

Low IR Drop and Low Power Parallel CAM Design Using Gated Power Transistor Technique

Anh Tuan Do, Shoushun Chen, Zhi-Hui Kong and Kiat Seng Yeo
School of EEE, Nanyang Technological University, Singapore

Abstract—In this paper we analyzed the IR drop problem in large scale content addressable memory (CAM) and proposed a simple yet efficient gated power transistor technique. Each row of CAM cells is powered by two metal rails, one for the memory element and another one for the comparison transistors and the match lines. The latter rail is powered by a row-based transistor, which presents a physical “gate” to limit the peak current during comparison. Smart control scheme is proposed to automatically turn the power transistor off using a feedback delay loop. Simulation reports 96% reduction in IR drop and 64% save in total energy consumption, for a conceptual 8K-word CAM macros based on Chartered 0.13 μ m CMOS technology.

I. INTRODUCTION

CAM is a type of widely used 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 matched word that is stored in its data-bank [1]. Since all available words in the CAM are compared at the same time, CAMs are faster than other hardware and software-based search systems. However, the nature of full array parallel processing leads to a significant reliability issue, namely IR drop on power rail or ground bounce in substrate, causing the sensitive analog portions to malfunction.

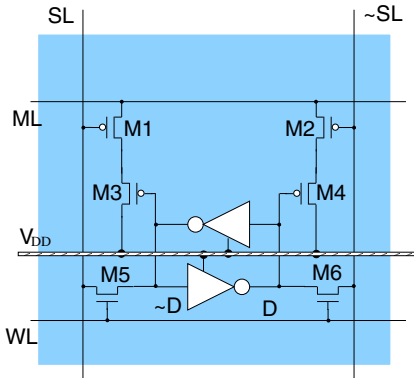


Fig. 1. A conventional 10-transistor NOR-type binary CAM cell.[1].

Fig. 1 shows a typical 10T NOR-type cell for high-speed parallel CAMs [1]. Four PMOS transistors $M1$ - $M4$ are used to compare the stored bit (D and $\sim D$) and the searched bit (SL and $\sim SL$). Each row of CAM cells share a match line (ML) bus, which is used to indicate whether the input search word matches the content of this row. During the pre-charge phase, the ML is initialized to ground voltage while all SL and $\sim SL$ are at V_{DD} . During the evaluation phase, complementary data

are broadcast to the SL and $\sim SL$. If mismatch occurs (for example, $D = '0'$ and $SL = '1'$), one of the two comparison branches ($M1 - M3$ or $M2 - M4$) will be enabled to charge up the ML to a higher voltage potential. Since all the CAM cells work in parallel and each mismatch will draw one charge up current from the power rail, the overall transient current is a linear function of the number of mismatches. Ironically, this incremental does not speed up the cycle time of the CAM as it is constrained by the worst case, i.e. one mismatch on a row. Instead, the large transient current leads to significant IR drop on the power rail in large scale CAMs. Fig. 2 illustrates the IR drop phenomenon along the main diagonal of the floor plan of a conventional 8K $\times$ 128 CAM, using a pure resistive power grid modeling.

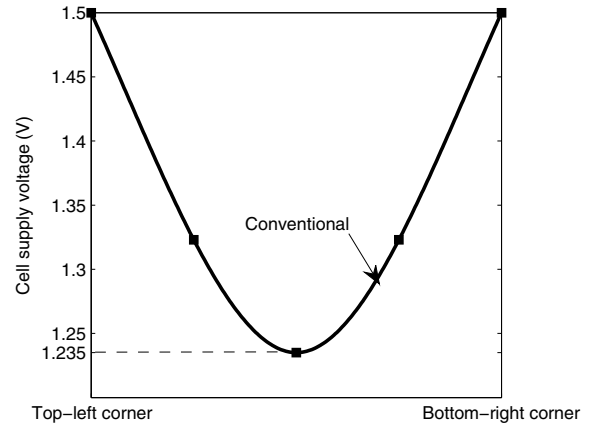


Fig. 2. Simulated IR drop phenomenon along the main diagonal a conventional 8K $\times$ 128 CAM. This corresponds to the worst case scenario, i.e. when mismatch happens to each and every word and bit. The peak current of the whole array can go up to 26 Amps.

Complete removal of the IR drop problem, however, involves appropriate trade-off between a number of design parameters such as average power consumption, delay and silicon area overheads. In the literature, most of the works [2–4] are dedicated to reduce the average power consumption on the ML , which is the main source of dissipation within the CAM. [5, 6] introduces a current-race scheme that supplies a limited current to each ML hence effectively eliminates the IR drop. However, this scheme has a low noise margin, not desirable for low voltage, nano-scale CMOS technologies. [7] proposes a segmented ML in which only a subset of ML is

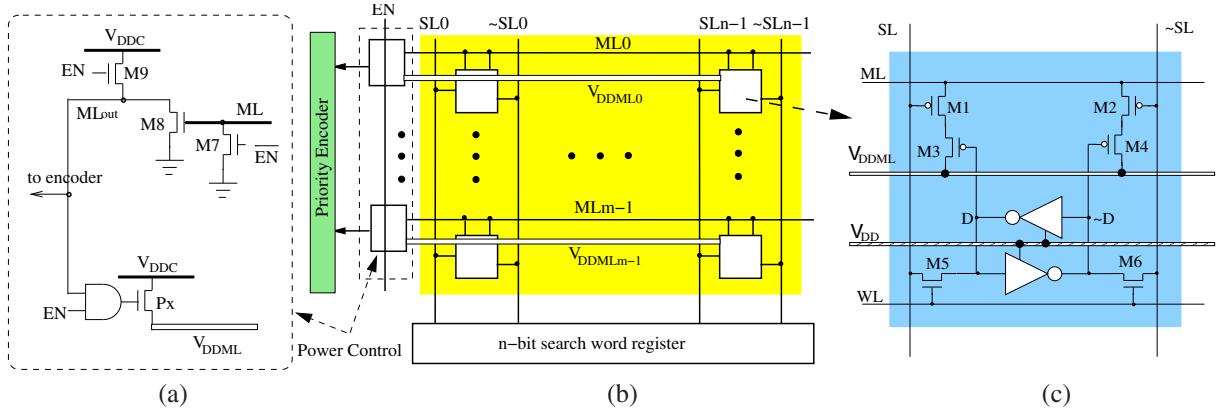


Fig. 3. (a) Row based power controller, which consists of a delay loop and a gated power transistor P_X . (b) Proposed CAM architecture. (c) The CAM cell, which is powered by two separate power rails, V_{DDML} for the compare transistors and ML , V_{DD} for the SRAM transistors.

pre-charged and hence reducing the peak current. However, it requires extra latency, which is not suitable in high speed CAMs.

In this paper, we propose an efficient gated power transistor technique. Each row of CAM cells is powered by two metal rails, one for the memory element and another one for the comparison transistors and the ML . The latter is connected to a row-based transistor, which presents a physical "gate" to limit the compare peak current. Average power consumption is lowered by reducing the voltage swing on ML buses. The main innovative contribution of this work is the gain of reliability against IR-Drop with minor area and timing overhead. The rest of paper is organized as follows: section II, we present the operation of the proposed CAM. Performance of the proposed circuit is evaluated in Section III. Section IV describes the chip's implementation. Section V concludes the paper.

II. PEAK CURRENT REDUCTION TECHNIQUE BASED ON GATED POWER TRANSISTOR

The proposed CAM architecture is depicted in Fig. 3. The CAM cells are organized into rows (word) and columns (bit). Each cell has the same comparison branches as the conventional NOR-type CAM (shown in Fig. 1). However, two separate metal rails are used to power the cell, namely V_{DDML} and V_{DD} . V_{DDML} is used to support the "comparison" unit, i.e., transistors $M1-M4$ and thus the ML . V_{DD} is another dedicated rail only to power the SRAM unit, i.e., the cross-coupled inverters. At the peripheral of each row, a transistor, P_X , is used as a gating device on the V_{DDML} rail. The gating transistor P_X , is further controlled by a feedback loop which will automatically turn off P_X once the voltage on the match line reaches a certain threshold.

At the beginning of each cycle, a global signal EN is used to initialize the ML as well as the gated power transistor P_X . At this time, EN is set to low. Transistor $M7$ is *ON* and thus the ML is discharged to the ground voltage. Transistor $M9$ is *ON* to charge up ML_{out} . The power transistor P_X is *OFF*. After that, signal EN turns high to turn on the power transistor P_X , thus initiates the comparison phase. When one or more

mismatches happen in the CAM cells, the ML will be charged up. When the voltage of ML reaches the threshold voltage of transistor $M8$ (i.e. V_{th8}), ML_{out} will be pulled down. After a certain but very minor delay, the power transistor P_X is turned off again. By engineering the size of the transistors and the delay interval, the ML will not be fully charged to V_{DD} , but limited to some voltage between V_{th8} and $(V_{DD} - V_{thPx})$.

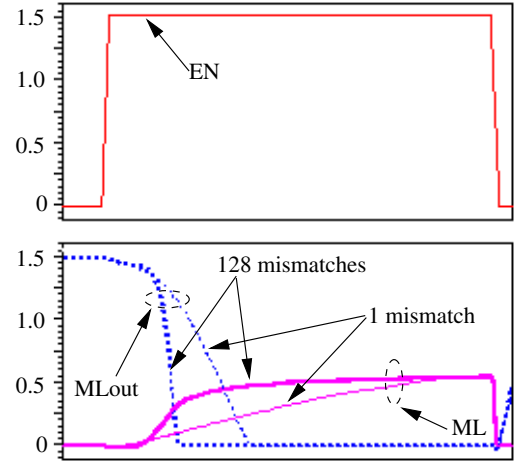


Fig. 4. Signal sequences in each comparison cycle. EN is first set to low to initialize the other signals. It then turns high to start comparison.

Fig. 4 shows the signal sequences by simulation. One can note that, the slopes of ML and ML_{out} depend on the number of mismatches. When more mismatches happen (128), the ML and the ML_{out} change faster. Less number of mismatches (1) slows down the transition of the ML_{out} and results in a longer delay to turn off the transistor P_X . Under both scenarios, ML is only charged to about 0.55V, far less than the 1.5V supply voltage.

III. PERFORMANCE COMPARISONS

A. Peak current and IR drop attenuation

The proposed power controller leads to a great reduction in the transient peak current. This can be explained by the

bottleneck effect of transistor P_X .

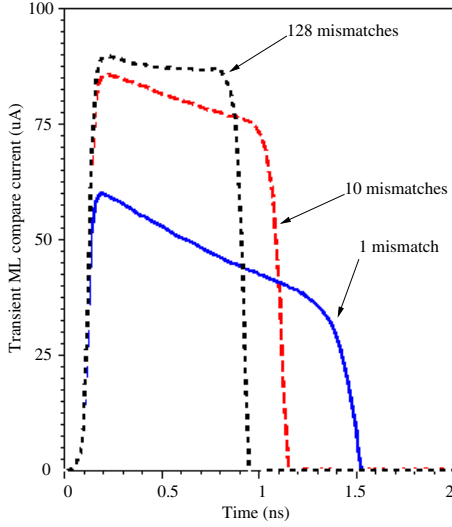


Fig. 5. Simulated transient current as function of the number of mismatches occurred on a row of 128 CAM cells.

Fig. 5 shows the simulation results of the transient current as function of the number of mismatches occurred on a row of 128 CAM cells during the comparison cycle. Though the overall transient ML charge up current increases with the number of mismatches, it will soon reach the limit due to the presence of the gating transistor P_X . One can note that, when one mismatch occurs the peak ML current is $60 \mu A$. When 128 mismatches happen, the current only increases to $85 \mu A$. This translates to a peak current of less than $0.7A$ of the whole array in the worst case scenario. Whereas in the conventional CAM, the peak current of the whole array can go up to $26A$.

In order to evaluate the efficiency of the the proposed technique against IR drop, we developed a power distribution model, which consists ideal voltage sources, power and ground wires modeled as a distributed RC network. Each segment of RC network powers one CAM cell. Each piece of power rail has an estimated resistance and capacitance based on technology parameters. To follow the common practice, a power source is placed at each corner of the RC network. This model is used to evaluate both the conventional architecture and the proposed one. When this network is applied to the new CAM structure, it will model the V_{DDC} , network, i.e., to power the comparison elements. While for the conventional architecture, both the comparison and data storage elements will be powered from this network. Fig. 6 records the supply voltage on the power grid of the two designs. The figure corresponds to the worse case IR drop scenario, i.e. all bits are mismatched. It is clearly shown that the supply voltage drops from the peripheral to the center. The proposed design permits to reduce the IR drop significantly. At the center of chip, a 96% reduction, from $0.265V$ ($=1.5-1.235$) to $0.01V$ ($=1.5-1.49$), is found in the simulation.

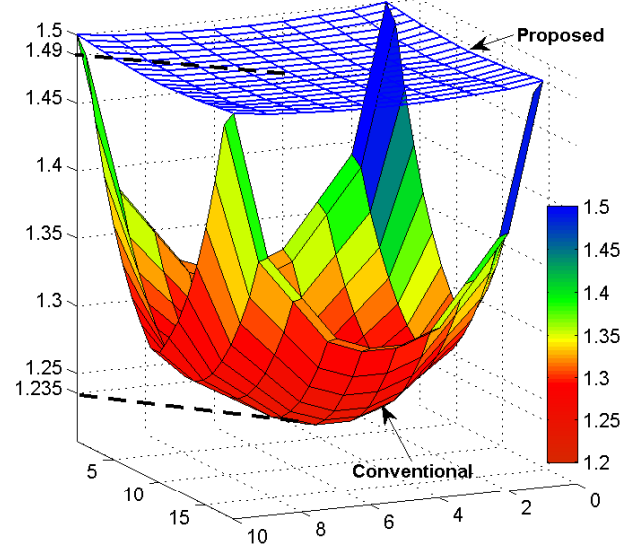


Fig. 6. IR drop on the V_{DD} (conventional) and the V_{DDC} (proposed) network in the CAMs. Simulation results are based on a Chartered $0.13\mu m/1.5V$ CMOS process. Each CAM macro has 8K words of 128 bits.

B. Average power consumption

Besides reduction in the transient current, the proposed design also achieves a lower average power consumption. Fig. 7 compares the total energy consumption between the proposed, two current-race [5, 6], and the conventional designs [1]. The total power dissipation is divided into two portions, on the ML and on the SL , respectively. One can note that the proposed design has the lowest ML power consumption. This is because the gated power transistor P_X will be turned off automatically as soon as the ML exceeds a threshold voltage. ML thus has a reduced voltage swing, slightly higher than the threshold voltage of NMOS. In the proposed design, since the global signal EN can turn off the power to the ML buses before comparison starts, the SL buses don't need to be pre-charged to V_{DD} . This leads to a statistical 50% reduction in power consumption compared to the conventional design. Similar scheme is also found in current-race designs [5, 6].

C. Noise margin

In the current-race designs, the critical ML potential race is between the matched and the 1-mismatch ML s. As a result, its noise margin is equal to the voltage difference between these two ML s. Also because of the process variations, sometimes the 1-mismatch ML even rises faster than the matched ML [5]. In our design, each ML is auto-turned off and the turn-off voltage of the ML only depends on the threshold voltage of the NMOS $M8$ in Fig. 3. Our design's noise margin is equal to the V_{th} of $M8$ which is $0.35V$, as the match ML stays at the ground level. Therefore, our design has a superior noise margin and has an advantage at low-supply voltage nano-scale CMOS processes.

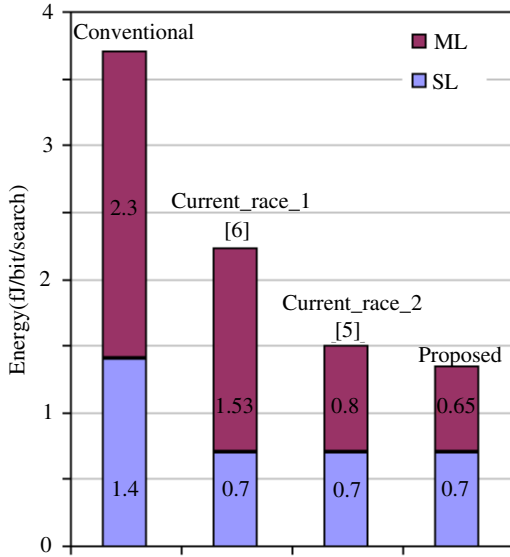


Fig. 7. Comparison of average power consumption of the proposed design with other three design reported in the literature. The total power is divided into two portions, that dissipated on the match lines (*ML*) and on the search lines (*SL*), respectively.

IV. VLSI IMPLEMENTATION

Two $8K \times 128$ CAM macros of the conventional and the proposed designs are implemented using a $0.13\mu m$ multi-threshold CMOS process. The choice of the size of the power transistor P_X is made so that the compare speed of the *ML* sensing is not significantly affected but at the same time it effectively limits the peak current of the row. This additional power transistors adds only a marginal area overhead when compared to the conventional design (less than 1%). Furthermore, the proposed design only incurs 10% additional sensing delay. Besides, since the *ML* voltage swing, and hence the average total power, depends on the threshold voltage of transistor *M8*, we use low-threshold *M8* to reduce the average power consumption.

At the system level, the CAM is divided into sixteen banks, each having $512\text{-words} \times 128\text{-bits}$. Each bank is placed between its corresponding row decoder and priority encoder. To achieve a compact layout, 1K-decoders and 1K-to-1-encoders are used in stead of 512-decoders and 512-to-1-encoders. Each encoder and decoder are shared by two adjacent CAM banks as shown in Fig. 8.

V. CONCLUSION

We report a gated power transistor technique to reduce the IR drop problem in large-capacity parallel CAM designs. On top of that, save in dynamic power consumption is achieved due to two reasons: the voltage swing on the *ML* is limited as the power transistor can be self-turned off; and the *SLs* no longer require to be pre-charge to V_{DD} . Our simulation have showed that 96% reduction in IR drop and 64% save in total energy consumption, compared to conventional architecture.

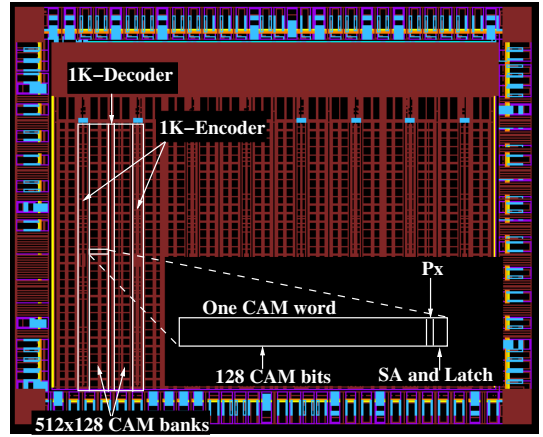


Fig. 8. Chip layout of the proposed $8K \times 128$ CAM macro

The proposed design can be used in conjunction with other low-power CAM cell architectures to further reduce the total energy consumption of the chip.

ACKNOWLEDGMENT

This work was supported by Nanyang Assistant Professorship (2009-2012).

REFERENCES

- [1] K. Pagiamtzis and A. Sheikholeslami, "Content-addressable memory (CAM) circuits and architectures: a tutorial and survey," *IEEE Journal of Solid-State Circuits*, vol. 41, pp. 712-727, 2006.
- [2] N. Mohan and M. Sachdev, "Low-Leakage Storage Cells for Ternary Content Addressable Memories," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 17, pp. 604-612, 2009.
- [3] A. R. Patwary, et al., "Low-power Ternary Content Addressable Memory (TCAM) array for network applications," in *International Conference on Communications, Circuits and Systems, 2009 (ICCCAS 2009)*, 2009, pp. 322-325.
- [4] M. Sultan, et al., "A low-power ternary content addressable memory (TCAM) with segmented and non-segmented matchlines," *IEEE Region 10 Conference in TENCON 2008*, pp. 1-5.
- [5] I. Arsovski and A. Sheikholeslami, "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.
- [6] I. Arsovski, et al., "A ternary content-addressable memory (TCAM) based on 4T static storage and including a current-race sensing scheme," *IEEE Journal of Solid-State Circuits*, vol. 38, pp. 155-158, 2003.
- [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.