SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

February 2026, Vol. 69, Iss. 2, 122403:1–122403:7
https://doi.org/10.1007/s11432-025-4507-0

Universal content-addressable memory with high-endurance flash for functionally complete logic-in-memory computing

Xinyi GUO[1], Yang FENG[1], Peng GUO[2], Shaoqi YANG[1], Xiaohuan ZHAO[1], Guangkuo YANG[1], Jixuan WU[1], Xuepeng ZHAN[1*] & Jiezhi CHEN[1*]

[1]*School of Information Science and Engineering, Shandong University, Qingdao 266200, China*
[2]*Shandong Sinochip Semiconductors Co., Ltd., Jinan 250000, China*

Received 21 January 2025/Revised 7 April 2025/Accepted 24 June 2025/Published online 4 January 2026

Abstract Multifunctional and long-lifespan memory is a promising element for implementing hardware in-memory computing (IMC) architectures. In this study, a universal content addressable memory (CAM) is proposed based on two high-endurance NOR flash (2F) memory cells, supporting binary CAM, ternary CAM, multibit CAM, and analog CAM modes. The proposed universal CAM unit facilitates 5-bit nonvolatile storage and 10^{10} program/erase cycles with minimal energy consumption of ∼0.27 fJ/bit/search. Furthermore, 16 types of functionally complete Boolean logic functions (e.g., AND, OR, NAND, NOR, XOR, and XNOR) are demonstrated through distinct mapping rules. These findings may offer considerable potential for developing energy-efficient, long-lifespan, and reconfigurable-logic IMC systems.

Keywords reconfigurable, multifunctional, universal content addressable memory, boolean logic, in-memory computing

Citation Guo X Y, Feng Y, Guo P, et al. Universal content-addressable memory with high-endurance flash for functionally complete logic-in-memory computing. Sci China Inf Sci, 2026, 69(2): 122403, https://doi.org/10.1007/s11432-025-4507-0

1 Introduction

The rapid development of artificial intelligence has led to an increase in the demand for data storage, transmission, and operation. Consequently, traditional von Neumann architecture is facing a series of challenges. To overcome the limitations imposed by the von Neumann bottleneck, the data-centric in-memory computing (IMC) architecture is regarded as a promising candidate, with the potential to significantly reduce operation latency and power consumption [1–3]. A substantial body of research has been dedicated to the development of emerging nonvolatile memory units, with a particular focus on hardware implementations of IMC architecture and systems [4–7].

Recently, content addressable memory (CAM), which enables in-memory search, has attracted considerable attention due to its suitability for large-scale, low-latency, and energy-efficient operations [8,9]. Depending on the number of match states, CAM operates in four different modes: binary CAM (BCAM, two states), ternary CAM (TCAM, three states), multibit CAM (MCAM), and analog CAM (ACAM). MCAM is capable of identifying multiple discrete states, while ACAM allows searches over adjustable successive ranges [10–15]. A standard static random-access memory-based CAM unit is composed of six transistors [10] and can support only fundamental BCAM operations, resulting in substantial area cost and power consumption. The advantages of high density and mature fabrication processes are demonstrated by MCAM units based on three-dimensional (3D) NAND flash, which exhibit 157 times higher storage density than those of complementary metal-oxide-semiconductor-based CAM units [11]. TCAM units based on ferroelectric field-effect transistors (FeFETs) exhibit low operating latency and power consumption [12]. In the context of ACAM units, emerging nonvolatile memories with multiple resistive states have gained notable traction, particularly in ACAM units supporting analog continuous matching ranges [13]. While FeFETs have been investigated for use in TCAM and BCAM, avenues remain for improving their multilevel characteristics and power consumption. Compared with other nonvolatile memories, flash memory offers several advantageous properties, including compatibility with different processes, high storage capacity and density, and reliable performance [16–19].

[*] Corresponding author (email: zhanxuepeng@sdu.edu.cn, chen.jiezhi@sdu.edu.cn)

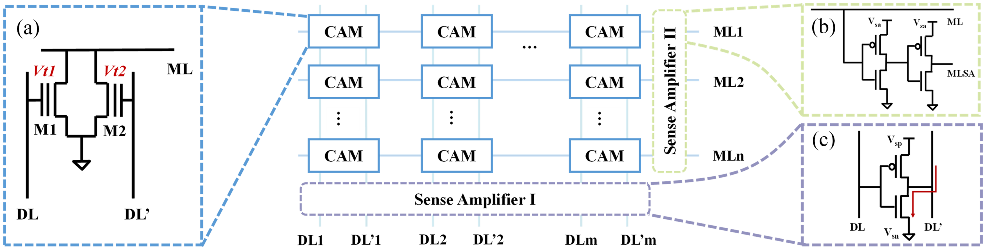


Figure 1 (Color online) UCAM consists of conventional CAM array, SA I and SA II. (a) shows the CAM unit; (b) shows the SA with two inverters in series; (c) illustrates an SA configuration with an inverter between DL and DL$'$.

Nevertheless, a universal CAM (UCAM) unit capable of supporting various modes is seldom reported, particularly in the context of functionally complete Boolean logic operations.

In this study, a UCAM unit supporting four different operating modes is proposed based on two high-endurance NOR flash memory cells. By using different mapping rules, the UCAM unit achieves a low energy consumption of 0.27 fJ/bit during search operations with a 5-bit storage capability, owing to its large threshold voltage range and high endurance. Furthermore, functionally complete Boolean logic functions are demonstrated, which can be further extended to many input logic operations. The findings of this study are of great importance in the development of energy-efficient, long-lifetime, and reconfigurable-logic IMC systems based on flash-based CAM units.

2 Results and discussions

2.1 UCAM unit and array characteristic

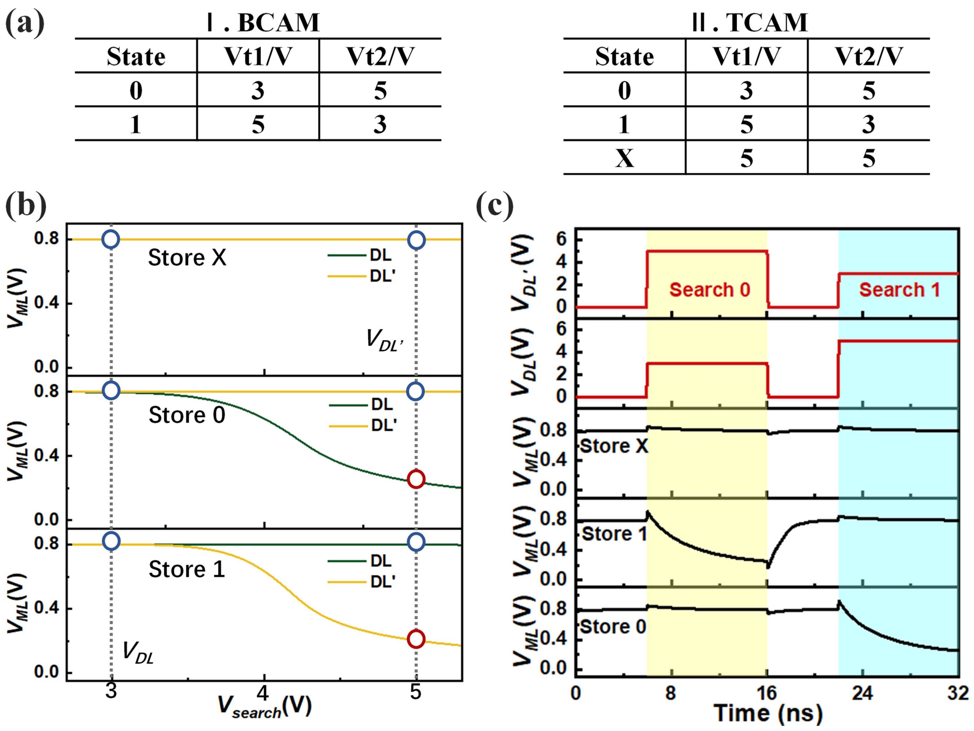
Figure 1 shows the schematic of the proposed UCAM structure. The array consists of multiple CAM units based on NOR flash memory, along with a sense amplifier (SA) for analog input (SA I) and output (SA II). As shown in the inset of Figure 1, the CAM cell consists of two flash memory cells, M1 and M2, connected in parallel. These cells possess different threshold voltages (V_t), namely V_{t1} and V_{t2}, representing different storage states. During the write operation, the device threshold voltage can be adjusted by applying different programming or erasing voltages. During the search operation, the corresponding voltage is applied to the gates of M1 and M2 through the data lines DL and DL$'$. If the input voltage is less than or equal to the storage cell's V_t, the cell is considered to have a matching state. Figure 1(b) presents the structural representation of SA II, which amplifies the match line (ML) signal by establishing a connection between two inverters in series at the output end of the ML.

Employing the multistate storage of flash memory cells enables the UCAM to execute the functions of BCAM, TCAM, and MCAM. The ACAM function can also be realized through simple peripheral circuits. Figure 2(a) shows the storage parameters of the CAM cell used as BCAM and TCAM. Specifically, V_{t1} and V_{t2} represent the threshold voltages of flash memory transistors, designated as M1 and M2, respectively. When V_{t1} and V_{t2} are set to 3 and 5 V, respectively, the cell stores the "0" bit. Conversely, it signifies the storage of "1". If both V_{t1} and V_{t2} are set to 5 V, it corresponds to the storage of the "X" symbol. In the search operation, applying a voltage of 3 V to DL and 5 V to DL$'$ initiates the search for the "0" state, while applying a voltage of 5 V to DL and 3 V to DL$'$ initiates the search for the "1" state.

Figure 2(b) shows the effect of different data line search voltages on the ML voltage ($V_{\rm ML}$) with storing "0" "1" and "X", respectively. A voltage difference of 2 V is sufficient for ML to differentiate between different states. Concurrently, the change of $V_{\rm ML}$ is simulated in the time domain, as illustrated in Figure 2(c). The operating time is measured in nanoseconds, allowing for the swift differentiation of matching results.

The MCAM mode can be realized by utilizing the rich threshold voltage range of flash memory cells. Figure 3(a) illustrates the corresponding relationship between V_{t1}, V_{t2}, and the corresponding storage state when operating in 2-bit MCAM mode. The adoption of a specific voltage $V_{\rm sn}$ at the ground terminal serves to impede the discharge path from DL$'$. To circumvent unnecessary power consumption, the voltage of SA I must be elevated to establish $V_{\rm sn} = V_{\rm sp} = V_{\rm DL'}$. It is evident that the sum of V_{t1} and V_{t2} remains constant. During the search, the DL voltage corresponds to the storage state. When the CAM stores different values, the DL query results for each value are illustrated in Figure 3(b). Based on $V_{\rm ML}$, the match result can be readily determined. By further subdividing ΔV_t, the discharge characteristics of the 16-state MCAM are shown in Figure 3(c).

The ACAM mode can also be realized using SA I and SA II, in which the input signal of DL$'$ is generated by inverting the DL signal. The upper and lower boundaries of the CAM cell are shown in Figures 3(d) and (e),



(a)

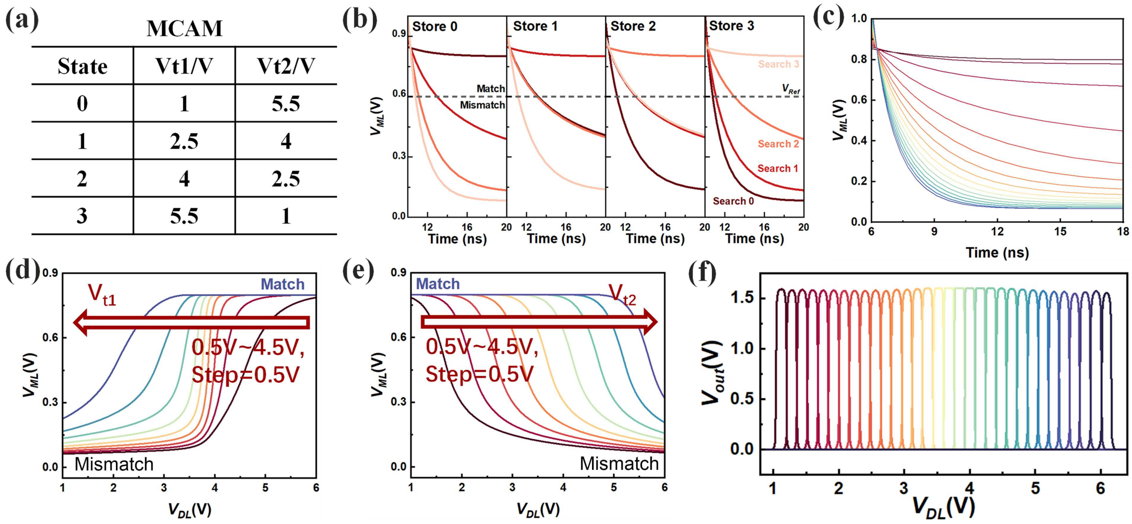
Ⅰ. BCAM

State	Vt1/V	Vt2/V
0	3	5
1	5	3

Ⅱ. TCAM

State	Vt1/V	Vt2/V
0	3	5
1	5	3
X	5	5

Figure 2 (Color online) (a) Schematic of CAM cell parameters in binary CAM (BCAM) and ternary CAM (TCAM) modes; (b) impact of different search voltages on matching results; (c) corresponding search results.



(a)

MCAM

State	Vt1/V	Vt2/V
0	1	5.5
1	2.5	4
2	4	2.5
3	5.5	1

Figure 3 (Color online) (a) CAM cell parameter diagram. (b) ML voltage distribution under matched and mismatched conditions in 2-bit multibit CAM (MCAM) mode. (c) Effect of different mismatch bits on ML voltage in 4-bit MCAM. (d) Upper boundary and (e) lower boundary distributions with V_t in ACAM. (f) Nonoverlapping matching intervals of 32.

respectively. By adjusting the V_t of the flash memory cell to modify the matching boundary, a finely tunable analog range can be achieved. As shown in Figure 3(f), up to 32 nonoverlapping matching intervals can be obtained. Thus, universal CAM functions (BCAM, TCAM, MCAM, and ACAM modes) can be realized.

2.2 Functionally-complete logic operations

The full-scale Boolean logic operation is realized based on the CAM structure by simply adjusting the mapping rules. For two-input logic operations, variable 1 (p) is represented by the threshold voltage (V_{t1} and V_{t2}), and variable 2 (q) is represented by the search voltage ($V_{\rm DL}$ and $V_{\rm DL'}$). If the input variable p is 1, $V_{t1} = 5$ V and $V_{t2} =$ 3 V are programmed; otherwise, $V_{t1} = 3$ V and $V_{t2} = 5$ V are programmed. The mapping rules of the input variable

Table 1 Mapping rules for various logical operations.

Function	Data	Variable q	
		V_{DL} (V)	$V_{\mathrm{DL}'}$ (V)
0	0	5	5
	1	5	5
1	0	3	3
	1	3	3
p	0	5	3
	1	5	3
q	0	5	5
	1	3	3
p'	0	3	5
	1	3	5
q'	0	3	3
	1	5	5
AND $(p \cdot q)$	0	5	5
	1	5	3
OR $(p + q)$	0	5	3
	1	3	3
NAND $(p \cdot q)'$	0	3	3
	1	3	5
NOR $(p + q)'$	0	3	5
	1	5	5
XOR $(p \cdot q' + p' \cdot q)$	0	5	3
	1	3	5
XNOR $(p \cdot q + p' \cdot q')$	0	3	5
	1	5	3
RIMP $(p + q')$	0	3	3
	1	5	3
IMP $(p' + q)$	0	3	5
	1	3	3
NIMP $(p' \cdot q)$	0	5	5
	1	3	5
RNIMP $(p \cdot q')$	0	5	3
	1	5	5

q change depending on the type of operation, as shown in Table 1. Taking the AND operation as an example, if the input variable 2 is 0, it indicates that the search voltages are $V_{\mathrm{DL}} = V_{\mathrm{DL}'} = 5$ V; otherwise, $V_{\mathrm{DL}} = 5$ V and $V_{\mathrm{DL}'}$ = 3 V. This method requires no modification of the circuit structure and imposes no additional overhead, making it extendable to multi-input variable logic operations, including AND, NAND, OR, NOR, XOR, and XNOR. For example, the mapping schemes of quaternary AND logic operations are shown in Tables 2 and 3. The values of the input variables X1, X2, and X3 are represented by the threshold voltages of the memory cells. The value of X4 is represented by the voltage values V_{DL} and $V_{\mathrm{DL}'}$, where $V_0 < V_1 < V_2 < \cdots < V_7$. Furthermore, the proposed logical mapping method is extended to n-bit input variables. The threshold voltages V_{t1} and V_{t2} stored in M1 and M2 are used to represent the input variables X1, X2, …, X$(n-1)$. Assuming that the decimal value corresponds to the binary number composed of X1X2$\cdots$X$(n-1)$ is i, then

$$V_{t1} = V_{i-1}, \tag{1}$$

$$V_{t2} = V_{2^{n-1}-i}. \tag{2}$$

Xn is represented by the pressurized value of DL and DL$'$, and the mapping method is shown in Table 4, where V_0 $<V_1$ $<V_2$ $<\cdots$ $<V_{2^{n-1}-1}$.

Table 2 Input 1, 2, 3 of four variables and operation.

Data	Input 1, 2, 3 (X1, X2, X3)	
	$V_{\rm DL}$	$V_{\rm DL'}$
000	V_0	V_7
001	V_1	V_6
010	V_2	V_5
011	V_3	V_4
100	V_4	V_3
101	V_5	V_2
110	V_6	V_1
111	V_7	V_0

Table 3 Input 4 of four variables and operation.

Function	Data	Input 4 (X4)	
		V_{t1}	V_{t2}
AND	0	V_7	V_7
	1	V_7	V_0

Table 4 Input n of n variables and operation.

Function	Data	Input n (Xn)	
		V_{t1}	V_{t2}
AND	0	$V_{2^{n-1}-1}$	$V_{2^{n-1}-1}$
	1	$V_{2^{n-1}-1}$	V_0

3 Device characterization analysis and benchmarking

Figure 4(a) illustrates the transmission electron microscopy (TEM) image and the device structure of the adopted flash unit. The device used is a floating-gate NOR flash memory fabricated using a 55 nm technology node. The functionality of UCAM relies on the device's ability to exhibit finely tunable with many discrete levels. By applying the incremental step pulse programming (ISPP) method, the threshold voltage of the flash memory can be precisely adjusted. Figures 4(b) and (c) show 170 discrete current-voltage (I-V) curves within a $\Delta V_{\rm th}$ of 5.4 V, demonstrating the device's capacity for finely tunable adjustment. The programming speed (write latency per cell) is a critical factor for CAM performance, particularly in applications requiring frequent updates. Figure 4(d) illustrates the threshold voltage shifts ($\Delta V_{\rm th}$) under varying programming times and voltages using the ISPP method. For a given memory window (MW), the write speed increases with increased programming voltages. In CAM modes, different combinations of programming voltage and time can be selected depending on the required MW. Figure 4(e) shows the device endurance under various MWs along with the corresponding number of discrete states. As the MW increases, the endurance of the flash memory cell gradually decreases. Nevertheless, the endurance remains consistently $\geqslant 10^5$, suggesting a long operational lifespan for the CAM unit. Conversely, as the MW decreases, the number of discrete states also reduces. For instance, at an MW of 1.28 V, 41 discrete states are still available, highlighting the robustness of the proposed UCAM based on flash memory. When the MW is larger than 1.5 V, the device can operate in MCAM mode (see green region in Figure 4(e)).

Table 5 [19–22] presents a summary of this study along with comparisons to related studies. The UCAM feature is realized by using the multivalue storage characteristics of flash memory and a simple inverter structure. This configuration achieves a maximum storage capacity of 32 bits with an energy consumption of 0.27 fJ/bit. For a 32 × 32 CAM array, the power consumption during matching, one-bit mismatch, and full mismatch conditions is 6.75, 114.75, and 117.57 pJ/search, respectively. Furthermore, bitwise operations and 16 types of functionally complete Boolean logic operations are demonstrated based on the CAM structure, with the potential for extension to logic functions involving multiple input variables.

4 Conclusion

Herein, a novel UCAM is proposed based on two NOR flash memory cells, featuring a large tunable threshold voltage range (∼5.0 V) and high endurance (up to 10^{10} program&erase cycles). The proposed UCAM is capable of

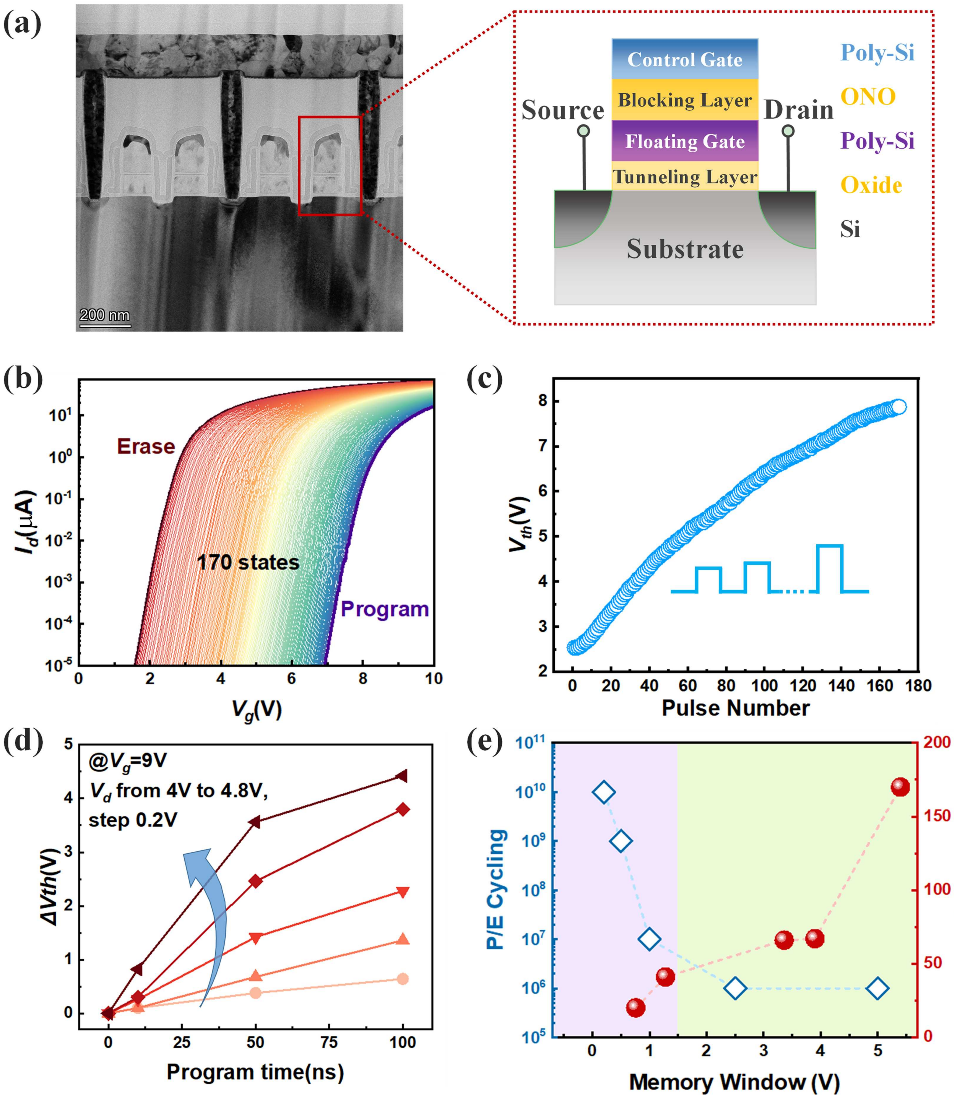


Figure 4 (Color online) (a) TEM image and structure diagram of the adopted NOR flash with a polysilicon gate and an oxide-nitride-oxide (ONO) barrier layer; (b) finely tunable adjustment of the flash V_{th}, programmed to 170 I-V curves; (c) fine tunability of the flash V_{th}; (d) V_{th} shifts (ΔV_{th}) under differ programming voltages and times; (e) decreasing endurance of flash memory cells with increasing memory window (MW).

Table 5 Benchmark of this work and other related studies.

	Ref. [19]	Ref. [20]	Ref. [21]	Ref. [22]	This work
CAM cell	2FeFET	2FeFET	2Flash+2T	2T2C	2Flash
TCAM mode	Yes	Yes	Yes	Yes	Yes
MCAM mode	Yes	Yes	Yes	No	Yes
ACAM mode	Yes	Yes	No	No	Yes
Bitwise operation	No	No	Yes	Yes	Yes
Functionally-complete Boolean operations	No	No	No	No	Yes
Multi-bit operations	No	No	No	No	Yes
Storage capacity (state)	1–8	1–4	1–16	1–2	1–32
Search energy (/search)	–	1.28 pJ/array	0.18 fJ/bit	0.29 W	0.27 fJ/bit

supporting BCAM, TCAM, MCAM, and ACAM modes, supporting 5-bit nonvolatile storage and energy consumption as low as ∼0.27 fJ/bit/search. Furthermore, the configurable UCAM demonstrates 16 types of functionally complete Boolean logic operations (e.g., AND, OR, NAND, NOR, XOR, and XNOR) through different mapping rules, which can be extended to multibit logic operations. This study offers considerable potential for the development of energy-efficient, long-lifetime, and reconfigurable-logic IMC systems based on high-endurance, multistate flash memory.

Acknowledgements This work was supported by National Key Research and Development Program of China (Grant Nos. 2023YFB4402500, 2023YFB4402400), National Natural Science Foundation of China (Grant Nos. U2441248, 62034006, 91964105), Natural Science Foundation of Shandong Province (Grant No. ZR2023LZH007), and MIND Project (Grant No. MINDXZ202407).

References

1 Haensch W, Raghunathan A, Roy K, et al. Compute in-memory with non-volatile elements for neural networks: a review from a co-design perspective. Adv Mater, 2023, 35: 2204944

2 Sebastian A, Le Gallo M, Khaddam-Aljameh R, et al. Memory devices and applications for in-memory computing. Nat Nanotechnol, 2020, 15: 529–544

3 Feng Y, Chen B, Liu J, et al. Design-technology co-optimizations for general-purpose computing in-memory based on 55 nm NOR flash technology. In: Proceedings of the IEEE International Electron Devices Meeting (IEDM), San Francisco, 2021

4 Wan W, Kubendran R, Schaefer C, et al. A compute-in-memory chip based on resistive random-access memory. Nature, 2022, 608: 504–512

5 Soliman T, Chatterjee S, Laleni N, et al. First demonstration of in-memory computing crossbar using multi-level cell FeFET. Nat Commun, 2023, 14: 6348

6 Feng Y, Chen B, Tang M F, et al. Near-threshold-voltage operation in flash-based high-precision computing-in-memory to implement Poisson image editing. Sci China Inf Sci, 2023, 66: 222402

7 Sun X, Khwa W S, Chen Y S, et al. PCM-based analog compute-in-memory: impact of device non-idealities on inference accuracy. IEEE Trans Electron Devices, 2021, 68: 5585–5591

8 Graves C E, Li C, Sheng X, et al. In-memory computing with memristor content addressable memories for pattern matching. Adv Mater, 2020, 32: 2003437

9 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

10 Jeloka S, Akesh N B, Sylvester D, et al. A 28 nm configurable memory (TCAM/BCAM/SRAM) using push-rule 6 T bit cell enabling logic-in-memory. IEEE J Solid-State Circ, 2016, 51: 1009–1021

11 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: 142402

12 Dutta S, Khanna A, Ye H, et al. Lifelong learning with monolithic 3D ferroelectric ternary content-addressable memory. In: Proceedings of the IEEE International Electron Devices Meeting (IEDM), 2021. 1–4

13 Li C, Graves C E, Sheng X, et al. Analog content-addressable memories with memristors. Nat Commun, 2020, 11: 1638

14 Yang L, Zhao R, Li Y, et al. In-memory search with phase change device-based ternary content addressable memory. IEEE Electron Device Lett, 2022, 43: 1053–1056

15 Kazemi A, Sharifi M M, Laguna A F, et al. FeFET multi-bit content-addressable memories for in-memory nearest neighbor search. IEEE Trans Comput, 2022, 71: 2565–2576

16 Tseng P H, Lee F M, Lin Y H, et al. A hybrid in-memory-searching and in-memory-computing architecture for NVM based AI accelerator. In: Proceedings of the Symposium on VLSI Technology, 2021. 1–2

17 Hsieh C C, Lue H T, Li Y C, et al. Chip demonstration of a high-density (43 Gb) and high-search-bandwidth (300 Gb/s) 3D NAND based in-memory search accelerator for ternary content addressable memory (TCAM) and proximity search of hamming distance. In: Proceedings of the IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), 2023. 1–2

18 Lin Y H, Tseng P H, Lee F M, et al. NOR flash-based multilevel in-memory-searching architecture for approximate computing. In: Proceedings of the IEEE International Memory Workshop (IMW), 2022. 1–4

19 Yin X, Li C, Huang Q, et al. FeCAM: a universal compact digital and analog content addressable memory using ferroelectric. IEEE Trans Electron Devices, 2020, 67: 2785–2792

20 Xu W, Luo J, Chen Z, et al. A novel ferroelectric FET based universal content addressable memory with reconfigurability for area- and energy-efficient in-memory-searching system. IEEE Electron Device Lett, 2024, 45: 1345–1348

21 Bai M Y, Wu S H, Wang H, et al. A 3D MCAM architecture based on flash memory enabling binary neural network computing for edge AI. Sci China Inf Sci, 2024, 67: 222403

22 Yavits L. DRAMA: commodity DRAM based content addressable memory. IEEE Comput Arch Lett, 2024, 23: 65–68