Integrated CAM-RAM Functionality using Ferroelectric FETs

Sumitha George[1], Nicolas Jao[1], Akshay Krishna Ramanathan[1], Xueqing Li[2],
Sumeet Kumar Gupta[3], John Sampson[1], Vijaykrishnan Narayanan[1]
[1]The Pennsylvania State University, University Park, PA;
[2]Tsinghua University, Beijing, China; [3]Purdue University, West Lafayette, IN;
[1]{sug241, naj5075, axr499, jms1257, vijay}@psu.edu; xueqingli@tsinghua.edu.cn; guptask@purdue.edu

Abstract

Our work proposes a new Ferroelectric FET (FeFET) based Ternary Content Addressable Memory (TCAM) with features of integrated search and read operations (along with write), which we refer to as TCAM-RAM. The proposed memory exploits the unique features of the emerging FeFET technology, such as 3-terminal device design, storage in the gate stack, etc., to achieve the proposed functionality. We also introduce Approximate CAM-RAM, which can quantize the bit vector similarity. All the proposed designs operate without negative voltages. We describe both NAND and NOR variants of CAM design. Our CAM design provides 31% area improvement over the previous FeFET 6T CAM design.

Keywords

FeFET, CAM, RAM, TCAM, Approximate CAM

1. Introduction

Recent advancements in big data analytics, data-intensive scientific applications and IoT demand systems with high performance and efficient data processing capabilities. Efforts to meet this increasing demand have relied on technology advancements as well as architectural innovations. Keys among these are technology-driven architecture designs to (a) extract more performance out of the co-designed systems or (b) achieve new functionalities which were not possible previously due to technology constraints [1][2][3]. One major stream of technology-architecture co-design research focuses on solving the "memory wall" by off-loading some portions of computation to various elements within the memory system itself in order to avoid or reduce data movement [1][4]. Examples of these logic-memory pairings include Compute-in-Memory, Compute-near-Memory, and Monolithic 3D approaches [1] [5] [6] [7].

Content addressable memories (CAMs) are crucial parts of such memory-oriented computing designs. CAMs support searches for a piece of data among the contents of a stored array, and ternary CAMs (TCAMs) have the added flexibility of ignoring certain bit positions and still producing a match. Prior in-memory computing work [2] has relied on these CAM functionalities for its operation. While CAMs have also been an integral part of traditional computing systems, the data processing needs of emerging applications increase their utility in beyond-von-Neumann architectures. CAM operations are inherently useful in tag matching, database operations, and high-speed searches while TCAMs can be found in associative memories and in routing nodes for forwarding packets. The inherent scaling of CAM parallelism with CAM capacity makes them appealing for memory-oriented computing models. CAMs can also play useful roles in specialized computations, such as hardware compression/ decompression and implementing reverse tagged directories for large-scale multiprocessor coherence [8].

One of the most common implementations of CAMs is based on SRAMs. However, SRAM-based CAMs suffer from high transistor count [9] leading to low density and leakage power issues. Nonvolatile technologies tend to overcome the aforementioned issues by eliminating standby current and usually provide more compact implementations [10]. Hence, several efforts have been directed towards building CAMs with nonvolatile technologies such as STTRAM [11], ReRAM [12], etc. Each of the emerging technologies comes with a set of unique properties, some of which can be utilized to enable new features which were not possible in CMOS implementations. Both ReRAM and STTRAM rely on low and high resistance states to store bits. However, technological constraints, such as small high-to-low resistance ratio in STTRAMs, endurance concerns in ReRAMs, and NVM elements with two terminal device structures requiring additional transistors to implement CAM with reasonable output swings [10] need to be addressed.

In this context, emerging devices such as Ferroelectric FETs (FeFETs)[13][15] have the potential to lead to innovative CAM designs. FeFET is a 3-terminal device with a high ON/OFF ratio (10^6) making it suitable for area and energy-efficient CAMs. Prior investigations into FeFET based CAMs have shown benefits over NVCAM technologies using SRAM, STTRAM, and ReRAM in terms of area and energy [10] while providing comparable features. This is primarily because of the inherent device features of the FeFET, such as the low-voltage operation, the unique polarization storage mechanism, the high ON-OFF resistance distinguishability, etc [13].

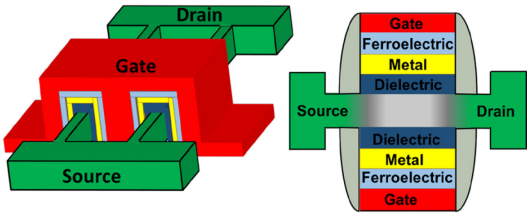
In this work, we propose compact FeFET based TCAM designs that reduce the number of devices by using the 3-terminal property of FeFETs and a novel bit cell organization compared to previous FeFET CAM designs. In addition, we explore potential functional features made feasible by our proposed CAMs which could lead to new computing paradigms. We choose a device-circuit co-design approach to achieve high efficiencies in terms of area, power and maintain nonvolatility. Our proposed design, referred to as the ***TCAM-RAM***, can be used both as CAM (search data in an array) and RAM (read/write data to a specified row based on address). A reconfigurable CAM-RAM not only has the potential to reduce the chip area but also decreases data movement, which

*This work was supported in part by the Semiconductor Research Corporation (SRC)-Center for Research in Intelligent Storage and Processing in Memory (CRISP) and in part by NSF award 1822923 and by the NSFC under grant 61874066.

is a critical issue in today's data-centric applications. Besides, a compact CAM-RAM facilitates local in-memory operations and thereby boost performance by reducing the data movement [2]. Also, we introduce Approximate CAM (ACAM) designs using FeFETs. Traditional CAMs search the exact match of the key, where our proposed ACAMs measure the relative match of the key in the array, which is a critical feature in pattern matching applications, hamming distance and sparse matrix data processing. Further, our proposed designs use fewer transistors (4 transistors) compared to a typical CAM (6 transistors) using the same FeFET technology and modeling. Note that, one of the previous 6T (4T-2Fe) CAM design [2] can potentially realize the abovementioned functionalities, it has not been explored. Our FeFET CAM-RAM/Approximate designs do not add any additional transistor or routing in the array compared to the CAM-only mode design.

Our contributions in this work include:

- Proposal and design of a FeFET-based ternary 4T (2T-2FE) CAM RAM.
- Introducing approximate compute in 4T (2T-2FE) array.
- Design of different topologies such as NAND and NOR variation for TCAM-RAMs.
- Comprehensive area/delay/energy analysis for NOR and NAND TCAM-RAM. NOR design provides 31% array area savings compared to previous FeFET CAM design [2], while showing 16% and 0.3% write and match delay improvement at the cost of 1.45x and 1.73x energy respectively in the CAM mode.



(a) (b)

Figure 1: (a) Ferroelectric FinFET (FeFET) device; (b) Cross section of FeFET;

2. Proposed Designs

2.1. FeFET Structure and Operation

Before we discuss the proposed design, let us briefly describe the operation of FeFETs. FeFETs are designed by adding a ferroelectric (FE) layer in the gate stack of a CMOS transistor as shown in Fig. 1. The charge-voltage relationship of FE materials and the couple of FE with the underlying transistor channel leads to the unique device characteristics of FeFETs, featuring steep-switching and/or hysteresis [13][14]. We focus on using FeFET as a memory device by utilizing the hysteretic property. The work in [15] explores the application of FeFETs for non-volatile memory by using the capability of FE to retain polarization in the absence of voltage biases. The direction of polarization is used to denote logic high or low. The behavior of FeFET memory is captured by connecting the capacitance associated with FE in series with that of the underlying transistor [16]. The behavior of FE is modelled with the time dependent LK equation

$E = \alpha P + \beta P^3 + \gamma P^5 + \rho dP/dt$ [16],

where P is the polarization, E is the electric field, α, β, γ and ρ are the coefficients.

Device-circuit co-design is essential for efficient functional operation, energy area optimization, optimal operating voltages, etc. For example, tuning parameters like the thickness of the FeFET determines whether or not FeFET exhibits non-volatile behavior. This also leads to the choice of suitable read and write voltages. The works in [10] [15] use negative voltages for writes, which leads to design overheads. However, by choosing a proper array–circuit design, negative voltages can be eliminated [14].

2.2. FeFET Based CAMs

In this section, we introduce our CAM design using FeFET. All described variants function as ternary CAMs as well. Traditionally, we have 2 basic versions of CAM implementation called NAND & NOR CAMs [8]. The important feature of NOR is its speed of operation [8]. In the case of a single bit mismatch in NOR implementation, the output path (match/mismatch) can establish a pull down path through a single transistor. For NAND CAM operation, a mismatch stops the propagation of signal from pre-charge. This indicates that after the encounter of the first mismatch the rest of the flow is broken. This could lead to better power saving.

In the next section, we introduce variations of NAND and NOR for our proposed FeFET CAM. Then we introduce the RAM operation which can be performed on both variants of the design we introduced above.

2.3. NAND CAM

In Fig. 2, we show a 2-bit NAND CAM. There is a pre-charge and inverter circuitry per row. The bits are stored in FeFETs (*F1* and *F2*) in complementary form. We name the stored state as '1' if *F1* is in low resistance and *F2* in high resistance. Subsequently, for state '0', *F1* is in high resistance and *F2* in low resistance. Throughout this paper, we follow the aforementioned naming assumption for stored "1" and "0" unless otherwise specified. To sense the match, we apply a bit voltage/complementary bit voltage at BL/ BL_ respectively. In the case of a match, we get a high voltage at the inverter output (OUT in Fig. 2). Suppose we store state "1" in the cell shown in Fig.1, which means *F1* is in low resistance and *F2* is in high resistance. To check for the stored "1", we apply high bit voltage (V_B) at BL and zero voltage (V_{GND}) at BL_. WL1 is grounded and WL2 is made high. Pre is made low to pass the V_{DD} to the inverter input, (INT point in Fig. 2). V_B is passed to the gate of C2 establishing a discharge path from INT to GND in ML leading to a high at the inverter output.

Write is a two-step operation. To write, ML is kept at ground and WL2 is kept low so that C1 is OFF. We adopt a two-step write process to eliminate negative voltage given to the gate of FeFET for writing a zero [10][15]. First, we transition both F1 and F2 to low resistance by applying the V_{WRITE1} at WL1 and by keeping both BL and BL_ at ground.

In the second step, WL1 is kept at GND. In order to write a "1", $V_{BLWRITE}$ is applied to BL_ and V_{GND} is applied to BL.

This switches the F2 to high resistance state. It is crucial to keep the C1 at OFF state so that we do not have a short current path through F1 to the BL source. Also, we keep WL1 of unselected rows to half V_{DD} to prevent the unintentional flipping of states. The biasing conditions are shown in Table 1.

In ternary operation, a *don't care* condition is included. For this, we make the resistance of FeFETs (both F1& F2) low by writing a 1 to both. This is done by applying V_{WRITE1} to WL1 and pulling both BL & BL_ to ground at the ignored bit positions (masked bit positions). BL & BL_ at the unmasked bit positions are kept at $V_{WRITE1}/2$ to prevent bit upsets. In the ternary operation, a logical high is applied at both BL and BL_ (i.e. V_B at both BL and BL_). A transistor G1 (Fig. 2) prevents the glitch induced changes of the output inverter. For example, while writing a "0", C2 might get turned ON and establishes a discharge path from INT to GND. This might lead to the inverter output flip if we had a high voltage at INT. Note that we don't care about the inverter output changes during writing. However, we decide to keep the G1 transistor to avoid glitches and output flipping. The write, match and don't care states of the proposed NAND CAM are shown in Fig. 3.

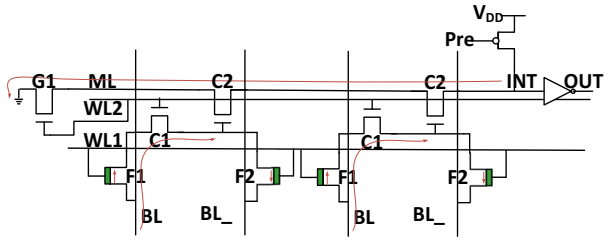


Figure 2: Proposed FeFET NAND CAM;

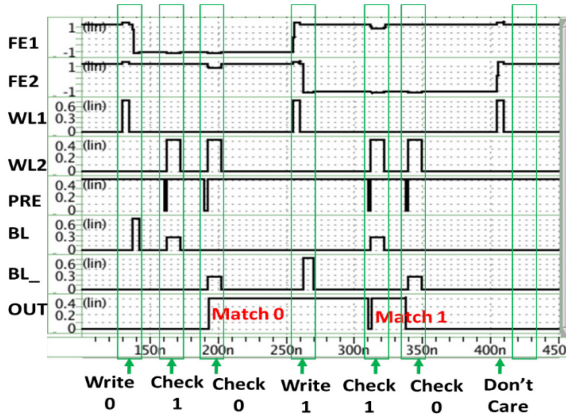


Figure 3: NAND CAM operation in write and match mode;

2.4. NOR CAM

Here we introduce a NOR variant of CAM design. In Fig. 4 we show a 2-bit NOR CAM. To sense the match, we apply bit voltage and complementary bit voltage at BL and BL_ respectively. The inverter output goes low if the match occurs. (Note the change of order of BL and BL_ in Fig. 4 compared to NAND CAM). Suppose we store state "1" in a cell (Fig. 4), which implies *F1* is in low resistance and *F2* is in high resistance. To check for the stored "1", we apply bit voltage (V_B) at BL and complementary bit voltage (V_{GND}) at BL_. We also apply GND to WL1 and V_{DD} to WL2. Pre has made low to pass the V_{DD} to the inverter input (INT point in Fig. 4). Since *F2* is in high resistance, V_B is unable to pass the voltage to the gate C2 and it remains turned OFF. As a result, there is no established discharge path from INT, which drives the inverter output low. The inverter output gets high with a mismatch. As opposed to NAND design, the ML discharge path contains only a single transistor making the response faster.

The write is similar to that of NAND CAM discussed in section **2.3**. ML and WL2 are kept at GND. We follow a two-step write process. In the first step, 1 is written to both F1 and F2 by applying V_{WRITE1} at WL1 and keeping BL_/BL at GND. In the second step, we apply V_{GND} / V_{WRITE2} to BL_/BL to write 1 and V_{WRITE2} /V_{GND} to BL_/BL to write 0 for changing the resistances of *F1* and *F2*. Unselected row's WL1 is kept at a raised voltage at the second step to prevent the flipping of bits. The biasing conditions are given in Table 2. In the don't care condition, logic zero (V_{GND}) is applied to both BL and BL_ at the required bit positions while checking the match. The internal states of *F1* and *F2* are changed to high resistance. This keeps C2 (Fig. 4) OFF preventing the discharge in ML. F1 and F2 are kept at high resistance states by applying V_{WRITE2} to both BL and BL_ of the masked bit positions and keeping all other lines grounded in the selected row. Raised voltage is applied at WL1 of unselected rows to avoid bit flipping. Fig.5 shows the functional states during write/match and also the don't care conditions. Note that polarization of both *F1* and *F2* is retained in the absence of input signal voltages due to nonvolatile design for both NAND and NOR TCAMs (Fig. 2 and Fig. 4).

Table 1: NAND CAM Biasing Conditions

	WL2 (R0)	WL1 (R0)	BL (C0)	BL_ (C0)	WL2 (R1)	WL1 (R1)	BL (C1)	BL_ (C1)
Match	VDD	GND	VBL	VBL_	VDD	GND	VBL	VBL_
Write0								
Step1	GND	V_{WRITE}	GND	GND	GND	GND	GND	GND
Step2	GND	GND	V_{WRITE1}	GND	GND	VDD1	V_{WRITE1}	GND
Write1								
Step1	GND	V_{WRITE}	GND	GND	GND	GND	GND	GND
Step2	GND	GND	GND	V_{WRITE1}	GND	VDD1	GND	V_{WRITE1}

* VDD1~=VDD/2

2.5. RAM Operation

Our CAM array also can perform READ operations, thus acting as a CAM-RAM. The read is a two-step process and is a destructive read. In step 1 of the read process (Fig. 6), we change the *F2* of all the stored bits in the specified row to low resistance. Note that, we keep the *F1* states unchanged. For this we apply V_{WRITE1} to the WL1 of the selected row, GND to BL_ and $V_{WRITE1}/2$ to BL. C1 is kept OFF during the first step by applying GND to WL2. In the second step, we apply a read voltage which is much lesser than the V_{WRITE1} to the BL and keep the C1 ON. This establishes a path inside each cell. Depending on the state of F1 we get low or high current at BL_. The C1s of unselected rows are kept off by applying GND to WL2, so that there is no current loop in unselected rows. The read operation and the corresponding waveforms are shown in Fig. 7 and 8. Since the storage mechanism is the same for NAND, NOR and approximate CAMs, the same read technique is applied to all.

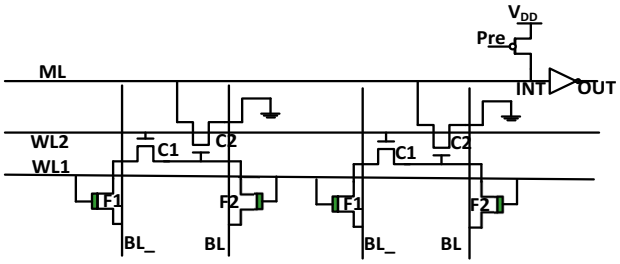


Figure 4: Proposed FeFET NOR CAM;

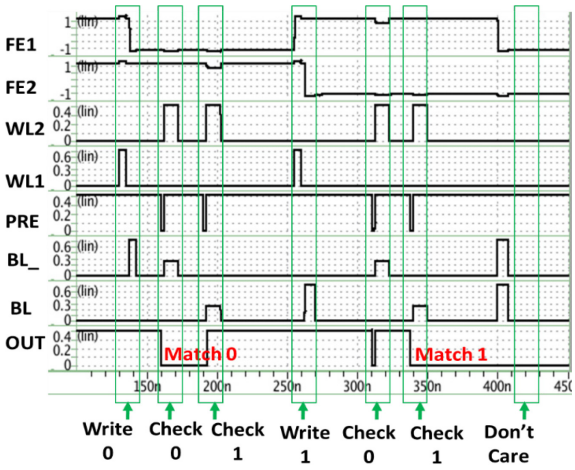


Figure 5: NOR CAM operation in write and match mode;

Table 2: NOR CAM Baising Conditions

	WL2 (R0)	WL1 (R0)	BL (C0)	BL_ (C0)	WL2 (R1)	WL1 (R1)	BL (C1)	BL_ (C1)
Match	VDD	GND	VBL	VBL_	VDD	GND	VBL	VBL_
Write0								
Step1	GND	V_{WRITE}	GND	GND	GND	GND	GND	GND
Step2	GND	GND	GND	V_{WRITE1}	GND	VDD1	GND	V_{WRITE1}
Write1								
Step1	GND	V_{WRITE}	GND	GND	GND	GND	GND	GND
Step2	GND	GND	V_{WRITE1}	GND	GND	VDD1	V_{WRITE1}	GND

We can restore the data after the read by a simple write back to the *F2*s in the selected row.F1s already retain the correct state as we haven't changed its state during read, so our read can be termed as a "*Half destructive read*". If the current sensed from the BL_ is high (which implies F1 is low resistance state) then we write back high resistance to *F2* by applying V_{WRITE2} to BL_. We keep the WL1 of unselected rows to $V_{WRITE1}/2$ during this time to prevent bit upsets. If the current sensed in the BL_ is low, then we do not need to write back to F2, as F2 already maintains the desired low resistance state.

2.6. Approximate CAM

In Fig. 9, we show 2x2 approximate CAM. Here we measure the degree of matching and categorize them into buckets. The operation is similar to NOR CAM. The main difference of NOR CAM from approximate CAM is the decision to keep the C2 (Fig. 9) ON or OFF when a query match occurs. In case of a match in NOR CAM, C2 is made

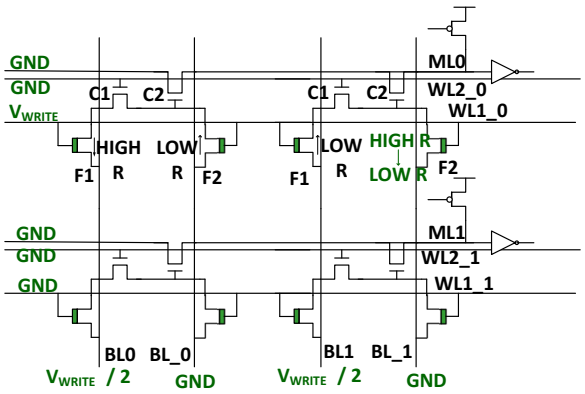


Figure 6: Read step1;

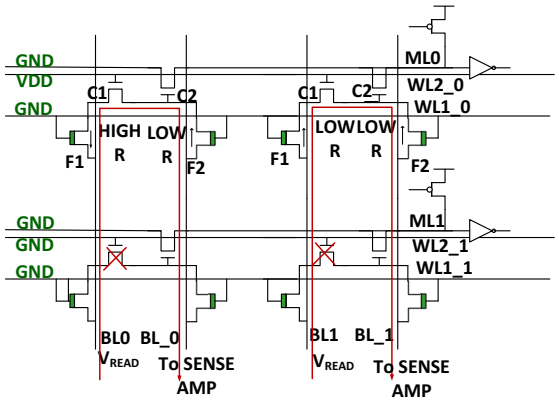


Figure 7: Read step 2;

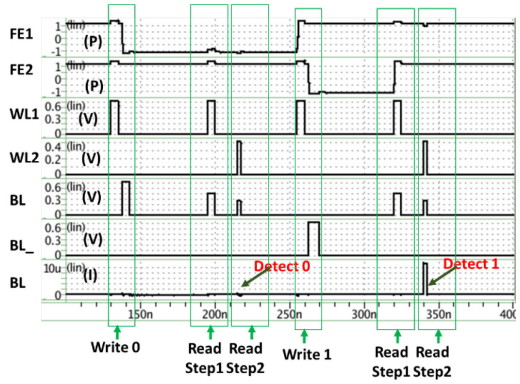


Figure 8: Read from CAM RAM; In the first step F2 is made low resistance. In step 2 current is measured. For bit state 0 low current is measured while for bit state 1 high current is measured;

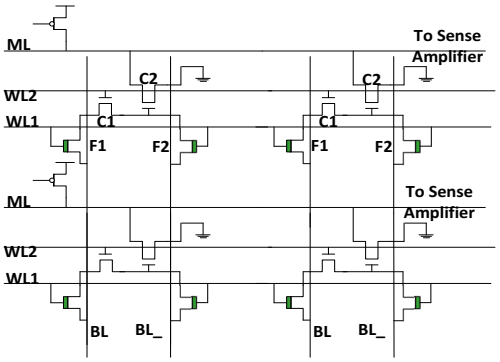


Figure 9: Proposed FeFET approximate NOR CAM;

OFF, such that there is no conducting path to GND. In approximate CAM, C2 is made ON establishing a discharge path from Match Line (ML) to GND. For the CAM query, we apply logic bit voltage to BL and complementary logic voltage to BL_. Suppose state 1 is stored in a cell (*F1* is low resistance and *F2* is in high resistance in Fig. 9). If the bit is a match, C2 gets turned ON creating a discharge path to ground. If we have more bit matches more discharge paths will be established. A sense amplifier with different sensing level will be able to quantify the number of matching cases into appropriate matching buckets. Fig. 10 shows the discharge rates when we have 0/2/4/8 bits matching in an 8 bit row. We see different discharge rates for different matching and a multi-level sense amplifier can categorize them to different levels.

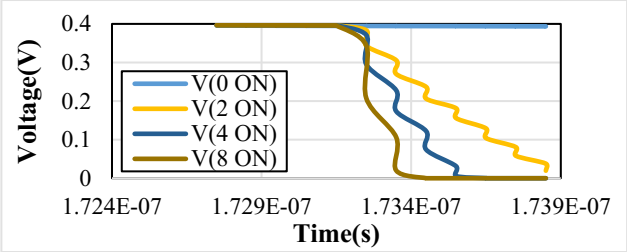


Figure 10: Approximate CAM discharge when there is 0/2/4/8 match in 8 bit array;

3. Performance Evaluation

3.1. Layout

The layouts for our 2x2 NAND and 2x2 NOR designs are shown in Fig. 11 and 12. Our 2x2 NAND design occupies 64 λ x 62 λ where λ is half the feature size. WL1, WL2 and ML run horizontally and BL &BL_ run vertically in the design. A single cell is shown within the dotted line. A 2x2 NOR cell array is given in Fig. 12. The contacts of C2 (Fig. 4) of adjacent cells are shared to reduce the area. The NOR design occupies 64 λ x 66 λ. We compare the 6T (4T-2Fe) FeFET CAM design from [2] (Fig. 4(b) in [2]) which doesn't use negative voltage in its operation. Our analysis shows 35% area improvement for our NAND design and 31% improvement for our NOR design compared to the CAM cell design in [2]. The area improvement is attributed to the elimination of two transistors and two global signal wires. Note that, in our design, we combine Bit line and Search line functions to the same signal wire which obviates the need for separate metal lines and reduces overall congestion.

Table 3: Simulation Parameters

Technology node		10nm	
Number of fins in FeFET	1	γ	6e11 m9/F/coul4
α	-1.05e9 m/F	ρ	0.1Ω-m
β	1e7 m5/F/coul2	Thickness of Ferro layer	9.3nm

Table 4: Energy delay comparison

	6T CAM[2]	4T NAND	4T NOR
Write delay(s)	1.82E-09	1.55E-09	1.52E-09
Avg Write Energy (J)	7.43E-15	1.07E-14	1.07E-14
Match Delay (s)	5.18E-11	3.79E-10	5.16E-11
Match Energy (J)	5.97E-15	1.03E-14	1.03E-14

Table 5: Common CAM Technologies.

Technology	Area	Storage mechanism	Different iability	Additional comments
SRAM	High, 16T tran[9]	Transistor ON/OFF	High	Volatile
STTRAM	Medium 9T-2MTJ [11]	Spin-polarized current applies torque on the magnetic moment	Low	Non volatile
ReRAM	Low 2T-2R[12]	Resistance change	Medium	Non volatile
FeFET	Low (6T)	Ferroelectric polarization	High	Non volatile
Proposed FeFET	Low (4T)	Ferroelectric polarization	High	Non volatile

3.2. Delay & Energy

We compare the delay energy numbers of our NAND/NOR CAM design against the 6T (4T-2Fe) CAM design in [2] (Fig. 4(b)) using the simulation parameters given in Table 3. We simulate a 64x8 array and write/match energy for a single cell is given in Table 4. We keep the ferro layer thickness at 9.3nm in all the designs to maintain nonvolatility. We observe write delay improvement of 14/16% for NAND and NOR designs at the cost of 1.45/1.45 times increased energy costs, respectively.

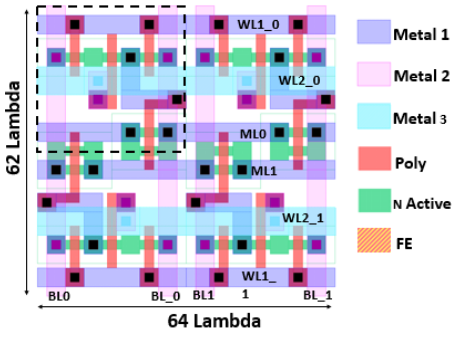


Figure 11: Layout of a 2x2 NAND cell

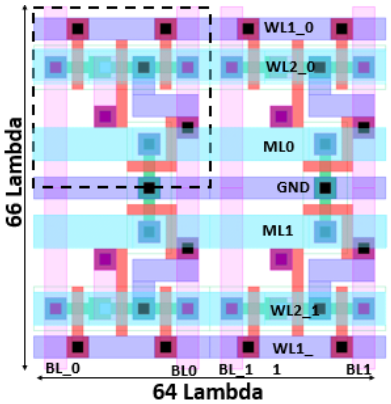


Figure 12: Layout of a 2x2 NOR cell

In the case of a match, we observe almost similar match time for the 6T design and our 4T NOR design (0.3% improvement). However, 4T NAND design incurs more delay (~7 times delay) compared to NOR design as the match occurs through the discharge path of sequentially connected NFETs in the ML line as shown in Fig. 2. Both 4T NAND and 4T NOR spends around 1.72/1.73 times energy compared to the 6T CAM-only design. Note that, in order to speed up the match time, we use an

increased V_B (Fig. 2), However, we need to use Half V_{DD} at WL1 in all rows to prevent bit upsets (Fig. 2). This increases Match energy. In Fig. 3, we show the match operation where we spend reduced energy with lower V_B for matching. Then we eliminate the need for Half V_{DD} in WL1 as we choose V_B value well below the write voltage of the cells.

The read delay in the RAM mode consists of two steps. In the first step, we write 1 to the F2 (Fig. 6), and in the second step, current sensing is done. The total delay is dominated by the write delay from the first part. The read delays from our experiment are $1.37e^{-09}$s and $1.26e^{-09}$s for the NAND and NOR versions, respectively.

4. Related Work

There have been a variety of CAM designs based on different devices. A glossary of standard designs is given in Table. 5. SRAM based designs use an increased number of transistors leading high to area consumption. Lower tunneling magnetoresistance ratio of MTJs and relatively smaller resistance ratios of ReRAMs coupled with 2 terminal device structure require additional circuitry for enhanced performance in CAMs [2][10][11][12] leading to exploration of FeFET designs. [2] and [10] compare FeFET based design (6T(4T-2FE)) with other NVM CAMs and show energy-delay benefit over MTJ/ReRAM based design. We propose a reconfigurable FeFET CAM–RAM having fewer transistors compared to the prior FeFET CAM designs. We also introduce new CAM features such as approximate compute in our design. In this section, we discuss prior FeFET-CAM designs.

[10] proposes a ternary NOR CAM with 6 transistors (4T-2FE). This design uses negative voltage for its write. [4] proposes a NAND CAM design based on FeFETs (5T) with the limitation of not having a ternary operation capability. [2] introduces a NOR TCAM design by eliminating negative voltage for writes which they term as "ws2" design. This ws2 CAM design also uses 6Ts (4T-2FE) and can conceivably support RAM functionality, though it has not been evaluated. [17] proposes an ultra-dense 2FeFET TCAM design based on a Multi-domain FeFET Model. [17] depends on the multi-domain effect to have CAM operations and uses high gate voltage swing to write 1 and 0. Since the bit line and sense line in [17] is connected to the gate of the transistors, RAM operation is not inherently supported in this design, which is true for most of the previous FeFET based CAM designs. Due to differences in modeling for multi-domain FeFETs, and in order to ensure a fair comparison, we only directly compare against the single-domain "ws2" design from [2] as it uses the same single domain mechanism to model the FeFET like our approach and also eliminates negative voltage in its design.

Our proposed array design inherently enables additional functionalities along with CAM function without adding any additional transistors or wiring inside the array. These approximate/RAM_CAM design will enable new applications to be executed without additional dedicated circuits making the overall design compact and application friendly by potentially eliminating the need for larger number of cycles for operations. One drawback of our design is the need to apply half VDD at write control lines (WL1 in Fig. 2&4) to increase the noise stability for faster searches leading to increased energy requirements.

5. Conclusion

We present a 4T (2T-2Fe) design and evaluation of a FeFET based T/CAM-RAM. Both the NAND and NOR variants of the CAM-RAM designs and approximate CAM-RAM design use only 4 transistors. We compare the performance of our design with a 6T (4T-2Fe) FeFET CAM design. Our NAND and NOR CAM designs shows 35% and 31% area improvements, respectively. We also show energy delay analysis for write and search operations, showing equivalent search speed with 1.73 times energy consumption and 16% faster write with 1.45 times energy for the proposed NOR CAM design.

References

[1] P. Chi, S. Li, et al. Prime: A novel processing-in-memory architecture for neural network computation in reram-based main memory. In 2016 ACM/IEEE 43rdAnnual International Symposium on Computer Architecture (ISCA), pages27–39, June 2016.

[2] X. Yin, X. Chen, M. Niemier and X. S. Hu. Ferroelectric fets-based nonvolatile logic-in-memory circuits. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 27(1):159–172, Jan 2019.

[3] S. George, MJ. Liao, et al. Mda cache: Caching for multi-dimensional-access memories. In Proceedings of the 51st Annual IEEE/ACM International Symposium on Microarchitecture, MICRO-51, pages 841–854, Piscat-away, NJ, USA, 2018. IEEE Press.

[4] X. Yin, et.al. Exploiting ferroelectric fets for low-power non-volatile logic-in-memory circuits. In 2016 IEEE/ACM International Conference on Computer-Aided Design, pages 1–8, Nov 2016.

[5] P. A. La Fratta and P. M. Kogge. Design enhancements for in-cache computations. In Workshop on Chip Multiprocessor Memory Systems and Interconnects, 2009.

[6] F. Duarte and S. Wong. Cache-based memory copy hardware accelerator for multicore systems. Computers, IEEE Transactions on, vol. 59, no. 11, 2010.

[7] S. Srinivasa, A. K. Ramanathan, et al. Robin: Monolithic-3d sram for enhanced robustness with in-memory computation support. IEEE Transactions on Circuits and Systems I: Regular Papers, 66 (7):2533–2545, July 2019.

[8] S. Beamer and M. Akgul. Design of a low power content addressable memory (cam). University of California, Berkeley, pages1–5, 2009.

[9] K. Pagiamtzis and A. Sheikholeslami. Content-addressable memory (cam) circuits and architectures: a tutorial and survey. IEEE Journal of Solid-State Circuits, 41(3):712–727, March 2006.

[10] X. Yin, M. Niemier and XS. Hu. Design and benchmarking of ferroelectric fet based tcam. In Design, Automation Test in Europe Conference Exhibition (DATE), 2017, pages 1444–1449, March 2017.

[11] S. Matsunaga et al. Design of a nine-transistor/two-magnetic-tunnel junction-cell-based low-energy nonvolatile ternary content-addressable memory. Japanese J. of Applied Physics, vol. 51, p. 02BM06, 2012

[12] J. Li et al. 1 mb 0.41 μm2 2t-2r cell nonvolatile tcam with two-bit encoding and clocked self-referenced sensing. JSSC, vol. 49, pp. 896–907, 2014.

[13] AI. Khan, et al. Negative capacitance in a ferroelectric capacitor. Nature materials 14.2 (2015): 182.

[14] D. Wang, et al. Ferroelectric transistor based non-volatile flip-flop. Proceedings of the 2016 ISLPD. ACM, 2016.

[15] S. George, K. Ma, et al. Nonvolatile memory design based on ferroelectric fets. In 2016 53nd ACM/EDAC/IEEE Design Automation Conference (DAC), pages 1–6, June 2016.

[16] A. Aziz, S. Ghosh, S. Datta and S. K. Gupta. Physics-based circuit-compatible spice model for ferroelectric transistors. IEEE Electron Device Letters, 37(6):805–808, June 2016.

[17] X. Yin, K. Ni, D. Reis, et al. An ultra-dense 2fefet tcam design based on a multi-domain fefet model. In IEEE Transactions on Circuits and Systems II: Express Briefs, 66(9):1577–1581, Sep. 2019.

[18] D. Patterson, T. Anderson, et al, A case for intelligent ram," Micro, IEEE, 1997.

[19] J Wang, H Li, C Chen, and C Yeh. An and type match-line scheme for energy-efficient content addressable memories. ISSCC, Feb 2005.