ELectronics EXpress

LETTER

IEICE Electronics Express, Vol.10, No.7, 1–9

High-throughput CAM based on a synchronous overlapped search scheme

Naoya Onizawa[1a)], Shoun Matsunaga[2], Vincent C. Gaudet[3], Warren J. Gross[1], and Takahiro Hanyu[2]

[1] *Department of Electrical and Computer Engineering, McGill University, 3480 University Street, Montreal, Quebec, H3A 0E9 Canada*

[2] *Research Institute of Electrical Communication, Tohoku University, 2–1–1 Katahira, Sendai, 980–8577 Japan*

[3] *Department of Electrical and Computer Engineering, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, N2L 3G1 Canada*

a) *naoya.onizawa@mail.mcgill.ca*

Abstract: A high-speed data-search mechanism, called a synchronous overlapped search mechanism (SOSM) that enables a next-word search after searching just a few bits of the current word by simple pre-computation in most cases, is introduced for a content-addressable memory (CAM). Since there are no delay elements in the proposed hardware based on the SOSM, the hardware is robust against timing variation, maintaining high-throughput computing under serious process variation. A 128×64-bit CAM is designed with considering 30% variations of threshold voltages under a 45 nm CMOS technology and operates at 4.2× faster throughput than that of a conventional CAM with 12.3% energy overhead.

Keywords: associative memory, cache, process variations

Classification: Integrated circuits

References

[1] K. Pagiamtzis and A. Sheikholeslami, "Content-addressable memory (CAM) circuits and architectures: a tutorial and survey," *IEEE J. Solid-State Circuits*, vol. 41, no. 3, pp. 712–727, March 2006.

[2] H.-Y. Li, C.-C. Chen, J.-S. Wang, and C. Yeh, "An AND-type match-line scheme for high-performance energy-efficient content addressable memories," *IEEE J. Solid-State Circuits*, vol. 41, no. 5, pp. 1108–1119, May 2006.

[3] Y.-J. Chang and Y.-H. Liao, "Hybrid-type CAM design for both power and performance efficiency," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 16, no. 8, pp. 965–974, Aug. 2008.

[4] A. Mupid, M. Mutyam, N. Vijaykrishnan, Y. Xie, and M. J. Irwin, "Variation Analysis of CAM Cells," *Proc. 8th ISQED*, pp. 333–338, March 2007.

[5] N. Onizawa, S. Matsunaga, V.C. Gaudet, and T. Hanyu, "High-throughput low-energy content-addressable memory based on self-timed overlapped search mechanism," *Proc. 18th ASYNC*, pp. 41–48, May 2012.

DOI: 10.1587/elex.10.20130148
Received February 28, 2013
Accepted March 07, 2013
Published April 10, 2013

[6] C. Zukowski and S.-Y. Wang, "Use of selective precharge for low-power on the match lines of content-addressable memories," *Proc. International Workshop on Memory Technology, Design and Testing, 1997*, pp. 64–68, Aug. 1997.
[7] Predictive Technology Model, [Online] http://www.eas.asu.edu/~ptm/
[8] A. Agarwal, S. Hsu, S. Mathew, M. Anders, H. Kaul, F. Sheikh, and R. Krishnamurthy, "A 128 × 128b high-speed wide-and match-line content addressable memory in 32 nm CMOS," *Proc. ESSCIRC 2011*, pp. 83–86, Sept. 2011.

1 Introduction

Content-addressable memories (CAMs) execute the lookup-table function that compares an input search word to a table of stored words, and return the match location at high speed in a fully parallel equality-search manner for several applications: cache memory, virus detection, and packet classification [1]. CAMs are designed by using dynamic NOR structures where CAM cells are connected to a match line (ML) in parallel, leading to high-speed search; however the CAMs cause large power dissipation because most of ML capacitances are discharged. In contract, a NAND-type structure is designed using a series of CAM cells in order to reduce the switching activity of MLs and hence it operates at medium speed [2]. A hybrid structure [3] using both NOR-type and NAND-type cells is to have benefits of high throughput and low power dissipation. The operation throughput is basically restricted by word lengths of CAMs because the worst-case delay of a match operation is usually proportional to the length. In addition, the worst-case delay is susceptible to process variations that would cause timing failures in emerging sub-65 nm technologies [4].

In this brief, we introduce a high-throughput CAM based on a synchronous overlapped search mechanism (OSM) under process variations in a 45 nm CMOS technology. The OSM eases the restriction of the throughput due to long word lengths based on pre-computation and overlaps precharge and evaluate phases in dynamic circuits using independent control of word circuits, eliminating the waste time of precharging. We have reported an earlier version of this mechanism in [5] and it has been designed using asynchronous circuits, which tend to be sensitive to process variations. In this brief, we introduce an alternative OSM and its hardware implementation using synchronous circuits in order to maintain high throughput under process variations.

The rest of the brief is organized as follows. Section 2 introduces the synchronous overlapped search mechanism. Section 3 describes the circuit implementation. Section 4 demonstrates comparison results of the proposed CAM with a conventional CAM and other related works. Section 5 concludes the paper.

DOI: 10.1587/elex.10.20130148
Received February 28, 2013
Accepted March 07, 2013
Published April 10, 2013

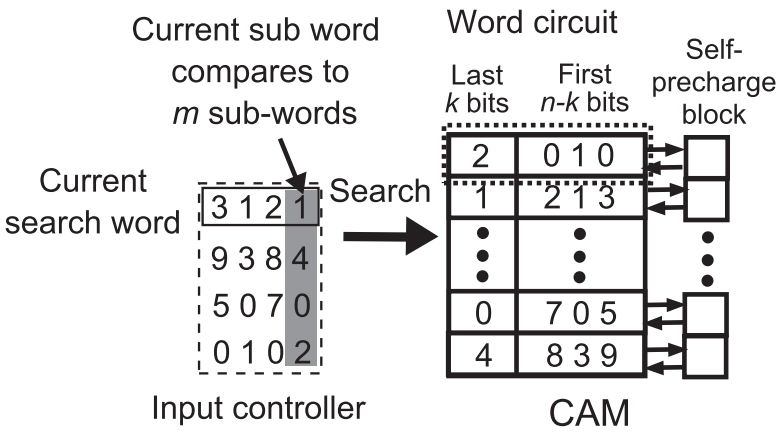

Fig. 1. Architecture of the proposed CAM (m=3).

2 Synchronous overlapped search mechanism

A synchronous overlapped search mechanism (SOSM) includes two approaches: a synchronous word-overlapped search (SWOS) and synchronous phase- overlapped processing (SPOP). Fig. 1 shows an architecture of the proposed CAM. The CAM contains large number of entries (word blocks). Each word block has its self-precharge block. An input controller stores input words and contains m comparators to check differences between the k-bit current input sub word and the k-bit subsequent m input sub words. The word block is partitioned into a small k-bit sub-word block and its subsequent large (n-k)-bit sub-word block using a segmentation method [6]. At first, k-bit input sub words are searched in the k-bit sub-word blocks of the first segment. Matched k-bit sub-word blocks enable the subsequent sub-word blocks in order to find a matched location of the search word. The other sub-word blocks in the second segment remain disabled and hence are not used because of the precedent mismatched k-bit sub-word blocks.

To enhance the throughput, we use the unused sub blocks. The ratio between the unused and the used sub blocks in the second segment is determined by a probability of matching in the first segment. The probability depends on applications. For some applications like caches, the last few bits of stored words tend to be random pattern. Suppose the last k bits of the stored words have a uniform random distribution. In the proposed CAM, the first small sub-word blocks store the last k bits of words and the second large sub-word blocks store the first (n-k) bits. Input words are searched from the Least Significant Bit (LSB). Hence, the probability of matching in the first segment is $(1/2)^k$. After the last k-bit search, most of the sub-word blocks in the second segment are not used.

Fig. 2 shows the SWOS scheme in the proposed CAM that operates at one of two modes: fast and slow. In the proposed SWOS scheme, a search word is processed at a rate based on the short delay of the first segment (T_{1st}) rather than the long delay of the whole word block. Before searching words in the CAM, the last k bits of the current search word are compared with the k-bit subsequent m sub words in the input controller shown in Fig. 2 (a).

DOI: 10.1587/elex.10.20130148
Received February 28, 2013
Accepted March 07, 2013
Published April 10, 2013

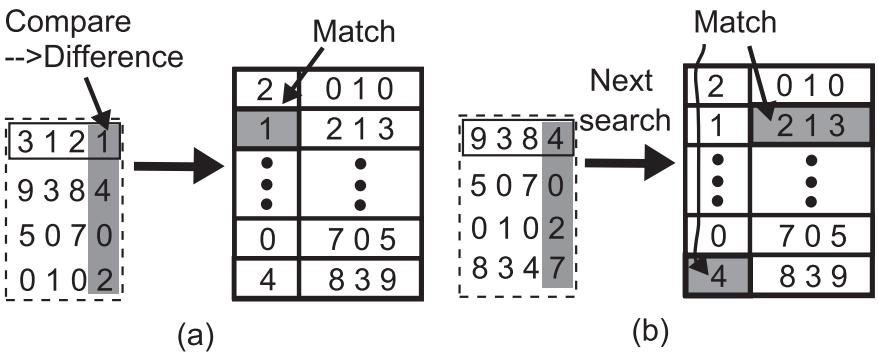

Fig. 2. Synchronous word-overlapped search (SWOS) (m=3). As long as the last k bits of the current search word are different from those of the subsequent m search words, the next search is initiated after the current k-bit sub-search is complete.

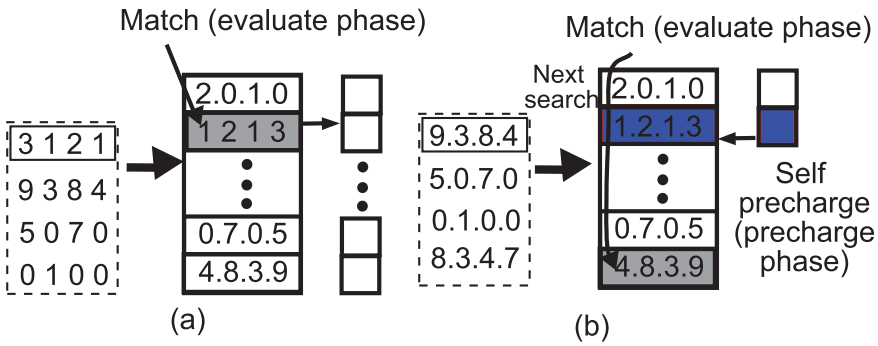

Fig. 3. Self precharging in synchronous phase-overlapped processing (SPOP). A match operation (evaluate) is initiated during precharging a match line of another word block (precharge) that operates as a match at the previous clock cycle.

If they are different, the next search is initiated as soon as the current k-bit sub-word search is complete shown in Fig. 2 (b). As long as the last k bits of the current search word are different from those of the subsequent m search words in the input controller, search words are assigned to unused different word blocks and hence are processed at a rate based on T_{1st}. This is the fast mode. When the consecutive sub-search words are found to be the same, the next sub-search word is assigned to the same sub-word blocks in which the current sub-search word matches. The matched sub-word blocks are initialized after the current search has completed in both segments. As the search time of the second segment is defined by mT_{1st}, the next search is initiated $(1+m)T_{1st}$ after the current search is initiated. This is the slow mode. The average search time based on the proposed SWOS scheme T_{avg} is given by

$$T_{avg} = T_{1st}\left(1-\left(\frac{1}{2}\right)^k\right)^m + (1+m)T_{1st}\left(1-\left(1-\left(\frac{1}{2}\right)^k\right)^m\right). \quad (1)$$

The word circuits are normally designed based on dynamic logic [1] and

DOI: 10.1587/elex.10.20130148
Received February 28, 2013
Accepted March 07, 2013
Published April 10, 2013

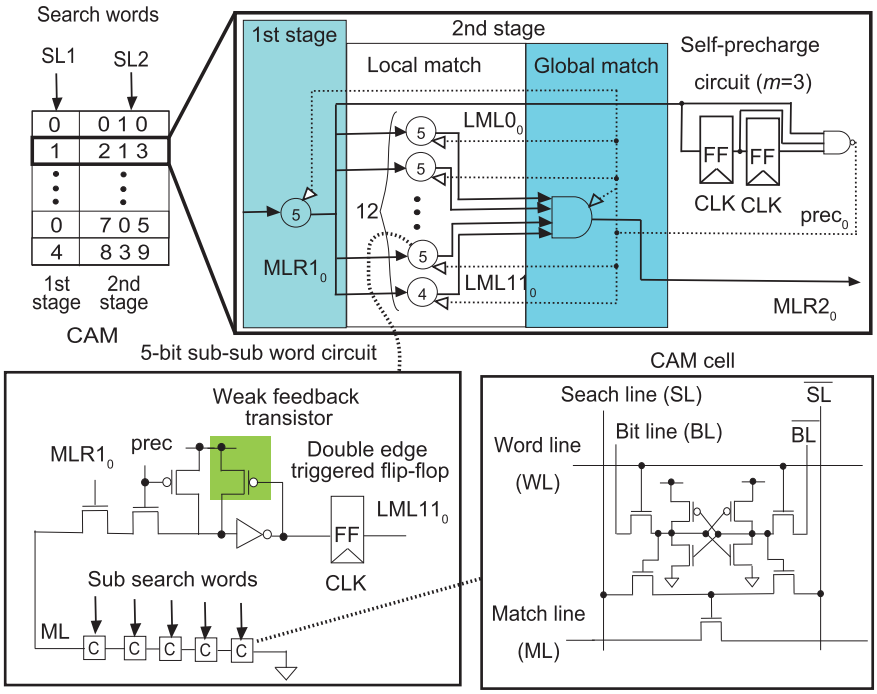

Fig. 4. Circuit diagram of 64-bit CAM word circuit.

operate in one of two phases: precharge and evaluate. In the proposed SPOP scheme, each word circuit is designed based on the NAND-type structure and is independently controlled using a self-precharge block shown in Fig. 3 (a). In the proposed word circuits, match line (ML) capacitances of only matched word circuits are discharged while ML capacitances of mismatched (unused) word circuits remain. The ML capacitances of the matched word circuits are then precharged by their self-precharge blocks in the precharge phase in Fig. 3 (b). In contrast, other word circuits that were previously mismatched still operate in the evaluate phase to keep the match operation of the CAM for the next search word. Using the proposed SWOS scheme, consecutive search words are assigned to unused different word circuits. As the ML capacitances of the unused word circuits have not been discharged, input search words can be processed without wasting the precharge time. Basically, the delay time of matching (evaluate phase) is larger than that of precharging (precharge phase). Hence, the cycle time of the proposed CAM is determined by the delay time of matching T_{avg}. Using the SPOP scheme, as a clock signal doesn't need to be assigned to the precharge phase (low) and the evaluate phase (high) unlike conventional CAMs, both transitions of a clock signal can be used.

3 Circuit implementation

Fig. 4 shows a circuit diagram of the proposed CAM word circuit. k is set to 5 and n is set to 64. It contains a 5-bit 1st-stage sub-word circuit, a self-precharge circuit and a 59-bit 2nd-stage sub-word circuit. Flip-flops segment the two sub-word circuits. An input controller sends a search word to the CAM at both transitions (high and low) of a clock signal. The 1st-stage

DOI: 10.1587/elex.10.20130148
Received February 28, 2013
Accepted March 07, 2013
Published April 10, 2013

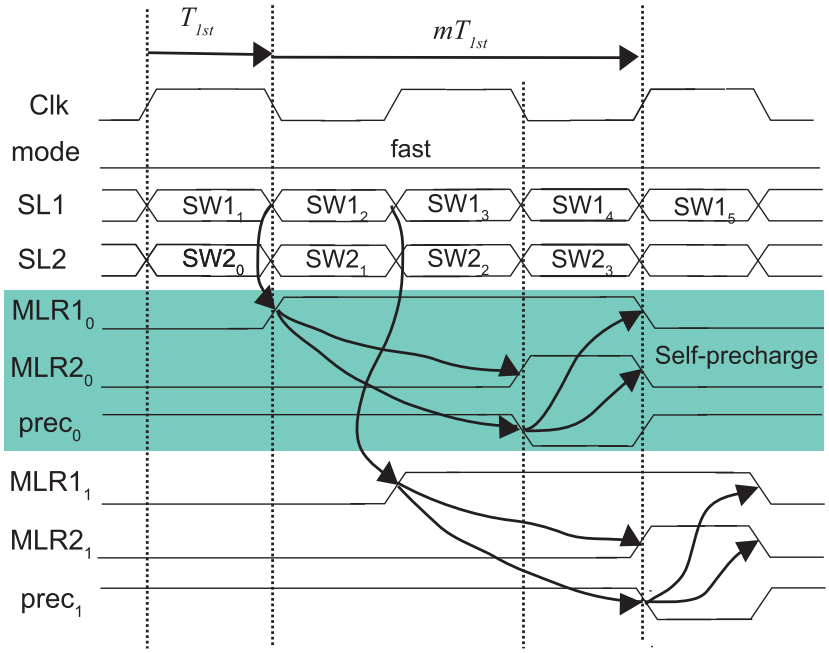

Fig. 5. Timing diagrams of the proposed CAM in fast mode (m=3).

segment operates at every transition, while the 2nd-stage segment requires $(m-1)$ transitions. Precharging MLs takes 1 transition. As a matched result needs to be stored at both transitions of a clock signal, double-edge triggered flip-flops are used.

The 5-bit 1st-stage segment is implemented using a series of 5 NAND-type cells, a precharge PMOS transistor, a weak feedback PMOS transistor, and a double-edge triggered flip-flop. A 5-bit sub word sent from the input controller is assigned and compared with the stored 5-bit sub word. The output $\mathrm{MLR1}_0$ is high when the input sub word matches and remains low when it mismatches. The ML is charged through the precharge PMOS transistor controlled by the self-precharge circuit after the current search is complete. The weak feedback PMOS transistor is used to solve a charge-sharing problem in the NAND-type cell [2]. $\mathrm{MLR1}_0$ is connected to the self-precharge circuit and the 2nd-stage segment.

The 2nd-stage segment contains 12 local match circuits and a global match circuit. We use a hierarchical design style where a word circuit is divided into local match circuits and a global match circuit [1]. Once $\mathrm{MLR1}_0$ is asserted, all 12 local match circuits can operate. Otherwise, they don't operate because the last 5 bits of an input word mismatch. The remaining 59-bit sub word is sent from the input controller in parallel and is partitioned into 5-bit 11 sub-sub words and a 4-bit sub-sub word. Each sub-sub word is processed in its local match circuit. Each output $\mathrm{LML_i}$ $(0 \leq i < 12)$ is high when its sub-sub word matches or low when it mismatches. Every output $\mathrm{LML_i}$ is connected to the global match circuit. When every output $\mathrm{LML_i}$ is asserted, the output $\mathrm{MLR2}_0$ is asserted. It means that a search word matches the stored word of the word circuit. Otherwise $\mathrm{MLR2}_0$ remains low.

DOI: 10.1587/elex.10.20130148
Received February 28, 2013
Accepted March 07, 2013
Published April 10, 2013

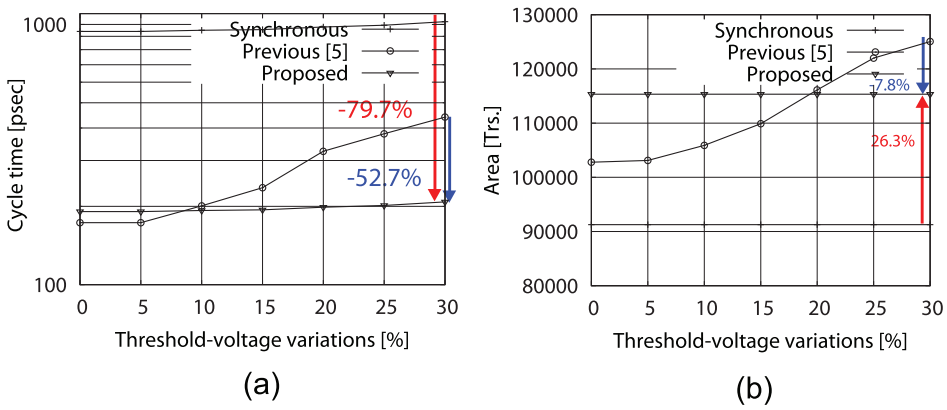

Fig. 6. Performance comparisons of 128×64 CAMs under threshold-voltage variations (3σ) on a 45 nm CMOS: (a) average cycle time and (b) area.

While the 2nd-stage segment operates, the self-precharge circuit also operates. It contains two double-edge triggered flip-flops and a three-input NAND gate. The number of flip-flops corresponds to $(m-1)$, where m is set to 3 in the figure. The output of the self-precharge circuit (prec_0) is deasserted to precharge all MLs of a matched word circuit m transitions after the 1st-stage segment is matched. The output (prec_0) charges all MLs of the word circuit and is then asserted at the next transition.

Fig. 5 shows the timing diagram of the input controller and the CAM block ($m = 3$) in fast mode. The input controller sends a k-bit sub word (SW1) and a $(n-k)$-bit sub word (SW2) onto SL1 and SL2, respectively, at different timing. First, $\mathrm{SW1}_0$ matches in a k-bit first segment ($\mathrm{ML1}_0$) and the matched result is stored in a flip-flop ($\mathrm{MLR1}_0$). Second, $\mathrm{SW2}_0$ matches using $\mathrm{MLR1}_0$ in the $(n-k)$-bit subsequent segment ($\mathrm{ML2}_0$). Concurrently, $\mathrm{SW1}_1$ is processed in another first segment ($\mathrm{ML1}_1$).

When consecutive SW1 values are the same, the input controller operates in slow mode. In the slow mode, the input controller stops sending a new search word until the current search is complete in the CAM block. In fact, once an input search word matches in a word block, the word block cannot be used during a period of the 2nd segment matching and the self pre-charging MLs. As that period corresponds to mT_{1st}, m needs to satisfy the following condition:

$$m = \lceil \frac{T_{2nd}}{T_{1st}} \rceil + \lceil \frac{T_{prec}}{T_{1st}} \rceil, \tag{2}$$

where T_{2nd} is the delay time of the 2nd segment and T_{prec} is the delay time of pre-charging MLs. After the slow mode, the input controller operates in the fast mode, again.

4 Evaluation

Fig. 6 shows performance comparisons under process variations (3σ) among a synchronous, a previous [5] and the proposed CAMs. These three CAMs are designed under a 45 nm PTM model [7]. The size of the CAMs is 128 × 64-bit words. The synchronous CAM is designed based on a combination

DOI: 10.1587/elex.10.20130148
Received February 28, 2013
Accepted March 07, 2013
Published April 10, 2013

Table I. Feature summary and comparisons.

	Synchronous V_{DD}=1.0 V	Hybrid [3]	HS-WA [8]	Proposed V_{DD}=0.9 V
Configuration	128×64	128×32	128×128	128×64
Technology	45 nm	0.13 μm	32 nm	45 nm
Cycle time (ns)	1.025	0.6	0.290	0.242
Energy metric (fJ/bit/search)	0.227	1.3	1.07	0.255
# of trs. in cell	9	9	11	9

of segmentation method [6] and the hierarchical design style that are used in the previous and proposed CAMs. m is set to 3 in the proposed CAM. The performance is evaluated using 100-time Monte Carlo simulations on HSPICE. The simulation condition is room temperature and V_{DD} is 1.0 V.

Under small variations of threshold voltages, the previous CAM designed using asynchronous circuits achieves the smallest cycle time, but the cycle time is increased as the variations are increased because it needs to use large delay elements that guarantee the CAM operates properly. The use of large delay elements degrades the throughput and increases the area. In contrast, the average cycle times of the synchronous and the proposed CAMs are slightly affected by the variations. The cycle time of the proposed CAM is 20.3% and 47.3% of the synchronous and the previous CAMs under 30% variations of threshold voltages. The number of transistors of the proposed CAM is reduced by 7.8% compared with that of the previous CAM under 30% variations of threshold voltages.

Table I summarizes the performance of the proposed and the synchronous CAMs with considering 30% variations of threshold voltages and other recently reported CAMs. A supply voltage of the proposed CAM is set to 0.9 V that leads to comparable energy metric of the synchronous CAM. The high-speed CAM designed in a 32 nm CMOS [8] achieves the cycle time of 290 ps using swapped CAM cells that reduces the search delay. However it increases the search power dissipation because the search line is connected to two gates of transistors in the cells while the search line of the proposed CAM is connected to the drain of a transistor in the cell. The energy metric of the proposed CAM is only 23.8% of the high-speed CAM in [8] while maintaining comparable cycle time. In addition, the number of transistors of the swapped cell is 11, causing large area because the number of transistors of a typical CAM cell used in the proposed CAM is 9.

5 Conclusion

In this brief, we have proposed a high-throughput content-addressable memory (CAM) based on a synchronous overlapped search mechanism (SOSM) under process variations. The SOSM eases the restriction of the throughput due to long word lengths based on pre-computation and overlaps precharge and evaluate phases, eliminating the waste time of precharging. The proposed hardware using synchronous circuits achieves better throughput than the pre-

DOI: 10.1587/elex.10.20130148
Received February 28, 2013
Accepted March 07, 2013
Published April 10, 2013

vious hardware using asynchronous circuits over 10% variations of threshold voltages. As a design example, a 128 × 64-bit CAM is designed with considering 30% variations of threshold voltages and evaluated by HSPICE simulation under a 45 nm CMOS technology. The proposed CAM achieves 245-ps cycle time that is just 23.7% of a conventional CAM with 12.3% energy overhead.

Acknowledgments

This research was supported by JST, CREST. This work was supported by JSPS KAKENHI (23700050) and JSPS KAKENHI (22360137).

DOI: 10.1587/elex.10.20130148
Received February 28, 2013
Accepted March 07, 2013
Published April 10, 2013