltage overshooting, it is induced by the coupling effect of capacitance between the gate and drain ($C_{\rm gd}$) of NAND transistors. When the search data are input to WLs, abrupt rising step biases are applied to the stacked gates simultaneously. Besides, due to the high density of grain boundary traps, the poly-Si channel of 3D-NAND flash has smaller mobility and poorer electrical property, thus a smaller drain current [32]. Therefore, the BL voltage will be pulled up in the initial stage of switching transient.

Considering the leakage problem, in the CAM design of NOR-type string, the total cell number connected to ML is equivalent to the length of the search word, which may lead to a very severe leakage problem for long search words (such as 144-bit or longer). While for the proposed CAM design (NAND-type string), there is only one cell connected to the ML in one search word. In the 3D-NAND-based CAM design, although there are multiple cells connected to one BL, which is known as DSL, the number of DSL is only 12 or 16 in the current mainstream 3D-NAND architecture, which has little influence on the leakage problem of BL voltage [30,33,34]. To verify this feature, we further simulate the fully match case and 1-bit mismatch case with 16 DSLs to evaluate the leakage current problem, as shown in Figure 5(b). It is found that we can still easily distinguish the worst case with 16 DSLs, where the slightly decreased

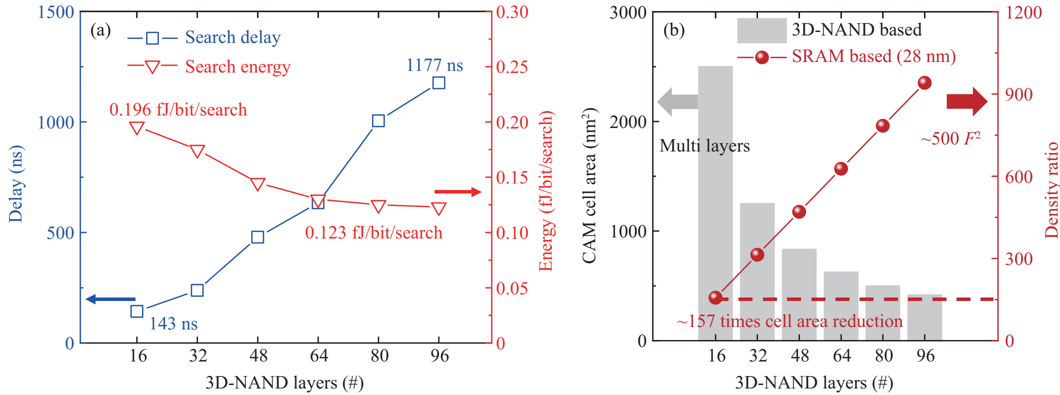


Figure 6 (Color online) (a) The impacts of 3D-NAND layers on search delay and energy consumption of the proposed CAM design; (b) CAM cell area of 3D-NAND-based with different layers and cell density ratio with conventional SRAM-based CAMs in the case of planar 28 nm. The BL number is set as 2 and the DSL number as 3.

sensing window is due to the parasitic effect.

Considering the device-to-device $V_{\rm th}$ variation within the 3D-NAND array may have a bad effect on the sensing window, further simulations are carried out to discuss this influence. At present, incremental step pulse programming (ISPP) is widely used in NAND flash, which replaces the constant voltage programming method to improve the tolerance to process and environmental variations [27]. The feature of ISPP is that the subsequent program pulse connected to the selected WL is incremented in a constant voltage step, which causes an automatic adjustment to different cells. After the ISPP programming operation, the $V_{\rm th}$ distribution of the 3D-NAND array nearly follows the Gaussian distribution [35]. Figure 5(c) shows the cumulative distributions of BL voltages for full match and 1-bit mismatch conditions with $V_{\rm th}$ of 200 mV standard deviation. We can see that the sensing window is still more than 700 mV under this perturbation. To explore the tolerance of the device-to-device $V_{\rm th}$ variations and find out the design space requirements, we carry out further simulations about the sensing window with different $V_{\rm th}$ standard deviations, as shown in Figure 5(d). It is found that the sensing window of our proposed CAM design can maintain a large voltage ($>$ 600 mV) in a wide range of device variations ($\leqslant$ 400 mV), which indicates its good robustness. Besides, the sensing window would degrade severely when the $V_{\rm th}$ standard deviation is up to 500 mV. This is because some transistors that were originally in OFF state, are now in ON state and some transistors that were originally in ON state, are now in OFF state, which leads to the incorrect BL discharge rate.

3.3 Property analysis

Search delay, energy consumption and cell area are three key factors of the performance of a CAM system. Since the 3D-NAND flash mainly boosts cell density by increasing layers, we will focus on the effects of 3D-NAND layer number in further simulations. As is shown in Figure 6(a), the search delay increases with the number of layers because of additional resistance and capacitance (RC) on the signal path. While the search energy drops off as the increasing layers, which is likewise due to the higher RC and therefore, lower discharge current. Besides, we can find that there exists a trade-off between search delay and energy consumption with respect to the selection of layers. Especially when we employ a 16-layer 3D-NAND array, its energy consumption (0.196 fJ/bit/search) is lower than the conventional SRAM-based TCAMs (0.58 fJ/bit/search in 32 nm technology [36]). Besides, it also has an invincible advantage on the cell density because of the 3D stacking feature as shown in Figure 6(b). Compared with the SRAM-based TCAM design under planar 28 nm technology (500 F^2), 3D-NAND-based TCAM saves around 157 times cell area only in 16 layers and the number of layers has already reached 128 so far [37]. Although the search delay of a single word in 3D-NAND-based CAM design is slower than that of SRAM (1 GHz), we can improve the overall throughput (the number of words being searched per second) of the whole system by employing multiple 3D-NAND blocks to perform search operation concurrently in one chip. Considering the application scenario of big-data searching, the data which need to be stored in CAM are much larger than the capacity of current SRAM-based CAM [38]. As a result, the stored contents will be constantly updated during the entire search task and therefore, the throughput is more important. In SRAM-based TCAM, the single search time is 1 ns and the search parallelism is 2K words

in one search cycle [36]. Hence the throughput is 2 trillion words per second. For the 3D-NAND-based TCAM, the single search time is 143 ns (16 layers) and the search parallelism depends on the page size of the 3D-NAND flash, which can achieve 16 KB in [28]. Therefore, the total throughput can achieve 0.896 trillion words per second. Due to the high density advantage, we can employ three or more 3D-NAND blocks simultaneously in one chip, which can exceed the SRAM-based CAM.

In comparison with the CAM designs based on emerging devices, such as RRAM, the proposed CAM design is based on the mature 3D-NAND flash technology, which is contained in the existing memory hierarchy, and hence is closer to commercial applications in large scale (terabits). Besides, benefited from the large memory window of the flash transistor and NAND-type string, the proposed CAM design possesses higher parallelism and throughput than that of the emerging devices. However, it should be noted that the proposed CAM design may not be suitable for the applications that need to frequently change the stored contents due to the long time required for programming and erasing flash cells.

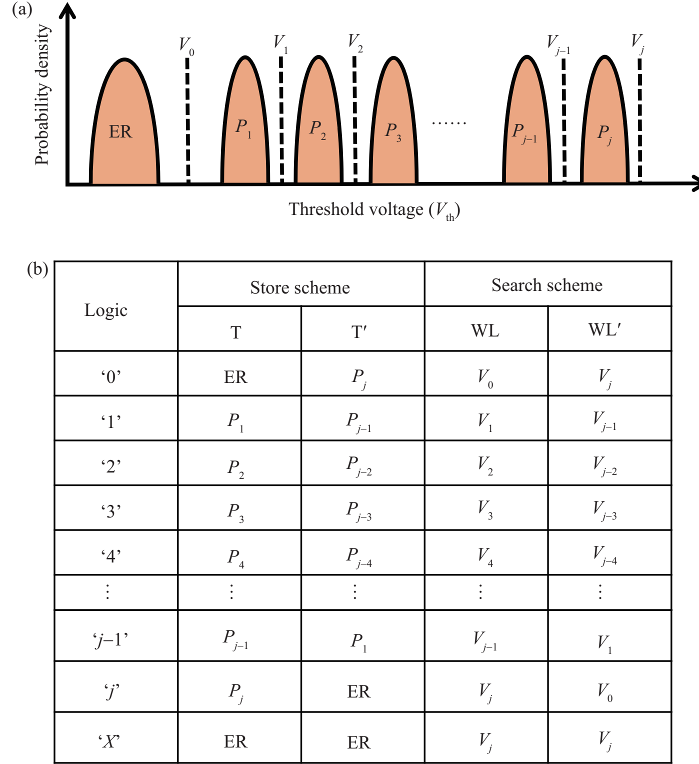
4 Multilevel CAM design based on 3D-NAND

As we know, 3D-NAND flash possesses a good and reliable multibit storage feature. On the basis of this, we further develop a multilevel CAM design based on 3D-NAND flash as shown in Figure 7, which stores several logic states in one CAM cell. This method significantly boosts cell density and expands the functionality as well. To store a logic “0”, the lower transistor, marked as T, is in erasing state and the upper transistor, marked as T', is in the maximum programming state. As the storage logic increases, $V_{\rm th}$ of T moves towards the right, which is in a higher programming state and $V_{\rm th}$ of T' moves towards the left, which is in a lower programming state. As for logic “X”, both of the transistors are in erasing state. To search logic “0”, the lowest read voltage, marked as V_0, is applied on WL and the highest read voltage, marked as V_j, is applied on WL'. As the search logic increases, the voltage applied on WL should increase and the voltage applied on WL' should decline, just as similar as the store scheme. To search “X”, voltages applied on both of the WLs are the highest read voltage level. Under this principle, when the stored contents are completely matched with the search input, two transistors of one CAM cell will be in ON state synchronously. Besides, when storing “X”, the CAM cell will be in ON state for any input data and in turn, when “X” is input, the CAM cell will be in ON state likewise for any stored contents. Because of the NAND architecture, only when all of the CAM cells are matched, the NAND string will be in ON state and discharge BL quickly, thus implementing the multilevel search functionality.

As a proof-of-concept illustration, we adopt the 2-bit (MLC) 3D-NAND flash to implement a 4-level CAM design. Considering the variation of $V_{\rm th}$ in multilevel mode may have a larger effect on the performance of the search operation, we take a more precise $V_{\rm th}$ distribution model including ISPP programming noise, WL-WL interference and random telegraph noise (RTN) effect as shown in Figure 8(a) [39]. The ISPP programming voltage is 8–18 V with incremental $V_{\rm step} = 0.5$ V. The reading voltages of four levels respectively are 0, 1.5, 3.5, 5 V. The BL voltage distributions of HSPICE simulations are shown in Figure 8(b), which represent fully match and only 1-bit mismatch conditions of the 4-level CAM design, respectively. This is the most difficult case for detection in practical applications. The large window verifies the feasibility of our proposed multilevel CAM design. To compare the performance of the proposed multilevel CAM design, we carry out additional simulations in 16-layer 3D-NAND with the above-mentioned variations. The average search energy is 0.073 fJ/bit/search, which is 8 times lower than the SRAM-based TCAM (0.58 fJ/bit/search) [36]. The superiority of cell density is further enhanced due to the multilevel property and achieves 314 times higher than the SRAM-based TCAM, which indicates the multilevel CAM design is highly promising in data-intensive search applications. Due to a smaller $V_{\rm gs} - V_{\rm th}$ of the flash transistors in NAND string and thus a larger RC delay in the discharge path, search delay is degraded to 427 ns.

Besides, to analyze the effects of 3D-NAND layers on the properties of multilevel CAM design, we perform additional simulations. The results are shown in Figure 9, from which we can see that with the aggravation of gate’s coupling effects due to increasing layers, the search delay rises faster and the energy consumption drops little after 32 layers. More optimization is still needed to cut down the aggravation.

For the parasitic effects of 3D-NAND flash mentioned in Subsection 3.1, the $R_{\rm sd}$ comes from the series resistance of polysilicon channel between adjacent WLs. The $C_{\rm dx}$ and $C_{\rm dy}$ come from the insulating layer between adjacent BLs and DSLs. The values are calculated based on 3D geometric size and material properties. To investigate the effects of parasitic RC, we scale all the values in the same ratio, from



Logic	Store scheme		Search scheme	
	T	T′	WL	WL′
‘0’	ER	P_j	V_0	V_j
‘1’	P_1	P_{j-1}	V_1	V_{j-1}
‘2’	P_2	P_{j-2}	V_2	V_{j-2}
‘3’	P_3	P_{j-3}	V_3	V_{j-3}
‘4’	P_4	P_{j-4}	V_4	V_{j-4}
⋮	⋮	⋮	⋮	⋮
‘j–1’	P_{j-1}	P_1	V_{j-1}	V_1
‘j’	P_j	ER	V_j	V_0
‘X’	ER	ER	V_j	V_j

Figure 7 (Color online) (a) The distribution of threshold voltage under the 3D-NAND multibit storage mode; (b) the multilevel CAM design based the on 3D-NAND flash which can further boost the cell density and expand the functionality.

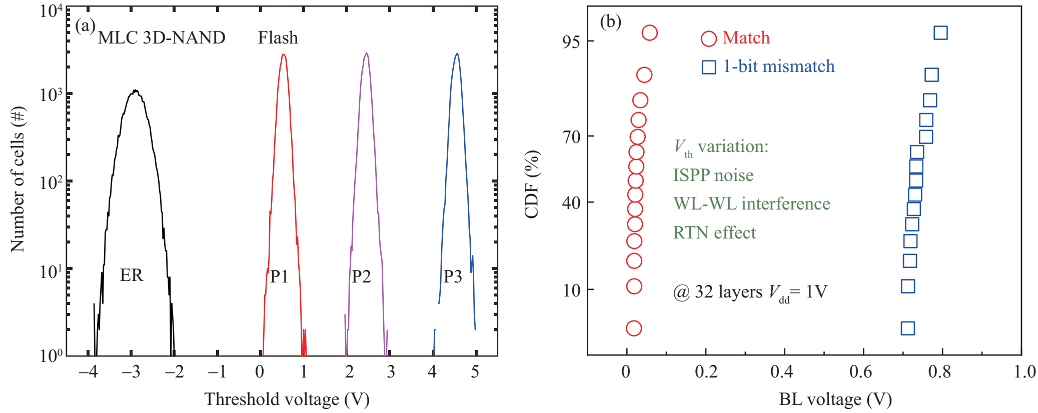


Figure 8 (Color online) (a) Threshold voltage distribution of the MLC mode with the programming noise; (b) cumulative distributions of BL voltages in 4-level CAM design for full match and only 1-bit mismatch conditions, respectively. The BL number is set as 2 and the DSL number as 3.

0.1 to 10 times, as shown in Figure 10. The enlarging parasitic increases the RC delay in the discharge path from BL to ground and thus increases the search delay. Due to the coupling effect of parasitic capacitance, the adjacent BL will influence each other. The BL voltage of the mismatch string will be pulled down a little by the match string when they are close, which reduces the sensing window of the search result. The simulation results indicate a requirement for improving processes and materials with smaller parasitic effects.

Benefited from the advantage of high cell density, the proposed CAM design can not only implement big data search, but also can be very suitable for data-intensive computing, such as deep random forest [40], nearest neighbor search [41], and one/few-shot learning with memory-augmented neural networks [42]. Take one-shot learning as an example to illustrate: the image database is first used to train a convolutional neural network for feature extracting. After that, all the extracted feature vectors of the database are

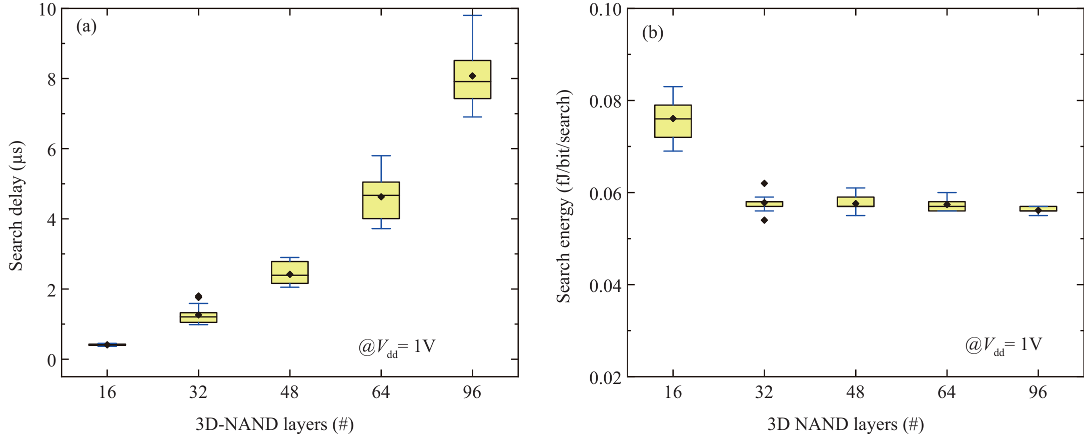


Figure 9 (Color online) The influence of 3D-NAND layers on (a) the search delay and (b) the energy consumption of the multilevel CAM design considering $V_{\rm th}$ variation. The BL number is set as 2 and the DSL number as 3.

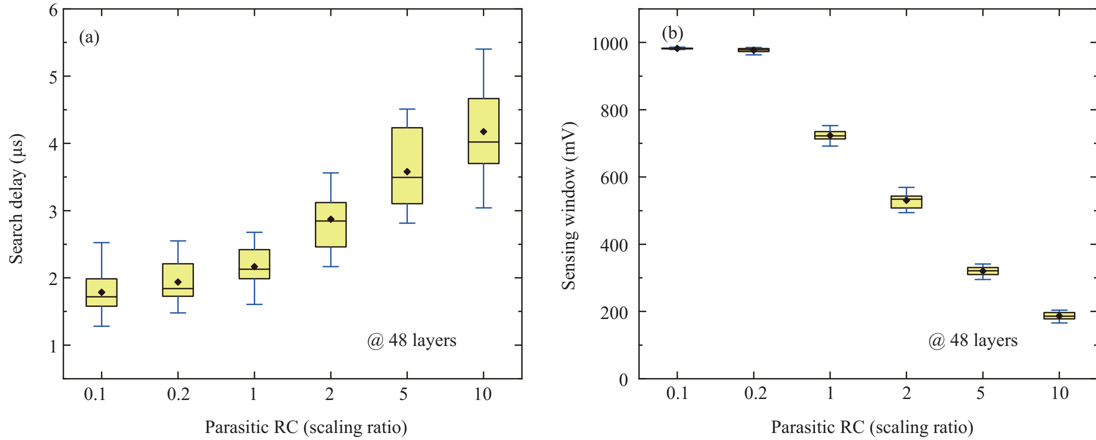


Figure 10 (Color online) The impacts of parasitic resistance and capacitance on (a) the search delay and (b) the sensing window of the multilevel CAM design with the $V_{\rm th}$ variation. The BL number is set as 2 and the DSL number as 16.

stored in CAM for the following search. The input images will be first applied to the network and then forward to the CAM to compare with stored contents for the nearest neighbor search. It should be noted that the number of feature vectors may be very large and exceed the capacity of current SRAM-based CAM. Thanks to the large parallelism, our proposed 3D-NAND-based CAM design can directly store the whole database and complete the search operation in one cycle, which largely reduces the time and energy consumption and indicates the great potential for data-intensive computing applications.

5 Conclusion

A novel CAM design based on 3D-NAND flash has been presented, in which two flash transistors are combined as one CAM cell. Simulation results show that the proposed design is capable of low-power and high-density CAM integrations, in which the energy consumption is 0.196 fJ/bit/search and the cell density of a 16-layer 3D-NAND is 157 times higher than the traditional CMOS-based CAMs. Furthermore, a multilevel CAM design has also been illustrated, which significantly boosts the cell density and expands the functionality. The proposed design will provide a powerful solution for data-intensive computing and is promising for search applications that require low energy dissipation and large volume storage.

Acknowledgements This work was supported in part by National Key Research and Development Program of China (Grant Nos. 2019YFB2205100, 2018YFA0701500), National Natural Science Foundation of China (Grant No. 62034006), and 111 Project Program (Grant No. B18001).

References

1 Thangam M S, Vijayalakshmi M. Data-intensive computation offloading using fog and cloud computing for mobile devices applications. In: Proceedings of International Conference on Smart Systems and Inventive Technology, Tirunelveli, 2018. 547–550
2 Jonathan A, Ryden M, Oh K, et al. Nebula: distributed edge cloud for data intensive computing. IEEE Trans Parallel Distrib Syst, 2017, 28: 3229–3242
3 Zhang X Y, Liu C L, Suen C Y. Towards robust pattern recognition: a review. Proc IEEE, 2020, 108: 894–922
4 Ye Q, Doermann D. Text detection and recognition in imagery: a survey. IEEE Trans Pattern Anal Mach Intell, 2014, 37: 1480–1500
5 Nielen L, Siemon A, Tappertzhofen S, et al. Study of memristive associative capacitive networks for CAM applications. IEEE J Emerg Sel Top Circ Syst, 2015, 5: 153–161
6 Zheng L, Shin S, Steve Kang S M. Memristor-based ternary content addressable memory (mTCAM) for data-intensive computing. Semicond Sci Technol, 2014, 29: 104010
7 Huang P T, Lai S L, Chuang C T, et al. 0.339 fJ/bit/search energy-efficient TCAM macro design in 40 nm LP CMOS. In: Proceedings of Asian Solid-State Circuits Conference, KaoHsiung, 2014. 129–132
8 Pagiamtzis K, Sheikholeslami A. Content-addressable memory (CAM) circuits and architectures: a tutorial and survey. IEEE J Solid-State Circ, 2006, 41: 712–727
9 El Baraji M, Javerliac V, Prenat G. Towards an ultra-low power, high density and non-volatile ternary cam. In: Proceedings of Non-Volatile Memory Technology Symposium, Pacific Grove, 2008. 1–7
10 Imani M, Peroni D, Rahimi A, et al. Resistive CAM acceleration for tunable approximate computing. IEEE Trans Emerg Top Comput, 2019, 7: 271–280
11 Karam R, Puri R, Ghosh S, et al. Emerging trends in design and applications of memory-based computing and content-addressable memories. Proc IEEE, 2015, 103: 1311–1330
12 Li J, Montoye R, Ishii M, et al. 1 Mb 0.41 μm^2 2T-2R cell nonvolatile TCAM with two-bit encoding and clocked self-referenced sensing. In: Proceedings of Symposium on VLSI Technology, Kyoto, 2013. 104–105
13 Han R, Shen W, Huang P, et al. A novel ternary content addressable memory design based on resistive random access memory with high intensity and low search energy. Jpn J Appl Phys, 2018, 57: 04FE02
14 Chen B, Zhang Y, Liu W, et al. Ge-based asymmetric RRAM enable $8F^2$ content addressable memory. IEEE Electron Device Lett, 2018, 39: 1294–1297
15 Zheng L, Shin S, Lloyd S, et al. RRAM-based TCAMs for pattern search. In: Proceedings of International Symposium on Circuits and Systems, Montreal, 2016. 1382–1385
16 Grossi A, Vianello E, Zambelli C, et al. Experimental investigation of 4-kb RRAM arrays programming conditions suitable for TCAM. IEEE Trans VLSI Syst, 2018, 26: 2599–2607
17 Yin X, Li C, Huang Q, et al. FeCAM: a universal compact digital and analog content addressable memory using ferroelectric. IEEE Trans Electron Dev, 2020, 67: 2785–2792
18 Tan A J, Chatterjee K, Zhou J, et al. Experimental demonstration of a ferroelectric HfO_2-based content addressable memory cell. IEEE Electron Device Lett, 2019, 41: 240–243
19 Ni K, Yin X, Laguna A F, et al. Ferroelectric ternary content-addressable memory for one-shot learning. Nat Electron, 2019, 2: 521–529
20 Wang C, Zhang D, Zeng L, et al. Design of magnetic non-volatile TCAM with priority-decision in memory technology for high speed, low power, and high reliability. IEEE Trans Circ Syst I, 2020, 67: 464–474
21 Gupta M K, Hasan M. Robust high speed ternary magnetic content addressable memory. IEEE Trans Electron Dev, 2015, 62: 1163–1169
22 Xu W, Zhang T, Chen Y. Design of spin-torque transfer magnetoresistive RAM and CAM/TCAM with high sensing and search speed. IEEE Trans VLSI Syst, 2009, 18: 66–74
23 Miwa T, Yamada H, Hirota Y, et al. A 1 Mb 5-transistor/bit non-volatile CAM based on flash-memory technologies. In: Proceedings of International Solid-State Circuits Conference, San Francisco, 1996. 40–41
24 Kramer A, Canegallo R, Chinosi M, et al. 55 GCPS CAM using 5b analog flash. In: Proceedings of International Solid-State Circuits Conference, San Francisco, 1997. 44–45
25 Fedorov V V, Abusultan M, Khatri S P. An area-efficient ternary CAM design using floating gate transistors. In: Proceedings of International Conference on Computer Design, Seoul, 2014. 55–60
26 Wang F, Feng Y, Zhan X, et al. Implementation of data search in multi-level NAND flash memory by complementary storage scheme. IEEE Electron Dev Lett, 2020, 41: 1189–1192
27 Suh K D, Suh B H, Lim Y H, et al. A 3.3 V 32 Mb NAND flash memory with incremental step pulse programming scheme. IEEE J Solid-State Circ, 1995, 30: 1149–1156
28 Kang D, Kim M, Jeon S C, et al. A 512 Gb 3-bit/Cell 3D 6th-Generation V-NAND flash memory with 82 MB/s write throughput and 1.2 Gb/s interface. In: Proceedings of International Solid-State Circuits Conference, San Francisco, 2019. 216–218
29 Yang H Z, Huang P, Han R Z, et al. A novel high-density and low-power ternary content addressable memory design based on 3D NAND flash. In: Proceedings of Silicon Nanoelectronics Workshop, Honolulu, 2020. 29–30
30 Lue H T, Du P Y, Chen W C, et al. A 128 Gb (MLC)/192 Gb (TLC) single-gate vertical channel (SGVC) architecture 3D NAND using only 16 layers with robust read disturb, long-retention and excellent scaling capability. In: Proceedings of International Electron Devices Meeting, San Francisco, 2017
31 Wang P, Xu F, Wang B, et al. Three-dimensional NAND flash for vector-matrix multiplication. IEEE Trans VLSI Syst, 2018, 27: 988–991
32 Lee S H, Kwon D W, Kim S, et al. Investigation of transient current characteristics with scaling-down poly-Si body thickness and grain size of 3D NAND flash memory. Solid-State Electron, 2019, 152: 41–45
33 Cho J, Kang D C, Park J, et al. A 512Gb 3b/cell 7th-generation 3D-NAND flash memory with 184 MB/s write throughput and 2.0 Gb/s interface. In: Proceedings of International Solid-State Circuits Conference, San Francisco, 2021. 426–428
34 Lee S, Kim C, Kim M, et al. A 1 Tb 4b/cell 64-stacked-WL 3D NAND flash memory with 12 MB/s program throughput. In: Proceedings of International Solid-State Circuits Conference, San Francisco, 2018. 340–342
35 Cai Y, Haratsch E F, Mutlu O, et al. Threshold voltage distribution in MLC NAND flash memory: characterization, analysis,

and modeling. In: Proceedings of Design, Automation & Test in Europe Conference & Exhibition, Grenoble, 2013. 1285–1290

36 Arsovski I, Hebig T, Dobson D, et al. A 32 nm 0.58-fJ/Bit/search 1-GHz ternary content addressable memory compiler using silicon-aware early-predict late-correct sensing with embedded deep-trench capacitor noise mitigation. IEEE J Solid-State Circ, 2013, 48: 932–939

37 Siau C, Kim K H, Lee S, et al. A 512 Gb 3-bit/cell 3D flash memory on 128-wordline-layer with 132 MB/s write performance featuring circuit-under-array technology. In: Proceedings of International Solid-State Circuits Conference, San Francisco, 2019. 218–220

38 Tseng P H, Lee F M, Lin Y H, et al. In-memory-searching architecture based on 3D-NAND technology with ultra-high parallelism. In: Proceedings of International Electron Devices Meeting, San Francisco, 2020

39 Wang K L, Du G, Lun Z Y, et al. Modeling of program $V_{\rm th}$ distribution for 3-D TLC NAND flash memory. Sci China Inf Sci, 2019, 62: 042401

40 Pedretti G, Graves C E, Serebryakov S, et al. Tree-based machine learning performed in-memory with memristive analog CAM. Nat Commun, 2021, 12: 1

41 Kazemi A, Sharifi M M, Laguna A F, et al. In-memory nearest neighbor search with FeFET multi-bit content-addressable memories. In: Proceedings of Design, Automation & Test in Europe Conference & Exhibition, Grenoble, 2021. 1084–1089

42 Li H T, Chen W C, Levy A, et al. One-shot learning with memory-augmented neural networks using a 64-kbit, 118 GOPS/W RRAM-based non-volatile associative memory. In: Proceedings of Symposium on VLSI Technology, Kyoto, 2021. 1–2