Design and Implementation of High Speed Low Power CAM with a Parity Bit and Power-Gated ML Sensing

Parameshwar Reddy (MTech), Krishna Mohan V S S S MTech.., Thyagaraj.S (MTech)

Department of Instrumentation Technology, Dayananda Sagar College of Engineering (Bangalore),
Vishvesharaya Technological University, Belgaum, Karnataka, India.

parameshec72@gmail.com   krishnamohan60@gmail.com   thyagraj.s.shekar@gmail.com

Abstract— Content addressable memory (CAM) is a type of solid-state memory in which data are accessed by their contents rather than physical locations. A Content-Addressable Memory compares input search data against stored data, and returns the address of the matching data. CAM offers high speed search function in a single clock cycle. In the conventional Content-Addressable Memory, equal power is consumed to determine if a stored word is matched to a search word or mismatched. A match line (ML) sensing scheme is used for match decision. Due to its parallel match-line (ML) comparison, CAM consumes more power. So, robust, high-speed and low power sense amplifiers are highly sought-after in CAM designs.

In this paper we introduce a parity bit that leads to delay reduction, area and power overhead. Furthermore, we propose an effective gated-power technique to reduce average power consumption and enhance the robustness of the design process variations. A feedback loop is employed to auto-turn off the power supply to the comparison elements and hence reduces the average power consumption. During the evolution stage ML is not fully charged to VDD. So sensing delay is also comparatively reduced. This paper compares the power dissipation and sensing delay of proposed CAM with existing structures.

Keywords: content addressable memory, Match line, parity bit, Power Gated ML sensing.

I.   INTRODUCTION

Most of the memory devices store and retrieve data by addressing specific memory locations. This path becomes the limiting factor for those systems that depend on fast memory access. The time required to find the data stored in memory can be reduced if the data can be identified by its content rather than by its address. A memory used for this purpose is Content Addressable Memory (CAM). CAM is used in applications where search time is very critical and very short. It is well suited for several functions like Ethernet address lookup, data compression, and security or encryption information on a packet-by-packet basis for high performance data switches. It can also be operated as a data parallel or Single Instruction/Multiple Data (SIMD) processor. Since CAM is an extension of RAM first, we have to know the RAM features to understand CAM [4]. In general RAM has two operations read and write i.e. the data stored in RAM can be read or written but CAM has three operations read, write and compare [2]. The compare operation of CAM makes it

useful in variety of applications like network routers. The network router is that which forwards the incoming packets from the sender port to the proper destination port by looking in to its routing table. Basically CAMs are used to design network routers for fast transfer or forwarding of packets.

We now take a more detailed look at CAM architecture. A small model is shown in figure 1. The figure 1 shows CAM consisting of 4 words, with each word containing 3 bits arranged horizontally (corresponding to 3 CAM cells). There is a match-line corresponding to each word (ML0, ML1, etc.) feeding into match line sense amplifiers (MLSAs), and there is a differential search line pair corresponding to each bit of the search word (SL0, $\overline{SL0}$, SL1, $\overline{SL1}$, etc.). CAM search operation begins with loading the search-data word into the search-data registers followed by pre-charging all match lines high, putting them all temporarily in the match state. Next, the search line drivers broadcast the search word onto the differential search lines, and each CAM core cell compares its stored bit against the bit on its corresponding search lines. Match lines on which all bits match remain in the pre-charged-high state. Match lines that have at least one bit that misses, discharge to ground. The MLSA then detects whether its match line has a matching condition or miss condition. Finally, the encoder maps the match line of the matching location to its encoded address.

Basic operation of CAM

Fig.1 shows the basic block diagram of a CAM, consisting of an array of storage elements, a search word register and a column of sense amplifiers. Each row of the array stores one word (144 bits) and has one associated match line (ML). This ML corresponding to each word feeding into match line sense amplifies. The ML is used to signal whether the stored word matches or mismatches the search word. There is a differential search line pair corresponding to each bit of the search word and compared bitwise against each stored word. CAM search operation begins with loading the search data word into the search data registers followed by pre-charging all match lines high, putting them all temporarily in the match state. Next, the search line drivers broadcast the search word onto the differential search lines and each CAM core cell compares its stored bit against the bit on its corresponding search lines. Match lines on which all bits

match remain in the pre-charged high state. Match lines that have at least one bit that misses, discharge to ground. The major portion of the CAM power is consumed during this parallel comparison, where all of the highly capacitive MLs are charged and discharged in every cycle.

One way to reduce CAM power is to reduce the switching capacitance on the MLs by using the NAND ML architecture. This architecture consists of a number of NAND-type CAM cells connected in series to create a long pass transistor network. NAND ML architecture suffers from unacceptably long search delays that grow quadratically with the number of CAM cells in series. To achieve a higher speed, NOR ML architecture is preferred. The NOR architecture consists of CAM cells that are connected in parallel, instead of in series.

II. METHODOLOGY

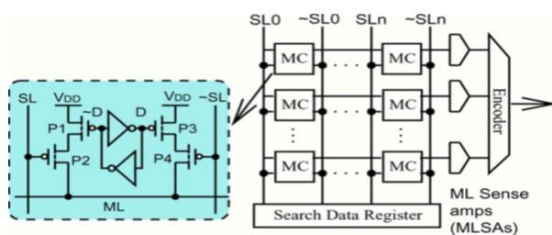*Conventional Block Diagram of CAM*

Fig.2: Block diagram of a conventional CAM.

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. Fig.1 shows a block diagram of a CAM core with an incorporated search data register and an output encoder. By loading an n bit word in to the search data register, compare operation will start. Then the search data are broadcast into the memory banks through n pairs of complementary search-lines (SL) 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. An encoder will identify the location of matched word, as shown in Fig.2. The MLs are held at ground voltage level while both SL and ~SL are at VDD during pre-charging phase. During evaluation stage, complementary search data is broadcast to the SL and ~SL. When mismatch occurs in any CAM cell (for example consider the first cell of the row D="1"; ~D= "0"; SL= "1"; ~SL="0"), transistor P3 and P4 will be turned on, charging up the ML to a higher voltage level. To detect the voltage change on the ML a sense amplifier (MLSA) is used and amplifies it to a full CMOS voltage output. If no charge up path unchanged. In this work, a power-gated ML sense amplifier is proposed to improve the performance of the CAM ML comparison in terms of power and robustness. Also reduces the peak turn-on current at the beginning of each search cycle of operation [1].

However, the full parallel search operation leads to critical challenges in designing a low-power system for high-speed high-capacity CAMs [4] : 1) the power hungry nature due to the high switching activity of the SL's and the ML's

and 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 ML's may cause a serious IR drop on the power grid, thus affecting the operational reliability of the chip [1] [3]. As a result, numerous efforts have been put forth to reduce both the peak and the total dynamic power consumption of the CAMs.

III. PROPOSED ML SENSE AMPLIFIER DESIGN

1. 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-computation schemes but in fact has a different operating principle. We first briefly discuss the Pre-computation schemes 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. 3(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. 3(a). These extra information are compared first and only those that have the same number of "1"s (e.g., the second and the fourth) 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 previously 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 later 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 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 I_on/I_off of the CAM by two times.

2) Parity Bit Based CAM: The parity bit based CAM design is shown in Fig. 3(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 these 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 has 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 I_on/I_off ratio of the design. We are going to propose a new sense amplifier that reduces the power consumption of the CAM. Fig.4 shows 1-mismatch ML transient waveforms of the original and the proposed architecture during the search operation.

2. Gated-Power ML Sense Amplifier Design

The proposed CAM architecture is depicted in Fig.5.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. The "COMPARISON" unit, i.e., transistors M1-M4, and the "SRAM" unit [1], i.e., the cross-coupled inverters, is powered by two separate metal rails, namely VDDML and the VDD, respectively. The VDDML is independently controlled by a power transistor (Px) a feedback loop that can auto turn-off the ML current to save power. The use of having two separate power rails of (VDD and VDDML) is to completely isolate the SRAM cell from any possibility of power disturbances during COMPARE cycle.

As shown in Fig. 5, the gated-power transistor Px, is controlled by a feedback loop, denote 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 match line is initialized by a global control signal EN. At this time, signal EN is set to low and the power transistor Px is turned OFF. Thus this will make the signal ML and C1 initialized to ground and VDD, respectively. Then, the signal EN turns HIGH and initiates the COMPARE phase.

If one or more mismatches happen, the ML will be charged up. All the cells of a row 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., Vth8), voltage at node C1 will be pulled down. After a very minor delay, the NAND2 gate will be toggled and so the power transistor Px is turned off again. As a result, the ML is not fully charged to VDD, but limited to voltage slightly above the threshold voltage of M8.

Fig. 6 shows the simulation results of the proposed power controller of CAM structure for 3 cases (matched case, mismatched case, and partial matched case).During matched case ML case has high output and VC1 charged to high voltage and during mismatched case ML and VC1 discharged to a low voltage. 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, the ML and node C1 change faster. Less number of mismatches will slow down the transition of node C1 and results in a longer delay to turn off transistor Px. With the introduction of the power transistor Px, the driving strength of the 1-mismatch case is about weaker than that of the conventional design. Thus it offers both low-power and high-speed operation.

IV. PERFORMANCE COMPARISONS

In this section, performance of the proposed design is evaluated and compared with the conventional circuit and those in, as references. In, the power consumption is limited by the amount of charge injected to the ML at the beginning of the search [5]. In, a similar concept is utilized with a positive feedback loop to boost the sensing speed. Both designs are very power efficient. The proposed design consumes slightly lower power when compared with the proposed design has comparatively low sensing delay hence increases the searching speed of CAM which is more favourable for VLSI design.

1. Power Consumption

Fig.7 illustrates the average power consumption of the proposed design and compared the performances with other three existing designs. The power-gated transistor is turned off after the output is obtained at the sense amplifier. So the proposed technique has lower average power consumption. Also because of reduced voltage swing on design. Because the EN signal turns off transistor Px of each row and column hence the SL buses do not need to be pre-charged, which in turn saves power on the SL buses.

2. Sensing Delay

The sensing delay comparison is shown in Fig. 8 where the proposed design improves the performance when compared with the conventional design. This figure also shows that sensing delay increases dramatically when supply voltage enters the near-sub threshold region. Sensing delay is defined as the sensing delay of the 1-mismatch.
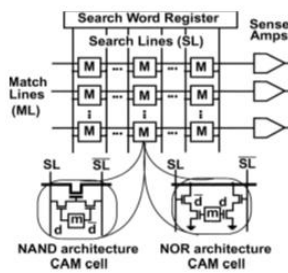
V. FIGURES

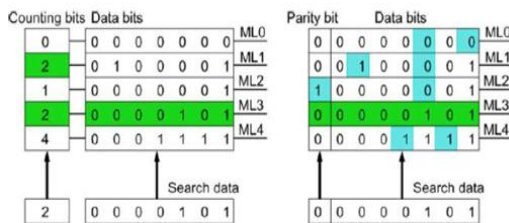Fig.1: Simplified CAM architecture.

Fig.3: Conceptual view of (a) conventional pre-computation CAM and (b) proposed parity-bit based CAM.
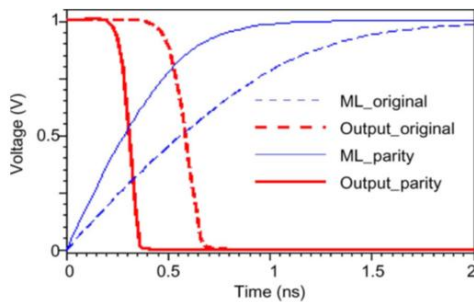

Fig.4: 1-mismatch ML waveforms of the original and the proposed architecture with parity bit during the search operation.
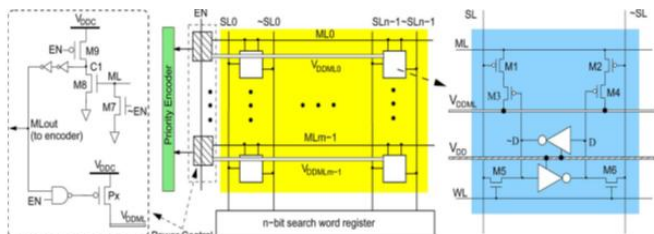

Fig.5: Proposed Content Addressable Memory (CAM) architecture. (b) Each CAM cell is connected to two power rails, for the compare transistors, for the SRAM transistors. The rail of a row is connected to the power network through a PMOS device, which is used to limit the transient current.
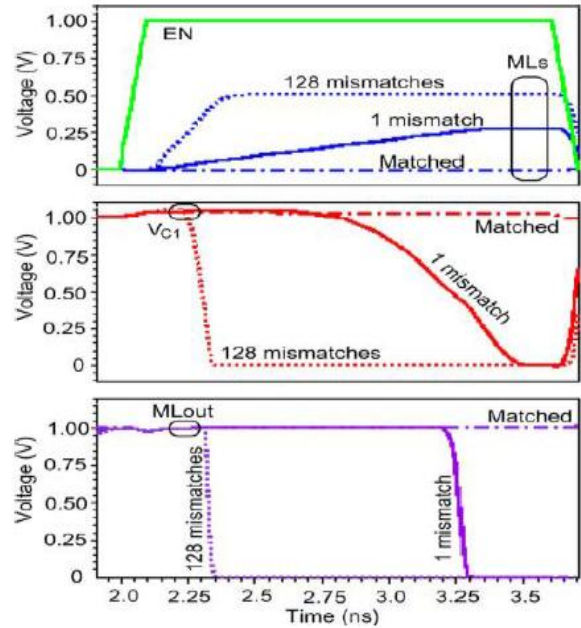

FIG.6: WAVEFORMS OF SOME IMPORTANT NODES DURING EVALUATION OF THE PROPOSED DESIGN.(A) OUTPUT WAVE FORM FOR MATCHED CASE.(B) OUTPUT WAVE FORM FOR PARTIAL MATCHED CASE.(C) OUTPUT WAVE FORM FOR MIS-MATCHED CASE.
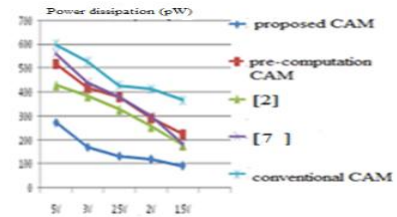

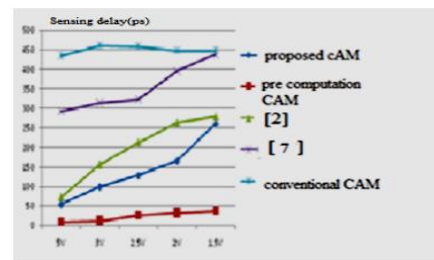Fig.7: Total average power consumption of the four designs.

Fig. 8 sensing delay of the four designs.

VI. CONCLUSION

An effective method for high speed and low power CAM is proposed in this paper. A power gated ML sensing technique offers several advantages like average power Consumption, boosted search speed when compare with conventional design. In this paper, we introduce a parity bit that leads to delay reduction, area and power overhead. Furthermore, we propose an effective gated-power technique to reduce average

power consumption and enhance the robustness of the design process variations. A feedback loop is employed to auto-turn off the power supply to the comparison elements and hence reduces the average power consumption. This paper compares the power dissipation and sensing delay of proposed CAM with existing structures.

VII. ACKNOWLEDGMENT

VIII. REFERENCES

1. Belma Joseph and Dr.S.Jayanthy "Power gated match line sensing content addressable memory" International Journal of Embedded Systems, Robotics and Computer Engineering. Volume 1, Number 1 (2014), pp. 1-6.
2. N. Mohan and M. Sachdev, "Low-leakage storage cells for ternary content addressable memories," IEEE Trans. Very Large Scale Integr. (VLSI) Syst., vol. 17, no. 5, pp. 604–612, May 2009.
3. S. Baeg, "Low-power ternary content-addressable memory design using a segmented match line," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 55, no. 6, pp. 1485–1494, Jul. 2008.
4. Kostas Pagiamtzis, and Ali Sheikholeslami, "Content-Addressable Memory (CAM) Circuits and Architectures: A Tutorial and Survey", IEEE Journal of Solid-State Circuits, Vol. 41, No. 3, March 2006.
5. Arsovski and A. Sheikholeslami, "A mismatch-dependent power allocation technique for match-line sensing in content-addressable memories," *IEEE J. Solid-State Circuits*, vol. 38, no. 11, pp. 1958–1966, Nov. 2003.