A Low-Power CAM with Efficient Power and Delay Trade-off

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

***Abstract*—In a Content Addressable Memory (CAM) architecture, both the match-line (ML) sensing circuit and the priority encoder (PE) contribute significantly large delays during a compare cycle. Meanwhile the priority encoder consumes significantly less energy when compared to the sensing circuits, i.e. $\sim$1% of the overall energy consumption. Based on this observation, we propose the use of dual-supply voltages to trade-off the power and delay budget between the comparison and priority encoding circuits. In this work, the memory array and priority encoder is powered by a low and a high supply voltage, respectively. On top of this, a self-power-off ML sense amplifier is employed to reduce the voltage swing on the ML buses. Simulation results show a 76% dynamic power reduction as compared to the conventional design without sacrificing the overall speed.**

I. INTRODUCTION

Content addressable memory (CAM) is widely used in applications where data is accessed by their contents rather than physical locations [1]. Since all words in the CAM are compared concurrently, CAM can return its output within one clock cycle and hence is faster than other hardware and software-based search systems. However, this advantage comes at a cost of high power consumption due to parallel charging and discharging of large capacitances associated with the ML and the search-line (SL) buses [1]. In the literature, several works have been proposed to reduce the power consumption on the ML buses. [2] introduced a segmented architecture where the ML in each CAM word is partitioned into segments and is selectively pre-charged. The partially charged MLs are evaluated to determine the final comparison result by sharing the charges deposited in various parts of the partitioned segments. C. A. Zukowski *et. al.* and K. Pagiamtzis *et. al.* presented a similar selective pre-charge and pipelined architecture [3] [4]. However, these techniques introduced speed penalty as well as increased circuit complexity . Another popular technique is to reduce the voltage swing on the MLs. [6] and [5] utilized the current race concept to differentiate between the matched and unmatched words. A self-power-off scheme will terminate the current supply when the reference ML (dummy) reaches the threshold voltage of the sense amplifier.

In this paper, we further explore the use of the self-power-off strategy and propose a dual-supply voltages scheme to efficiently trade-off the power and delay budget between the two major components of a CAM device, namely, the comparison circuit (which is the memory array with associated MLs) and priority encoding circuits. The comparison operation, i.e, the ML, is in fact the main contributor of the overall CAM power consumption, but it only contributes to part of the overall clock cycle time. On the other hand, the priority encoder consumes very little power while incurring almost the same delay when compared to its comparator counterpart. Based on this observation, we propose to power the memory array and priority encoder using a low and a high supply voltage, respectively. This leads to an extra delay of about 500 ps on ML buses but this is easily compensated by the now faster priority encoder due to the use of higher supply voltage. Therefore, a 25%-30% saving in dynamic power consumption was achieved without any compromise in the overall clock speed. The rest of paper is organized as follows: section II introduces the system architecture and the operation principle of the self-power-off scheme. Section III discusses the proposed timing and power budget scheme based on simulation result. Section IV presents the simulation results and additional performance analysis. Section V concludes the paper.

II. SYSTEM OVERVIEW

Fig. 1 illustrates the proposed CAM architecture. It consists of row-based ML sense amplifiers, array of CAM cells and a priority encoder (PE). We use NOR-type CAM cell since it is preferred for high-speed applications. Each cell has the same number of transistors as the conventional NOR-type CAM and use the same 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 V_{DDL}. A row of CAM cells will share the V_{DDML} rail which is in turn gated by a power device Px. The purpose of having transistor Px is to self-turn-off the ML supply current after the comparison result is achieved. This function is implemented within the row-based sense amplifier, which consists of four transistors, a delay chain of two inverters, one NAND gate and a D-latch.

The operation principle is as follows. At the beginning of each cycle, the ML is first initialized to ground. At this time, EN is low, $M7$ and $M9$ are ON and the power transistor Px is OFF. After that, signal EN turns HIGH and initiates the $COMPARE$ phase. When one or more mismatches happen in the CAM cells, the ML will be charged up. However, the charge up current will be cut off when ML reaches the threshold voltage of $M8$ and in turn triggers signal $C1$ and turns off transistor Px. Therefore ML is not fully charged to V_{DDL}, but limited to some voltage slightly above the threshold voltage of $M8$. A clock signal ($LaEN$), produced

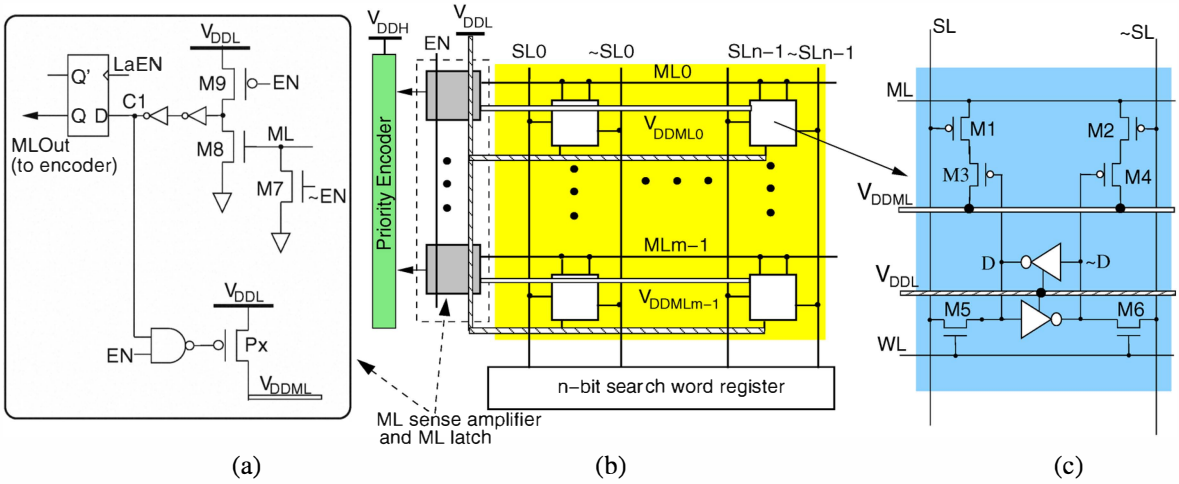

Fig. 1. (a) Row-based ML sense amplifier, which consists of a delay loop and a gated power transistor P_X. (b) Proposed CAM architecture in which the priority encoder is powered by a high supply voltage V_{DDH} while the rest are powered by a low supply voltage V_{DDL}. (c) The CAM cell, which is powered by two separate power rails, V_{DDML} for the compare transistors and V_{DDL} for the SRAM transistors.

by a dummy row, is used to latch the ML comparison result and pass it to the priority encoder.

Fig. 2 further illustrates the timing waveforms of a few important signals. The slope of the MLs depends on the number of mismatches. When more mismatches happen (e.g. 64 or 128 in the simulation), the ML changes faster. Less number of mismatches (e.g. 1 in the simulation) will slow down the transition of the ML and result in a longer delay. Once node $C1$ is triggered, it will automatically turn off the gated power transistor Px and hence preventing its corresponding ML from rising further, which in turn saves a significant amount of power. It is important to note that the delay of the $COMPARE$ phase is constrained by the scenario of 1-bit mismatch. In fact, the latch enable signal ($LaEN$) is generated from a dummy row that contains a 1-bit mismatch. Therefore upon arrival of $LaEN$, all the ML buses have completed the comparison process and are ready for sampling.

III. Power and Speed Budgeting

The CAM array and the priority encoder are respectively powered by V_{DDL} and V_{DDH}. As mentioned earlier, the purpose of introducing dual-supply voltages approach here is to trade-off the power and delay partitioning between the CAM array and the priority encoder for a reduced overall power consumption while maintaining the same speed.

In order to evaluate the contribution factor of each building block to the overall power and delay, we simulate the proposed system using a multi-threshold 65 nm CMOS process. When a single supply voltage (i.e, V_{DDL}=V_{DDH}=1 V) is used, the ML delay is about 450 ps while that of the priority encoder is 890 ps, which is more than 60% of the overall delay. Meanwhile, 99% of the power is dissipated on the array. Based on this observation, we use a higher supply voltage to boost the speed of the priority encoder while a lower voltage to reduce the power consumption of the array.

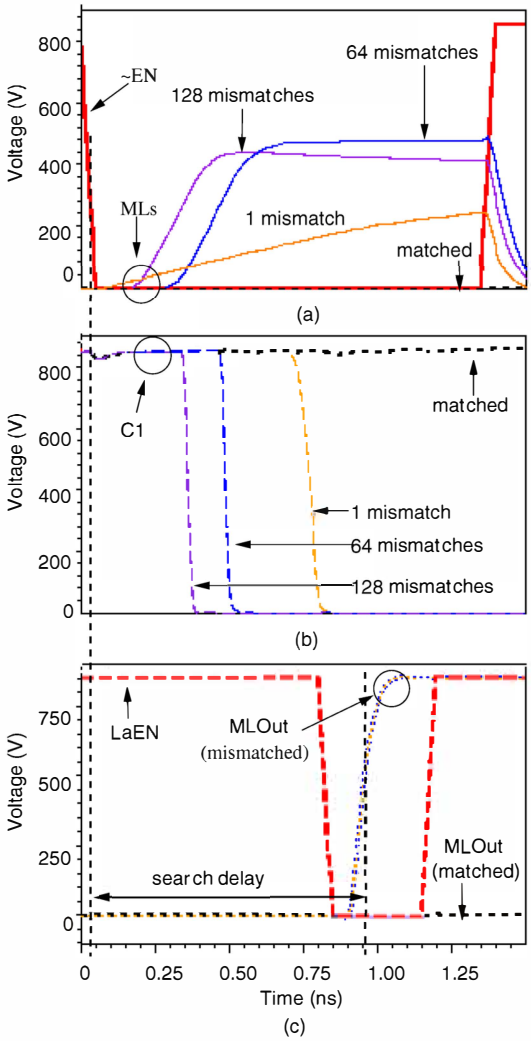

Fig. 2. Compare operation waveforms of the proposed design at different number of mismatches. (a) Compare enable signal (EN) and MLs. (b) Node $C1$ (c) ($LaEN$) and ML latch outputs. ($LaEN$)is produced by a dummy row.

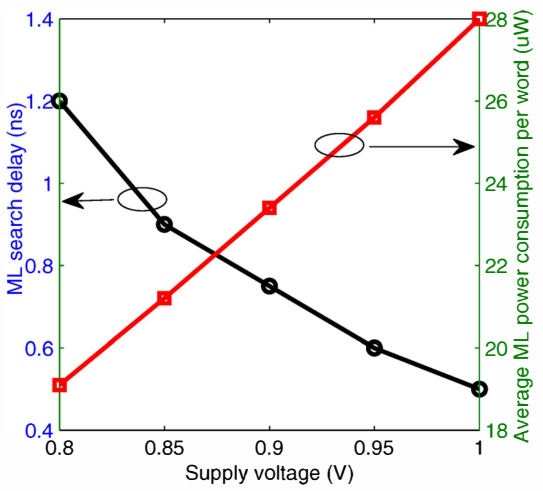

Fig. 3. ML sensing delay and power consumption per row of the proposed design as a function of different supply voltages. By lowering the V_{DDL} from 1V to 0.8 V, its power consumption is reduced from 28 to 19 μW. However, this is achieved at the expense of increased delay from 450 ps to 1.2 ns.

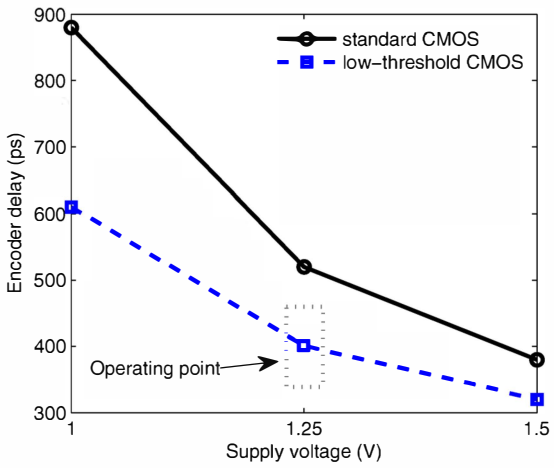

Fig. 4. Delay of the two 1k priority encoders implemented in standard- and low-threshold transistors, respectively. Their delay is lowed with increased supply voltage and a voltage of 1.25V is selected as a operating point.

Fig. 3 illustrates the ML sensing delay and power consumption per row at different supply voltages. It clearly shows the timing-power relationship as explained earlier. By lowering the V_{DDL} from 1V to 0.8 V, its power consumption is reduced from 28 to 19 μW. However, this is achieved at the expense of increased delay from 450 ps to 1.2 ns. Similarly, Fig. 4 reports the delay of 1K-input priority encoder at different supply voltages. To attain an even better delay margin, we propose to build the priority encoder using low threshold transistors. As shown in the simulation, the low-threshold priority encoder is about 25% faster than the standard-threshold priority encoder. At 1 V supply, the standard CMOS priority encoder has a delay of 890 ps while at 1.5V, the low-threshold CMOS priority encoder has a delay of only 320 ps.

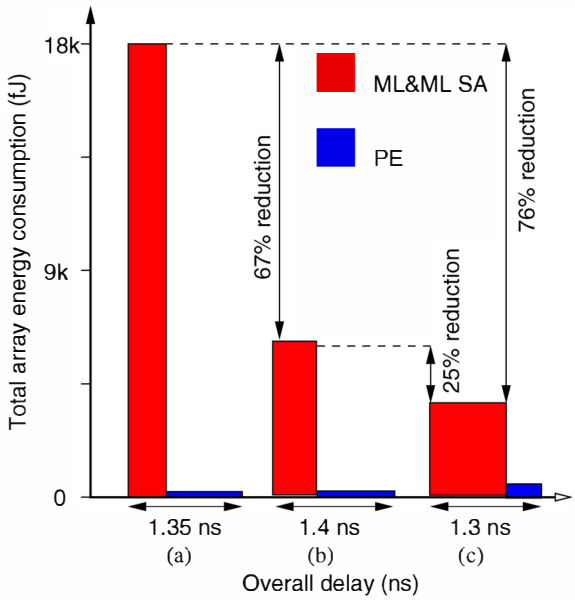

Fig. 5. Overall delay and dynamic energy consumption of a 1K-word CAM. (a) Conventional design. (b) Auto turn-off ML sense amplifier without using dual supply voltage. (c) Auto-turn off ML sense amplifier coupled with dual supply voltage. The heights of the bars represent the energy consumption while their widths represent the corresponding delay. The three designs feature almost the same speed. The proposed work achieves an overall reduction of 76% in power consumption.

Careful engineering of the two supply voltage, V_{DDH} and V_{DDL} will therefore achieve an optimized timing budget at a lower power consumption. Fig. 5 intuitively summarizes the overall performance of the proposed design in comparison to the conventional implementation. The heights of the bars represent the energy consumption while their widths represent the corresponding delay. As shown in Fig. 5(a), most of the energy dissipated by the conventional design is from the array while most of its delay caused by the priority encoder. With the use of the self-power-off scheme, the array power consumption is dramatically reduced by 67%. This is because the voltage swing on the ML buses is limited. The dual-supply voltages scheme further achieves another 25% reduction in the array power consumption at the cost of a slight increase in the priority encoder. The delay in the array is increased but eventually get compensated by the priority encoder, to maintain the same speed performance.

IV. Implementation Results and Discussions

A 1k-input priority encoder in conjunction with a 1k-word × 128-bit CAM array was implemented using a multi-threshold 65 nm CMOS process. The main design trade-off is the choice of the size of the power transistor. On one hand, the larger the transistor, the larger charge up current it can provide and hence the faster comparison speed. On the other hand, the larger current will leads to increased ML voltage swing and hence higher power consumption

In addition to improved dynamic power efficiency, the proposed technique brings forth enhancement in yet another

TABLE II
POWER EFFICIENCY COMPARISON OF THE PROPOSED CIRCUIT WITH RECENTLY PUBLISHED DESIGNS

	Conventional	JSSC'03 [5]	JSSC'03 [6]	TCAS-I'09 [7]	This work
Process and supply	65nm/1V	180nm/1.2V	130nm/1.2V	180nm/1.8V	65nm/0.85V/1.25V
Word length	128	144	144	144	128
Delay Time (ML+PE) (ns)	1.3	3	2.5	1.7	1.35
Normalized Energy/bit/search (fJ)	1.4	0.95	0.5	0.4	0.33

TABLE I
SIMULATION RESULTS OF LEAKAGE CURRENTS OF DIFFERENT BUILDING BLOCK AS COMPARED TO CONVENTIONAL DESIGN. THE CAM ARRAY FEATURES 1K-WORD × 128-BIT.

	Proposed	Conventional
CAM array (μA)	249	432.5
Sense amplifier(μA)	98	135
Encoder (μA)	35	14
Total (μA)	382 (34% reduction)	581

important design parameter, namely the static power consumption due to leakage. As the technology scales down, leakage power consumption plays a significant role and hence it is important to reduce the leakage current on the ML buses. In the proposed design, the gated-power transistor Px provides a bottleneck effect to reduce the leakage. This together with the lowered supply voltage leads to a 34% reduction in the leakage power consumption. Table I reports the leakage current of different building blocks.

Since the supply voltage to CAM array is lowered, it is important to ensure that the SRAM cells are sufficiently robust against noise disturbances. Fig. 6(a) shows the SRAM cell SNM simulation results when powered at different voltages. It is obvious that the cell can maintain a positive SNM even at supply voltage as low as 0.4 V. Furthermore, our Monte-Carlo simulation results show that at a supply voltage of 0.85 V and with process variations taken into account, the proposed design is able to demonstrate a competitive SNM of 138 mV.

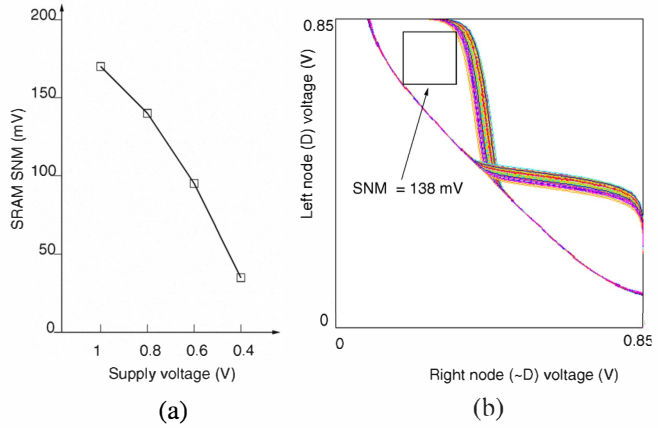

Fig. 6. SNM of the SRAM cell. (a) SNM at different supply voltages. The cell maintains a positive SNM even at supply voltage as low as 0.4 V. (b) Butterfly curve of the SRAM cell using Monte-Carlo simulations to verify that it has a good SNM of 138 mV at 0.85V supply voltage.

V. CONCLUSION

In this paper, we proposed a dual-supply voltages scheme coupled with an auto-power-off technique. By engineering the two supply voltages, one for the array and another for the priority encoder, a balanced timing budget was achieved at a lower power consumption. Simulation results show that the dynamic energy consumption is reduced by 76%. Additional advantage such as reduction in leakage power consumption is achieved. Table II summarizes the power efficiency comparison of the proposed circuit with recently published designs.

ACKNOWLEDGMENT

This work was supported by Nanyang Assistant Professorship (M58040012,2009-2012).

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

[3] C. A. Zukowski and W. Shao-Yi, "Use of selective precharge for low-power content-addressable memories," IEEE International Symposium on Circuits and Systems (ISCAS), vol.3, pp. 1788-1791, 1997.

[4] 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.

[5] 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.

[6] 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.

[7] 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.