2024 IEEE 24th International Conference on Nanotechnology (NANO)
July 8-11, 2024. Gijón, Spain

The Impact of Device Technologies on the Design of Non-Volatile Content Addressable Memories

Sabrina Hassan Moon*, Prayash Dutta†, Parsa Khorrami*, Sanjukta Bhanja†, Dayane Reis*
*Department of Computer Science and Engineering, University of South Florida, Tampa, FL, 33620
†Department of Electrical Engineering, University of South Florida, Tampa, FL, 33620
{ms38, prayash, parsak, bhanja, dayane3}@usf.edu

2024 IEEE 24th International Conference on Nanotechnology (NANO) | 979-8-3503-8624-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/NANO61778.2024.10628710

***Abstract*—Content Addressable Memories (CAMs) are employed in the design of computing-in-memory (CiM) accelerators for data-intensive applications due to their ability to perform massively parallel searches. This paper presents a study of different device technologies, i.e., resistive RAMs (ReRAMs), ferroelectric field-effect transistors (FeFETs), and magnetoresistive random access memory (MRAMs) that are leveraged in the development of dense and energy-efficient non-volatile content addressable memories (NVCAMs). We leverage data from our own research and that of others to present a comprehensive evaluation of different NVCAMs and compare their power consumption, area efficiency, speed, and endurance with respect to CMOS-based CAM counterparts. Additionally, we explore potential application scenarios that map to the unique strengths of the various NVCAMs. Our discussion highlights pathways for future research on the application mapping of NVCAM designs for CiM architectures.**

I. Introduction

Computing-in-Memory (CiM) performs logic and arithmetic operations within the memory. CiM may be an alternative to reduce the latency and energy overheads of data transfers in data-intensive application scenarios such as artificial intelligence (AI), and can offer high throughput through parallel computation. A component of CiM, content-addressable memories (CAMs) allow for the retrieval of data based on its content rather than its address [1]. Designing and implementing CAM-based CiM systems using silicon-based complementary metal oxide semiconductor (CMOS) technology, both TCAM and BCAM, require 16 and 10 transistors (16T and 10T), respectively, which leads to a high standby power and large area footprints [2]. In data-intensive application scenarios, minimizing the power consumption of the memory system is essential for efficiently storing and processing large volumes of data. Consequently, most CAM designs utilized in CiM architectures incorporate non-volatile memories (NVMs) such as resistive random access memories (ReRAMs) [1], ferroelectric field-effect transistors (FeFETs) [3], and magnetoresistive random access memories (MRAMs) [4].

This paper presents a study of three device technologies, i.e., ReRAMs, FeFETs, and MRAMs, highlighting their impact on the design of non-volatile content-addressable memory (NVCAM) architectures. We leverage a combination of results reported in existing work and new SPICE simulations to evaluate different NVCAM designs based on these devices. We report the power consumption, area efficiency, speed, and reliability of NVCAMs and compare them with CMOS-based CAM counterpart designs. We delve into the specifics of device characteristics, circuit design considerations, and application mapping, underlining the feasibility and effectiveness of the different NVM-based CAM designs for practical applications. Additionally, we explore potential challenges and opportunities for future research, focusing on the integration of these technologies into CiM systems.

II. Background

A. Content-Addressable Memory

CAMs have emerged as a crucial component in the architecture of CiM accelerators due to their support for massively parallel content-based searches. As demonstrated in [5], [6], CAM significantly improves performance in applications like few-shot learning, bioinformatics, etc. CAM configurations vary to suit different search tasks: Binary CAMs (BCAMs) store bits as "0" or "1". Ternary CAMs (TCAMs) introduce a "don't care" state "X", allowing a bit to match either "0" or "1", broadening the application usage scenarios for CAMs. Finally, multi-bit CAM (MCAMs) (e.g., [7]) enable each CAM cell to store and search across multiple bits simultaneously, improving the system's density and energy efficiency.

A CAM search is executed through a series of systematic steps. It starts with the parallel comparison of input data (i.e., the query) against all stored data (i.e., entries), utilizing CAM's parallel computation feature. Then, a sense amplifier circuit is activated to identify matches between the query and stored data. In cases of mismatches, methodologies like those discussed in [8] offer techniques to assess and use the degree of mismatch, which can be particularly useful in applications such as nearest neighbor computations for classification tasks in the domain of AI applications.

B. Device Technologies

The limitations of conventional CMOS technology, including scaling challenges, leakage currents, and power consumption issues, have prompted a shift of focus toward emerging device technologies in recent years. Emerging devices have been specially employed in the design of NVMs, due to their ability to retain stored data even when power is turned off. NVMs offer several advantages over conventional flash memory, including faster read/write speeds, lower power consumption, and improved scalability. Additionally, some technologies may exhibit enhanced endurance and retention characteristics, making them suitable for a wider range of applications. Fig. 1 illustrates ReRAM, FeFET, and MRAM devices, discussed in the next paragraphs.

The ReRAM device, illustrated in Fig. 1(a), exploits the resistance-switching behavior of oxides, which can transition

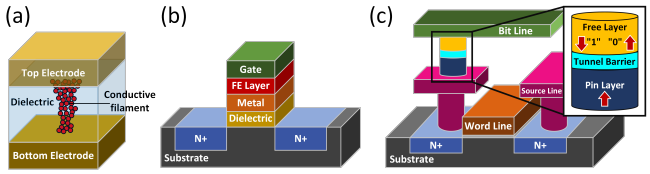

Fig. 1: A physical representation of the studied device technologies: (a) A ReRAM, (b) An FeFET, and (c) A MRAM.

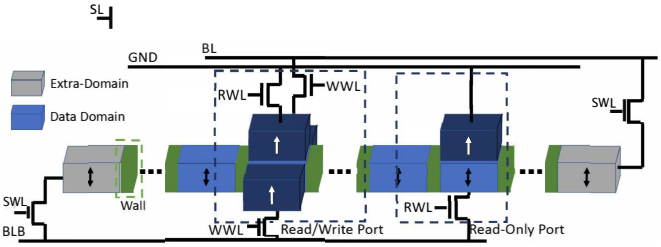

Fig. 2: Anatomy of a RTM nanowire [10].

between high-resistance and low-resistance states. ReRAM cells are written by applying a voltage across the cell through the bottom and top electrodes, which results in a change in resistance state due to the movement of ions within the conductive filament material. Reading a ReRAM cell involves measuring the resistance state, typically by applying a low voltage at the electrodes and sensing the resulting current.

The FeFET device has a ferroelectric (FE) layer within the metal-oxide-semiconductor (MOS) gate stack (Fig. 1(b)). This FE layer (illustrated in red) is typically located under the gate terminal, and precedes the standard metal and dielectric layers, with a semiconductor substrate at the base. The FE layer introduces hysteresis to the device, enabling it to preserve a programmed memory state even without power supply.

The MRAM device, illustrated in Fig. 1(c), leverages the orientation of ferromagnetic materials to encode data. It comprises a magnetic tunnel junction (MTJ) with a fixed ferromagnetic layer, an insulation layer, and a variable-magnetization-free layer. Racetrack memory (RTM), or nanowire memory (NWM), extends the free layer into a nanowire for storing multiple bits, marked by domain walls in inverse magnetic polarizations. Crafted notches prevent unintended domain wall shifts. Fig. 2 depicts a RTM nanowire; SWL, WWL, RWL control the shift, read, and write operations. Both left and right shifts can be performed with a small shift current. To read, the nanowire is shifted to a precise amount so that the target cell reaches the access point. Then, the read current is sent from one side of the MTJ, tunneled through the device and collected from the other side [9]. The read current and the high and low resistance states (RH and RL, respectively) create a voltage difference. Then, a sense amplifier is used to sense that bit and declare whether it is "0" or "1". For the write operation, a higher write current is required to alter the magnetization of the free layer.

C. NVCAMs based on Device Technologies

The fast-paced research in device technologies has fuelled a significant interest in the design of NVCAMs that leverage the intrinsic characteristics of the devices to address the growing demands for faster, more energy-efficient CiM accelerators. This summary highlights recent NVCAM advancements using ReRAM, FeFET, and MRAM.

Reference [11] presents a 86×12 memristor-based NV TCAM array for key-value lookups and pattern matching, which requires less area and power than CMOS alternatives, but faces ReRAM-related reliability and integration issues. Alternatively, FeFET-based NVCAMs discussed in [12], [7] leverage FeFETs for the design of dense and low static power BCAMs/TCAMs and MCAMs, respectively. While FeFET NVCAM designs offer fast content-based searches and parallel processing capabilities, their implementation requires additional silicon area for drivers that can provide the negative, high-magnitude write voltages required by FeFET cells. This issue, which could potentially lead to dynamic power consumption and chip area overheads, has not been discussed in [12], [7]. Furthermore, the low endurance and long write times of FeFETs may make them unamenable to applications with a high volume of memory writes (e.g., DNN training).

For MRAMs, [13] introduces the AM^4 architecture using Samsung MRAM crossbars for integrated CAM functions, optimizing performance with STT-MTJ despite TMR limitations. Data coding strategies have been implemented to improve reliability and efficiency. Additionally, [14] explores RTM-based TCAMs for bit-serial processing, although scalability remains a challenge. Research in [15] combines RTMs with CMOS for better write/search operations and single-event-upset (SEU) resistance, an improvement over phase change memories (PCMs) and NAND flash, though SEU mitigation may require further study.

TABLE I: RTM-based TCAM search operation

Operation	SL/SL'	Stored Data	Result
Search '0'	0/1	'1' (RTM1=RL, RTM2=RH)	mismatch
	0/1	'0' (RTM1=RH, RTM2=RL)	match
Search '1'	1/0	'0' (RTM1=RH, RTM2=RL)	mismatch
	1/0	'1' (RTM1=RL, RTM2=RH)	match
Search 'X'	0/0	'0' (RTM1=RH, RTM2=RL)	match
	0/0	'1' (RTM1=RL, RTM2=RH)	match

TABLE II: Multi-domain MTJ Properties and Dimensions [9]

Property		Values	
Fixed and Free Layer Material		CoFeB	
Fixed Layer Size		$80\,\text{nm} \times 40\,\text{nm} \times 2\,\text{nm}$	
Free Layer (Domain) Size		$80\,\text{nm} \times 40\,\text{nm} \times 2\,\text{nm}$	
Notch dimensions		$12\,\text{nm} \times 10\,\text{nm} \times 2\,\text{nm}$	
MgO Layer Size		$80\,\text{nm} \times 40\,\text{nm} \times 1\,\text{nm}$	
CoFeB Parameters		K_u	$99\,999\,\text{erg/cc}$
AMR Ratio	0.014	Ms	$1200\,\text{emu/cc}$
Resistivity	$15\,\mu\Omega\,\text{cm}$	Exchange Stiffness	$2.2\,\mu\text{erg/cm}$
MgO TMR Ratio	0.8	J_C	$3.21 \times 10^{10}\,\text{A/m}^2$

III. NVCAM Design Space Exploration

In this section, we discuss the different technology alternatives for ultra-compact, energy-efficient NVCAM designs that can potentially replace CMOS-based CAM designs for data-intensive applications. We contrast the different NVCAM design choices with their CMOS counterparts.

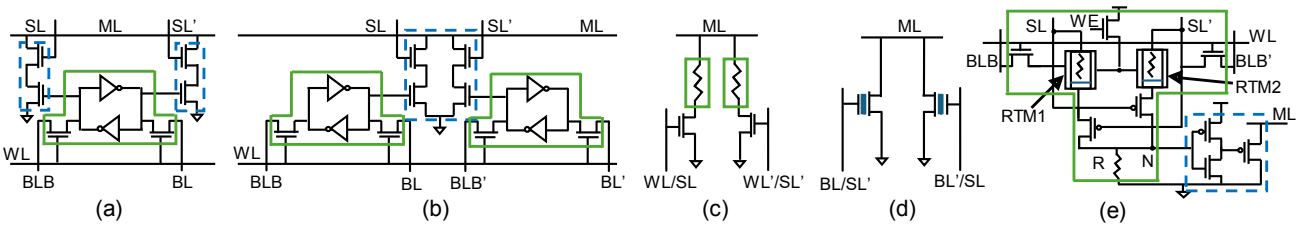

Fig. 3: NVCAM cell designs based on: (a) 10T (BCAM) [2], (b) 16T (TCAM) [2], (c) 2T-2R TCAM [16], (d) 2FeFET BCAM/TCAM/MCAM [7], and (e) RTM TCAM [17].

TABLE III: Cell-level* comparison of different NVCAM designs

Figure-of-Merit (FoM)	CMOS [18], [19]	ReRAM [16], [18]	FeFET [20], [12]	STT-MRAM [18], [21]	RTM [9]
Read Time	<1 ns	<1 ns	<1 ns	<1 ns	<1 ns
Write Time	<1 ns	∼10 ns	10 to 500 ns	2 to 20 ns	∼2 ns
Read Energy	Low	Low	Low	Low	Low
Write Energy	Low	High	Low	High	Low
Leakage Power	High	Low	Low	Low	Low
Endurance Cycles	$> 10^{16}$	$10^6 - 10^{12}$	$10^6 - 10^{12}$	$> 10^{15}$	$> 10^{16}$
CAM Cell Size	$> 100F^2$	$\sim 50F^2$	$\sim 20F^2$	$> 100F^2$	$> 100F^2$
Potential Application	Network Router, Pattern Matching	Internet of Things, Big-Data Processing	Few Shot Learning, Genome Sequencing	Image Processing, Feature Extraction	Associative Processor, Deep Neural Network

* Energy and access time differences between different technologies tend to amplify at the array, as the parasitic component of NVCAM arrays are usually smaller in arrays with denser cells.

A. CMOS vs. NVCAM Design

In CMOS-based CAM architectures, a CAM cell has two blocks of transistors to accomplish bit-wise comparison: (i) a storage block and (ii) a XOR block. The storage block of transistors of a CMOS-based BCAM is illustrated in Fig. 3(a), encircled in a green solid line. It comprises a 6T SRAM with a pair of cross-coupled inverters to store the bit, and two access transistors to control read/write operations in the cell. The XOR block of the same CMOS BCAM cell is depicted inside blue dashed rectangles in Fig. 3(a). The XOR block of transistors establish a pathway from ML to GND, so the search result can be interpreted based on the ML discharge. Fig. 3(b) depicts a CMOS TCAM cell with the same design concept as the 10T BCAM, but with doubled storage blocks to accommodate the storage of a ternary state, in addition to the logic "0" and "1" states. The high transistor count of CMOS-based BCAM (10 transistors) and TCAM (16 transistors) becomes a constraint in terms of scalability and makes the design power-hungry even in standby mode, which motivates the design of NVCAM alternatives.

In NVCAMs, devices can either store 1-bit per cell (e.g., ReRAM, FeFET, MRAM) or multi-bit in a single cell (e.g., ReRAM, FeFET). [16] proposed the 2T-2R structure, illustrated in 3(c), to implement a ReRAM-based TCAM design. This structure requires two access transistors to write complementary bits into the ReRAMs (marked within green rectangles) and to facilitate the search operation. Alternatively, [12] utilized 3-terminal, CMOS-compatible FeFET devices to implement a BCAM/TCAM based on FeFET technology. With just 2 FeFETs per cell, this design is illustrated in Fig. 3(d)). Subsequent work [7] demonstrated that this same structure could also implement MCAMs. Note that both the ReRAM and FeFET NVCAM cell designs do not require separate storage and comparison blocks of transistors like their CMOS counterparts. In the ReRAM-based design, the query data is applied at the gate of the two access transistors, and a current flows from ML to GND through the two memristor branches. The resulting voltage at the ML is sensed by a sense amplified as match/mismatch. In FeFET-based BCAM/TCAM/MCAM, the query data is directly applied at the gates of the FeFET, eliminating the need for access transistors.

Magnetic devices such as RTMs have also been used to implement NVCAMs [15], [17]. Despite recent advancements in these devices, the design of NVCAMs with RTMs is still high in transistor count. Fig. 3(e) shows an RTM-based TCAM design comprising two blocks: (i) a storage block and (ii) a sensing block. In the storage block, highlighted in green solid line, complementary states are stored within two RTMs (RTM1 and RTM2). To perform a search operation, query data is searched via searchlines (SL and SL') with appropriate voltage as listed in Table I. In the sensing block (highlighted within a blue dashed rectangle), the output of the inverter will be set to "1" when the search result is a match, preventing the pMOS from conducting and ML remains charged. However, if the search result is a mismatch the output of the inverter will be "0". This activates the pMOS and ML discharges through it. Despite the large cell design, this pMOS is the only transistor responsible for effective capacitance per cell and reduces ML capacitance compared to previous work [2].

B. Cell-level Comparison of NVCAMs

Here, we conduct a comparative analysis of various CMOS CAM and NVCAM designs, focusing on technologies such as ReRAM, FeFET, MRAM (STT-MRAM and RTM). We evaluate these designs based on metrics that include read/write times and energies, leakage power, cell size, and endurance cycles. Our comparison draws upon data from existing literature on CMOS, ReRAMs, FeFETs, and STT-MRAMs. Furthermore, we implement a TCAM design, as proposed in [17], integrating the read/write capabilities of RTM devices examined in our earlier work [9]. The simula-

tion parameters for the RTM-based TCAM cell are detailed in Table II.

While all technologies demonstrate low read energy, and read times that are less than 1 ns, write figures vary due to the different material properties and writing mechanisms of each device. CMOS has an efficient writing mechanism, which leads to low write time and energy for the CMOS-based CAM design. ReRAM and STT-MRAM, in turn, require a high current passing through the memristor and MTJ, respectively, to change their resistance, which results in high energy. Write times for ReRAM are reported to be around 10 ns [16], [18]. STT-MRAMs display a broader range of write times, from 2 to 20 ns [18], [21]. To switch the polarization of the ferroelectric material, FeFET requires a high write voltage and long pulse, from 10 to 500ns [20], [12]. However, these estimates may significantly fluctuate due to the ongoing development and rapid advancements in research on each of these devices.

ReRAM and FeFET exhibit lower endurance, typically ranging from 10^6 to 10^{12}, and wear out faster compared to their CMOS and MRAM counterparts. For instance, STT-MRAM and RTM exceeds 10^{15} and 10^{16} write cycles for endurance, respectively. Additionally, RTM can store hundreds of bits in a single nanowire with the use of multiple domain walls, which further motivates its use in NVCAM design as an ultra-dense storage option.

C. Application Mapping

Software-based search engines are limited by the size of available physical memory and bandwidth, which results in increased latency due to frequent external memory accesses. This limitation undermines their efficiency for a wide range of AI applications. However, despite its lower endurance, ReRAM-based NVCAM stands out in scenarios characterized by long idle times and infrequent data updates, thanks to its low read energy and low static power consumption. Similarly, FeFETs, although also affected by low endurance, offer benefits such as low read energy, minimal static power consumption, and high density, positioning them as a promising alternative for the inference phase in machine learning applications, such as few-shot learning [5], and in genome sequencing [6]. Traditional CAM systems face challenges in processing large patterns of 2D images, but an effective solution for efficient feature extraction is found in utilizing the XNOR architecture of STT-MRAM-based NVCAM. In the context of CNN inference, where the device configuration occurs post-offline training, the high write energy of STT-MRAM becomes a lesser concern. Associative processors (AP), operating on a bit-serial and word-parallel basis, see increased efficiency with larger datasets, making them suitable for data-intensive SIMD tasks. The write operation in NVM-based APs, including ReRAM, FeFET, and STT-MRAM, is typically slow, energy-intensive, and harmful to device longevity. Addressing this, RTMs, with their inherent shift operations, emerge as a potential alternative for bit-serial processing in APs, representing a significant area for future research. The use of APs can also be extended to DNN inference, further enhancing their relevance in advanced computing tasks. Nevertheless, challenges such as fabrication complexity, scalability constraints, and integration difficulties need to be addressed to enable a broader adoption of RTMs.

IV. Conclusion

In this study, we present an exploration of ultra-compact, energy-efficient NVCAM designs employing various device technologies, including ReRAM, FeFET, and MRAM. Through experimental validation and SPICE simulations, we analyze the power consumption, latency, and energy efficiency of NVCAMs in comparison to CMOS-based CAM designs. Additionally, we highlight the feasibility and efficacy of leveraging different NVM-based CAM designs for practical applications, thereby laying a foundation for significant advancements in the design of CiM accelerators.

References

[1] Y. Chen, et al. Reconfigurable 2t2r reram with split word-lines for tcam operation and in-memory computing. In *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE, 2020.

[2] K. Pagiamtzis et al. Content-addressable memory (CAM) circuits and architectures: A tutorial and survey. *IEEE journal of solid-state circuits*, 41(3):712–727, 2006.

[3] X. Yin, et al. An ultracompact single-ferroelectric field-effect transistor binary and multibit associative search engine. *Advanced Intelligent Systems*, 5(7):2200428, 2023.

[4] C. Zhuo, et al. Design of Ultra-Compact Content Addressable Memory Exploiting 1T-1MTJ Cell. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2022.

[5] A. F. Laguna, et al. Invited paper: Algorithm/hardware co-design for few-shot learning at the edge. In *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, pages 1–9, 2023.

[6] A. F. Laguna, et al. Seed-and-vote based in-memory accelerator for dna read mapping. *2020 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*, pages 1–9, 2020.

[7] A. Kazemi, et al. Fefet multi-bit content-addressable memories for in-memory nearest neighbor search. *IEEE Transactions on Computers*, 71(10):2565–2576, 2022.

[8] K. Ni, et al. Ferroelectric ternary content-addressable memory for one-shot learning. *Nature Electronics*, 2(11):521–529, 2019.

[9] P. Dutta, et al. A multi-domain magneto tunnel junction for racetrack nanowire strips. *IEEE Transactions on Nanotechnology*, 22:581–583, 2023.

[10] R. Venkatesan, et al. Dwm-tapestri-an energy efficient all-spin cache using domain wall shift based writes. In *Proc. of DATE*, pages 1825–1830, 2013.

[11] C. E. Graves, et al. In-memory computing with memristor content addressable memories for pattern matching. *Advanced Materials*, 32(37):2003437, 2020.

[12] X. Yin, et al. An ultra-dense 2fefet tcam design based on a multi-domain fefet model. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 66(9):1577–1581, 2018.

[13] E. Garzón, et al. Am 4: Mram crossbar based cam/tcam/acam/ap for in-memory computing. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 13(1):408–421, 2023.

[14] J. P. C. de Lima, et al. Efficient associative processing with rtm-tcams. 2023.

[15] P. Junsangsri, et al. A non-volatile low-power tcam design using racetrack memories. In *2016 IEEE 16th International Conference on Nanotechnology (IEEE-NANO)*, pages 525–528. IEEE, 2016.

[16] J. Li, et al. 1 mb 0.41 μm² 2t-2r cell nonvolatile tcam with two-bit encoding and clocked self-referenced sensing. *IEEE Journal of Solid-State Circuits*, 49(4):896–907, 2014.

[17] K. P. Gnawali, et al. Low power spintronic ternary content addressable memory. *IEEE TNANO*, 17(6):1206–1216, 2018.

[18] S. Yu et al. Emerging memory technologies: Recent trends and prospects. *IEEE Solid-State Circuits Magazine*, 8(2):43–56, 2016.

[19] M.-T. Chang, et al. Technology comparison for large last-level caches (L3Cs): Low-leakage SRAM, low write-energy STT-RAM, and refresh-optimized eDRAM. In *IEEE 19th International Symposium on High Performance Computer Architecture (HPCA)*, pages 143–154. IEEE, 2013.

[20] S. Narla, et al. Modeling and design for magnetoelectric ternary content addressable memory (tcam). *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, 8(1):44–52, 2022.

[21] F. Oboril, et al. Evaluation of hybrid memory technologies using sot-mram for on-chip cache hierarchy. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 34(3):367–380, 2015.