A Low-Complexity Sensing Scheme for Approximate Matching Content-Addressable Memory

Esteban Garzón, *Member, IEEE*, Roman Golman, Marco Lanuzza, *Senior Member, IEEE*, Adam Teman, *Member, IEEE*, and Leonid Yavits, *Member, IEEE*

***Abstract*—The need for approximate rather than exact search arises in numerous compare-intensive applications, from networking to computational genomics. This brief presents a novel sensing approach for approximate matching content-addressable memory (CAM) designed to handle large Hamming distances (HDs) between the query pattern and stored data. The proposed matchline sensing scheme (MLSS) employs a replica mechanism and a 12-transistor positive feedback sense amplifier to effectively resolve the approximate match operation. The MLSS was integrated into a 4 kB approximate CAM array and fabricated in a 65 nm CMOS technology. With an overall area footprint of 0.0048 mm^2, which includes 512 sense amplifiers and the replica mechanism, the MLSS allows a flexible and dynamic adjustment of the HD tolerance threshold via several design variables. Experimental measurements demonstrate the efficiency of our sensing scheme in tolerating very large HDs with the highest sensitivity.**

***Index Terms*—HD-CAM, hamming distance, content-addressable memory, approximate CAM, approximate match, matchline sense amplifier.**

I. Introduction

CONTENT-ADDRESSABLE memories (CAMs) are widely used in many applications requiring high-speed parallel search operations between an input query pattern and the complete dataset stored within the memory [1], [2]. Due to the wide demand for similarity search in numerous emerging applications such as compare-intensive big data workloads, machine learning, and pattern recognition in various domains, including images, DNA sequencing, and biomedical data, there is a growing need for CAMs that can perform approximate search operations [3], [4], [5], [6]. Conventional CMOS-based or emerging resistive memory-based CAMs typically support only exact search operations [7], [8]. Other memory sensing schemes have been proposed [1], [2], [9], [10].

Several techniques have been proposed in literature to enable approximate matching in CAMs by leveraging customized sensing schemes and other solutions (i.e., involving redundancy). For instance, error-correction codes have been suggested for Ternary CAMs (TCAMs) and NAND-type CAMs, which employ parity bits and a dedicated matchline scheme [11], [12]. However, these methods can only handle a small Hamming distance (HD) of 1 to 4 bits between the input query pattern and stored data, and their implementations require large area overhead and increased design complexity. Tunable sampling time techniques have also been explored [13]; however, their implementation is challenging due to the strong dependency on precise device and circuit sizing, susceptibility to jitter, and higher probability of generating false results (matches instead of mismatches and vice versa). A recent solution, proposed in [14], presents a large HD-tolerant approximate CAM (HD-CAM) based on matchline charge redistribution. Unfortunately, the sensing scheme presented in this brief also suffers from a high degree of design complexity and large area overhead.

This brief proposes a low-complexity, scalable, and area-efficient sensing scheme for approximate CAMs with a tunable matchline discharge rate [14], [15]. Our sensing scheme consists of a 12-transistor positive feedback sense amplifier along with a replica mechanism that provides control of the sampling time during approximate match operations. Specifically, the replica line enables the sensing of the sense amplifier that further resolves the compare result. Additional design variables allow adjusting the HD tolerance threshold and the sensitivity of the proposed sensing scheme. A 4 kB HD-CAM design [14] integrating the proposed matchline sensing scheme, was fabricated in a commercial 65 nm CMOS technology. The sensing scheme of the HD-CAM design has a silicon footprint of 0.0048 mm^2. The effectiveness of the suggested approximate match sensing scheme (i.e., its sensitivity as a function of HD and its susceptibility to variations) is evaluated through experimental measurements.

This brief provides the following main contributions:

- To our knowledge, this is the first sensing scheme for approximate search CAM, enabling highly sensitive

Manuscript received 4 April 2023; revised 22 May 2023; accepted 12 June 2023. Date of publication 14 June 2023; date of current version 25 September 2023. This work was supported in part by the European Union's Horizon Europe Programme for Research and Innovation under Grant 101047160; in part by the Israeli Ministry of Science and Technology under Lise Meitner Grant for Israeli-Swedish Research Collaboration; and in part by the Italian Ministry of University and Research (MUR) through the Project PRIN under Grant 2020LWPKH7. The work of Esteban Garzón was supported by the Italian MUR under the call "Horizon Europe 2021–2027 Programme under Grant H25F21001420001." This brief was recommended by Associate Editor Z. Di. *(Corresponding author: Esteban Garzón.)*

Esteban Garzón and Marco Lanuzza are with the Department of Computer Engineering, Modeling, Electronics and Systems, University of Calabria, 87036 Rende, Italy (e-mail: esteban.garzon@unical.it; m.lanuzza@dimes.unical.it).

Roman Golman, Adam Teman, and Leonid Yavits are with the EnICS Labs, Faculty of Engineering, Bar-Ilan University, Ramat Gan 5290002, Israel (e-mail: roman.golman@live.biu.ac.il; adam.teman@biu.ac.il; leonid.yavits@biu.ac.il).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSII.2023.3286257.

Digital Object Identifier 10.1109/TCSII.2023.3286257

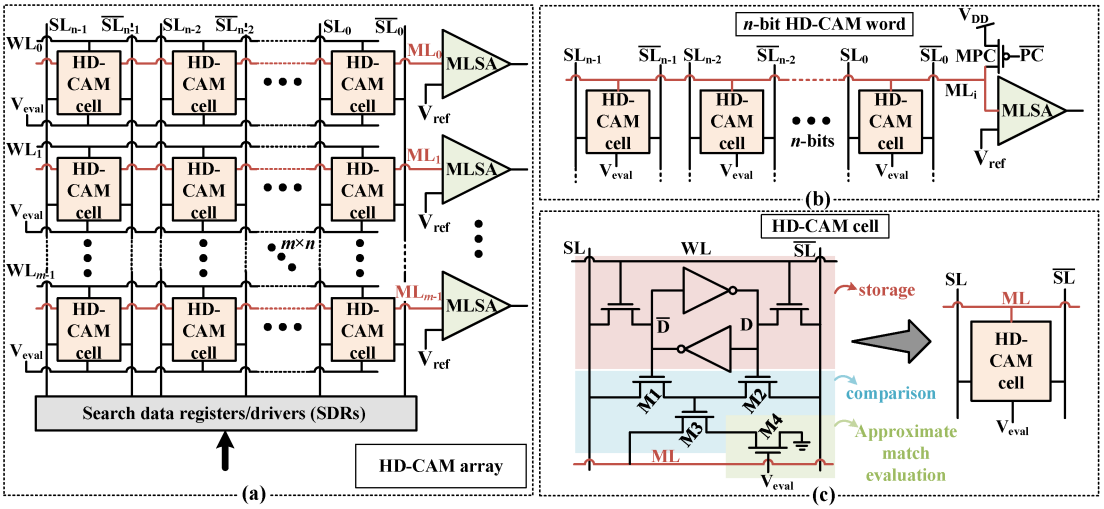


Fig. 1. Overview of the Hamming Distance (HD) tolerant CAM (HD-CAM) based on tunnable matchline discharge rate. (a) HD-CAM array. (b) n-bit HD-CAM word. (c) HD-CAM cell highlighting the storage, comparison and approximate match evaluation blocks. For the sake of simplicity, wordline (WL) is not shown in the HD-CAM cell block of (b) and (c).

approximate match sensing capabilities, that has been fabricated and evaluated in silicon.
- Our sensing scheme supports a very wide range of HD tolerance through user-configurable design variables.
- The proposed sensing scheme presents low susceptibility to sampling time, temperature, and process variations.
- Unlike state-of-the-art matchline sensing schemes, the proposed design utilizes the charge redistribution of a replica line to control the sense amplifier sampling time.

II. Background: Hamming Distance Tolerant CAM

The HD tolerant CAM (HD-CAM), proposed in [14], is capable of both exact and approximate matching; the latter tolerating HDs of up to 60% of the length of the query pattern. HD-CAM design is based on the observation that the matchline voltage drop is proportional to the HD between the query pattern and a data word. To evaluate the efficiency of HD-CAM approximate matching, it was tested as a real-time DNA classifier programmed to detect Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) DNA in a metagenomic sample (i.e., containing the DNA of multiple organisms). Noteworthy attributes of HD-CAM include its ability to tolerate large HDs with high sensitivity and precision, its resilience to DNA sequencing errors and sampling time variation, and its reduced area overhead and design complexity.

Fig. 1(a) shows the top-level schematic view of an $m \times n$ HD-CAM [14]. Each row in the CAM has its own matchline (ML), which is connected to a ML sense amplifier (MLSA). A pair of searchlines (SLs) are connected to all the bitcells in a column, thereby forming an n-bit HD-CAM word, as shown in Fig. 1(b). The precharge ($\overline{\text{PC}}$) transistor (MPC) is used to precharge the ML. The MLSA senses the state of the matchline against a reference voltage (V_{ref}). Fig. 1(c) shows the NOR-type HD-CAM bitcell, which is built upon the conventional NOR-type CAM bitcell [1]. Similar to a standard six-transistor static random access memory (6T-SRAM) cell, it is based on a pair of cross-coupled inverters for storing data and accessed for write and read by enabling row access through the word line (WL) and driving SL and $\overline{\text{SL}}$ to opposite logic values for write or pre-charging them for read. The associative search operation involves two steps: precharge and evaluation. During the precharge step, the ML is precharged to V_{DD} by enabling the MPC transistor ($\overline{\text{PC}}$ = '0'). This is followed by the evaluation step, where the MPC transistor is cut off ($\overline{\text{PC}}$ = '1'), and the query data is loaded onto the SLs. The comparison between the query pattern and the data word is performed by the M1-M3 transistors. An evaluation transistor (M4) is integrated into the HD-CAM cell to regulate the discharge rate of the ML according to the evaluation voltage level (V_{eval}). By controlling the voltage level on M4, HD-CAM can perform approximate matching when $V_{\text{eval}} < V_{\text{DD}}$, while a conventional exact match CAM operation is executed when M4 is driven by a full voltage level, $V_{\text{eval}} = V_{\text{DD}}$.

III. Proposed Matchline Sensing Scheme (MLSS)

A. Design and Operating Principle

The proposed matchline sensing scheme (MLSS) is based on a positive feedback sense amplifier (SA) that is controlled by a replica line, as illustrated in Fig. 2 (a). The ML replica line (MLRL) is composed of n replica transistors (M_{n-1} to M_0) that are connected in parallel. The gate terminals of these devices are grounded, their drain terminals are connected to V_{DD}, and their source terminals are connected to the MLRL. The MLSS includes three additional transistors (MN1, MN2, and MP), along with an inverter (I1). MN2 and MP are controlled by the $\overline{\text{PC}}$ signal. The gate of MN1 is connected to the replica voltage (V_{rep}), which in the final design is the same as V_{eval} (or V_{DD}), i.e., does not require a separate voltage source (or MN1). However, for the sake of evaluating the MLSS susceptibility to sampling time variations, we enable the V_{rep} to be biased separately to adjust the MLRL discharge, as presented hereafter.

The MLRL emulates the capacitance of a ML. Its output is the **Sen** signal that timely enables the sensing of the positive feedback SA. Fig. 2(b) details the schematic of the positive feedback SA. It comprises a pair of cross-coupled inverters with four enable transistors (MEN_1-MEN_4), whose gates are driven by **Sen**. MEN_1 (MEN_2) acts as header (footer) to connect the latch to (down to) V_{DD} (ground). The last two enable transistors, MEN_3 and MEN_4, are connected to the output

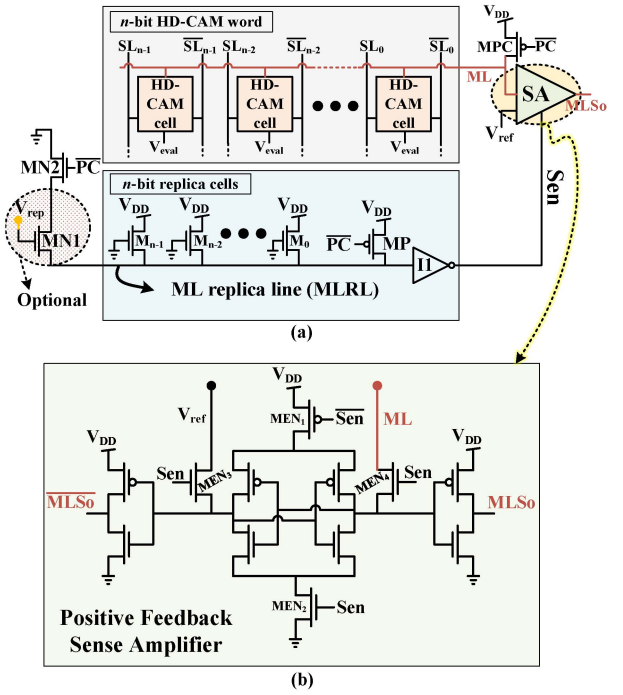


Fig. 2. Proposed matchline sensing scheme (MLSS) for HD tolerant CAM based on tunable matchline discharge rate. (a) Schematic of the MLSS. (b) schematic of the positive feedback sense amplifier.

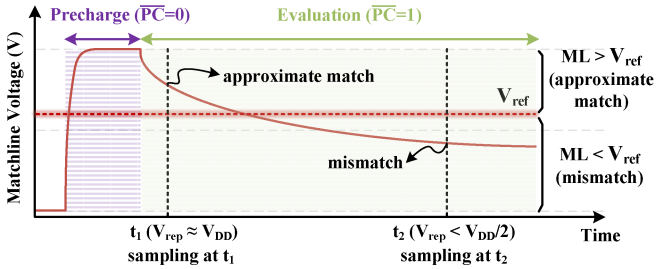


Fig. 3. Matchline sensing scheme (MLSS) behavior during search operation.

terminals of the cross-coupled inverters. The sources of the MEN_3 and MEN_4 devices are connected to V_{ref} and ML, respectively, while the drains are connected to the inverters that output $\overline{MLSo}$ and MLSo respectively.

Overall, the MLSS operation is controlled by the $\overline{PC}$ signal. When $\overline{PC}$ = '0', MLRL is charged to V_{DD}, avoiding any sensing activity by the SA. The comparison starts when $\overline{PC}$ = '1', enabling the MLRL to discharge through transistors MN1 and MN2. Once the MLRL voltage level drops below the threshold of the I1 inverter, the **Sen** triggers the start of the SA sampling. The MEN_3 and MEN_4 transistors will transfer the V_{ref} and ML signals to the terminals of the cross-coupled inverters. The highest voltage level will raise a terminal to V_{DD}, while the other terminal will drop down to ground. The role of the positive feedback SA is to compare the ML and V_{ref} signals. When the ML voltage level is above V_{ref}, the SA signals a match (exact or approximate). On the other hand, if the ML voltage falls below V_{ref}, the SA signals a mismatch, as illustrated in Fig. 3. This figure also shows two particular cases: when V_{rep} is close to V_{DD} and when V_{rep} is less than half V_{DD}. These two examples are labeled as sampling time at t_1 and t_2, and correspond to approximate match and mismatch responses, respectively. Note that lowering V_{rep} slows down the MLRL discharge, which in turn delays the **Sen** signal assertion. Therefore, the design variable (V_{rep}) may provide an additional level of flexibility, enabling the fine-tuning of the sensitivity response of the approximate match.

B. MLSS in Approximate CAM Memory Array

The top-level architecture of the HD-CAM memory array, including the MLSS and peripherals, is illustrated in Fig. 4(a). A 32-kbit HD-CAM array, organized as 512 64-bit words, comprises two 256×64 memory blocks. Two replica circuits are built in the HD-CAM array: the replica row and the replica column. The replica row has 64 transistors connected in parallel. Note that to balance the IR drop between the MLRL and the 512 SAs, the MLRL is placed at the middle of the memory array. The replica row generates the self-timed **Sen** signal to control the positive feedback SAs. Every evaluation of the SA is preceded by a precharge of the MLRL to initialize the sensing of the 512 HD-CAM words. This ensures that the evaluation phase will start only after the precharge to achieve a correct search operation across Process-Voltage-Temperature (PVT) variations. The replica column serves to synchronize the delays of the **Sen** signal and the SL/$\overline{SL}$ lines. This replica column is connected to the **Sen** line, and comprises 512 delay cells connected in parallel, as shown in Fig. 4. The delay cells represent the capacitance of the SL/$\overline{SL}$ lines that are connected along the column of the memory cell.

The layout of the HD-CAM memory array is shown in Fig. 4(b). The height of three rows matches the height of the SA. Therefore, three SAs are placed next to each other in a single row, as shown in the sub block view at the top of the memory array. The inset shows the layout of the positive feedback SA, exhibiting an area footprint of 13.4 μm^2. The total area of the MLSS including all 512 SAs, as well as the replica row and column, is 0.0048 mm^2.

IV. Experimental Results

A. Test Chip and Measurement Setup

Fig. 5(a) shows the test board with the fabricated test chip, nicknamed "LEO-II". The layout of the fabricated chip is provided on the right side of Fig. 5(a) with the approximate search CAM arrays highlighted among the various SoC components and other research projects integrated within the chip. Fig. 5(b) provides the main features of the 4 mm^2 chip, fabricated in 65 nm CMOS technology. The SoC features the operating frequency of 300 MHz at a supply voltage of 1.2 V. Fig. 5(c) shows our experimental setup, with an Intel Cyclone-V FPGA used for control and testing support during measurements.

B. Methodology and Measurement Results

Offline setup: First, we create a random dataset and store it in the HD-CAM array. Second, we build a query data set, which is the same dataset as the one stored in HD-CAM, but overlaid with random errors at a certain predefined rate (i.e., a certain number of bit errors in random positions per memory row). The error rate defines the Hamming distance (HD) between queries and the data stored in HD-CAM. Third, the MLSS is configured by setting its HD tolerance threshold using the design variables V_{eval} and V_{ref}.

Online test: The query datawords are searched one by one in the HD-CAM, and the number of matches is recorded. Since the MLSS HD threshold is configured to tolerate said HD,

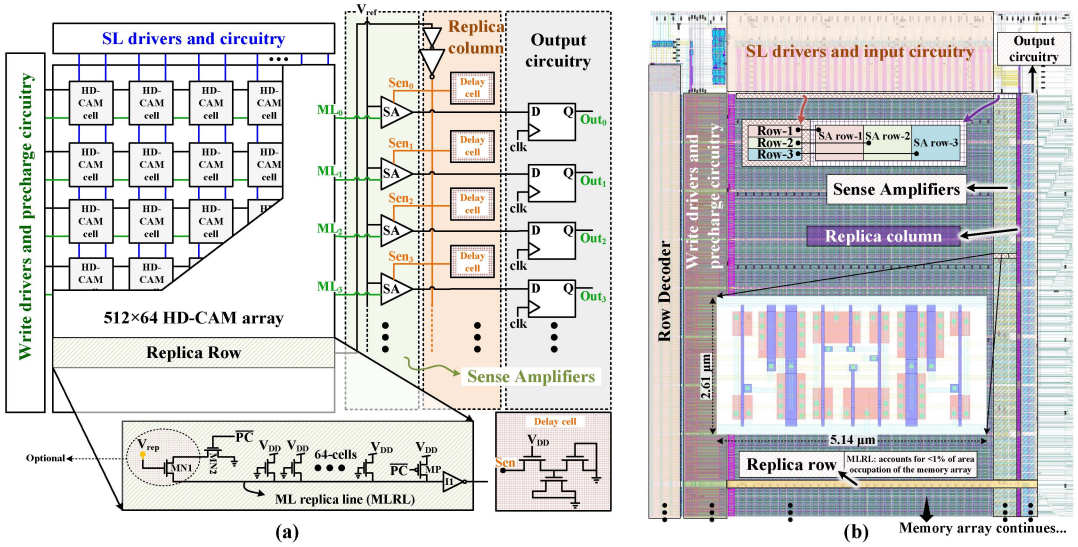
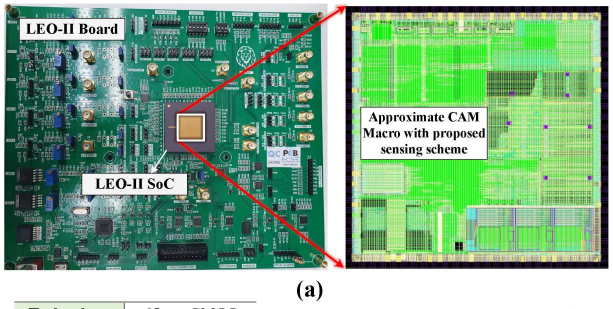


Fig. 4. (a) Top-level architecture of the HD-CAM memory array along with the matchline sensing scheme and peripherals. (b) Layout of the HD-CAM array highlighting the replica row and sense amplifiers. In the inset: the positive feedback sense amplifier layout.



Technology	65 nm CMOS
Die area	2 × 2 mm²
Sense amplifier scheme	Replica word + positive feedback SA
HD-CAM array size	4 kB
MLSS area*	0.0048 mm²
Power	~3 mW @ (1.2 V, 150 MHz)
Max. Frequency	300 MHz

* measured as the total area of the replica lines, and the SAs

(b)

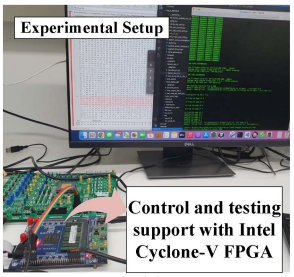


Fig. 5. (a) LEO-II SoC board along with a top-level view of the SoC layout highlighting HD-CAM, the approximate search CAM equipped with the proposed sensing scheme. (b) Main features of the test chip. (c) Photo of the experimental setup. For the purposes of testing and control, an Intel Cyclone-V FPGA is connected to the prototyping board.

we expect each query to match in HD-CAM. Therefore, all matches are true positive (*TP*) results, while all mismatches are false negative (*FN*) ones. Using these results, we can calculate the sensitivity of the MLSS as $TP/(TP + FN)$.

Silicon measurement results of the MLSS sensitivity, for different V_{ref}, V_{eval}, and V_{rep}, temperature, and different silicon samples are provided in Fig. 6. Sensitivity as a function of V_{ref} is shown in Fig. 6(a) for a HD of 4 (i.e., 4 bit errors in random positions in each HD-CAM row). $V_{\text{rep}} = V_{\text{eval}}$, and V_{eval} is varied between 0.6 V and 1.0 V. For lower values of V_{ref}, the MLSS sensitivity is 100%, meaning that MLSS tolerates the HD of 4 and correctly resolves the compare results. With increasing V_{ref}, the MLSS sensitivity diminishes, meaning some matches are falsely registered as mismatches.

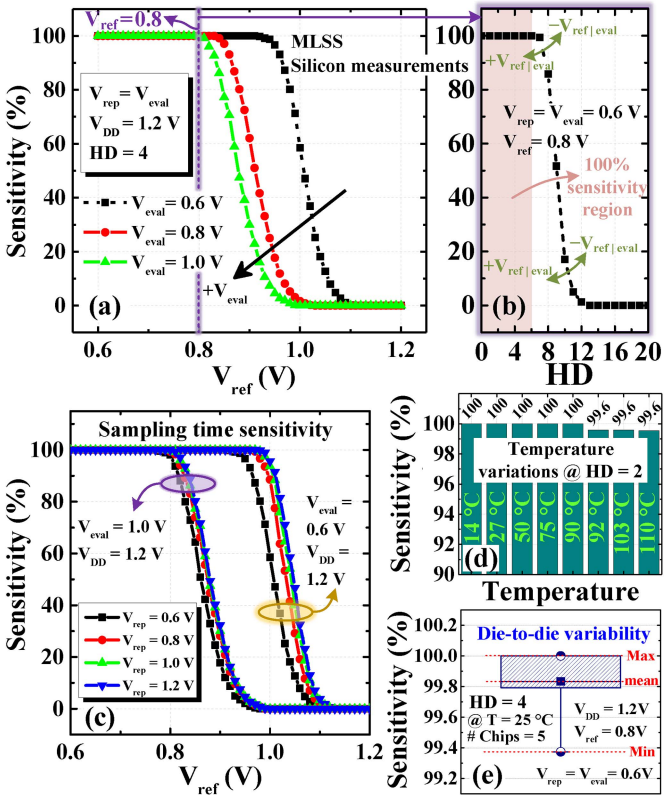


Fig. 6. Measurement results of the MLSS sensitivity for different voltages and temperature variations. Sensitivity as a function of: (a) V_{ref} and V_{eval}, (b) HD for $V_{\text{eval}} = 0.6$ V and $V_{\text{ref}} = 0.8$ V, (c) V_{ref} for different evaluation V_{rep}, (d) temperature. (e) Die-to-die variability box plot of sensitivity.

The threshold V_{ref} (i.e., the highest V_{ref} at which the MLSS sensitivity is still 100%) depends on V_{eval}. The lower the V_{eval}, the higher the threshold V_{ref}.

Fig. 6(b) shows the MLSS sensitivity as a function of HD for $V_{\text{ref}} = 0.8$ V and $V_{\text{eval}} = 0.6$. The highest HD for which the MLSS sensitivity is still 100% (refer to the highlighted region), marks the HD tolerance threshold of the MLSS. The

TABLE I
COMPARISON BETWEEN THE PROPOSED DESIGN AND OTHER POSSIBLE SENSING SCHEMES COMPATIBLE WITH HAMMING DISTANCE CAM

Design	[13]	[12]	[11]	[14]	This work
Fabricated	No	No	No	No	Yes (65 nm)
HD Tolerated	Low	Low	Low	High	High
Sensing Scheme	Counters + comparators	NAND match-line organization	Comparator	StrongARM comparator	Replica circuitry + positive feedback SA
Design Complexity	High (Precise device and circuit sizing, and sampling time)	High (Encoder, dedicated ML scheme and comparator)	Medium–High (Error-correcting code scheme)	High (Sizing, control circuitry, and precise sampling	Low (Replica circuitry controlling a SA)
Area overhead	Large	Large	Large	Large	Small
PVT Stability	Low	Medium–High	–†	High*	High*

† Not enough circuit information was reported. ∗ Flexibility to adjust the HD tolerance threshold.

higher (lower) the V_{eval} or V_{ref}, the lower (higher) the HD tolerance threshold.

We also analyze the MLSS susceptibility to sampling time variation, to model which we vary the V_{rep}. Fig. 6(c) shows the MLSS sensitivity as function of V_{ref} for V_{DD} of 1.2 V and different V_{rep} values. Two sets of measurement results are shown: for V_{eval} of 1 V and 0.6 V. For V_{eval} of 1 V, the sample timing variation shows a very little effect on the MLSS sensitivity. For V_{eval} of 0.6 V, this variation is higher mainly at V_{rep} of 0.6 V. Overall, the susceptibility to the sampling time variation is limited over a wide range of V_{eval} and V_{ref}.

Finally, we also analyze the MLSS temperature and process variability (shown in Fig. 6(d, e)), where about 100% sensitivity is maintained for a wide range of temperatures, and across 5 different chips, respectively. Note that dynamic adjustment of the MLSS design variables effectively resolves the issue when significant PVT variations adversely affect the target HD tolerance.

C. Related Work and Comparison With State-of-the-Art

Table I qualitatively compares the proposed matchline sensing scheme with other sensing approaches compatible with approximate search CAMs. These approximate matching MLSSs have a large area footprint, can only tolerate a limited HD, and present a high degree of design complexity. Garzón et al. [14] require a complex sizing process and extra peripherals. Krishnan et al. [11] use an analog comparator and additional circuitry. Efthymiou [12] requires an encoder for parity bits, a dedicated ML scheme, and an embedded comparator in each cell. Imani et al. [13] use delay lines at the clock inputs of four SAs per match line, as well as precise device and circuit sizing. The StrongARM comparator, used in [14], as well as the proposed MLSS, assure better PVT stability, mainly due to the flexibility to adjust the tolerance threshold to a wide range of HDs [14]. To our knowledge, none of these state-of-the-art designs have been silicon-proven. In contrast, the proposed MLSS has been demonstrated, by means of silicon prototyping and measurements, to provide an efficient low-complexity and low-cost solution for the tunable matchline discharge rate-based approximate search CAM.

V. CONCLUSION

This brief introduced a low-complexity, scalable, and area-efficient matchline sensing scheme for approximate search CAM based on tunable matchline discharge. Our circuit was fabricated as part of a 65 nm test chip and evaluated through post-silicon testing and measurements. The proposed sensing scheme exhibits high sensitivity over a wide range of HDs between the queries and stored data. Testing results show that the proposed design can flexibly adjust the tolerance threshold, while exhibiting very limited susceptibility to sampling time, temperature, and process variations. The proposed design offers an efficient, low-complexity and robust alternative to state-of-the-art approximate search CAM sensing approaches.

REFERENCES

[1] K. Pagiamtzis and A. Sheikholeslami, "Content-addressable memory (CAM) circuits and architectures: A tutorial and survey," *IEEE J. Solid-State Circuits*, vol. 41, no. 3, pp. 712–727, Mar. 2006.
[2] K. Pagiamtzis, N. Azizi, and F. N. Najm, "A soft-error tolerant content-addressable memory (CAM) using an error-correcting-match scheme," in *Proc. IEEE Custom Integr. Circuits Conf.*, 2006, pp. 301–304.
[3] M. Imani, Y. Kim, A. Rahimi, and T. Rosing, "ACAM: Approximate computing based on adaptive associative memory with online learning," in *Proc. Int. Symp. Low Power Electron. Design*, 2016, pp. 162–167.
[4] M. Ali, A. Agrawal, and K. Roy, "RAMANN: In-SRAM differentiable memory computations for memory-augmented neural networks," in *Proc. ACM/IEEE Int. Symp. Low Power Electr. Design*, 2020, pp. 61–66.
[5] M. M. Taha and C. Teuscher, "Approximate memristive in-memory hamming distance circuit," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 16, no. 2, pp. 1–14, 2020.
[6] R. Kaplan, L. Yavits, and R. Ginosasr, "BioSEAL: In-memory biological sequence alignment accelerator for large-scale genomic data," in *Proc. 13th ACM Int. Syst. Stor. Conf.*, 2020, pp. 36–48.
[7] B. Song, T. Na, J. P. Kim, S. H. Kang, and S.-O. Jung, "A 10T-4MTJ nonvolatile ternary CAM cell for reliable search operation and a compact area," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 64, no. 6, pp. 700–704, Jun. 2017.
[8] E. Garzón, M. Lanuzza, A. Teman, and L. Yavits, "AM^4: MRAM crossbar based CAM/TCAM/ACAM/AP for in-memory computing," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 13, no. 1, pp. 408–421, Mar. 2023.
[9] Z. Yang et al., "A novel computing-in-memory platform based on hybrid Spintronic/CMOS memory," *IEEE Trans. Electron Devices*, vol. 69, no. 4, pp. 1698–1705, Apr. 2022.
[10] N. Mohan, W. Fung, D. Wright, and M. Sachdev, "A low-power ternary CAM with positive-feedback match-line sense amplifiers," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 56, no. 3, pp. 566–573, Mar. 2009.
[11] S. C. Krishnan, R. Panigrahy, and S. Parthasarathy, "Error-correcting codes for ternary content addressable memories," *IEEE Trans. Comput.*, vol. 58, no. 2, pp. 275–279, Feb. 2009.
[12] A. Efthymiou, "An error tolerant CAM with nand match-line organization," in *Proc. 23rd ACM Int. Conf. Great Lakes Symp. VLSI*, 2013, pp. 257–262.
[13] M. Imani, A. Rahimi, D. Kong, T. Rosing, and J. M. Rabaey, "Exploring hyperdimensional associative memory," in *Proc. IEEE Int. Symp. High Perform. Comput. Architect. (HPCA)*, 2017, pp. 445–456.
[14] E. Garzón et al., "Hamming distance tolerant content-addressable memory (HD-CAM) for DNA classification," *IEEE Access*, vol. 10, pp. 28080–28093, 2022.
[15] R. Hanhan, E. Garzón, Z. Jahshan, A. Teman, M. Lanuzza, and L. Yavits, "EDAM: Edit distance tolerant approximate matching content addressable memory," in *Proc. 49th Annu. Int. Symp. Comput. Architect.*, 2022, pp. 495–507.

Open Access funding provided by 'Università della Calabria' within the CRUI CARE Agreement