A High Speed Low Power CAM With a Parity Bit and Power-Gated ML Sensing

Anh-Tuan Do, Shoushun Chen, *Member, IEEE*, Zhi-Hui Kong and Kiat Seng Yeo, *Senior Member, IEEE*

***Abstract*—Content addressable memory (CAM) offers high-speed search function in a single clock cycle. Due to its parallel match-line (ML) comparison, CAM is power-hungry. Thus, robust, high-speed and low-power ML sense amplifiers are highly sought-after in CAM designs. In this paper, we introduce a parity bit that leads to 39% sensing delay reduction at a cost of less than 1% area and power overhead. Furthermore, we propose an effective gated-power technique to reduce the peak and average power consumption and enhance the robustness of the design against process variations. A feedback loop is employed to auto-turn off the power supply to the comparison elements and hence reduce the average power consumption by 64%. The proposed design can work at a supply voltage down to 0.5 V.**

I. INTRODUCTION

CAM is a type of solid-state memory in which data are accessed by their contents rather than physical locations. It receives input search data, i.e. a search word, and returns the address of a similar word that is stored in its data-bank [1].

In general, a CAM has three operation modes: READ, WRITE and COMPARE, among which ”COMPARE” is the main operation as CAM rarely reads or writes [4]. Fig. 1(a) shows a simplified block diagram of a CAM core with an incorporated search data register and an output encoder. It starts a compare operation by loading an n-bit input search word into the search data register. The search data are then broadcast into the memory banks through n pairs of complementary search-lines (SLs) and directly compared with every bit of the stored words using comparison circuits. Each stored word has a ML that is shared between its bits to convey the comparison result. Location of the matched word will be identified by an output encoder, as shown in Fig. 1(a). During a pre-charge stage, the MLs are held at ground voltage level while both SL and $\sim SL$ are at V_{DD}. During evaluation stage, complementary search data is broadcast to the SLs and $\sim SLs$. When mismatch occurs in any CAM cell (for example at the first cell of the row D = ’1’; $\sim D$ = ’0’; SL = ’1’; $\sim SL$ = ’0’), transistor $P3$ and $P4$ will be turned on, charging up the ML to a higher voltage level. A sense amplifier (MLSA) is used to detect the voltage change on the ML and amplifies it to a full CMOS voltage output. If mismatch happens to none of the cells on a row, no charge up path will be formed and the voltage on the ML will remain unchanged, indicating a match.

Since all available words in the CAMs are compared in parallel, result can be obtained in a single clock cycle. Hence, CAMs are faster than other hardware- and software-based search systems [1]. They are therefore preferred in high-throughput applications such as network routers and data compressors. However, the full parallel search operation leads to critical challenges in designing a low-power system for high-speed high-capacity CAMs [1]: (1) The power hungry nature due to the high switching activity of the SLs and the MLs; (2) A huge surge-on current (i.e. peak current) occurs at the beginning of the search operation due to the concurrent evaluation of the MLs may cause a serious IR drop on the power grid, thus affecting the operational reliability of the chip [1]. As a result, numerous efforts have been put forth to reduce both the peak and the total dynamic power consumption of the CAMs [2–8]. For example, C. A. Zukowski *et. al.* and K. Pagiamtzis *et. al.* introduced selective pre-charge and pipeline architecture, respectively to reduce the peak and average power consumption of the CAM [8]. [5], [6] and [3] utilized the ML pre-charge low scheme (i.e. low ML swing) to reduce the average power consumption. These designs however are sensitive to process and supply voltage variations. As will be shown later in section IV, they can hardly be scaled down to sub-65 nm CMOS process.

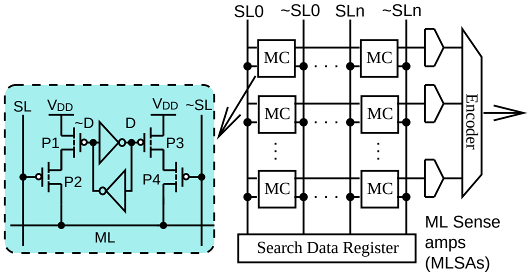


Fig. 1. Block diagram of a conventional CAM

In this work, a parity-bit is introduced to boost the search speed of the parallel CAM with less than 1% power and area overhead. Concurrently, a power-gated ML sense amplifier is proposed to improve the performance of the CAM ML comparison in terms of power and robustness. It also reduces the peak turn-on current at the beginning of each search cycle. The rest of paper is organized as follows: Section II introduces parity-bit based CAM architecture. In section III, the gated-power technique is proposed. Performance analysis are presented in Section IV. Section V concludes the paper.

II. SEARCH SPEED BOOST USING A PARITY BIT

We introduce a versatile auxiliary bit to boost the search speed of the CAM at the cost of less than 1% area overhead and power consumption. This newly introduced auxiliary bit at a glance is similar to the existing $Pre-computationschemes$ but in fact has a different operating principle. We first briefly

A.T. Do is with Virtus, IC Design Centre of Excellent, School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore (e-mail: atdo@ntu.edu.sg). S.S. Chen, K.S. Yeo and Z.H. Kong are with the School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore (e-mail: eechenss, eksyeo, and zhkong@ntu.edu.sg).

discuss the $Pre - computationschemes$ before presenting our proposed auxiliary bit scheme.

1) Pre-computation CAM design: The pre-computation CAM uses additional bits to filter some mismatched CAM words before the actual comparison. These extra bits are derived from the data bits and are used as the first comparison stage. For example, in Fig. 2(a) number of ”1” in the stored words are counted and kept in the $Counting\ bits$ segment. When a search operation starts, number of ”1”s in the search word is counted and stored to the segment on the left of Fig. 2(a). These extra information are compared first and only those that have the same number of ”1”s (e.g. the 2^{nd} and the 4^{th}) are turned on in the second sensing stage for further comparison. This scheme reduces a significant amount of power required for data comparison, statistically. The main design idea is to use additional silicon area and search delay to reduce energy consumption.

The above mentioned pre-computation and all other existing designs shares one similar property: The ML sense amplifier essentially has to distinguish between the matched ML and the 1-mismatch ML. This makes CAM designs sooner or latter face challenges since the driving strength of the single turned-on path is getting weaker after each process generation while the leakage is getting stronger. This problem is usually referred to as $\frac{I_{on}}{I_{off}}$. Thus, we propose a new auxiliary bit that can concurrently boost the sensing speed of the ML and at the same time improve the $\frac{I_{on}}{I_{off}}$ of the CAM by two times.

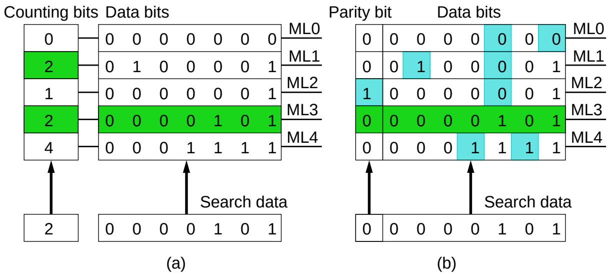


Fig. 2. A conceptual view of (a) conventional pre-computation CAM. and (b) proposed parity-bit based CAM

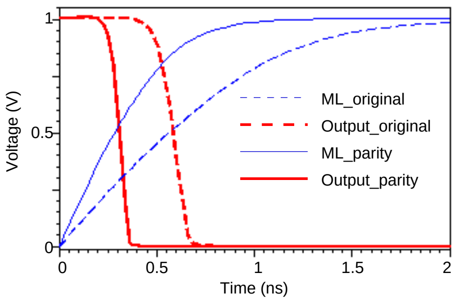


Fig. 3. 1-mismatch ML waveforms of the original and the proposed architecture with parity bit during the search opearation

2) Parity bit based CAM: The parity bit based CAM design is shown in Fig. 2(b) consisting of the original data segment and an extra one-bit segment, derived from the actual data bits. We only obtain the parity bit, i.e. odd or even number of ”1”s. The obtained parity bit is placed directly to the corresponding word and ML. Thus the new architecture has the same interface as the conventional CAM with one extra bit. During the search operation, there is only one single stage as in conventional CAM. Hence, the use of this parity bits does not improve the power performance.

However, this additional parity bit, in theory, reduces the sensing delay and boosts the driving strength of the 1-mismatch case (which is the worst case) by half, as discussed below.

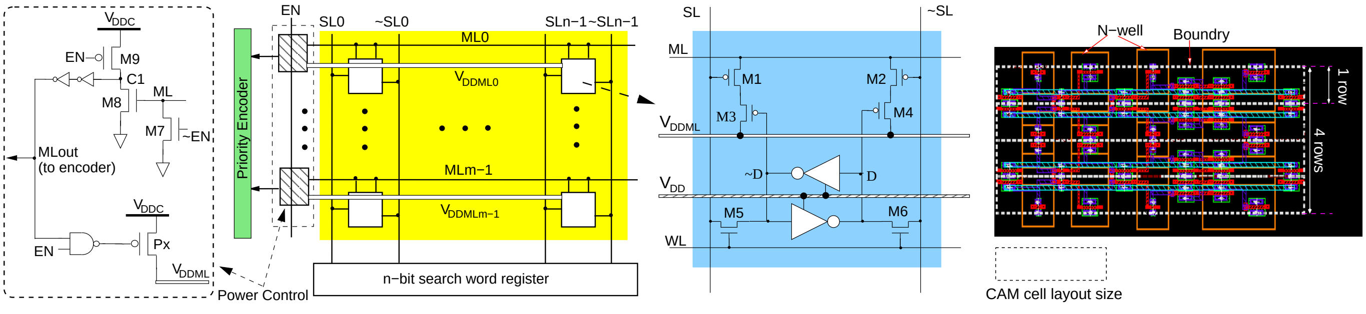
In the case of a matched in the data segment (e.g. $ML3$), the parity bits of the search and the stored word is the same, thus the overall word returns a match. When 1 mismatch occurs in the data segment (e.g. $ML2$), numbers of ”1”s in the stored and search word must be different by 1. As a result, the corresponding parity bits are different. Therefore now we have two mismatches (one from the parity bit and one from the data bits). If there are two mismatches in the data segment (e.g. $ML0$, $ML1$ or $ML4$), the parity bits are the same and overall we have two mismatches. With more mismatches, we can ignore these cases as they are not crucial cases. The sense amplifier now only have to identify between the 2-mismatch cases and the matched cases. Since the driving capability of the 2-mismatch word is twice as strong as that of the 1-mismatch word, the proposed design greatly improves the search speed and the $\frac{I_{on}}{I_{off}}$ ratio of the design. Fig. 3 shows the 1-mismatch ML transient waveforms of the original and the proposed architecture during the search operation. In the next section, we are going to proposed a new sense amplifier that reduces the power consumption of the CAM.

III. THE GATED-POWER ML SENSE AMPLIFIER DESIGN

A. Operating principle

The proposed CAM architecture is depicted in Fig. 4. The CAM cells are organized into rows (word) and columns (bit). Each cell has the same number of transistors as the conventional P-type NOR CAM (shown in Fig. 1) and use a similar ML structure. However, the ”COMPARISON” unit, i.e transistors $M1$-$M4$, and the “SRAM” unit, i.e the cross-coupled inverters, are powered by two separate metal rails, namely V_{DDML} and the V_{DD}, respectively. The V_{DDML} is independently controlled by a power transistor (Px) and a feedback loop that can auto turn-off the ML current to save power. The purpose of having two separate power rails of (V_{DD} and V_{DDML}) is to completely isolate the SRAM cell from any possibility of power disturbances during COMPARE cycle.

As shown in Fig. 4, the gated-power transistor Px, is controlled by a feedback loop, denoted as ”Power Control” which will automatically turn off Px once the voltage on the ML reaches a certain threshold. At the beginning of each cycle, the ML is first initialized by a global control signal EN. At this time, signal EN is set to low and the power transistor Px is turned OFF. This will make the signal ML and $C1$ initialized to ground and V_{DD}, respectively. After that, signal EN turns $HIGH$ and initiates the COMPARE phase. If one or more mismatches happen in the CAM cells, the ML will be charged up. Interestingly, all the cells of a row



(a) (b) (c)

Fig. 4. (a) Proposed CAM architecture(b) Each CAM cell is powered by two power rails, V_{DDML} for the compare transistors, V_{DD} for the SRAM transistors. The rail V_{DDML} of a row is connected to the power network V_{DDC} via a PMOS device Px, which is used to limit the transient current. All the cells of a row will share the limited current offered by the transistor Px, despite whatever number of mismatches. (c) Layout of four power control blocks of four rows. Each block has the same height as the CAM cell to fit in the row pitch and 2.5x longer the length of one CAM cell

will share the limited current offered by the transistor Px, despite whatever number of mismatches. When the voltage of the ML reaches the threshold voltage of transistor $M8$ (i.e. V_{th8}), voltage at node $C1$ will be pulled down. After a certain but very minor delay, the NAND2 gate will be toggled and thus the power transistor Px is turned off again. As a result, the ML is not fully charged to V_{DD}, but limited to some voltage slightly above the threshold voltage of M8, V_{th8}.

Fig. 5 shows the simulation result of the proposed power controller. One can see that, the slopes of the ML, node $C1$ and node ML_{out} depend on the number of mismatches. When more mismatches happen (e.g. 128 in the simulation), the ML and node $C1$ change faster. Less number of mismatches (e.g. 1 in the simulation) will slow down the transition of node $C1$ and results in a longer delay to turn off transistor Px. The voltage on the ML is finally charged to only around 0.5 V which is far below V_{DD} and hence the power consumption is reduced.

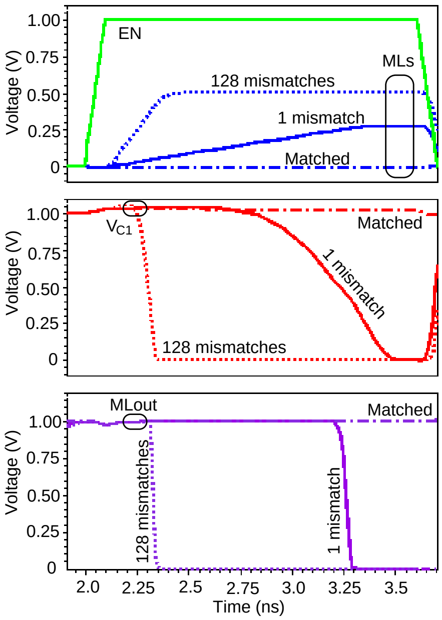


Fig. 5. Waveforms of some important nodes during evaluation of three rows of 128-bit of the proposed design.

With the introduction of the power transistor Px, the driving strength of the 1-mismatch case is about 10% weaker than that of the conventional design and thus slower. However, as we combine this sense amplifier with the parity bit scheme mentioned in Section II, the overall search delay is improved by 39%. Thus the new CAM architecture offers both low-power and high-speed operation.

B. CAM cell layout

Fig. 6 shows the layout of the CAM cell using 65 nm CMOS process. Since the new CAM cell has a similar topology of that of the conventional design (except the routing of V_{DDML}), their layouts are also similar. These two cell layouts have the same length but different heights. In the new architecture, V_{DDML} cannot be shared between two adjacent rows, resulting in a taller cell layout, which incurs about 11% area overhead, as shown in Fig. 6.

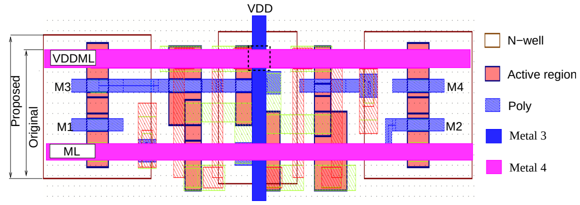


Fig. 6. Layout of the proposed CAM cell. Nets ML and V_{DDML} are routed horizontally on $Metal4$ - (i.e. purple) while net V_{DD} is routed vertically (i.e. blue)

IV. PERFORMANCE COMPARISONS

In this section, performance of the proposed design will be evaluated using the conventional circuit and those in [5], [6] as references. In [5], the power consumption is limited by the amount of charge injected to the ML at the beginning of the search. In [6], a similar concept is utilized with a positive feedback loop to boost the sensing speed. Both designs are very power efficient. As will be shown latter, the proposed design consumes slightly higher power consumption when compared with [5] and [6] but is more robust against PVT variations.

A. Peak current and IR drop attenuation

The proposed power controller demonstrates a great reduction in the transient peak current. This can be explained by the bottleneck effect of transistor Px.

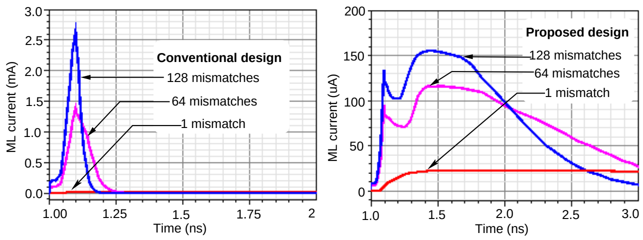


Fig. 7. Simulated transient current occurred on a row of 128 CAM cells during the compare cycle of the conventional (left) and the proposed (right) designs.

Fig. 7 shows the transient current as a function of the number of mismatches occurring in a row of 128 CAM cells during the COMPARE cycle of the proposed and the conventional designs. The conventional design's peak current increases almost linearly from 25 μA (1 mismatch) to 1.45 mA (64 mismatches) and finally 2.8 mA (128 mismatches). Although the overall transient ML charge up current of the proposed design also increases with the number of mismatches, it will soon reach its limit due to the presence of the gated-power transistor Px. For instance, when 128 mismatches occurs, the peak current is capped at 155 μA, which is less than eight times as compared to the case when only one mismatch occurs (i.e. 21 μA). This drastic reduction in the peak current translates to a vast improvement in operation reliability. Our simulation result has shown that for a 8K × 128 CAM array implemented in a 65nm CMOS process, the worst-case IR drop at the center of the conventional CAM can be as large as 0.18 V (i.e. 18 % V_{DD}) while that of the proposed design is only 8 mV (i.e. 0.8 % V_{DD}). Also, it only requires the V_{DDML} net to have a width of only 150 nm instead of 2 μm vertical V_{DD}. The new vertical V_{DD} now only supply the leakage current to the SRAM cell and thus does not require a large metal width.

B. Dynamic power consumption

Because the power-gated transistor is turned off after the output is obtained at the sense amplifier, the proposed technique renders a lower average power consumption. This is mainly due to the reduced voltage swing on the ML bus. Another contributing factor to the reduced average power consumption is that the new design does not need to precharge the SL buses because the EN signal turns off transistor Px of each row and hence the SL buses do not need to be pre-charged, which in turn saves 50% power on the SL buses.

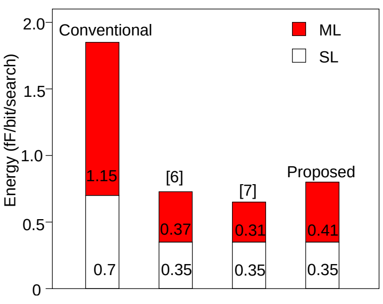
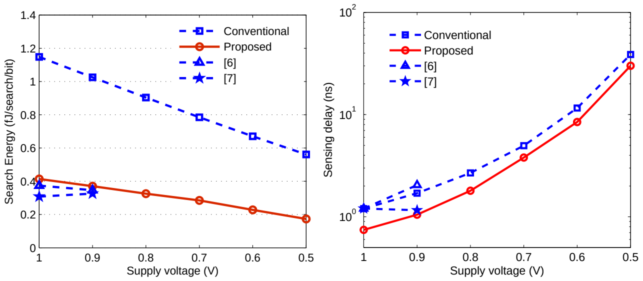


Fig. 8. Total average energy consumption of the four designs in consideration.

Fig. 8 illustrates the average energy consumption (divided into ML power and SL power) of the proposed design as compared to other three benchmark designs, including all the power overhead of the control circuitry. Since [5], [6], and the proposed design do not pre-charge the SLs before each compare cycle, their SLs energy consumption is only half of that of the conventional circuit. As for the ML energy, at 1 V supply voltage the proposed design only dissipates 0.41 fJ/search/bit while that of the conventional design is 1.148 fJ/search/bit. Our ML energy consumption is higher than that of [5] (10.8%) and [6] (32%) but as will be shown below, our proposed design is much more robust against process and environment variations.

C. Supply voltage scaling analysis

We investigate the ability of the four designs to work at low supply voltage, by re-implementing the designs in [5], [6] and the conventional one into the same 65 nm technology. Designs in [5] and [6] demonstrate poor adaptability to voltage scaling. They can not operate at a supply voltage lower than 0.9 V. On the contrary, when the supply voltage scales to 0.5 V, both the proposed and the conventional design can work well.



(a) (b)

Fig. 9. (a) Search energy per bit (b) ML sensing delay of the four designs in consideration against supply voltage scaling from 1V to 0.5 V. Sensing delay is defined as the sensing delay of the 1-mismatch ML, i.e. the worst-case scenario.

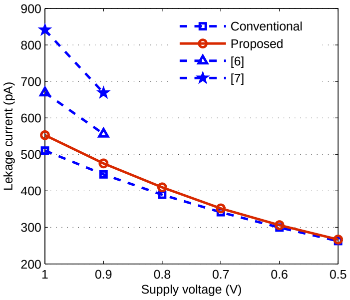


Fig. 10. Standby leakage current of the four designs in consideration against supply voltage scaling from 1V to 0.5 V.

First, the search energy of the four designs in consideration is presented in Fig. 9(a). It can be seen that at 1 V supply voltage, [5] and [6] have the lowest energy consumption per search, followed by the proposed design. However, they cease to work when the supply voltage scales down to be low 0.9 V. Between the conventional and the proposed design, the proposed design consumes 62% less power consumption at any supply voltage value. Second, the sensing delay comparison is shown in Fig. 9 where the proposed design has 39% improvement when compared to the conventional design and is the fastest design. This figure also suggests that sensing delay increases dramatically when supply voltage enters the near-subthreshold region. Finally, the corresponding leakage currents of the four designs against voltage scaling is are shown in Fig. 10. The proposed design is the second-best circuit after the conventional design. Both of them have about 20% and 37% lower leakage current when compared to [5] and [6] at 1 V, respectively. This feature confirms that the proposed design is more suitable for ultra-low power applications in 65 nm CMOS process and beyond.

D. Temperature variation analysis

We also carry out the temperature variation analysis on the four designs (Fig. 11). It can be seen that the [6] is the most vulnerable design and thus can only work in a narrow range of temperature variation. [5] can work through out the whole temperature range but having more than 30% speed fluctuation. In contrast, the proposed and the conventional design are much more stable with less than 4% sensing delay variation.

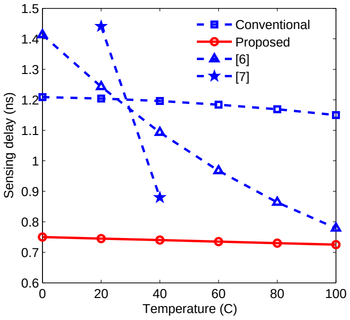


Fig. 11. ML sensing delay of the four designs in consideration against temperature variations. Sensing delay is defined as the sensing delay of the 1-mismatch ML, i.e. the worst-case scenario.

E. Process variation analysis

Process variation is a critical issue in nano-scale CMOS technologies. We simulate the performance of the proposed design against empirical process variation data from the foundry. It is worth mentioning here that the feedback loop to turn off the gated-power transistor Px operates digitally and hence is almost insensitive to process variations. Similar to the conventional design, there are two scenarios where the proposed design may sense the results wrongly: (1) The sense amplifier is enabled too early, the 1-mismatch ML has not been pulled up to a voltage higher than the threshold value and thus trigger the output inverter. (2) The delay of the enable signal is too long, resulting in the matched ML to be pulled up by the leakage current, indicating wrong miss. We use 50000-cycle Monte-carlo simulations on these designs at different supply voltages and count the number of errors accordingly. The [5] and [6] are very sensitive to process variations with more than 1000 and 10000 errors count, respectively. Also, they stop working at 0.9 V supply. On the contrary, the proposed and the conventional design has no sensing error even if V_{DD} scales down to 0.7 V. At lower supply voltage, the conventional design continues to work 100% correctly while the proposed design has 51 and 298 error counts at 0.6 and 0.5 V, respectively. This is because both designs operates at the same frequency but the proposed design has a smaller pull-up current due to the gated-power transistor Px and hence some times error happens. We have carried out a separate simulation for the proposed design with a slightly slower frequency and has confirmed that no error occurs. It is worth mentioning here that extending the period for [5] and [6] does not result in any error count reduction since these designs are based on feedback loop structure and decisions are made at the very beginning of the sensing cycle.

V. CONCLUSION

We proposed an effective gated-power technique and a parity-bit based architecture that offer several major advantages, namely reduced peak current (and thus IR drop), average power consumption (36%), boosted search speed (39%) and improved process variation tolerance. It is much more stable than recently published designs while maintain their low-power consumption property. When compared to the conventional design, its stability is degraded by 0.6% only at extremely low supply voltages. At 1 V operating condition, both designs are equally stable with no sensing errors, according to our Monte-carlo simulations. Its area overhead is about 11%. It is therefore the most suitable design for implementing high capacity parallel CAM in sub-65 nm CMOS technologies.

REFERENCES

[1] K. Pagiamtzis et. al. ”Content-addressable memory (CAM) circuits and architectures: a tutorial and survey,” IEEE Journal of Solid-State Circuits, vol. 41, pp. 712-727, 2006.

[2] W. Chao-Ching, et al., ”High-Speed and Low-Power Design Techniques for TCAM Macros,” IEEE Journal of Solid-State Circuits, vol. 43, pp. 530-540, 2008.

[3] I. Arsovski et. al. ”A mismatch-dependent power allocation technique for match-line sensing in content-addressable memories,” IEEE Journal of Solid-State Circuits, vol. 38, pp. 1958-1966, 2003.

[4] N. Mohan et. al. , ”Low-Leakage Storage Cells for Ternary Content Addressable Memories,” IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 17, pp. 604-612, 2009.

[5] O. Tyshchenko, et al., ”Match sensing using match-line stability in content addressable memorys (CAM),” IEEE Journal of Solid-State Circuits, vol. 43, pp. 1972-1981,2008.

[6] N. Mohan, et al., ”A Low-Power Ternary CAM With Positive-Feedback Match-Line Sense Amplifiers,” IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 56, pp. 566-573, 2009.

[7] S. Baeg, ”Low-Power Ternary Content-Addressable Memory Design Using a Segmented Match Line,” IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 55, pp. 1485-1494, 2008.

[8] K. Pagiamtzis and A. Sheikholeslami, ”A low-power content-addressable memory (CAM) using pipelined hierarchical search scheme,” IEEE Journal of Solid-State Circuits, vol. 39, pp. 1512-1519, 2004.