Design and Benchmarking of Ferroelectric FET based TCAM

Xunzhao Yin, Michael niemier and X. Sharon Hu
Department of Computer Science and Engineering, University of Notre Dame
Notre Dame, IN 46556, USA, Email: {xyin1, mniemier, shu}@nd.edu

***Abstract*—We consider how emerging transistor technologies, specifically ferroelectric field effect transistors (or FeFETs), can realize compact and energy efficient ternary content addressable memories (TCAMs). As Moore's Law-based performance scaling trends slow, and many computational tasks of interest are now more data-centric than compute-centric, researchers are looking to improve performance/save energy by integrating efficient and compact logic/processing elements into various levels of the memory hierarchy. Potential benefits include reduced I/O traffic, energy/delay from data transfers, etc. A TCAM is an example of a logic-in-memory element that is ubiquitous in routers, caches, databases, and even neural networks. Not surprisingly, researchers continue to study how emerging technologies could lead to improved TCAMs. Recent work has considered how non-volatile (NV) memory technologies (e.g., resistive random access memory (ReRAM) or magnetic tunnel junctions (MTJs)) could best be used to construct low energy, NV TCAMs. However, acceptable R$_{on}$-R$_{off}$ ratios and the two terminal nature of these devices introduce energy and area overheads. Due to hysteresis in a device's I-V curve, an FeFET-based NV TCAM, offers low area overhead, as well as search energies and search speeds that are superior to other TCAM designs (i.e., based on MTJ, ReRAM and CMOS in array- and architectural-level evaluations.)**

I. Introduction

It is becoming increasingly difficult for CMOS technology scaling to provide the high performance and energy efficiencies that emerging applications demand. Furthermore, information processing tasks related to data mining, scientific computing, video/image streaming, etc. will continue to stress the processor-memory hierarchy. To address these challenges, researchers are looking to emerging devices, innovative circuits and architectures, and combinations thereof.

In this paper, we study how emerging technologies can impact the performance, energy, and area efficiencies of ternary content addressable memories (TCAMs). TCAMs perform parallel searches for a given piece of data against a table of stored data, and return information as to whether a match occurred. TCAMs have obvious utility in networking hardware/applications – i.e., in routers, database search applications, and associative memories [1]. More recently, [2] has also proposed using TCAMs for more energy efficient, in memory data processing by reducing the amount of redundant data associated with traditional von-Neumann processing.

We are especially interested as to how ferroelectric field effect transistors (or FeFETs) that are compatible with current process technologies, can lead to more efficient TCAMs. While CMOS-based TCAMs have obviously been implemented, they frequently suffer from low density and high energy consumption when compared to static random access memories (SRAMs) [3], or dynamic random access memories (DRAMs). Researchers have also been investigating TCAM designs based on resistive RAM (ReRAM), and spin torque transfer RAM (STT-RAM) based on magnetic tunnel junctions (MTJs) [4], [5]. Both devices use high resistance states (HRS) and low resistance states (LRS) to encode binary states. However, these technologies face challenges. For example, spin-based memories may have low variable resistance (from 10Ωs to 100kΩs in general [6]), low HRS/LRS ratios, and two-terminal structures. This can lead to relatively high energy consumption and extra transistors for write operations and to maintain acceptable output swings.

We present/study an FeFET-based NV TCAM design that can offer superior energy/area efficiencies when compared to CMOS- [7], MTJ- [8], and ReRAM- [4] based TCAMs. Each cell consists of four transistors and two FeFETs (i.e., we have a 4T-2FeFET TCAM). Binary state is not stored by variable resistance. Rather, by tuning the thickness of the ferroelectric material at the gate, hysteresis can be introduced into a device's I-V characteristic allowing for a 1T storage element. FeFETs can also exhibit high on-off ratios (I_{ON}/I_{OFF}=10^6) and provide inherent gain as its structure is otherwise similar to a MOSFET. Only one access transistor per FeFET is needed to facilitate writes, and write/sense ciruitry can be reduced.

Building off of a preliminary FeFET TCAM circuit design from the authors of [9], we consider an FeFET TCAM in terms of its structure, operations and layout at the cell level, and evaluate TCAM arrays with varying word widths as well as row numbers against other TCAM arrays (i.e., based on CMOS, MTJs and ReRAM). We also examine the energy efficiency of FeFET-based TCAMs in the context of a enhanced GPU architecture which utilizes TCAMs as an associative memory [10]. Our results show that FeFET-based TCAM requires 42% less area than a conventional 16T CMOS-based TCAM with similar search energies and delays. Additionally, when studying TCAM arrays, an FeFET-based approach can offer a maximum benefit of 7.5X/149X in energy delay product (EDP) versus a ReRAM/MTJ design (assuming 64 rows). Other benefits will also be discussed.

II. Background

Below, we describe FeFETs and the device model, and review other TCAM designs based on NV memory technologies.

A. The FeFET device

The transistors that form the basis of our work are built by stacking a ferroelectric (FE) layer on the gate of a MOSFET as shown in Fig. 1(a). The equivalent circuit is also shown in Fig. 1(a), where the FE capacitance (C_{FE}) couples with the capacitance of the underlying MOSFET (C_{MOS}). Per [11] there is a negative change in polarization of the FE layer with

978-3-9815370-8-6/17/$31.00 ©2017 IEEE

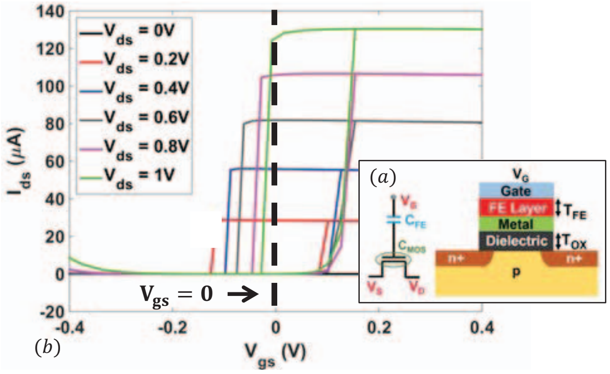


Fig. 1. (a) FeFET structure and its equivalent circuit representation showing ferroelectric capacitance (CFE) and the capacitance of the underlying MOSFET (CMOS). (b) FeFET I-V curves with tunable hysteresis (from [9]).

respect to the electric field, leading to negative capacitance ($C_{FE} < 0$). A large relative capacitance C_{FE}/C_{MOS} stabilizes the FE layer in the negative capacitance region and therefore the FE layer does not retain remnant polarization. This leads to a voltage step-up action in the device, which can result in steep-switching behavior (SS < 60mV/decade at room temperature) as well as higher ON-OFF current ratios than the standard MOSFETs (i.e., due to the inherent gain of the underlying MOSFET). This type of device is referred to as a negative capacitance FETs (NCFET), and is being explored by both academia and industry [12]–[14].

As FE layer thickness increases, and the C_{FE}/C_{MOS} ratio is sufficiently low, the polarization in the FE layer can be retained, leading to hysteretic behavior in an NCFET's transfer characteristic, and hence non-volatility. (An NCFET with hysteresis is an FeFET.) Per Fig. 1(b), device hysteresis can span over positive and negative gate-source voltages (V_{gs}), and remains at high or low current in the absence of a gate-source voltage ($V_{gs} = 0$). Per [15] electrostatic coupling between an FeFET's channel and drain on C_{FE} and C_{MOS} can alter the position and width of the hysteresis loop.

The I_{ON}/I_{OFF} ratio of FeFETs corresponding to the two logic states (I_{ON}, I_{OFF} represent logic '0', '1' respectively) can be up to 10^6 due to the inherent gain of the underlying MOSFET. This allows an FeFET to be used as a switch instead of a variable resistor. Also, the FeFET's three terminal structure separates the writing or polarization switching path (by applying sufficient positive/negative V_{gs} voltage) from the reading or state sensing path (via the drain-source current). This provides for more flexibility and less complexity in the design space when considering circuit/application-driven device optimization versus other two-terminal memory devices.

B. FeFET simulation models

We use a SPICE model for FeFETs based on time-dependent Landau Khalatnikov (LK) equations [16] that describes the polarization-electric field behavior of a FE layer.

$$E = \alpha P + \beta P^3 + \gamma P^5 + \rho P/dt \tag{1}$$

where α, β, γ are the static coefficients and ρ is the kinetic coefficient associated with the FE material, which in our model is calibrated to the experimental data on hafnium zirconium oxide (HZO). For HZO, $\alpha = -7 \times 10^9$ m/F, $\beta = 3.3 \times 10^{10}$ m^5/F/coul^2, $\gamma = -2 \times 10^9$ m^9/F/coul^4, and the FE thickness is 5.7 nm. The FeFET characteristics are obtained by combining self-consistent LK equations with the 45 nm predictive technology models [17]. Fig. 1(b) illustrates a representative, simulation-based example of an FeFET with tunable hysteresis given this model.

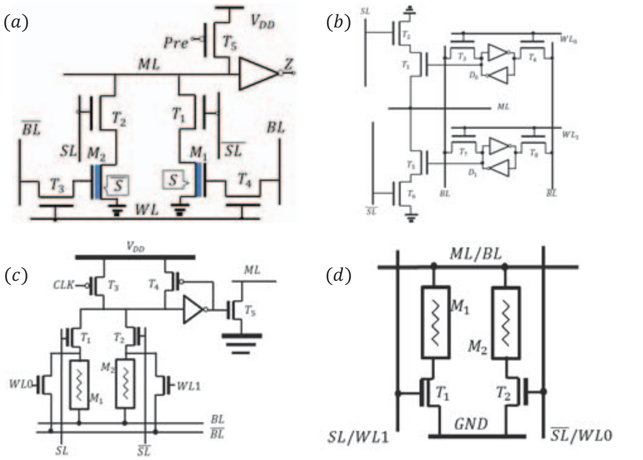


Fig. 2. TCAMs: (a) 4T-2FeFET (a precharge transistor and a sense amplifier (inverter) included); (b) 16T CMOS; (c) 9T-2MTJ MTJ; (d) 2T-2R ReRAM.

C. Related work: Other TCAM designs

Here, we review other TCAM designs. A NV, 4T-2MTJ cell circuit was proposed and fabricated [18], which consumes 40%/14% of the area of a 12T/16T SRAM-based TCAM (Fig. 2(b)) essentially due to the fact that the MTJs were placed on top of MOSFETs. Due to its small output swing stemming from an MTJ's low tunneling magnetoresistance ratio (TMR), a sense amplifier and separate access transistors are added into the above cell to achieve full output swing, resulting in a 9T-2MTJ TCAM design [8] (Fig. 2(c)). ReRAM-based TCAM designs have also been considered [4], [19], [20]. A 2T-2R TCAM (two transistors, two ReRAMs, Fig. 2(d)) has a compact structure and has been utilized as an example to overcome the existing challenges of degraded sensing margins caused by low HRS/LRS ratios [4]. This design has also been used in an enhanced GPU architecture [21].

In essence, MTJs, ReRAMs, etc. have different device characteristics when compared with FeFETs, and these characteristics introduce design limitations. First, MTJs/ReRAMs exhibit rather low resistance ratios (between 100% - 250% for MTJs and 10 - 10^4 for ReRAMs), and thus are used as variable resistors with weak drive capability. Low resistance values can also lead to leakage power. Second, these devices have two terminals, which can require additional transistors to facilitate a read/write operation. However, given that an FeFET has a sufficiently high resistance ratio (10^6) and three terminals (for separate writing/reading paths) it can serve as both a switch and a NV storage element. In subsequent sections we compare the FeFET-based TCAM design against the conventional 16T CMOS, 9T-2MTJ and 2T-2R TCAM designs to quantitatively capture the advantages of FeFETs.

III. FeFET TCAM Design

We now present the FeFET-based TCAM cell design that employs FeFETs as both a switch and a storage element.

We first review the cell structure at the circuit-level, and discuss how search and write operations are performed. We then present a layout of the FeFET TCAM cell, and use the layout to compare the area of an FeFET-based TCAM cell to TCAM cells based on CMOS/other emerging technologies. Finally, we discuss an array architecture that will be used to compare our work to other technologies.

A. *FeFET TCAM cell structure*

A FeFET-based TCAM design was briefly discussed in the context of other FeFET-based logic-in-memory (LiM) structures in [**?**]. Per Fig. 2(a), this FeFET design consists of two parallel FeFETs that are connected to the matchline (ML) via two transistors. The two FeFETs can both store logic '0' in addition to complementary bits, which facilitates the "don't care" state. In the cell schematic, transistors M_1/T_1 and M_2/T_2 serve as two pull down paths for the ML to ground. The inputs to the transistors T_1 and T_2 – SL and $\overline{SL}$ – together with the memory state stored in M_1 and M_2 (S and $\overline{S}$) determine whether the pull down paths are ON/OFF, and provide an XNOR output $\overline{S \oplus SL}$ at the ML.

B. *Search and write modes*

While search and write operations for the FeFET-based TCAM were briefly discussed in [9], no simulations and discussions for the claim of NV or write functionality were presented. Here, we briefly review write and search operations, and also show that the structure will indeed be NV.

To perform word-wise write operation, the wordline is activated for the word to be written (WL to V_{DD}), and write voltages are applied to bitlines ($BL/\overline{BL}$) per the input data – i.e., V_{write} (0.4V) for logic '1' and $-V_{write}$ for logic '0' – to switch the FE polarization within an FeFET. Negative V_{DD} is applied to wordlines of the words that are not to be written to ensure that the gate-source voltages of the access transistors in these words remain less than 0 or at 0 during the write, and no unexpected write occurs to these words. Searchlines $SL/\overline{SL}$ are driven to ground during the write to eliminate static current. The write time for one write operation is 0.57ns.

Fig. 3 shows the simulation waveforms of the FeFET-based TCAM cell. When there is a match, both pull-down paths are OFF, and the match line ML is not discharged and stays high. When there is a mismatch, at least one of the pull-down paths is ON, resulting in the discharge of the ML. In the '$don't$ $care$' state the ML always stays high regardless of the input data. To illustrate that (i) FeFETs can retain state and (ii) the TCAM cell still functions as intended given a power supply interruption, we periodically set $V_{DD} = 0$ in our simulations. In all cases the cell functions exactly as intended/as it did before the power supply interruption.

C. *Layout and area*

To determine whether or not FeFET-based TCAMs can truly be competitive with functional equivalents based on CMOS and/or other emerging technologies, it is necessary to make "apples-to-apples" comparisons to other approaches in terms of area, latency, and energy. For the area metric, we completed a layout of the FeFET-based approach. The FeFET design requires 6 transistors per TCAM cell, and an FeFET also requires similar area as compared to a conventional MOSFET (see Fig. 1(a)). The layout of a 2X2 TCAM (that follows MOSIS Scalable CMOS design rules for 45 nm – i.e., SCMOS DEEP rules) is shown in Fig. 4.

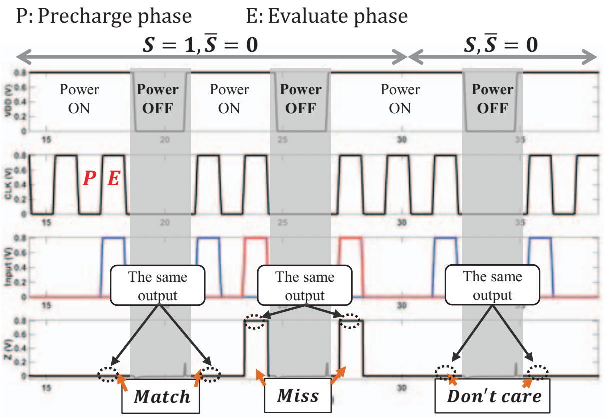


Fig. 3. FeFET-based TCAM cell simulation waveforms.

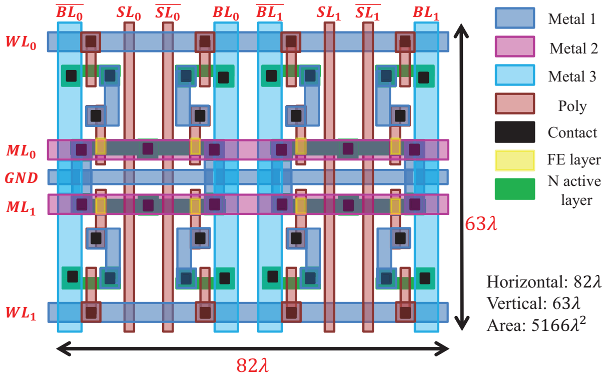


Fig. 4. 2X2 TCAM cell layout. Note that λ represents half feature size F

Fig. 5 compares the FeFET TCAM cell size with other TCAM designs from the literature. Based on the "push rule" SRAM scaling trend [22], [23] – i.e., $124F^2$ at 65 nm and $171F^2$ at 45 nm – the area of a 16T CMOS TCAM is projected to be $1.12\mu m^2$. Based on the layout in Fig. 4, the FeFET-based TCAM's cell size is estimated to be 58% of that of 16T CMOS design. When compared to other emerging technologies (e.g., ReRAM), besides relatively large area of MTJ-based TCAMs, we expect ReRAM-based TCAMs to have slightly lower areas due to reduced transistor counts (i.e. 2T-2R TCAMs) when compared to the FeFET design point in Fig. 5. However, FeFET designs can be superior in terms of energy and delay.

D. *TCAM array architecture*

For energy/delay analysis, we evaluate all technologies in the context of the same TCAM array architecture (Fig. 6). The array consists of the TCAM core, the input buffer/drivers, the output sense amplifier (SA – an inverter in our case), the clock signal and the output encoder. The TCAM core contains M words with a word width of N bits. The match and word lines (MLs and WLs) are placed horizontally, while search and bit lines ($SL/\overline{SL}$s and $BL/\overline{BL}$s) are placed vertically within the TCAM cell grid. The search and bit lines are driven by the input buffer and at the end of each match line, a sense amplifier detects the voltage of the match line, and outputs the indicator

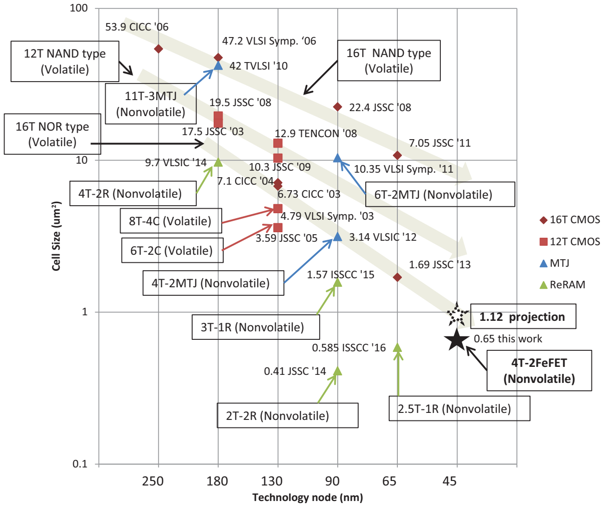


Fig. 5. Comparisons of TCAM cell sizes. The CMOS TCAM area projection is based on the scaling trend of push-rule SRAM according to ISSCC trends [23] since CMOS TCAM uses two 6T SRAMs plus 4 transistors.

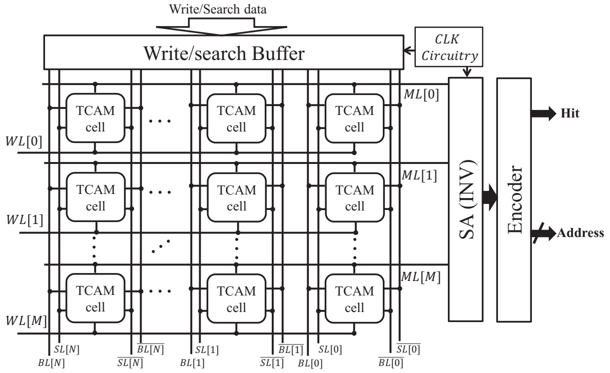


Fig. 6. Architecture of an $M \times N$ TCAM array

of match/mismatch to the encoder, which sends a "hit" signal and the corresponding address of the matched entry.

IV. Evaluation

In this section, we present a performance and energy study of our FeFET-based TCAM array, and compare it with alternative TCAMs based on CMOS, MTJ and ReRAM technologies. We will use a TCAM-based GPU processor [24] as a case study to evaluate the energy efficiency of these TCAMs in the context of an associative memory-based computing system.

A. Evaluation setup

As noted above, we replace the TCAM cell in Fig. 6 with technology specific designs. All delay/energy evaluations are conducted via HSPICE simulation. A simple inverter-based SA is used for all TCAMs. Whenever possible, the same or similar technology nodes are assumed. Device parameters associated with other technologies are obtained from published work.

We compare four different TCAM designs: CMOS, FeFET, ReRAM and MTJ, with the last three being non-volatile. For FeFET TCAM, the FeFET model discussed in Sec. II-B as well as a 45 nm PTM model [17] are used. We assume the minimum sized transistors for the TCAM cell and SA in order to reduce power. For the conventional CMOS-based TCAM (Fig. 2(b)), we use the same 45nm PTM model and minimum transistor sizes that were used for the FeFETs. For the ReRAM-based TCAM, we adopt the 2T-2R design (Fig. 2(d)), a common example used in the literature [4]. We assume 2MΩ for HRS and 20kΩ for LRS [25] to implement the ReRAM-based TCAM. For MTJ-based TCAMs, we employ the 9T-2MTJ TCAM which uses a single-end sense amplifier and a single pass transistor to achieve full swing output per Fig. 2(c) [8]. Though the dual-rail TCAM cell in [26] eliminates the static current that always exists in MTJ-based TCAM designs, the output feedback that cuts the conducting path leads to a significant increase in search delay (>1 ns), which is not applicable for the enhanced GPU architecture which operates at 1 ns cycle time. We assume MTJs with a parallel resistance (R_p) of 3kΩ, and a magnetoresistance ratio (MR) of 120% [27] in the 9T-2MTJ TCAM. Note that while the TMR of an MTJ can be as high as 500% [28], it effectively will not influence array evaluation data.

B. TCAM array evaluation

Here we summarize the delay and energy comparisons for the 4 different TCAMs assuming a 64-bit word with different numbers of rows. (We choose a 64-bit word as it is of sufficient size for many applications such as network switches and routers [29].) Fig. 7 shows the search delays of 64-bit TCAMs based on different technologies at different sizes. The delay is measured for the worst case, where only one bit is mismatched. At small sizes, the MTJ-based TCAM had the lowest delay due to the small load capacitance (one transistor per bit) at the match line. However, at large sizes, the buffer delay which is used to distribute the input across the array increases, and the in-cell sense amplifier slows down, resulting in a rapidly growing total delay versus other TCAM arrays. The reason that the FeFET-based TCAM is faster than CMOS- and ReRAM-based TCAMs is that the FeFET has a larger I_{ON} as well as a better I_{ON}/I_{OFF} ratio, which leads to a larger discharging current upon a mismatch. For application-level case studies, our TCAM designs can work with the original GPU execution as the delays are smaller than the nominal cycle time.

The total search energy per operation consists of two parts: the buffer energy and the cell energy. As TCAM size increases, the buffer sizes grow as well to drive the large TCAM array, thus the buffer energy increases too. The cell energy depends on the schematics of the TCAM designs. For FeFET- and ReRAM-based TCAMs, the cell consumes precharging energy; For CMOS- and MTJ-based TCAMs, the cell consumes precharging energy plus static energy that comes from the leakage of SRAM and conducting current associated with constantly ON paths. Fig. 8 shows the 64-bit TCAM search energies for different TCAMs. Note that MTJ-based TCAM is always conducting large static current due to its schematic and low resistance values. (For clarity, it is not included in Fig. 8.) FeFET-based TCAMs have similar latencies and energies when compared to CMOS-based TCAMs as they have similar capacitance at the match line and search lines. (They are also denser and non-volatile.) When comparing with ReRAM- and MTJ-based TCAMs, FeFET-based TCAM have EDPs that are 2.8X better (1-row) to 7.5X better (64-row) and 14X (1-row) better to 149X (64-row) better for ReRAM and MTJ-based designs respectively, as shown in Fig. 8.

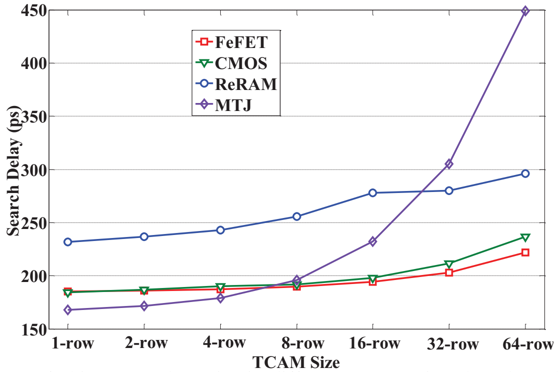


Fig. 7. 64-bit TCAM latencies in different TCAM sizes based on different technologies.

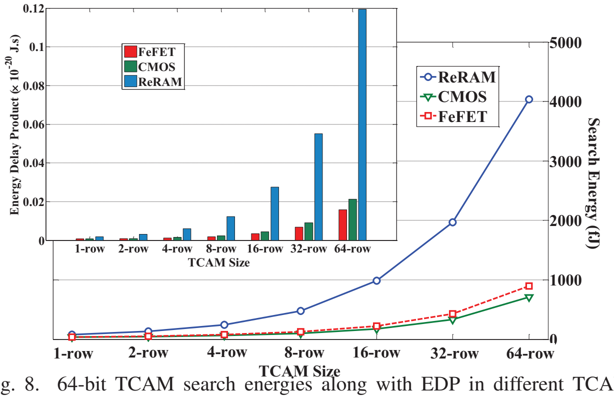


Fig. 8. 64-bit TCAM search energies along with EDP in different TCAM sizes based on different technologies.

C. Case study of a TCAM-based GPU implementation

To further demonstrate the benefit of improved TCAMs, we evaluate the TCAM-based associative memory employed in an AMD Southern Island GPU device for energy reduction in the context of GP-GPU applications. The TCAM-based GPU architecture introduced in [24] integrates a TCAM array with each of the four main floating point units (FPUs) as shown in Fig. 9. Several OpenCL applications including 3 image processing and 3 general applications are run on the GPU platform, the data for the applications are from Caltech 101 dataset [30], being partially trained (10%) and totally tested (100%) respectively. The TCAM arrays store the pre-trained data as the application starts. The system sends the input operands to both the FPU and TCAM block simultaneously. If a match happens, the TCAM array disables the corresponding FPU computation, and provides the actual result. With low-power TCAM designs, such a TCAM-based GPU architecture can provide unique advantages in many energy-conscious applications ranging from mobile devices to data centers.

We compare the four TCAM arrays discussed in Sec. IV-B in this GPU architecture. The evaluation process adopted here follows that proposed in [31]. Details are omitted due to page limit. The TCAM arrays use the minimum transistor sizes for buffer, cell and SA to save energy while satisfying the basic search functionality within the same cycle time of the FPUs. We assume 32-bit words for SQRT, 64-bit words for ADD/MUL and 96-bit words for MAC, respectively, as they require varying numbers of input operands.

For architecture level evaluation, we need to evaluate the average energy consumption per operation for all the FPUs and TCAMs at different sizes. Table I summarizes the energy consumptions of TCAMs based on the four technologies. The individual FPU energy consumptions are obtained from the synthesized 6-stage FPU design [24]. From the table it can be observed that MTJ-based TCAMs consume much more energy than even the individual FPU energy especially for large row numbers. Thus, we will not show the MTJ-based TCAM plot in later discussions. For the other three TCAM designs, their energy numbers suggest the potential for improved energy efficiency when they are integrated in the enhanced GPU architecture. Using ADD as an example, FeFET-, CMOS- and ReRAM-based TCAMs achieve 75X, 62X and 19X energy efficiency at 4 rows compared with the corresponding FPU energy number. It is also shown that FeFET-based TCAMs consume the least energy among all designs due to FeFET's high I_{ON} and high distinguishability.

Fig. 10 shows the normalized total energy consumption of the GPU with different TCAM sizes for representative applications. All the energy curves have a similar energy trend – at small TCAM sizes, the total energy consumption of the enhanced GPU architecture decreases as TCAM size increases, as more frequently referenced input patterns can be pre-stored in the TCAM, and higher hit rates lead to fewer FPU operations. After reaching the minimum energy points, increasing TCAM size does not improve the hit rate enough

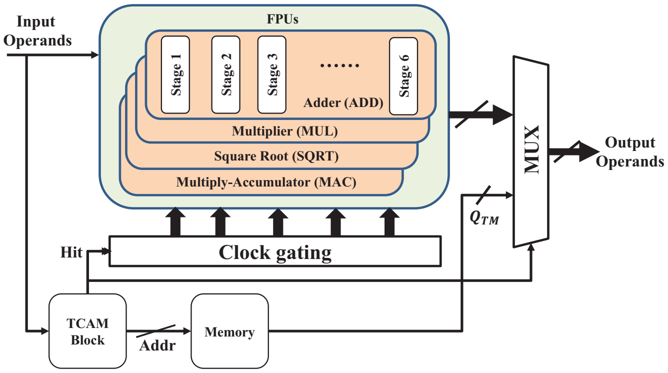
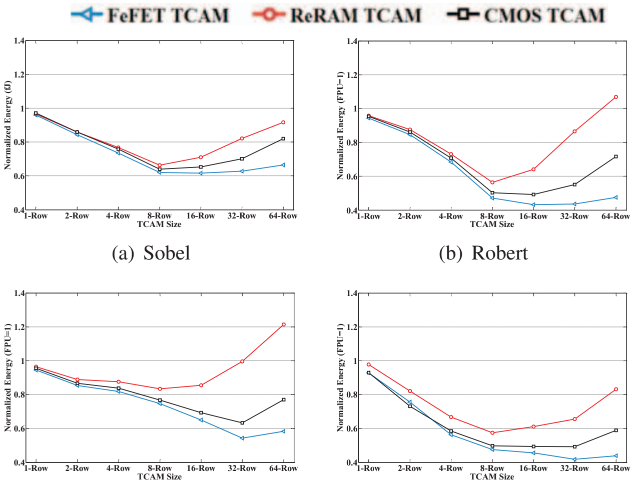


Fig. 9. The framework integrating FPUs with TCAM as associative memory systems. The framework originates from [31].

TABLE I
ENERGY CONSUMPTION (FJ) PER OPERATION FOR FPUS AND TCAM ARRAYS

Module	FPU (1.0V)	device	TCAM⋆		
			4-row	16-row	64-row
ADD (64-bit)	4742	FeFET	63.6	175.6	714.0
		CMOS	75.9	223.8	895.8
MUL (64-bit)	9891	MTJ	2149	9368	52488
		ReRAM	244.5	987.5	4034
SQRT (32-bit)	9983	FeFET	33.4	95.8	442.7
		CMOS	39.5	120.5	524.5
		MTJ	1078	4699	26337
		ReRAM	122.4	491.5	2006
MAC (96-bit)	12051	FeFET	93.5	255.0	984.7
		CMOS	110.8	321.7	1234
		MTJ	3220	14039	78623
		ReRAM	343.0	1300	6060

⋆: 3 row numbers out of 7 are shown, others are omitted for space.



(a) Sobel

(b) Robert

(c) DwHaar1D

(d) QuasiRandom

Fig. 10. Normalized energy consumption of enhanced GPU integrating different technology-based TCAMs in different sizes.

to compensate for higher energy consumption associated with TCAMs, and the total energy starts to increase. Because of this we omitted the data of larger (>64-row) sizes From the figure we conclude that FeFET-based TCAM can achieve more energy efficiency than CMOS- and ReRAM-based TCAMs in the enhanced GPU architecture. On average, FeFET-based TCAM achieves 45% energy savings over the six applications, while the energy savings of CMOS- and ReRAM-based TCAM are 39% and 32% respectively.

V. Conclusion

We exploited the unique properties of FeFET devices to build and evaluate a compact and energy efficient TCAM array. Due to the FeFET's three terminal structure, high on current and high I_{ON}/I_{OFF} ratio, the design requires 42% less area overhead than a CMOS-based TCAM, and requires less energy consumption compared with other existing NV TCAM designs. We further examined the energy efficiency of FeFET-based TCAMs over other designs in the context of a TCAM array as well as an enhanced GPU architecture where TCAMs are used as the associative memory. From the delay and energy plots we observe that FeFET-based TCAM achieves better energy-delay efficiency than other emerging technology-based TCAM designs (i.e. 7.5X over ReRAM-based TCAM and 149X over MTJ-based) which indicates potential benefit in TCAM related applications such as network routers and switches. As just seen above, improvements also extend to the GPU architecture. Overall, FeFET-based TCAMs are a promising candidate for next generation electronics.

Acknowledgment

This work was supported in part by the Center for Low Energy Systems Technology (LEAST), one of six SRC STARnet centers sponsored by MARCO and DARPA. The authors would like to thank Mohsen Imani of University of California, San Diego for helpful discussions.

References

[1] R. Karam *et al.*, "Emerging trends in design and applications of memory-based computing and content-addressable memories," *Proceedings of the IEEE*, vol. 103, pp. 1311–1330, 2015.

[2] T. Kohonen, *Associative memory: A system-theoretical approach*. Springer Science & Business Media, 2012, vol. 17.

[3] M. Imani *et al.*, "Hierarchical design of robust and low data dependent finfet based sram array," in *NANOARCH*. IEEE, 2015, pp. 63–68.

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

[5] L. Xue *et al.*, "Odesy: a novel 3t-3mtj cell design with optimized area density, scalability and latency," in *ICCAD*. ACM, 2016, p. 118.

[6] S. Yu *et al.*, "A neuromorphic visual system using rram synaptic devices with sub-pj energy and tolerance to variability: Experimental characterization and large-scale modeling," in *IEDM*. IEEE, 2012, pp. 10–4.

[7] K. Pagiamtzis and A. Sheikholeslami, "Content-addressable memory (cam) circuits and architectures: A tutorial and survey," *JSSC*, vol. 41, pp. 712–727, 2006.

[8] 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.

[9] X. Yin *et al.*, "Exploiting ferroelectric fets for low-power non-volatile logic-in-memory circuits," in *ICCAD*. ACM, 2016, p. 121.

[10] M. Imani *et al.*, "Resistive configurable associative memory for approximate computing," in *DATE*. IEEE, 2016, pp. 1327–1332.

[11] S. Salahuddin and S. Datta, "Use of negative capacitance to provide voltage amplification for low power nanoscale devices," *Nano Letters*, vol. 8, pp. 405–410, 2008.

[12] S. George *et al.*, "Nonvolatile memory design based on ferroelectric fets," in *DAC*. ACM, 2016, p. 118.

[13] ——, "Device circuit co design of fefet based logic for low voltage processors," in *ISVLSI*. IEEE, 2016, pp. 649–654.

[14] E. Times, "Finfet's father forecasts future," 2016 (April 1). [Online]. Available: http://www.eetimes.com/document.asp?doc_id=1329333

[15] A. I. Khan, "Negative Capacitance for Ultra-low Power Computing," Ph.D. dissertation, University of California at Berkeley, 2015.

[16] T. Song, "Landau-khalatnikov simulations for ferroelectric switching in ferroelectric random access memory application," *Journal of the Korean Physical Society*, vol. 46, pp. 5–9, 2005.

[17] R. Vattikonda *et al.*, "Modeling and minimization of PMOS NBTI effect for robust nanometer design [Online]http://ptm.asu.edu/," in *DAC*, 2006, pp. 1047–1052.

[18] S. Matsunaga *et al.*, "A 3.14 um 2 4t-2mtj-cell fully parallel tcam based on nonvolatile logic-in-memory architecture," in *VLSIC*. IEEE, 2012, pp. 44–45.

[19] M.-F. Chang *et al.*, "A 3t1r nonvolatile tcam using mlc reram with sub-1ns search time," in *2015 ISSCC DIGEST OF TECHNICAL PAPERS (ISSCC)*, vol. 58, 2015, pp. 318–U449.

[20] C.-C. Lin *et al.*, "7.4 a 256b-wordlength reram-based tcam with 1ns search-time and 14 improvement in wordlength-energyefficiency-density product using 2.5 t1r cell," in *ISSCC*. IEEE, 2016, pp. 136–137.

[21] M. Imani *et al.*, "Masc: Ultra-low energy multiple-access single-charge tcam for approximate computing," in *DATE*. IEEE, 2016, pp. 373–378.

[22] S. Jeloka *et al.*, "A 28 nm configurable memory (tcam/bcam/sram) using push-rule 6t bit cell enabling logic-in-memory," *JSSC*, vol. 51, pp. 1009–1021, 2016.

[23] S. G. Narendra *et al.*, "Through the looking glass? the 2015 edition: Trends in solid-state circuits from isscc," *ISSC*, vol. 7, pp. 14–24, 2015.

[24] A. Rahimi *et al.*, "Approximate associative memristive memory for energy-efficient gpus," in *DATE*. IEEE, 2015, pp. 1497–1502.

[25] M. A. Lastras-Montano *et al.*, "Architecting energy efficient crossbar-based memristive random-access memories," in *NANOARCH*. IEEE, 2015, pp. 1–6.

[26] T. Hanyu *et al.*, "Spintronics-based nonvolatile logic-in-memory architecture towards an ultra-low-power and highly reliable vlsi computing paradigm," in *DATE*. IEEE, 2015, pp. 1006–1011.

[27] C. Lin *et al.*, "45nm low power cmos logic compatible embedded stt mram utilizing a reverse-connection 1t/1mtj cell," in *IEDM*. IEEE, 2009, pp. 1–4.

[28] Y. Lee *et al.*, "Effect of electrode composition on the tunnel magnetoresistance of pseudo-spin-valve magnetic tunnel junction with a mgo tunnel barrier," *Applied Physics Letters*, vol. 90, p. 2507, 2007.

[29] H. J. Chao and B. Liu, *High Performance Switches and Routers*. Wiley-IEEE Press, 2007.

[30] http://www.vision.caltech.edu/Image_Datasets/Caltech1.

[31] M. Imani *et al.*, "Approximate computing using multiple-access single-charge associative memory," 2016.