0.339fJ/bit/search Energy-Efficient TCAM Macro Design in 40nm LP CMOS

Po-Tsang Huang[1], Shu-Lin Lai[2], Ching-Te Chuang[2], Wei Hwang[2], Jason Huang[3], Angelo Hu[3], Paul Kan[3], Michael Jia[3], Kimi Lv[3] and Bright Zhang[3]

[1]Dept. of Electrical and Computer Engineering, [2]Dept. of Electronics Engineering, National Chiao Tung University, Taiwan
[3]Faraday Technology Corporation, Hsinchu, Taiwan

***Abstract*—In this paper, a 256x40 energy-efficient ternary content addressable memory (TCAM) macro is designed and implemented in 40nm low power (LP) CMOS. Due to the thicker gate oxide in LP process, a 16T TCAM cell with p-type comparison circuits is proposed to increase the I_{on}/I_{off} difference of the dynamic circuitry. To further improve energy efficiency, don't-care-based ripple search-lines/bit-lines are used to reduce both the switching activities and wire capacitance. Moreover, column-based data-aware power control is employed for leakage power reduction and write-ability improvements. The experimental results show a leakage power reduction of 28.9%, a search-line power reduction of 31.74% and an energy efficiency metric of the TCAM macro of 0.339 fJ/bit/search.**

Keywords—Embedded memory, energy-efficient, TCAM

I. INTRODUCTION

Emerging internet of things (IoT) brings enormous opportunity for new types of applications. Ubiquitous devices and facilities construct IoT systems, and the demand of ternary content addressable memory (TCAM) macro is rapidly increasing for data routing in energy-efficient network chips. TCAM examines address-lookup functions for selecting the corresponding routing paths for packets. Furthermore, TCAM should be integrated with network processors in single chip duo to energy-limitation and small form factor in IoT gateways. Accordingly, as TCAM applications grow, energy consumption becomes one of the critical challenges. Moreover, leakage currents increasingly dominate the overall power consumption in nano-scale technologies.

Previously, we proposed a 256x144 TCAM macro with several low-power design techniques in 65nm standard performance (SP) CMOS process for network routers [1]. Accordingly, hierarchal search-lines, XOR-based conditional keepers, butterfly match-lines and super cut-off power-gating techniques are utilized to realize low-power TCAM design. However, it is difficult to directly migrate the 65nm SP design to 40nm low power (LP) CMOS for the integration with network processors. The critical challenges are the thicker gate oxide and weak mobility of NMOS in UMC 40nm LP process. In view of this, an energy-efficient 256x40 TCAM macro is proposed in 40LP CMOS using 16T TCAM cell with p-type comparison circuits, ripple search-lines/bit-lines (SLs/BLs) and column-based data-aware power control (DAPC) for achieving both dynamic/leakage power reduction and write-ability improvements.

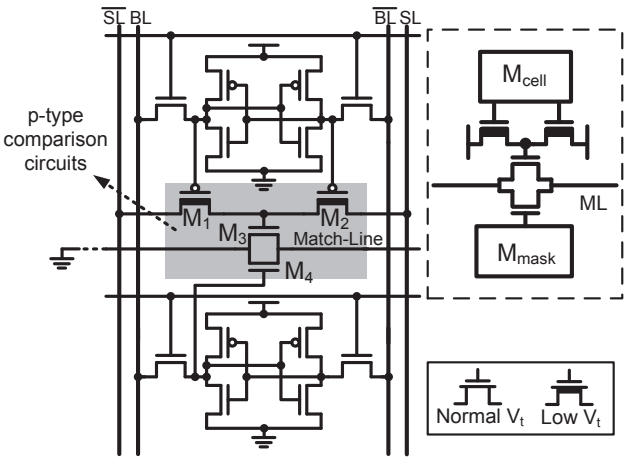


Fig. 1. 16T TCAM cell with p-type comparison circuits.

II. ENERGY-EFFICIENT TCAM MACRO DESIGN

The proposed 256x40 TCAM is implemented in 40LP CMOS process via pseudo-footless clock-and-data pre-charge (PF-CDPD) dynamic circuits [2], XOR-based conditional keeper and butterfly match-line scheme [1]. For further increasing the sizing boundary of the keeper, the butterfly match-line scheme is cascaded with AND gates instead of with XOR gates. Furthermore, three energy-efficient design techniques are employed, including 16T TCAM cell with p-type comparison circuits, ripple SLs/BLs and column-based DAPC. The details of these four design techniques are described below.

A. 16T AND-type TCAM Cell with P-type Comparison

In LP processes, the thicker gate oxide decreases both I_{on} and I_{off} to reduce the standby power for low-power applications. However, the I_{on} (V_{DD}-V_t) and I_{off} (0V) difference in traditional 16T AND-type TCAM cell is also reduced significantly that induces erroneous evaluations due to conflict between the keeper current and I_{on}. Hence, a TCAM Cell with P-type comparison circuits with dual-V_t is proposed as shown in Fig. 1. This TCAM cell provides full V_{DD} operation by the comparison PMOS M_1/M_2, and matching NMOS M_3. M_1 and M_2 are low-V_t PMOSs to reduce I_{off} of M3. Fig. 2 presents the drain current versus gate voltage of M3 in linear scale for 40SP and 40LP processes, respectively, and the gray regions are the upper bound and lower bound of keeper current (I_{Keeper}). For maintaining the noise margin, I_{on}

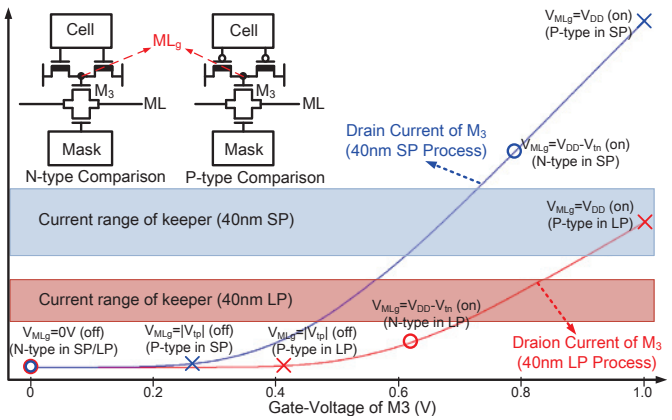


Fig. 2. Drain current versus gate voltage of M3 in linear scale for 40nm SP and LP processes.

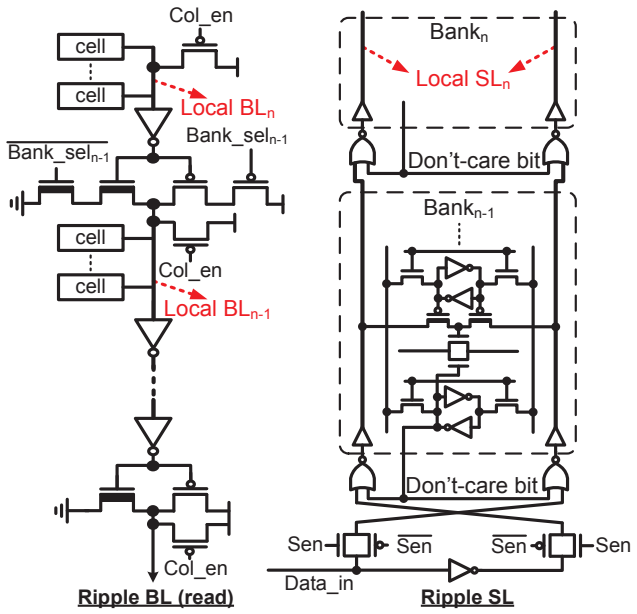


Fig. 3. Don't-care-based ripple SLs/BLs.

and I_{off} of M3 should be larger and smaller than the range of I_{Keeper}, respectively. In 40nm SP CMOS process, the I_{on}/I_{off} difference of N-type comparison circuits is larger than that of P-type comparison circuits, and both I_{Keeper} are between I_{on} and I_{off} of M3. However, the I_{on}/I_{off} difference of N-type comparison circuits is much smaller than that of P-type comparison circuits in 40nm LP CMOS process, and thus the TCAM cell with N-type comparison circuits cannot survive under PVT variations.

B. Don't-Care-Based Ripple Search-lines and Bit-lines

Hierarchical SLs/BLs have been commonly used to reduce the switch activities and loading capacitance of local SLs/BLs [1, 3]. However, the increasing power and delay overheads of long global SLs/BLs reduce the overall performance of hierarchical SLs/BLs due to increasing wire RC in deeply scaled technologies. Hence, segmented ripple SLs/BLs are proposed to decrease both the power consumption and search-time by reducing the wire capacitance, switching activities and metal tracks of vertical lines as shown in Fig. 3.

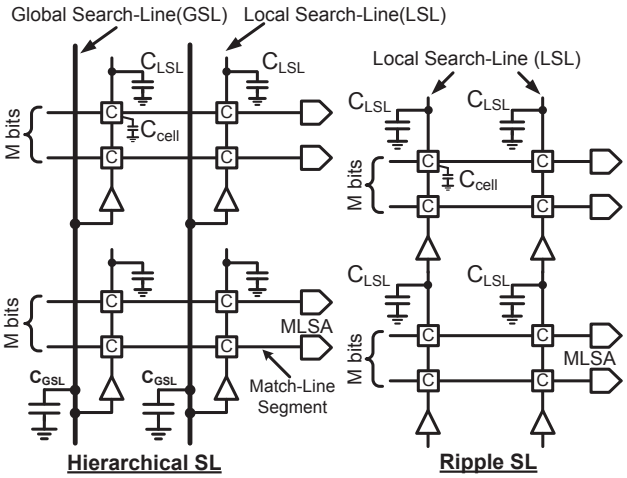


Fig. 4. Hierarchical SL versus ripple SL.

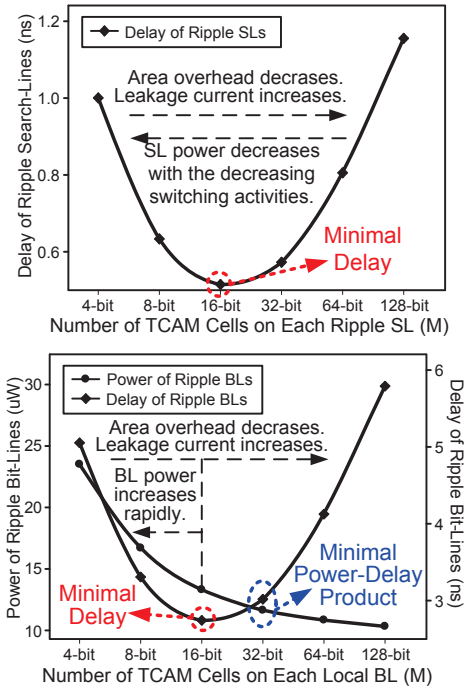
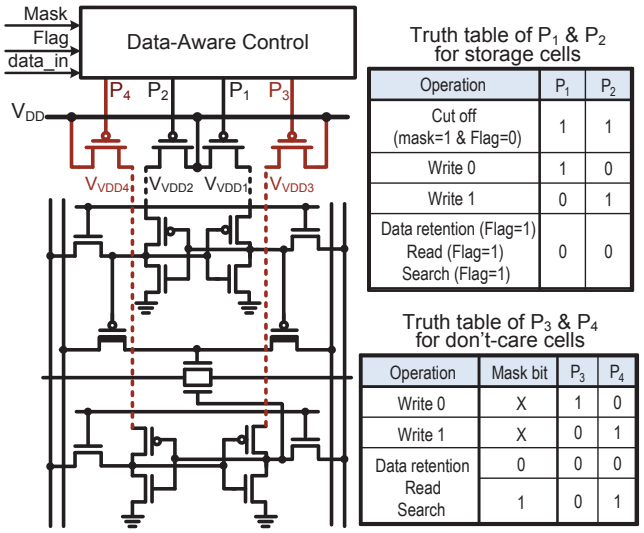


Fig. 5. Analysis of different segmentations of ripple SLs/BLs.

For the read operations, all local BLs in the selected column are precharged by Col_en. Accordingly, rail-to-rail large signal sensing on the selected local BL is utilized, and the readout data is propagated to the output port through ripple BLs. And thus, Bank_sel is used to activate the corresponding word-line in this bank and to block the signal from the previous local BL preventing read-disturb. Moreover, low-V_t NMOSs are utilized in ripple BLs for decreasing the propagation delay.

For the don't-care-based ripple SLs, each SL is divided into several local SLs (LSLs). A LSL is activated based on the mask data on the lowest word of this bank. Fig. 4 presents the concept of don't-care-based hierarchal SL [1] and ripple SL. If the don't-care state is true, the following local SL pairs are discharged to ground for reducing the switching activities of local SLs. Fig. 6 illustrates the delay and power analysis of different segmentations of ripple SLs/BLs, respectively. After



Truth table of P_1 & P_2 for storage cells

Operation	P_1	P_2
Cut off (mask=1 & Flag=0)	1	1
Write 0	1	0
Write 1	0	1
Data retention (Flag=1) Read (Flag=1) Search (Flag=1)	0	0

Truth table of P_3 & P_4 for don't-care cells

Operation	Mask bit	P_3	P_4
Write 0	X	1	0
Write 1	X	0	1
Data retention	0	0	0
Read Search	1	0	1

Fig. 6. Column-based data-aware power control.

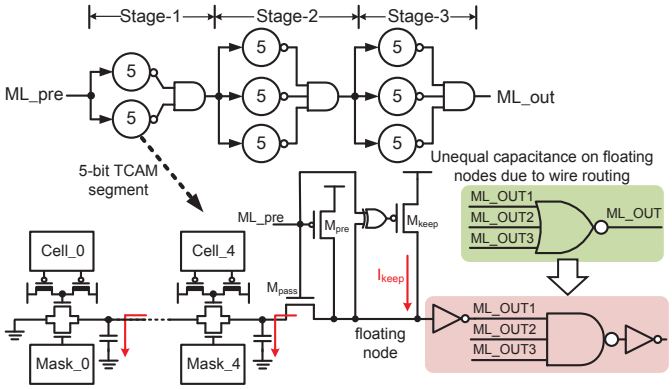


Fig. 7. Butterfly match-line with AND gates.

the delay and power tradeoff, 16-bit local BL/SL is realized in this TCAM macro, and each bank contains 16x40 TCAM cells.

C. *Column-based Data-aware Power Control (DAPC)*

Instead of row-based power-gating in [1], column-based DAPC is proposed to reduce both the leakage and write power and to enhance the write-ability, write margin, and time-to-write as shown in Fig. 7. In each TCAM bank, the storage and don't-care SRAMs are controlled by two power switch pairs based on the write data and don't-care patterns of the lowest word in each bank. During write operations, the data-in from BL turns off one of the half-cell supply to enhance write-ability and time-to-write. The write pulse width is generated by a write replica circuit to ensure the stability of half-selected cells on the selected column with PVT tracking for variation tolerance.

In other modes, the 4 power switch PMOSs are controlled by the don't-care patterns to reduce the leakage power as shown in Fig. 4. The flag indicates whether the storage data can be destroyed or not when the don't-care state is true. Hence, the flag is defined by users for different applications. If the flag=0 and mask =1, the storage cells are in the cut-off mode by turning off P1 and P2.

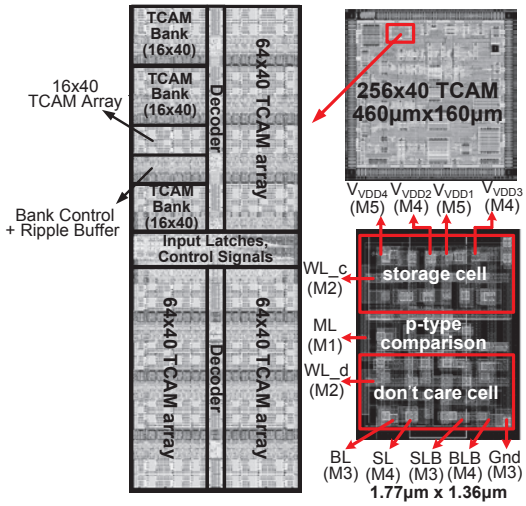


Fig. 8. Floorplan and test chip of TCAM macro and layout view of 1 TCAM cell.

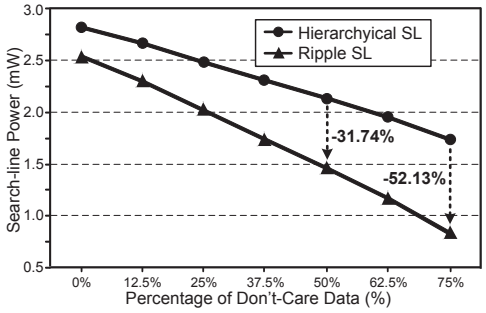


Fig. 9. Comparison of hierarchical SLs and ripple SLs.

D. *Butterfly Match-Line with AND Gates*

A 40-bit match-line is divided into 8 segments within 3 stages using butterfly match-line scheme as shown in Fig. 7. The butterfly match-line scheme in [1] uses NOR gates with XOR conditional keeper to construct their butterfly connection. However, the capacitances on the floating nodes of different segments are large and unequal due to the complex routing wires. Furthermore, the sizing range of keepers is quite small because the small I_{on}/I_{off} difference in LP CMOS process. In view of these, the butterfly match-line is implemented with AND gates by DeMorgan laws. Although six inverters are inserted in the critical path, the delay of match-lines is decreased since the capacitances on the floating nodes are reduced and equal.

III. Implementation & Experimental Results

The proposed 256x40 TCAM macro is implemented in UMC 40LP CMOS, and divided into 4 columns with 16 banks as shown in Fig. 8. In this TCAM macro, 16T TCAM cell with p-type comparison circuits, don't-care-based ripple search-lines/bit-lines and column-based data-aware power control are employed. Additionally, the NOR gates in the butterfly match-line scheme are replaced via inverters and AND gates to reduce the capacitance on floating nodes and to guarantee the same capacitance in each match-line segment. The layout view, test chip and metal tracks of a TCAM cell are also shown in Fig. 8. Word-lines, global match-lines and

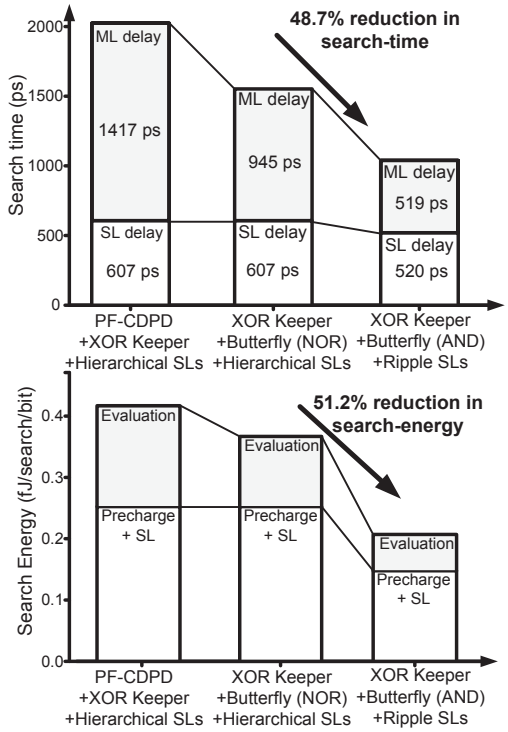


Fig. 10. Analysis on search time and search energy of 256x40 TCAM.

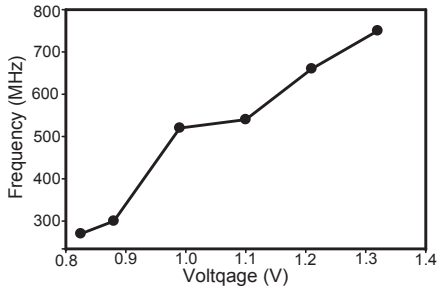


Fig. 11. Measured frequency versus voltage of the proposed TCAM macro.

local match-lines are routed in Metal-2 and Metal-1 in horizontal direction. SLs, BLs and power lines are routed in Metal-3 to Metal-6 in vertical direction. Fig. 9 compares SL power between hierarchal SLs and ripple SLs under different percentages of don't-care data by pre-layout simulation (including the line model). As the percentage of don't-care data increases, the power of hierarchical SLs and ripple SLs are reduced due to the decreasing switching activities. The ripple SLs realizes 31.7% power reduction compared to hierarchical SLs. Fig. 10 shows the analyses on search time and search energy of 256x40 TCAM by pre-layout simulation. The proposed ripple SLs and butterfly match-line with AND gate can achieve both delay and power reductions. Fig. 11 presents the measured frequency versus voltage of the proposed TCAM, and Fig. 12 shows the measured leakage power with/without DAPC under different percentages of don't-care patterns. 28.9% leakage reduction is realized by the proposed column-based DAPC with 50% don't care patterns. Table I summarizes the characteristics of the proposed TCAM and comparison with other recently proposed TCAMs. The proposed TCAM can be integrated in network chips in LP CMOS process and provide both search and read operations.

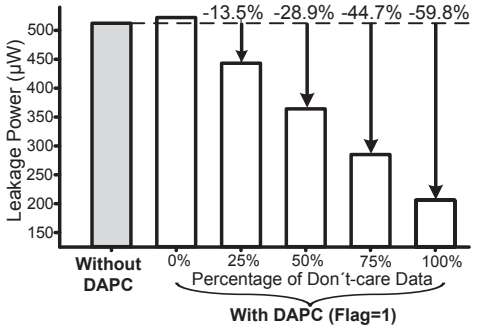


Fig. 12. Measured leakage power with/without DAPC under different percentages of don't-care patterns of 256x40 TCAM macro.

TABLE I. Feature summary and comparisons.

	Tree-style [2]	EPLC [4]	Butterfly [1]	This Work
Configuration	256x128	2048x460	256x144	256x40
Technology	0.18µm SP CMOS	32nm High-K Metal Gate SOI	65nm SP CMOS	40 nm LP CMOS
Supply voltage	1.8V	0.95V	1.0V	1.0V
Search time	1.56ns	1GHz	400MHz	400MHz
Energy metric (fJ/bit/search)	1.420	0.290	0.165	0.339
Operation	W+S	W+S	W+S	RW+S

IV. CONCLUSION

This work presents an energy-efficient TCAM design approach in 40nm LP CMOS process with the size of 256x40. Due to the thicker gate oxide in LP process, a 16T TCAM cell with p-type comparison circuits is proposed to increase the I_{on}/I_{off} difference of the dynamic circuitry. Additionally, the butterfly match-line scheme with AND gates is implemented to decrease the match delay by reducing the capacitances on floating nodes. To further improve energy efficiency, don't-care-based ripple SLs/BLs are utilized to reduce both the switching activities and wire capacitances. Moreover, the column-based DAPC is employed by power gating devices for leakage power reduction and write-ability improvement. The experimental results show a leakage power reduction of 28.9%, a search-line power reduction of 31.74% and an energy efficiency metric of the TCAM macro of 0.339 fJ/bit/search.

ACKNOWLEDGMENT

This work is supported by the Ministry of Science and Technology in Taiwan under MOST 102-2218-E-009-025 and MOST 102-2220-E-009-062.

REFERENCES

[1] P.-T. Huang and W. Hwang, "A 65 nm 0.165 fJ/Bit/Search 256x144 TCAM Macro Design for IPv6 Lookup Tables," *IEEE Journal of Solid-State Circuits*, Vol.46, No.2, pp.507-519, Feb. 2011.

[2] C.-C. Wang, et al., "High-Speed and Low-Power Design Techniques for TCAM Macros," *IEEE Journal of Solid-State Circuit*, Vol.43, No.2, pp.530-540, Feb. 2008.

[3] K. Pagiamtzis, A. Sheikholeslami, "Content-addressable memory (CAM) circuits and architectures: a tutorial and survey," *IEEE Journal of Solid-State Circuits*, Vol.41, No.3, pp. 712- 727, March 2006.

[4] I. Arsovski, et al. , "A 32 nm 0.58-fJ/Bit/Search 1-GHz Ternary Content Addressable Memory Compiler Using Silicon-Aware Early-Predict Late-Correct Sensing With Embedded Deep-Trench Capacitor Noise Mitigation," *IEEE Journal of Solid-State Circuit*, Vol.48, No.4, pp.932-939, April 2013.