Computational Associative Memory Based on Monolithically Integrated Metal-Oxide Thin Film Transistors for Update-Frequent Search Applications

Zijian Zhao[1*], Jorge Gomez[2*], Huacheng Ye[2], Mohsen Imani[3], Xunzhao Yin[4], Shan Deng[1], Bryan Melanson[1], Jing Zhang[1], Xiao Gong[5], Angel Abusleme[6], Suman Datta[2], and Kai Ni[1]

[1]Rochester Institute of Technology, Rochester, NY, USA; [2]University of Notre Dame, Notre Dame, IN, USA; [3]University of California, Irvine, CA, USA; [4]Zhejiang University, Hangzhou, China; [5]National University of Singapore, Singapore; [6]Pontificia Universidad Católica de Chile, Santiago, Chile. *Equal contribution; email: kai.ni@rit.edu

***Abstract*—Associative memory based on ternary content addressable memory (TCAM) is typically used in a static inference mode where the update is occasional. However, for update-frequent associative search applications (e.g., clustering), existing TCAM designs have a fundamental gap between low-density, high write performance SRAM, and high-density, poor write performance nonvolatile memories. In this work, we demonstrate: i) monolithic 3D TCAM designs based on thin film transistors (TFT) that can simultaneously achieve high density and excellent write performance, thus bridging the performance gap among existing TCAM designs for update-frequent search applications; ii) the necessity of 6T TFT TCAM design to restore the desired independence among the TCAM words, which is destroyed in the 4T TFT TCAM design; iii) excellent write performance with logic-compatible write voltage (<1.5V), <20ns write latency, >10^{10} endurance; iv) up to 14x/35x improvement in speed/energy over GPU in executing the K-Means clustering algorithm, showing great promise of the TFT TCAM.**

I. Introduction

Computational associative memory, notably TCAM, is gaining popularity in accelerating pattern matching tasks because it can search across the whole memory in parallel for matched entries directly in-memory. In addition, its capability of computing the hamming distance enables its application as a distance kernel to accelerate various kinds of machine learning applications [1]. Among many of them, the TCAM works in the static inference mode, where it is mostly used as a distance kernel with only once or occasionally update (Fig.1(a)). This opens up a large design space of TCAM by exploiting the dense nonvolatile memories, e.g., resistive random-access memory (ReRAM) and ferroelectric FET (FeFET) [1], while avoiding their drawbacks of poor write performance (e.g., limited write endurance, high write voltage and latency).

However, there are many other important applications, such as unsupervised clustering (Fig.1(c)), where frequent update to the stored entries in the distance kernel is a must while the long term retention can be relaxed [2]. As a rule of thumb, we estimated that about 0.13 update is required per search operation with the classical K-means clustering algorithm. Given a large dataset and many iterations required (Fig.1(b), (c)), the required updates could exceed the endurance limits of ReRAM or FeFET. SRAM based TCAM, with almost infinite endurance and excellent write performance, provides an alternative solution. However, each SRAM TCAM consumes 16 transistors, significantly limiting the TCAM capacity to accommodate big data (Fig.2(d)). In this work, we propose new TCAM designs based on the TFT gain-cell embedded DRAM (eDRAM) (Fig.2(b) and (c)), which has great write performance and high memory density.

Two transistor gain cell eDRAM, by utilizing a TFT as the access device, significantly extends the retention time over its silicon counterpart by harnessing the ultralow leakage current in the TFT [3]. For update-frequent search applications, retention is required only up to the next update, which greatly benefit TFT eDRAM. A great advantage of metal oxide TFT is that it can be integrated in the back-end-of-line (BEOL) in a multi-tier stack, thereby significantly increasing the memory capacity (Fig.2(a)) [3]. Utilizing the TFT eDRAM, monolithic 3D TCAM is also enabled. Together with its excellent write performance, TFT TCAM could be an excellent candidate for update-frequent associative search applications.

II. IWO TFT & 4T TFT TCAM Demonstration

TFT with tungsten (W) doped In_2O_3 (IWO) channel is fabricated at less than 250°C, well within the BEOL thermal budget for monolithic integration (Fig.3(a)). It features a 20 nm thick Palladium (Pd) bottom gate, 10nm HfO_2 gate dielectric through atomic layer deposition (ALD) at 250°C, and about 3 nm IWO channel by RF magnetron sputtering in the presence of 0.02 Pa excess O_2 at room temperature. Fig.3(b) shows the scanning electron microscopy (SEM) images of a TFT with L_G of 50nm. The transmission electron microscopy (TEM) (Fig.3(c)) and energy-dispersive X-ray (EDX) elemental mapping (Fig.3(d)) show the 10nm HfO_2 and 3nm IWO in the gate stack. Fig.3(e) and (f) show the measured I_D-V_G and I_D-V_D characteristics for a TFT with W_G/L_G=100nm/50nm, exhibiting about 150μA/μm at 1.5V, well suited for high-speed write and read operations. The TFT characteristics are used to calibrate the virtual source transistor model [4], which will be used for the TCAM design exploration and performance evaluation.

For the proposed TFT TCAM design, the 4T cell is first introduced, its critical issue of destroying the independence among TCAM words is identified, and then a 6T cell is proposed to address this issue. The 4T TCAM cell operation is shown in Fig.4(a) and (b), corresponding to the match and mismatch conditions, respectively. Complementary storage node voltages are written into the two branches and then search pulses are applied such that the match line (ML) current (I_{ML}) is low/high when the search matches/mismatches the stored information. As an example, bit '0' is stored as V_L and V_H on the left and right branches and a matched bit '0' search applies

2021 IEEE International Electron Devices Meeting (IEDM) | 978-1-6654-2572-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/IEDM19574.2021.9720705

a high/low voltage on the search line (SL) on the V_L/V_H branch, respectively (Fig.4(a)). In this way, I_{ML} is negligible because the ON state TFT has a zero V_{DS} and a high V_{DS} is applied to the OFF state TFT. If instead a bit '1' is stored, the storage node voltages become V_H and V_L on the left and right branch, respectively (Fig.4(b)). Applying the same search bit '0' will cause a high I_{ML} flowing through the ON TFT as both its V_{GS} and V_{DS} are high. Therefore, sensing the I_{ML} allows to detect whether match or mismatch happens. To verify its operation, a 4T TCAM cell, composed of 2 eDRAM cells, is fabricated (Fig.4(c) and (d)) and tested. The bit '0'/'1' is written into the TCAM cell (Fig.4(e)/(f)) by applying 0V (V_L)/1.5V (V_H) on the bit line (BL) and activating the word line (WL), then a search bit '0' applies a 1.5V on the SL. Measurement clearly shows the low/high I_{ML} for the match/mismatch case, demonstrating its successful operation. Note that the 'don't care' case in TCAM is also realized by storing both branches a low node voltage such that the I_{ML} remains low, irrespective of the search.

III. The Issue of 4T TFT TCAM & The Solution of 6T TFT TCAM

In the 4T TCAM design, the TCAM cells on the drain connected SL draw current from the SL along its propagation. Due to the presence of the interconnect metal wire resistance R_i and the SL driver launch resistance R_L [5], [6], the SL voltage (V_{SL}) in the last TCAM word will fluctuate between the best scenario, where all the cells on the SL match, and the worst scenario, where all the cells mismatch (Fig.5(a)). Fig.5(b) and (c) simulates the V_{SL} at the last TCAM word at different R_L and R_i for two scenarios: 1 mismatch and 64 mismatches. In the former case, due to small I_{ML}, the V_{SL} can still correctly propagate to the last word, despite the wide range of parasitic resistances (Fig.5(b)). However, when all the cells mismatch, R_L and R_i cause significant V_{SL} degradation (Fig.5(c)). The impact of the SL length on the V_{SL} (Fig.5(d)) and the cell I_{ML} (Fig.5(e)) shows that longer SL causes more V_{SL} degradation.

That the V_{SL} depends on the cell matching conditions on the same SL destroys the independence among TCAM words. Depending on the number of mismatches ($N_{Mismatch}$) along the SL, the V_{SL} at the last word can fluctuate between 1V and 0.5V (Fig.6(a)) and the I_{ML} follows the same trend (Fig.6(b)). The V_{SL} fluctuation on each column will cause hamming distance sensing almost impossible because each TCAM cell on the ML has a different current. On a 64x64 TCAM array, I_{ML} for 6 bits mismatch and 7 bits mismatch overlaps (Fig.6(c)) and becomes worse for more mismatched bits (Fig.6(d)). With a longer SL (128x64 in Fig.6(e)), even 1 bit mismatch overlaps. Correct operation can only be realized in a short SL (8x64 in Fig.6(f)).

The solution is to augment each 4T TCAM branch with a TFT and apply the SL on the gate (Fig.7(a) and (b)). Similar to the 4T TCAM case, successful TCAM operations can be realized through information encoding such that upon a match, 1 TFT on both sides is cut off, causing a low I_{ML} (Fig.7(a)). Otherwise, when mismatch happens, 2 TFTs on the same side are turned ON, causing a large I_{ML} (Fig.7(b)). Experimental verification (Fig.7(d) and (e)) on fabricated 6T TCAM cell (Fig.7(c)) exhibits a low/high I_{ML} for the match/mismatch case shown in Fig.7(a)/(b), respectively. With the SL connected to a TFT gate (Fig.8(a)), the V_{SL} can propagate without degradation through any practical word length because of the ultra-high gate resistance. As shown in Fig.8(b), the V_{SL} at the last word is constant, irrespective of the $N_{Mismatch}$ along the SL, unlike the 4T TCAM case (Fig.6(a)). With this cell, the hamming distance can be correctly detected in a 64x64 TCAM array (Fig.8(c) and (d)) and even a 256x64 TCAM array (Fig.8(e)). Fig.8(f) compares the Hamming distance that can be correctly detected as a function of SL length. As expected, 6T TCAM is independent of SL length while 4T TCAM significantly degrades with the SL length.

IV. Performance Benchmarking of TFT TCAM

To evaluate the write performance of TFT TCAM, the write speed (Fig.9(a) and (b)) is characterized on a TFT eDRAM cell (Fig.9(c)). It shows that with logic-compatible 1.5V, 20ns (instrument limitation), a 5μA read current can be obtained, demonstrating great write performance. The retention of the intrinsic eDRAM cell, i.e., no external storage capacitor, is measured as a function of hold voltage, V_{hold} (Fig.9(d)) and temperature (Fig.9(e)). Longer retention up to 1000s can be obtained by reducing V_{hold} to cut off the leakage current of the access transistor. Temperature dependence shows an activation energy of 0.27eV for the retention (Fig.9(f)).

The write performance is also predicted for scaled write transistor width (experimentally 5μm) and load capacitance on a 64x64 TCAM array. It shows around 10ns write latency (Fig.10(a)) and excellent write energy of 1-4 fJ/word/bit (Fig.10(b)). The search performance is evaluated as a function of the R_L and the metal wire parasitic capacitance (C_i). The delay (Fig.11(a)) and energy (Fig.11(b)) increase with the capacitance. Fig.11(c) presents a performance comparison of TCAM arrays using different technologies [7]. It is apparent that TFT based TCAM shows a logic-compatible write voltage (1.5V), high write speed (20ns), high endurance (10^{10} cycles), good search performance, making it a leading candidate for update-frequent search applications.

The usage of TCAM in accelerating the K-Means algorithm is shown in Fig.12(a). The original real value datapoints for clustering are first converted to binary high dimensional vectors, which allow to harness the Hamming distance calculation capability of TCAM. To implement K-Means algorithm, the clusters centers are stored in TCAM and then datapoints are searched through the cluster centers stored in the TCAM array. The minimum distance is identified, based on which the cluster centers need to be updated in the TCAM array. System level benchmarking shows that TFT TCAM design provides on average the second highest speedup over GPU (14x), only second to SRAM TCAM and on average highest energy saving (35x) considering the overall search and update applications.

V. Conclusions

In summary, monolithic 3D TFT TCAMs are proposed for update-frequent associative search applications. Two designs are evaluated and 6T TFT TCAM design are robust against the search voltage degradation, a serious issue in 4T TFT TCAM design. Excellent write performance with logic-compatible write voltage (<1.5V), high speed (<20ns), and high endurance (>10^{10} cycles) are demonstrated. Together with its high density through 3D stacking, TFT TCAM can fill in the gap of existing TCAM designs for update-frequent search applications.

References

[1] K. Ni et al., Nature Electronics 2019. [2] D. Singh et al., J. Big Data 2015. [3] H. Ye et al., IEDM 2020. [4] W. Chakraborty et al., VLSI Symp. 2020. [5] H. Lee et al., IEEE Magneticc Lett. 2017. [6] B. Gopireddy et al., ISCA 2019. [7] X. Yin et al., IEEE TCAS II 2019.

Metal Oxide TFT eDRAM Based Content Addressable Memory for Update-Frequent Associative Search

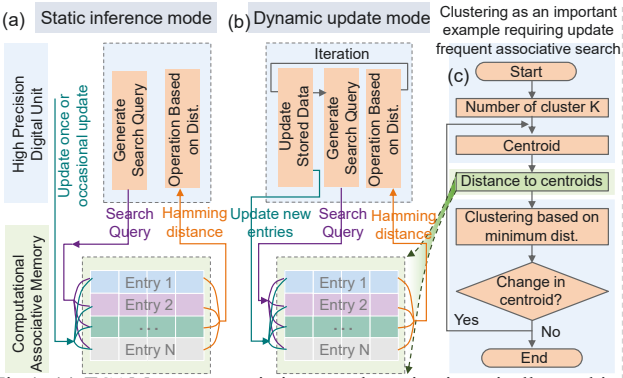


Fig.1. (a) TCAM as an associative search engine is typically used in a static inference mode without frequent content update. (b) Being able to calculate Hamming distance while also dynamically updating content is desirable for important applications such as (c) clustering, which uses TCAM as a distance kernel in a iterative loop.

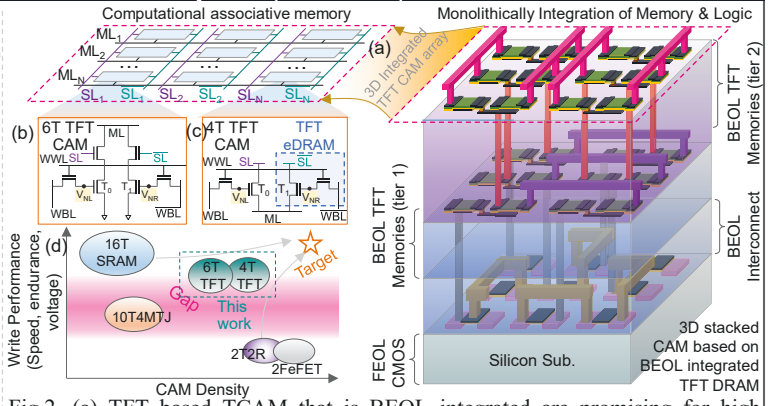
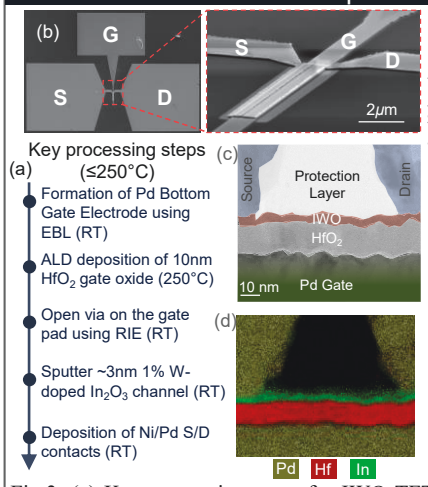


Fig.2. (a) TFT based TCAM that is BEOL integrated are promising for high density and update frequent applications. (d) It fills the gap between low-density, high write performance SRAM and high-density, poor write performance NVM TCAM designs. Both (b) 6T and (c) 4T TFT TCAM designs are proposed.

Metal Oxide TFT Development and Model Calibration



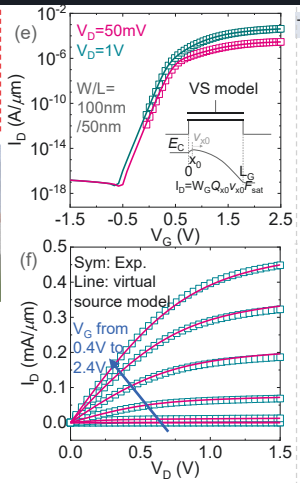


Fig.3. (a) Key processing steps for IWO TFT TCAM. (b) SEM of the TFT device, (c) TEM and (d) EDX elemental mapping of the gate stack. (e) I_D-V_G and (f) I_D-V_D characteristics and calibration with the virtual source model.

4T TFT TCAM Principles & Experimental Verification

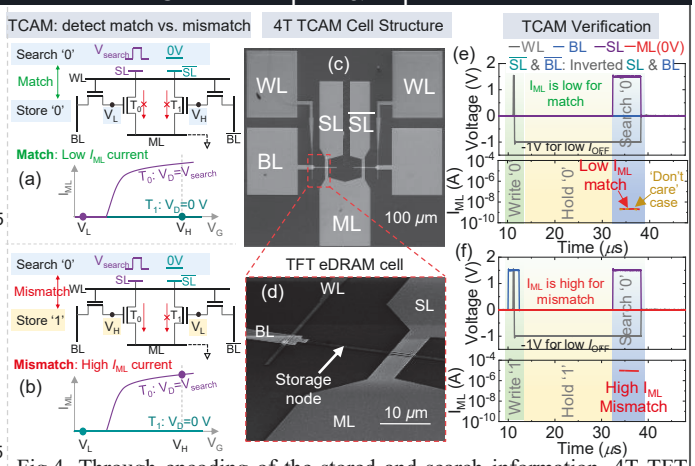
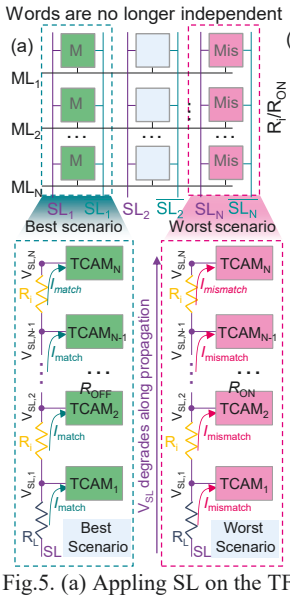
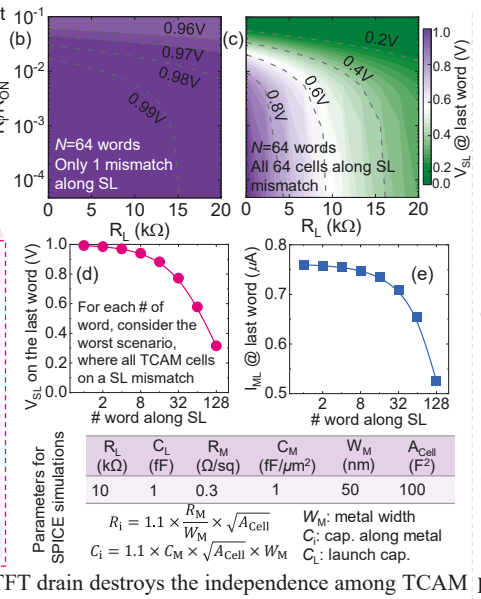


Fig.4. Through encoding of the stored and search information, 4T TFT TCAM can detect (a) match v.s. (b) mismatch. (c) SEM of the TCAM cell and (d) SEM of the eDRAM on one branch. Experiments show correct TCAM behaviors, i.e., I_{ML} is (e) low/(f) high for match/mismatch.

The Issue of Search Voltage Delivery in a 4T TFT TCAM Array Destroys the Independence Among CAM Words





R_L (kΩ)	C_L (fF)	R_M (Ω/sq)	C_M (fF/μm²)	W_M (nm)	A_{Cell} (F^2)
10	1	0.3	1	50	100

Parameters for SPICE simulations

$$R_i = 1.1 \times \frac{R_M}{W_M} \times \sqrt{A_{Cell}}$$

$$C_i = 1.1 \times C_M \times \sqrt{A_{Cell}} \times W_M$$

W_M: metal width
C_i: cap. along metal
C_L: launch cap.

Fig.5. (a) Appling SL on the TFT drain destroys the independence among TCAM words as the V_{SL} depends on the conditions of other cells on the SL. (b)/(c) simulated V_{SL} on the last word shows strong differences between best/worst scenarioes. (d)/(e) The V_{SL}/I_{ML} on the last word degrades more for a longer SL.

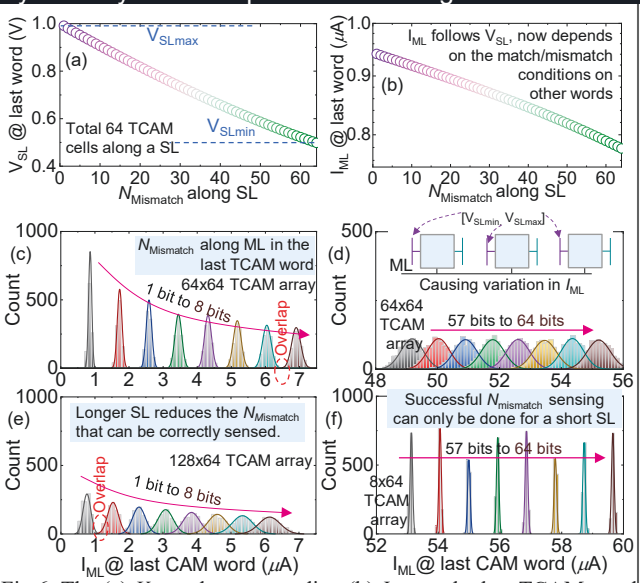


Fig.6. The (a) V_{SL} and corresponding (b) I_{ML} on the last TCAM word depends on the $N_{Mismatch}$ cells along a SL. This creates V_{SL} fluctuation among cells of a word, resulting in (c)/(d) variation in I_{ML}. (e) The overlap in I_{ML} can only be minimized by (f) reducing the SL length.

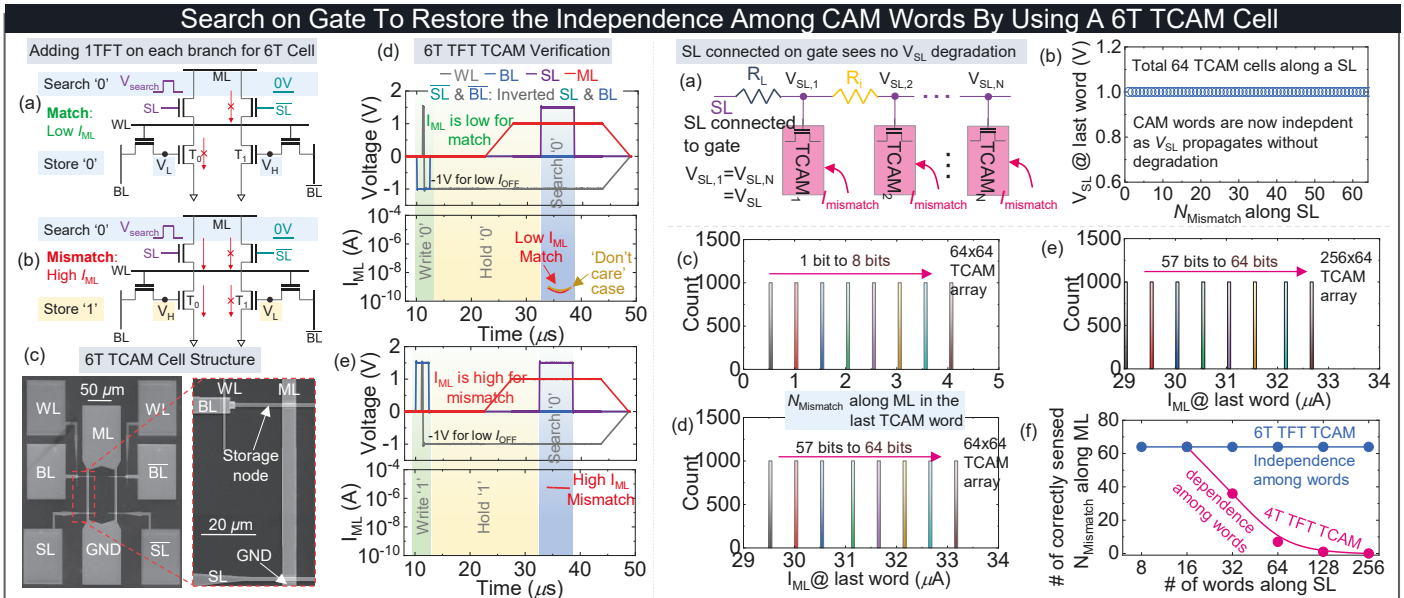


Fig.7. SL is applied on the gate in a 6T TFT TCAM cell, avoiding previous issue. It also differentiates the (a) match from (b) mismatch. (c) SEM of the TCAM cell. Measurements show correct cell behaviors, i.e., a (d) low/(e) high I_{ML} for the match/mismatch.

Fig.8. (a) Connecting the SL on TFT gate restores the independence among words by (b) delivering V_{SL} with no degradation. Successful detection of (c) and (d) all the $N_{Mismatch}$ along a ML is shown. (e) 256 words also works. (f) Detectable $N_{Mismatch}$ along a ML degrades with the # of words in 4T TCAM.

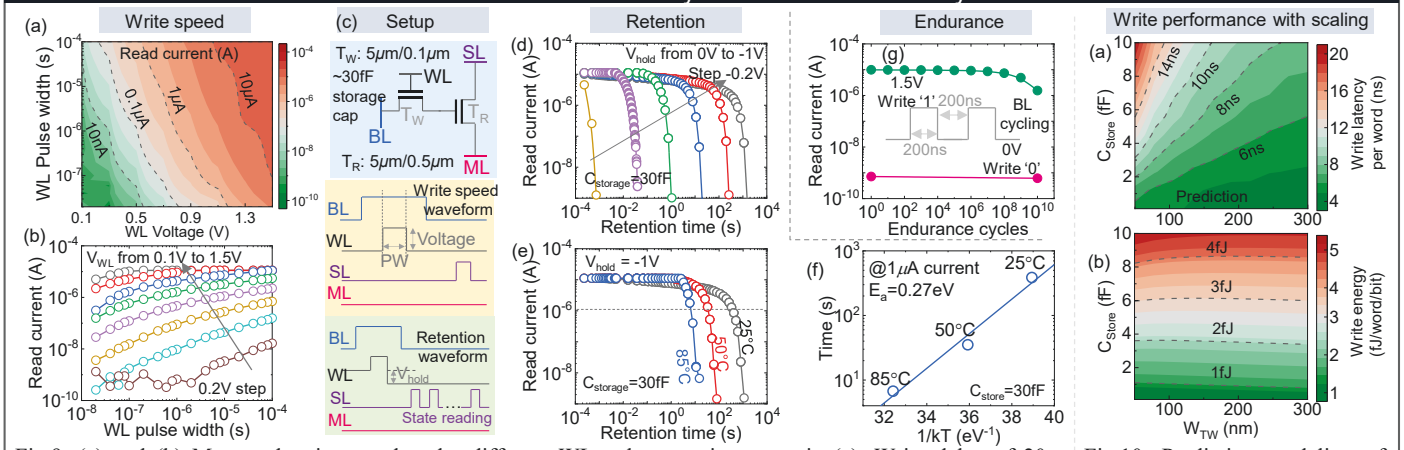
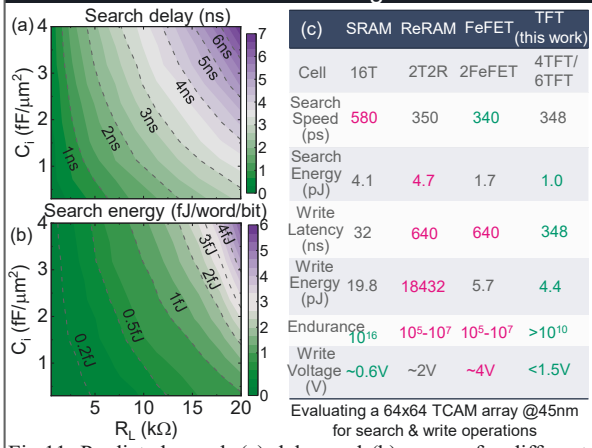


Fig.9. (a) and (b) Measured write speed under different WL voltages using setup in (c). Write delay of 20ns (instrument limit) is demonstrated with <1.5V. (d) Retention up to 1000s @RT is demonstrated with V_{hold} of -1V relying solely on the intrinsic storage capacitance. (e) Retention degrades at high temperatures, with (f) an activation energy (E_a) of 0.27eV. (g) Endurance cycling shows negligible degradation up to 10^{10} cycles.

Fig.10. Predictive modeling of write (a) latency and (b) energy with the calibrated virtual source model. External C_{Store} is added.

Benchmarking of TFT TCAM and Its Applications in Accelerating Clustering Applications



(c)	SRAM	ReRAM	FeFET	TFT (this work)
Cell	16T	2T2R	2FeFET	4TFT/ 6TFT
Search Speed (ps)	580	350	340	348
Search Energy (pJ)	4.1	4.7	1.7	1.0
Write Latency (ns)	32	640	640	348
Write Energy (pJ)	19.8	18432	5.7	4.4
Endurance	10^{16}	10^5-10^7	10^5-10^7	>10^{10}
Write Voltage (V)	~0.6V	~2V	~4V	<1.5V

Evaluating a 64x64 TCAM array @45nm for search & write operations

Fig.11. Predicted search (a) delay and (b) energy for different configurations. (c) Benchmarking of TCAM arrays shows that TFT TCAM has excellent write and search performance.

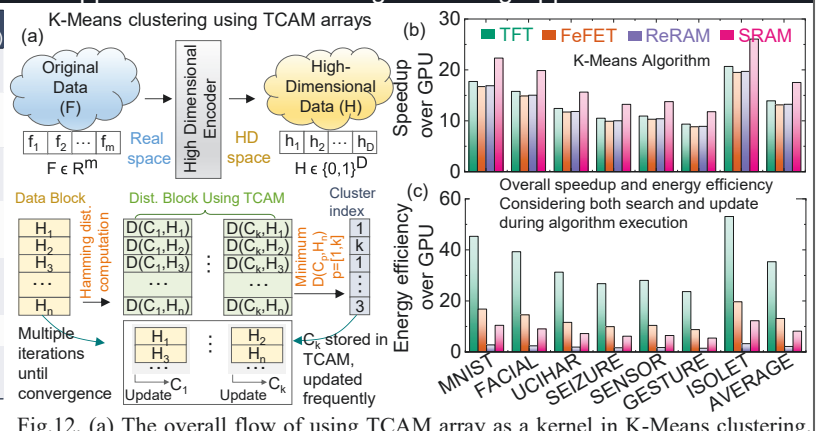


Fig.12. (a) The overall flow of using TCAM array as a kernel in K-Means clustering. The cluster centers are stored in TCAM and the distances between data and cluster centers are computed using TCAM. It requires about 0.13 update per search. (b) and (c) Overall latency and energy show 14x/35x improvement on average over GPU.