In-depth Characterization of Resistive Memory-Based Ternary Content Addressable Memories

D. R. B. Ly, B. Giraud, J-P Noel, A. Grossi, N. Castellani, G. Sassine, J-F Nodin, G. Molas, C. Fenouillet-Beranger, G. Indiveri, et al.

In-depth Characterization of Resistive Memory-Based Ternary Content Addressable Memories

D. R. B. Ly[1], B. Giraud[1], J-P Noel[1], A. Grossi[1], N. Castellani[1], G. Sassine[1], J-F Nodin[1], G. Molas[1], C. Fenouillet-Beranger[1], G. Indiveri[2], E. Nowak[1] and E. Vianello[1]
[1]Univ. Grenoble Alpes, CEA, LETI, 38000 Grenoble, France, email: denys.ly@cea.fr ; elisa.vianello@cea.fr
[2] Institute of Neuroinformatics, University of Zurich and ETH Zurich

Abstract—Resistive Memory (RRAM)-based Ternary Content Addressable Memories (TCAMs) were developed to reduce cell area, search energy and standby power consumption beyond what can be achieved by SRAM-based TCAMs. In previous works, RRAM-based TCAMs have already been fabricated, but the impact of RRAM reliability on TCAM performance has never been proven until now. In this work, we fabricated and extensively tested a RRAM-based TCAM circuit. We show that a trade-off exists between search latency and reliability in terms of match/mismatch detection and search/read endurance, and that a RRAM-based TCAM is an ideal building block in multi-core neuromorphic architectures. These ones would not be affected by long latency time and limited write endurance, and could greatly benefit from their high-density and zero standby power consumption.

I. Introduction

Ternary Content Addressable Memories (TCAMs) provide a way of searching large data set using masks that indicate ranges. Therefore, they are very attractive for complex routing and big data applications, where an exact match is not often necessary [1]. As opposed to classic memory systems, where a memory cell stored information is retrieved by its physical address, TCAM circuits allow to search a stored information by its content. Fig. 1 depicts the search principle in a TCAM word of 3 bits. The Match Line (ML) is first precharged at a voltage VDD_ML. During the search phase, the ML is left floating and starts discharging. If the stored data in the TCAM word matches with the input searched data (Fig. 1 (a)), the ML slowly discharges through leakage currents $I_{ML,m}$. If at least one bit of the stored data mismatches the input searched data (Fig. 1 (b)), the ML quickly discharges through the mismatching bit with a high discharge current $I_{ML,mis}$.

Conventional SRAM-based TCAM circuits are usually implemented with 16 transistors (16T) [2]. This limits storage capacity of TCAM circuits to tens of Mbs [2-3] in standard memory structures, and takes up valuable silicon real-estate in neuromorphic computing spiking neural network chips [4-5]. In order to increase storage density, RRAM-based TCAM cells have been proposed [6-10]. However, one drawback of RRAM-based with respect to SRAM-based TCAMs is the relatively small ON/OFF current ratio of the memory elements ($\sim 10^5$ for MOSFET compared to 10-100 for RRAMs). Therefore, for long TCAM words, the sum of leakage currents $I_{ML,m}$ in case of a match can become comparable to a mismatching current $I_{ML,mis}$. Another challenge in designing TCAM circuits with RRAMs is the limited endurance of RRAMs. While SRAMs can sustain an endurance up to 10^{16} cycles, it is difficult to reach endurance higher than 10^6 write cycles for RRAMs. However, endurance in SRAM refers to both write and search/read operations while in RRAM-based TCAMs, write and search/read operations are well distinguished, and must be characterized separately [11].

In this paper, an extensive characterization of search/read and write operations in a RRAM-based TCAM circuit is presented. To date, only the impact of the search/read voltage on the search latency time has been presented (Table 1). The expected impact of RRAM has been evaluated only by simulations [11]. Strong RRAM programming conditions associated with low search/read voltage allow to improve TCAM reliability (Time Ratio>5 and search/read endurance> 10^6 cycles) at the expense of lower performance (longer latency time and lower write endurance). Therefore, we propose to use the RRAM-based TCAM for routing in neuromorphic circuit where long match times (from few tens to hundreds of µs) are required to be compatible with spike length [4-5]. Moreover, this application features long idle times, frequent search and few write operations, thus taking full advantage from the zero standby power consumption.

II. Fabricated RRAM-Based TCAM Circuit

Fig. 2 (a) presents the schematic of the fabricated RRAM-based TCAM circuit, composed of a Search word register (SL), a bit cell matrix and a read circuit (Sense Amplifier, SA). The TCAM bit cell is composed of two HfO_2-based RRAMs and two transistors in a 2T2R structure (Fig. 2 (b)). Fig. 3 (a) and (b) show a photo of the fabricated circuit and a SEM cross section of the integrated RRAMs, respectively. RRAMs are integrated in the Back End Of Line of a 130nm CMOS process, on top of the fourth metal layer.

Forming, Set and Reset operations are performed as in single 1T1R RRAM cell by applying the required top electrode voltage on the ML, the bottom electrode voltage on the BL while activating the gate voltage with SLT (resp. SLF). SLF (resp. SLT) signal must be at 0V in order to activate each 1T1R structure independently. During a search operation (Fig. 4), ML voltage is sensed with the SA. In order to evaluate the discharge time of the ML, the applied voltage is compared with a reference voltage V_{TRIP}. The Time Ratio (TR), defined as the ratio between the discharge time when all the word bits are matching ($t_{discharge,m}$) and when only 1 bit of the word is mismatching ($t_{discharge,mis1b}$), has to be maximized to guarantee a sufficient search margin. Different

capacitances can be added on the ML signal with the signal CAP_CALIB to slow down the discharge of the ML and facilitate its measurement. A capacitance of 315pF is used in the following. Fig. 5 (a) and (b) show the impact of the ML capacitance on the discharge time $t_{discharge}$, in case of match (green) and mismatch (red). The discharge time increases for higher capacitance values. However, the TR is independent of the ML capacitance (Fig. 5 (c)).

III. Time Ratio Dependencies

The ideal TCAM should have a short search latency time (short discharge time) while maximizing the Time Ratio (TR). Fig. 6 (a) presents the pristine, Low Resistance State (LRS) and High Resistance State (HRS) cumulative distributions directly measured on the TCAM cells. HRS distributions are obtained using the soft and strong programming conditions in Fig. 6 (b). The pristine resistance distribution can be used if the TCAM is programmed only one time. The Memory Window (MW) is defined as the ratio between the HRS and LRS values at 2.5σ of the distributions. Fig. 7 shows the impact of the search/read voltage ΔV, applied between the ML and the BL across the bit cells during the search phase, on the discharge time for soft and strong HRS. Higher ΔV enables lower $t_{discharge}$ and thus better performance (lower latency). However, TR is independent of ΔV (Fig. 8). Fig. 9 shows the impact of the MW on the TR. Stronger programming conditions (larger MWs) increase the TR. As expected, TR decreases with the TCAM word length WDL (Fig. 10). To ensure a reliable search taking into account spatial and transient variability, TR cannot go below a sensing limit of 2. Thus, a TCAM row of 128 bit cannot be programmed with soft HRS and it is necessary to use strong HRS. Fig. 11 shows discharge time as a function of the number of mismatching bits, *n*. Since the discharge time decreases with *n*, it can be used as an estimation of the Hamming distance between searched and stored data. This block is useful for hardware implementation of brain-inspired hyperdimensional computing systems [12].

IV. Search/Read Reliability

During a search operation, a positive voltage ΔV is applied on the RRAM top electrode in the same polarity as a Set operation (Fig. 12 top). Therefore, unwanted switching from HRS to LRS can occur as shown in Fig. 12 bottom. Constant Voltage Stress (CVS) measurements have been performed on a 4kbit 1T1R RRAM array to extract the Set switching time $t_{switching}$ for different ΔV. Fig. 13 presents the measured $t_{switching}$ as a function of ΔV (black lines). The different curves correspond to different percentages of switched cells. For comparison, we reported the discharge time $t_{discharge,m}$ as a function of the search voltage ΔV in case of a match (green lines), for different word lengths. To avoid unwanted switching during search operations, the search voltage must be diminished so that the discharge time is lower than $t_{switching}$ (green region in Fig. 13). As shown in Fig. 14 (a) top, a reduction of ΔV from 0.6V to 0.4V increases from 90k to 450k the number of reliable searches (defined as the number of search operations before a RRAM device switches from the HRS to the LRS, Fig. 14 (b)). Another way to improve search reliability is to adopt strong HRS. For ΔV=0.6V, 400k cycles are performed (Fig. 14 (a) bottom). These results are obtained in the worst case scenario, since the 315pF line capacitance artificially increases the search time. By using a capacitance of 90pF (Fig. 14 (a), bottom), we have an endurance higher than 10^6 cycles.

V. Programming Reliability

Fig. 15 shows a write endurance characterization measured on a 4kbit 1T1R RRAM array, for soft HRS. This case allows to improve endurance performance [11]. After 10^4 Set/Reset cycles, some cells remain stuck in HRS with a probability $p_{HRS\ stuck}$. At 10^6 cycles, breakdown failures (cells stuck in LRS) occur with a probability $p_{breakdown}$. For a TCAM circuit, HRS stuck failures increase the discharge time and therefore they have no impact on matches. However, this can lead to a mismatch failure with a probability depending on the word length (Fig. 16 (a)). On the other hand, breakdown failures decreases the discharge time. This leads to a match failure whatever the impacted matching TCAM cell, *i.e.* with a probability independent of the word length (Fig. 16 (a)). Fig. 16 (b) shows the probability of failures for a match (red) and a mismatch (blue) as a function of the number of write operations. An endurance in programming of 10^6 cycles can be reached, as the probability of mismatch failures is negligible ($\sim 10^{-38}$).

VI. Discussion and Conclusion

In this work, we experimentally show the impact of RRAM reliability on a TCAM circuit. To improve the search margin (Time Ratio) and search/read endurance, strong RRAM programming conditions (high HRS), low search voltage and limited word length have to be adopted. This comes at the expense of lower performance in terms of longer search latencies and lower write endurance (Table 2). The performance reduction can be a critical limiting factor in standard TCAM-based applications. However, multi-core neuromorphic computing architectures would not be affected by these problems and could greatly benefit from their high density. For example, the processing cores of the NeuRAM3 DYNAP-SEL neuromorphic chip recently proposed in [5] comprise multiple TCAM cells per neuron, to implement memory-optimized source-address routing schemes (see [4] for details). These TCAM cells are typically small in size (e.g. 22 bit in the DYNAP-SEL chip) and are programmed only at network configuration time. Assuming future neuromorphic computing architectures of this type will have thousands of cores, the non-volatility feature of the proposed TCAM circuits will provide an additional crucial benefit, as it will require the user to upload all the configuration bits only the first time the network is configured, and will be able to skip this potentially time-consuming process every time the chip is reset or power-cycled. Finally, long match times are required to be compatible with spike length.

ACKNOWLEDGMENTS: This work has been partially supported by the European H2020 NeuRAM 687299 project.

I. Introduction

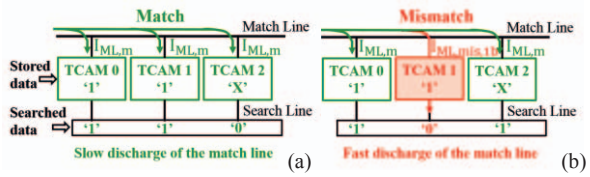


Fig. 1: Example of search/read operation. The Match Line (ML) is initially precharged, then it is left floating. If the stored data matches with the searched data, the ML slowly discharges (a). If the stored data mismatches with the searched data, the ML quickly discharges (b).

	[6] - 2T2R	[7] – 2.5T1R	[8] – 5T2R	[9] – 4T2R	[10] – 3T1R	This Work – 2T2R
TCAM circuit	8×2048×64 bit 90 nm CMOS	64×256 bit 90 nm CMOS	128×64 bit 90 nm CMOS	128×32 bit 180 nm CMOS	2×64×64 bit 90 nm CMOS	3×128 bit 130 nm CMOS
Search Latency	1.9 ns @ 0.75 V	1 ns @ 0.45 V	1.9 ns @ 0.75 V	1.2 ns @ 1.4 V	0.96 ns @ 0.48 V	90 ns @ 0.6 V
Measured results	Impact of *Search Voltage* on **Search Latency**					• Search Latency • Match/mismatch search margin • Search/Read endurance (>10^6) • Programming endurance (> 10^6)

Table 1: Silicon verified RRAM-based TCAM circuits presented in the literature. Few electrical characterization results have been presented and the impact of RRAM reliability on the TCAM performance has not been studied before.

II. Fabricated RRAM-based TCAM circuit

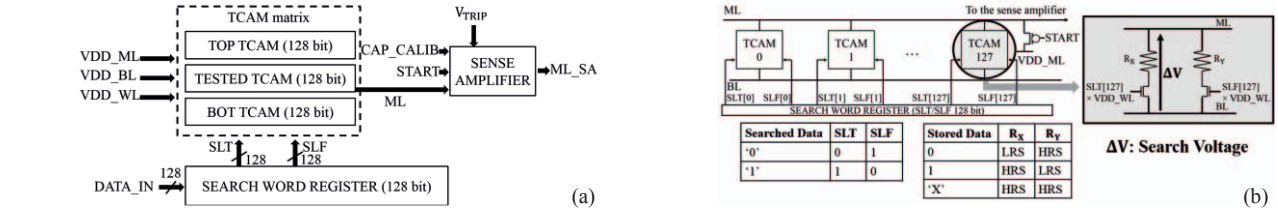


Fig.2: (a) RRAM-based TCAM circuit schematic. The Search word registers (SLs) take as input the searched data DATA_IN and send it to the TCAM matrix via the signals SLT=DATA_IN and SLF. The TCAM matrix comprises 3 rows of 128 bits. (b) RRAM-based TCAM row and bit cell schematics, and states definition.

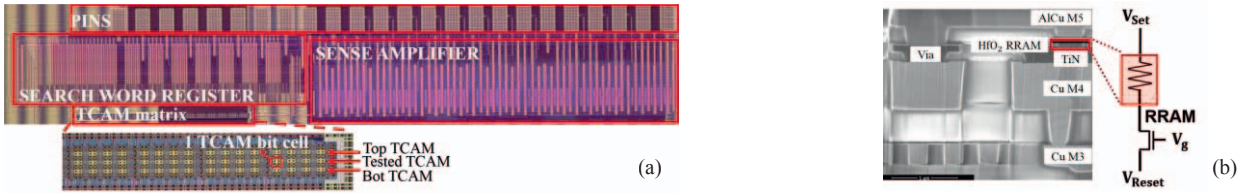


Fig. 3: (a) Die photo. (b) SEM cross section of the integrated TiN/HfO_2/Ti/TiN RRAM. Both HfO_2 and Ti layers feature a 10 nm thickness.

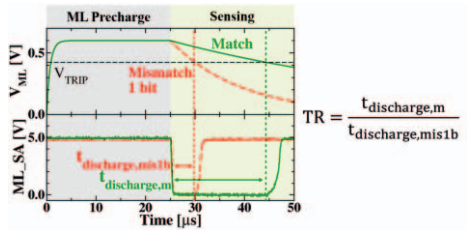


Fig. 4: Waveforms of the search operation for match (green) and mismatch (red) cases. The Match Line (ML) is initially precharged. During sensing, the ML is left floating and its voltage V_{ML} decreases. The output signal of the Sense Amplifier (SA), ML_SA, stays at 0V as long as V_{ML} remains higher than V_{TRIP}.

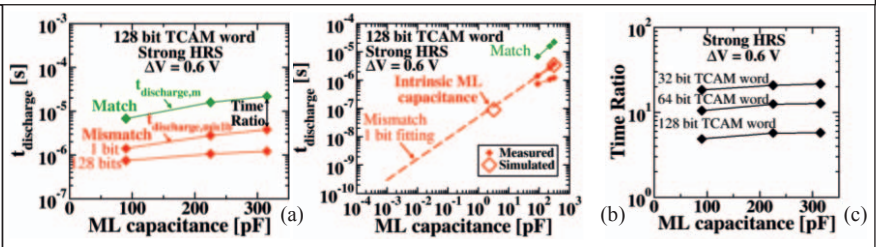
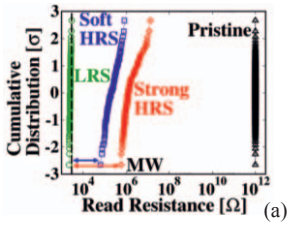


Fig. 5: (a) Discharge time as a function of the Match Line (ML) capacitance for the match (green) and mismatch (red) cases for a 128 bit TCAM word and (b) comparison with simulated results. (c) Time Ratio (TR) as a function of the ML capacitance for different word lengths. TR is independent of the ML capacitance. Strong programming conditions defined in Fig. 6 are used for (a), (b) and (c).

III. Time Ratio Dependencies



	LRS	Soft HRS	Strong HRS
ML	2.0 V	GND	GND
BL	GND	2.5 V	2.5 V
VDD_WL	1.3 V	2.5 V	3.0 V
R_{median} (Ω)	2.5 k	198 k	1.27 M
$R_{\pm 2.5\sigma}$ (Ω)	3.0 k	66 k	594 k

(b)

Fig. 6: (a) Pristine, LRS and HRS cumulative distributions and (b) associated programming conditions. Memory Window (MW) is defined as the ratio between the HRS and LRS values at $\pm 2.5\,\sigma$ of the distributions.

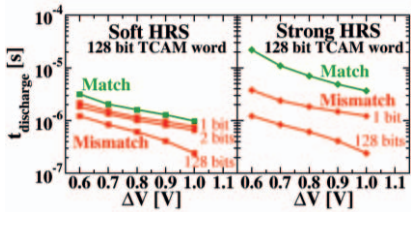


Fig. 7: Discharge time as a function of the search voltage (ΔV in Fig. 2 (b)) for soft (a) and strong (b) programming conditions.

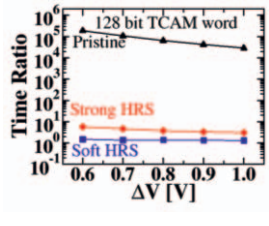


Fig. 8: Time Ratio as a function of the search voltage.

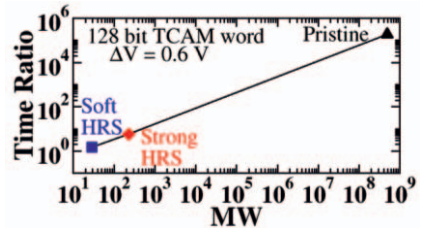


Fig. 9: Time Ratio as a function of the Memory Window. A large time ratio guarantees the correct detection of matched and mismatched words.

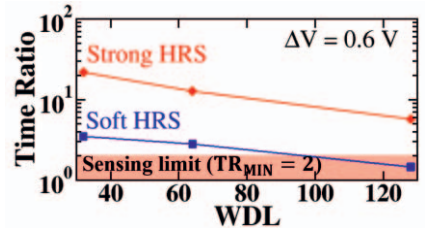


Fig. 10: Time Ratio as a function of the TCAM Word Length (WDL).

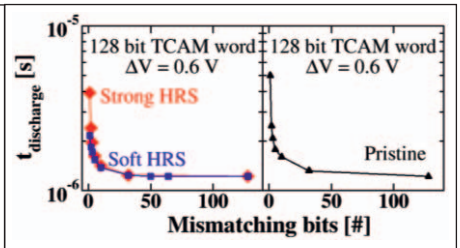


Fig. 11: Discharge time as a function of the number of mismatching bits for a 128 bit TCAM word.

IV. Search/Read Reliability

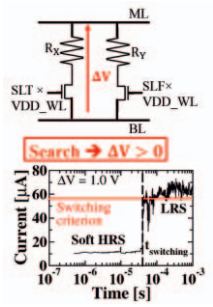


Fig. 12: During a search operation, ΔV is applied on the two 1T1R structures (top). This can cause an unwanted switching from HRS to LRS (bottom).

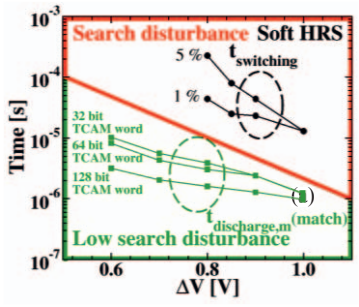


Fig. 13: Set switching time as a function of the applied voltage (black lines) extracted from CVS measurements on a 4kbit 1T1R array. The discharge time in a match case is reported for comparison (green lines). The discharge time has to be lower than the switching time.

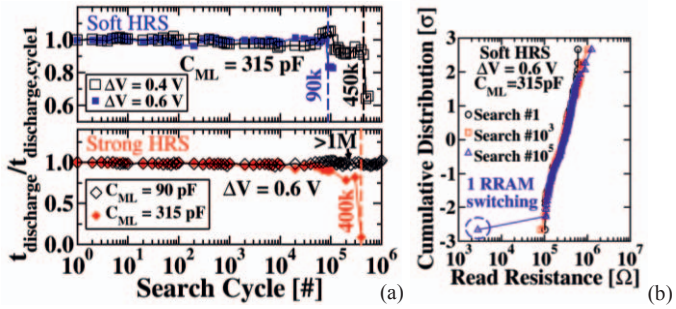


Fig. 14: (a) (Top plot) Read/search endurance with soft HRS and a match line capacitance $C_{ML} = 315$ pF, for $\Delta V = 0.6$ V and $\Delta V = 0.4$ V. (a) (Bottom plot) Search/read endurance with strong HRS and $\Delta V = 0.6$ V, for $C_{ML} = 315$ pF and $C_{ML} = 90$ pF. (b) HRS cumulative distribution after 1, 1k and 100k search operations, for the soft HRS and $\Delta V = 0.6$V. Measurements are performed on a 128 bit TCAM word in a match case configuration (all the cells in HRS).

V. Programming Reliability

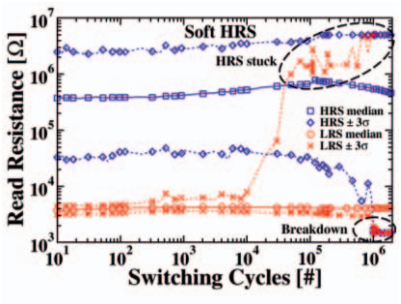


Fig. 15: Endurance characterization measured on a 4 kbit 1T1R array, for soft HRS. HRS stuck and breakdown failures are observed. HRS stuck failures occur much earlier than breakdown failures.

HRS stuck ➔ longer $t_{discharge}$
➔No impact on match failures
➔Mismatch failure with a probability
$\frac{WDL}{2^{WDL}-1} \times p_{HRS\ stuck}$ (WDL = 128 bit)

Breakdown cells ➔ shorter $t_{discharge}$
➔No impact on mismatch failures
➔Match failure with a probability
$1 \times p_{breakdown}$

(a)

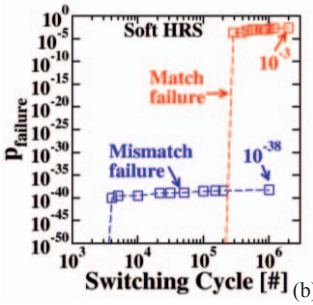


Fig. 16: (a) Impact of HRS stuck and breakdown failures on RRAM-based TCAM circuits. (b) Probability of mismatch and match failures due to the HRS stuck and breakdown failures respectively, for soft HRS.

VI. Conclusion

	ΔV↑	R_{HRS}↑	WDL↑
Search Latency (Discharge Time)	↓ ☺	↑ ☹	↓ ☺
Search Margin (Time Ratio)	↓ ☹	↑ ☺	↓ ☹
Comments	Search/read endurance ↓ ☹	➢ Search/read power ↓ ☺ ➢ Search/read endurance ↑ ☺ ➢ Programming endurance ↓ ☹	➢ Memory capacity ↑ ☺

Table 2: Summary of the study. There is a trade-off between Search Latency (Discharge Time) and search margin reliability (Time Ratio).

REFERENCES

[1] R. Karam et al., *Proc. IEEE*, 2015, vol. 103, no. 8, pp 1311-1330
[2] I. Hayashi et al., *A-SSCC*, 2012, pp. 65-68
[3] L. Nii et al., *ISSCC*, 2014, pp. 240-241
[4] Moradi et al., *TBCAS*, 2018, vol. 12 , no. 1, pp.106-122
[5] Qiao et al., *Proc. IEEE BioCAS*, 2016, pp. 552-555
[6] J. Li et al., *JSSC*, 2014, vol. 49, no. 4, pp 896-907
[7] C-C Lin et al., *ISSCC*, 2016, pp. 136-137
[8] M-F. Chang et al., *ISCAS*, 2016, pp. 1142-1145
[9] M-F. Chang et al., *JSSC*, 2016, vol. 51, no. 11, pp. 2786-2798
[10] M-F. Chang et al., *JSSC*, 2017, vol. 52, no. 6, pp. 1664-1679
[11] A. Grossi et al., *TVLSI*, 2018
[12] H. Li et al., *VLSI-TSA*, 2017, pp. 1-2