Novel 1T2R1T RRAM-based Ternary Content Addressable Memory for Large Scale Pattern Recognition

J-P. Noel, J -F. Nodin, D. R B Ly, J-P Noel, B. Giraud, P. Royer, E. Esmanhotto, N. Castellani, T. Dalgaty, J-F Nodin, et al.

Novel 1T2R1T RRAM-based Ternary Content Addressable Memory for Large Scale Pattern Recognition

D. R. B. Ly, J-P. Noel, B. Giraud, P. Royer, E. Esmanhotto, N. Castellani, T. Dalgaty, J-F. Nodin, C. Fenouillet-Beranger, E. Nowak and E. Vianello
CEA-LETI, Minatec Campus, Grenoble, France, email: Denys.Ly@cea.fr ; Elisa.Vianello@cea.Fr

***Abstract*—**Resistive Random Access Memories (RRAMs) are a promising solution to implement Ternary Content Addressable Memories (TCAMs) that are more area- and energy-efficient with respect to Static Random Access Memory (SRAM)-based TCAMs. However, RRAM-based TCAMs are limited in the number of bits per word due to the low ratio between the resistances of the high and low resistance states (HRS/LRS) and resistance variability of RRAM. Such a limitation on the word length hinders the parallel search of a very large number of data bits for data-intensive applications. To overcome this issue, for the first time, we propose a new TCAM cell composed of two transistors and two RRAMs in a 1T2R1T configuration, where a RRAM voltage divider (2R) biases a transistor gate (1T) and an additional transistor is used to program the RRAMs (1T). A 3x128bits 1T2R1T TCAM macro were designed, integrated and extensively characterized. We experimentally demonstrate that the sensing margin of the proposed structure is insensitive to HRS/LRS RRAM resistance ratio and variability. With respect to the most common type of 2T2R RRAM-based TCAM [1-3], the proposed circuit improves the sensing margin by >5000x while reaching search times of 0.93ns. This allows the search of large volumes of data in parallel. In addition, the proposed structure improves programming and search endurance by 100x and >10x, respectively.

I. PROPOSED 1T2R1T TCAM CIRCUIT

TCAM performs parallel searches by comparing input searched data with data stored in the TCAM and returning the data address when a match occurs [1]. They provide a lookup response in a single clock cycle making them faster and more energy-efficient than random access memory-based search systems. High pattern matching speeds make TCAM a key function for data-intensive applications, such as Internet Protocol (IP) lookup, word search, and routing in neuromorphic circuits [1, 4-5]. Conventional TCAMs based on SRAMs suffer from low area density and static power consumption. Resistive Memory (RRAM)-based TCAMs have been demonstrated [2, 6-10] enabling to overcome these challenges without degrading performance (search energy and time) (**Fig.1**). The most common RRAM-based TCAM is composed of two pairs of access transistors and RRAMs (2T2R) connected in parallel on a Match Line (ML) [1-3]. However, these structures are constrained in word length due to the low ON/OFF current ratio (<100), determined by the RRAM resistance ratio between the High Resistance State (HRS) and the Low Resistance State (LRS). A large ON/OFF current ratio comparable to that of SRAMs (>10^5) is required to enable parallel searches of longer words. Here we propose a new TCAM cell composed of two transistors and two RRAMs in a 1T2R1T configuration (**Fig.2a**), in which two RRAMs (2R) compose a voltage divider that biases the transistor gate of N2 (1T) for search operations. An additional transistor N1 (1T) works as an access transistor to program the RRAMs. The searching scheme is shown in **Fig.2b**. The ML is first pre-charged at VDD (transistors N1 and N2 are OFF). The ML is then left floating and a search voltage V_{search} is applied across the RRAM voltage divider. When the search bit is '1', V_{search} is applied on the top of the voltage divider (SLT), while maintaining SLF to 0V. When the search bit is '0', V_{search} is applied on the bottom (SLF). The internal voltage V_{INT} depends on R_X and R_Y values. If the stored and searched data match (V_{INT}~0V) transistor N2 remains OFF and the ML stays high. If we have a mismatch (V_{INT}~V_{search}) N2 turns ON and the ML is pulled down to ground. **Fig.3a** and **b** show a photo of the fabricated 3x128bits 1T2R1T RRAM-based TCAM and a SEM cross section of the integrated RRAMs, respectively. HfO_2-based RRAMs are integrated in the Back End of Line of a 130nm CMOS process [11].

During ML sensing, in the most common 2T2R RRAM-based TCAM (**Fig.4a**), the top electrodes of both RRAMs are connected to the ML. Current flows in the 1T1R branches with the selector transistor in the ON state, discharging the ML. In case of match, the ML slowly discharges through RRAMs in HRS (**top**) whereas in case of mismatch the ML discharges quickly through RRAMs in LRS (**bottom**). Since the leakage currents of the TCAM cells on the same ML add together, the limited HRS/LRS resistance ratio makes difficult to distinguish between a match and 1-bit mismatch state (the hardest mismatch to detect) for long TCAM words. In the proposed 1T2R1T structure (**Fig.4b**), the ML is connected to transistors controlled by the RRAM voltage divider (N2). Therefore, the sensing margin no longer depends on HRS/LRS ratio (~30 at V_{search}=0.6V) but on the MOSFET current ratio (~10^6 at V_{search}=0.6V as shown on the transistor characteristic in **Fig.5**), leading to the possibility of longer word search.

II. SENSING MARGIN AND SEARCH CAPACITY

We performed extensive circuit-level electrical characterization of the fabricated 1T2R1T TCAM array. Measurements are performed on the 128-bits middle TCAM of **Fig.3a**. The transistor N1 is used to perform Forming, Set and Reset operations using the standard 1T1R scheme. The configuration with both cells in LRS is forbidden (always miss case). The programming sequence of **Fig.6 top** is adopted in

order to form all the RRAM devices. **Fig.6 bottom** shows the pristine, LRS and HRS cumulative distributions and their corresponding programming conditions. The Memory Window (MW) is defined as the ratio between HRS and LRS values at -2σ and $+2\sigma$, respectively. Increasing HRS values from Soft HRS (WL=3V) to Strong HRS (WL=3.5V) allows an increase in the MW at the cost of a lower programming endurance [11].

During a search operation (**Fig.7**), the ML voltage is sensed by a Sense Amplifier (SA) circuit. The SA compares the ML voltage to a reference voltage VREF to measure the ML discharge time, t_{search}. In [1], we defined the sensing margin as the time ratio (TR) between the discharge time in match state and 1-bit mismatch state (slowest mismatch case). In the proposed TCAM, the ML discharge time in the match state is longer than the limit of measurement of 1s. For the sake of a fair comparison, we fixed t_{search} in match at 1s for the proposed TCAM.

The ideal TCAM should minimize the discharge time of the 1-bit mismatch state for fast search operations while maximizing the sensing margin (TR) to improve the parallel search capability. **Fig.8** shows the impact of V_{search} (voltage applied across the RRAM voltage divider) on t_{search} in case of match (green) as well as 1-bit and 128-bits mismatch (red). RRAMs are programmed using the conditions in **Fig.6 bottom**. In the match state, t_{search} remains higher than 1s for any V_{search} as the ML does not discharge. In the case of a 1-bit mismatch the ML discharges if $V_{search} \geq 0.5V$ because the N2 transistor of the mismatching bitcell turns ON. t_{search} is slightly dependent on the HRS resistance value as it mostly depends on the current flowing through the transistor N2. The higher V_{search}, the lower the search time. As the ML does not discharge in match, TR also increases with V_{search} (**Fig.9**). **Fig.9** compares the sensing margin obtained in the proposed structure with the one measured in the 2T2R structure in [1], which has been fabricated using the same CMOS and RRAM technologies. The TR improves by >2000x/>5000x for the Strong/Soft programming conditions, respectively. This structure enables the use of Soft HRS whereas in [1] TR was below the sensing limit at any V_{search} in the same programming conditions. This improves programming endurance by 100x (cf. **Fig.6 bottom**). **Fig.10** shows the impact of the MW on the TR for a given V_{search}=0.6V. In the proposed TCAM, the sensing margin is insensitive to the MW whereas that of [1] could not operate for a MW<50. **Fig.11** shows the TR as a function of the TCAM Word Length (WDL). The sensing margin of 2T2R TCAMs decreases with an increase in WDL, thereby limiting the maximum word length to 97bits for Soft HRS and ~256bits for Strong HRS. In the proposed structure WDL has a minimal impact on the sensing margin, thus allowing for an increase in the maximum word length (>2kbits) for both Soft and Strong HRS. This permits the parallel search of large volumes of data.

III. SEARCH ENDURANCE CHARACTERIZATION

During a search operation a positive voltage is applied on the top electrode of one RRAM cell (R_X for search '1', R_Y for search '0') (**Fig.12a**). Therefore, in the match state, undesired switching from HRS to LRS can occur after a certain number of search operations. This causes a match failure. We define the search endurance as the maximum number of search operations before a match failure occurs (**Fig.12b**). Decreasing V_{search} or using Strong HRS improves the search endurance. **Fig.13** reports the search endurance as a function of the search voltage value for the proposed 1T2R1T TCAM structure as well as the 2T2R TCAM of [1]. The proposed structure allows an improvement in the search endurance at any V_{search}. At V_{search}=0.6V and Strong HRS, we observe no degradation for more than 10^7 searches, improving on [1] by >10x.

IV. SEARCH TIME AND SEARH ENERGY CONSUMPTION

To reduce the search time, the ML sensing circuit has to be as fast as possible. This is the reason why we used an analog circuit to sense the ML voltage and compare it to a reference voltage VREF (**Fig.14 top**). In this paper, the search time is defined as the time taken to discharge the ML of a given voltage (ΔV) from the pre-charged value (VDD). **Fig.15** shows measured (symbols) and simulated (lines) search times as a function of V_{search}, for a mismatch state of 128bits (**Left**) and 1bit (**Right**). First, by changing ΔV from 3V to 80mV (accurate activation of the comparator) the search time improves by 96x at V_{search}=0.6V. Second, replacing the thick oxide MOS, used for transistor N2, by a thin oxide MOS speeds up searches by 300x. A thin oxide transistor, with minimum permitted gate length, can be adopted since the transistor N2 is not involved in RRAM programming. At V_{search}=0.6V, we reach a search time of 0.93ns. Since, in the proposed TCAM, the ML discharges only in the mismatch state, the sensing circuit can also be simplified by the use of a digital inverter (**Fig.14 bottom**), thereby reducing design complexity [12]. This is not possible for the common 2T2R TCAM cell.

Energy consumption during search is estimated as the integral of power over the search time. Therefore, a cell with a thin oxide MOS requires 13x less search energy (**Fig.16**).

V. SUMMARY

In this work, we proposed a new 1T2R1T TCAM cell based on a RRAM voltage divider biasing a transistor gate. We experimentally demonstrated large sensing margin, comparable to that of SRAM (>10^4), even for RRAM technologies with reduced HRS/LRS ratio (~30). This allows a large volume of data to be searched in parallel thanks to the long word length (**Table 1** and **2**). We also showed that the relaxed requirements on the HRS/LRS ratio allow programming of RRAM devices in a low voltage regime, which implies better programming endurance (100x) without degrading the sensing margin and the maximum word length as in previously reported 2T2R RRAM-based TCAMs. This makes this bitcell suitable for applications requiring long pattern matching (IPv6 packet routing, DNA sequence matching, and active control list management [4]). In addition, we prove better search endurance, improving the common 2T2R cell by >10x which could greatly benefit event routing in neuromorphic processors [5].

ACKNOWLEDGMENT: This work has been supported by the **French ANR via the Carnot funding** and the **European H2020 NeuRAM3 687299 project**.

REFERENCES: [1] D.R.B. Ly et al., *IEDM*, 2018 [2] J. Li et al., *JSSC*, 2014, vol.49, no.4 [3] R. Yang et al., *Nature Electronics*, 2019, vol.2, no.3 [4] Q. Guo et al., *Micro*, 2015, vol. 35, no. 5 [5] S. Moradi et al., *BioCAS*, 2018, vol.12, no.1 [6] M-F. Chang et al., *JSSC*, 2016, vol.51, no.11 [7] M-F. Chang et al., *JSSC*, 2017, vol.52, no.6 [8] C-C. Lin et al., *ISSCC*, 2016 [9] M-F. Chang et al., *ISCAS*, 2016 [10] C. Kim et al., *TCAS I*, 2019, vol.66, no.2 [11] A. Grossi et al., *TVLSI*, 2018, vol.26, no.12 [12] K. Pagiamtzis et al., *JSSC*, 2006, vol.41, no.3

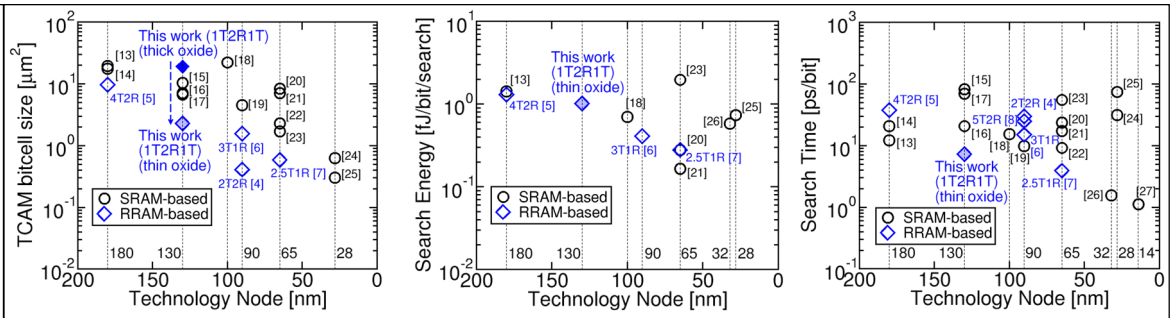
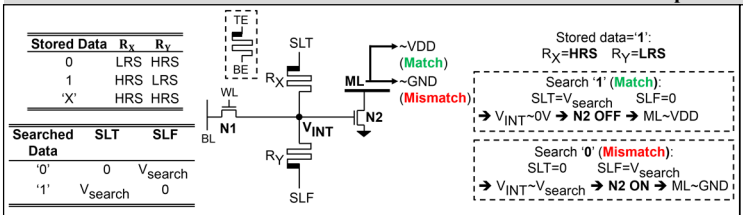


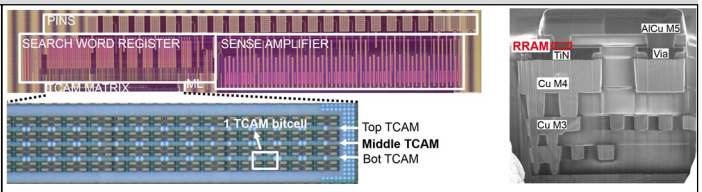
Fig.1: Comparison of reported silicon-proven SRAM- (circle) and RRAM-based (diamond) TCAMs in terms of bitcell size **(Left**, adapted from [28]), search energy **(Middle)** and search time **(Right)** as a function of technology node.

I. Proposed 1T2R1T TCAM circuit



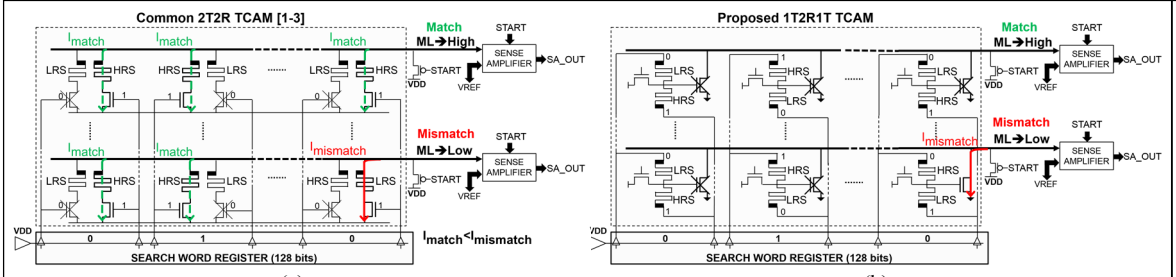
(a) (b)

Fig.2: **(a)** Circuit diagram of the proposed 1T2R1T TCAM bitcell. RRAM Top (TE) and Bottom (BE) Electrodes are indicated. **(b)** Match and mismatch states definition.



(a) (b)

Fig.3: **(a)** Photo of the fabricated 1T2R1T TCAM circuit and **(b)** SEM cross-section of the integrated TiN/HfO_2/Ti/TiN RRAM cells. Both HfO_2 and Ti layers are 10nm thick.



(a) (b)

Fig.4: Circuit diagrams of **(a)** the common 2T2R TCAM array fabricated in our previous work [1] compared with **(b)** the proposed 1T2R1T TCAM array during a search operation. Match and mismatch states are illustrated. The sensing margin (~$I_{mismatch}/\sum I_{match}$) of the proposed 1T2R1T structure is independent of the resistance value of the HRS state; it only depends on the N2 transistor characteristic.

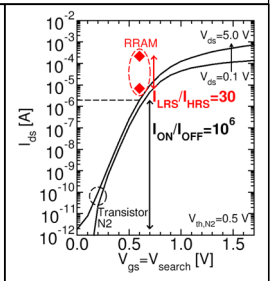


Fig.5: I_{ds}-V_{gs} characteristic of the N2 transistor. The RRAM I_{LRS}/I_{HRS} ratio is highlighted (red) for comparison.

II. Sensing margin and search capacity

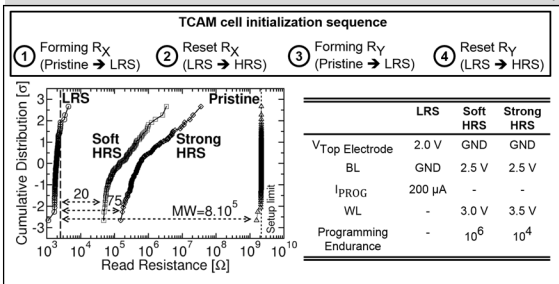


Fig.6: **(top)** TCAM cell initialization sequence and **(bottom)** RRAM pristine, LRS and HRS cumulative distributions with associated programming conditions. HRS distributions are obtained using the Soft HRS and Strong HRS programming conditions.

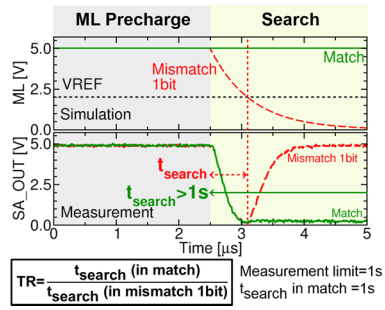


Fig.7: Waveforms of the search operation for a match (green) and mismatch (red). The search time is the time required to discharge the ML to VREF from the pre-charged value (VDD).

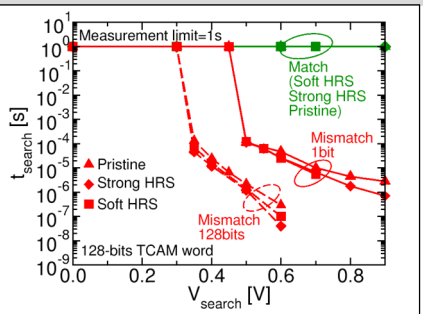


Fig.8: Measured search times at different search voltages, in case of match (green) and mismatch (red).

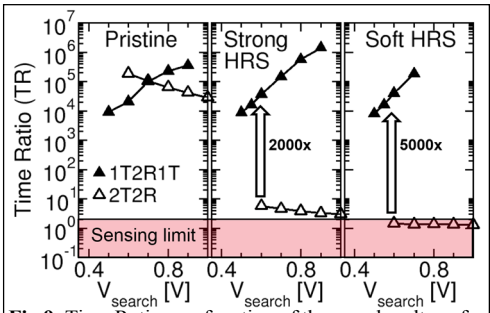


Fig.9: Time Ratio as a function of the search voltage for a 128-bits TCAM word. Results obtained in our previous work [1] (open symbols) on common 2T2R TCAMs are reported for comparison with the proposed 1T2R1T TCAM (filled symbols).

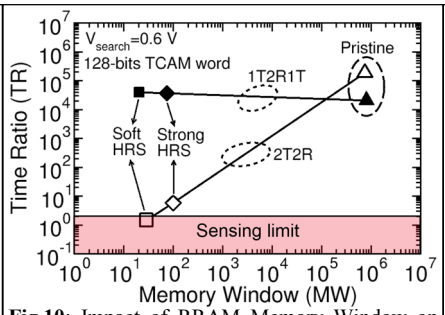


Fig.10: Impact of RRAM Memory Window on Time Ratio for the 1T2R1T (filled symbols) and 2T2R (open symbols) TCAMs. The sensing margin (Time Ratio) of the 1T2R1T TCAM cell is independent of the RRAM Memory Window.

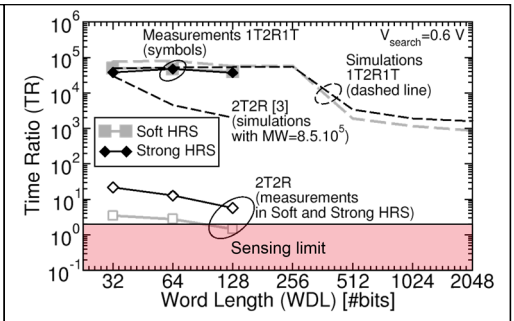


Fig.11: Impact of TCAM Word Length (WDL) on Time Ratio (TR). The proposed 1T2R1T TCAM bitcell with large TR can enable very large search capacity even with Soft HRS (grey lines).

III. Search endurance characterization

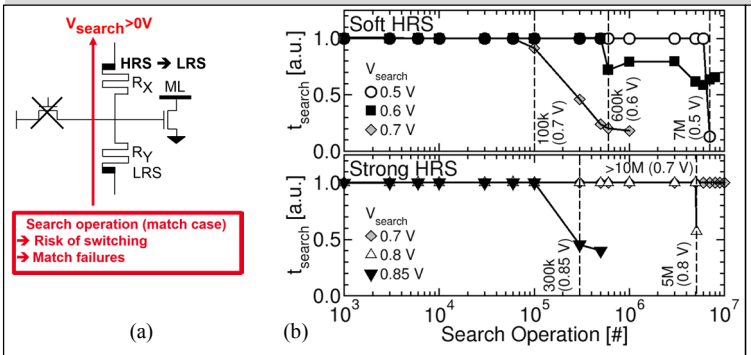


Fig.12: **(a)** Search endurance explanation. **(b)** Search endurance characterization (Soft HRS **top**, Strong HRS **bottom**) at different search voltages V_{search}.

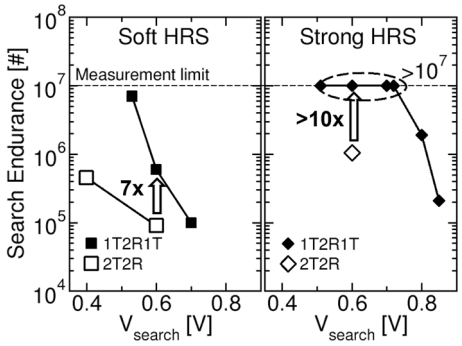


Fig.13: Search endurance as a function of V_{search}. The proposed structure improves the search endurance in both Soft **(Left)** and Strong **(Right)** HRS.

IV. Search time and search energy consumption

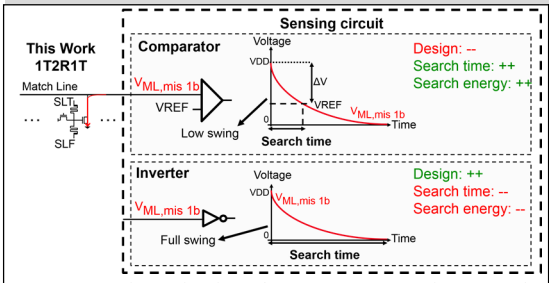


Fig.14: Sensing circuit scheme. **(top)** Sensing can be performed faster by tuning the comparator reference voltage VREF. **(bottom)** As the match line discharges only in the mismatch state, the sensing circuit can be simplified by a digital inverter.

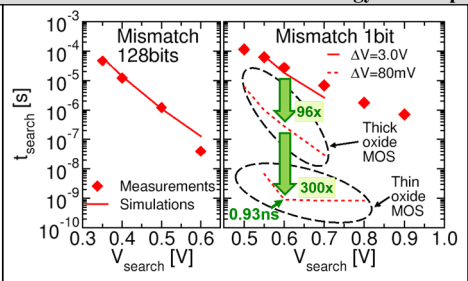


Fig.15: Measured (filled symbols) and simulated (lines) search times as a function of V_{search}. Search times can be improved by optimizing the sensing circuit (ΔV and MOS oxide thickness).

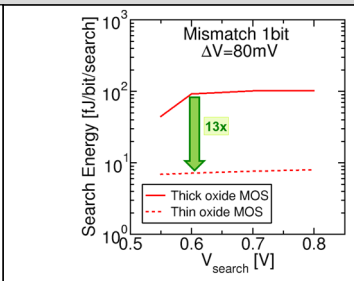


Fig.16: Simulated search energy in the case of 1-bit mismatch, when transistor N2 is implemented with thick oxide MOS (full line) and thin oxide MOS (dashed line).

V. Summary

		Previous work [1] 2T2R		This work 1T2R1T	
		HRS/LRS		HRS/LRS	
		~30	~100	~30	~100
Reliability (Higher is better)	Sensing Margin	1.5	5.7	4.10^4	4.10^4
	Search Endurance	9.10^4	10^6	6.10^5	$>10^7$
	Prog. endurance	10^6	10^4	10^6	10^4
Capacity (Higher is better)	Word Length max	**97 bits**	**~256 bits**	**>2 kbits**	**>2 kbits**

Table1: Impact of RRAM HRS/LRS ratio on reliability and maximum capacity of common 2T2R TCAMs and the proposed 1T2R1T TCAM. The proposed TCAM is less sensitive to RRAM HRS/LRS ratio, which allows use of longer words.

	[2] 2T2R	[6] 4T2R	[7] 3T1R	[8] 2.5T1R	[9] 5T2R	Previous work [1] 2T2R	This work 1T2R1T
Technology node	90 nm	180 nm	90 nm	65 nm	90 nm	130 nm	130 nm
TCAM capacity	16k×64bits	128×32bits	128×64bits	64×256bits	128×64bits	3×128bits	3×128bits
Prog. Endurance [#cycles]	-	-	-	-	-	10^4	10^6
Search Endurance [#searches]	-	-	-	-	-	10^6 (V_{search}=0.6V)	$>10^7$ (V_{search}=0.6V)
Word Length max	-	-	-	-	-	**97 bits** (Soft HRS)	**>2 kbits** (Soft HRS)
Normalized Search Time [ps/bit]	30	38	15	4	25	700 (thick oxide MOS)	2180* (thick oxide MOS) 7.3* (thin oxide MOS)

*With ΔV=80mV

Table2: Comparison with silicon-proven RRAM-based TCAM circuits presented in the literature. Search times have been normalized by the TCAM word length.