Digital Associative Memories Based on Hamming Distance and Scalable Multi-Chip Architecture

Yusuke Oike[†], Makoto Ikeda[†‡], and Kunihiro Asada[†‡]

[†]Dept. of Electronic Engineering, University of Tokyo
[‡]VLSI Design and Education Center (VDEC), University of Tokyo
7–3–1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
Phone: +81-3-5841-6719, Fax: +81-3-5841-8912
E-mail: {y-oike, ikeda, asada}@silicon.u-tokyo.ac.jp

Abstract

In this paper, we present a new concept and its circuit implementation for high-speed associative memories based on Hamming distance. A chained search logic embedded in a memory cell enables a word-parallel search operation based on Hamming distance. A hierarchically chained search architecture maintains a high-speed operation with faultless precision in a large input number. We also propose a scalable multi-chip architecture which attains fully chip- and word-parallel Hamming distance search by taking advantage of the digital associative memories. We have designed and fabricated a prototype chip with 64 bit × 32 word memories using a 0.18 μm CMOS process. The measurement results show that the operation achieves a speed of 411.5 MHz at a supply voltage of 1.8 V. The worst-case search time is 158.0 ns for 64-bit 32-word stored data. In a low-voltage operation, the operation speed achieves 40.0 MHz at a supply voltage of 0.75 V.

Introduction

Some applications, such as data compression, pattern recognition, multi-media and intelligent processing, require considerable memory access and data processing time. Therefore, context addressable memories (CAMs) [1]–[2] have been developed to reduce the access and data processing time for detection of complete-match data. In recent years, many advanced applications require a search operation for not only complete-match data but also near/nearest-match data. Conventional associative memories that employ analog circuit techniques have been proposed for quick nearest-match detection [3]–[6]. Their circuit implementations are generally compact, however there are difficulties in operating them with faultless precision in a deep sub-micron (DSM) process and at a low-voltage supply. Moreover, the feasible data capacity is limited by the analog operation. Therefore, they are not suitable for a system-on-a-chip VLSI in DSM process technologies.

In this paper, we report digital associative memories based on Hamming distance and a scalable multi-chip architecture. We have proposed a hierarchically chained search architecture embedded in memories [7]. It has four principal advantages as follows. The first advantage is that the hierarchical search architecture enables a high-speed search in a large database. The search cycle time is limited by O($\sqrt{N}$) and O(log M) at an N-bit M-word data capacity. Although the total search time increases in proportional to the bit length, it reduces the number of clocks for nearest-match detection in practical use since it provides a result for the data close to the input with a fewer number of clocks. The second advantage is no theoretical limitations on the data patterns M, the bit length N, and the data distance D due to the synchronous search operation. Moreover, the proposed multi-chip architecture using pipelined binary-tree priority decision circuits attains high capacity scalability. The third advantage is a low-voltage operation in a DSM process. The circuit implementation has a tolerance for device fluctuation and allows a low-voltage operation of less than 1.0 V, which is difficult to attain using the conventional analog approaches. The fourth advantage is that it provides additional functions for associative processing. The present architecture provides data addresses in the sorted order of the exact Hamming distance. Therefore, it enables high-speed data sorting in addition to nearest-match detection for conventional use. We have designed 64-bit 32-word associative memories using a 1P5M 0.18 μm CMOS process and have successfully demonstrated the high-speed distance computation and the low-voltage operation with faultless precision.

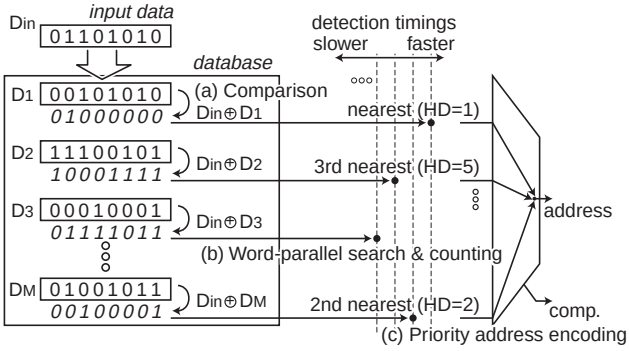

Fig. 1. Operation diagram of the fully digital and word-parallel associative memories.

Word-Parallel and Hierarchical Search Architecture

We propose a logic-in-memory architecture using word-parallel search signal propagation via chained search circuits. The Hamming distance (HD) search operation includes data comparison and search signal propagation with mismatch masking as shown in Fig.1. First, an input (D_{in}) is compared with all the stored data ($D_0, D_1, ...D_M$) using an XOR gate in bit parallel. Then, the number of mismatch bits is counted by the chained search circuits in word parallel. The search circuit is also embedded in a memory cell and controls the search signal propagation based on the comparison results ($D_{in} \oplus D_M$). Stored data are divided into blocks and connected by hierarchical nodes as shown in Fig.2 since the search cycle time is limited by the search signal propagation via the chained search circuits. A hierarchical node provides a permission signal to the next block and the next hierarchical node. The permission signal makes a mismatch bit maskable.

Search signals (SS) are simultaneously injected to all blocks.

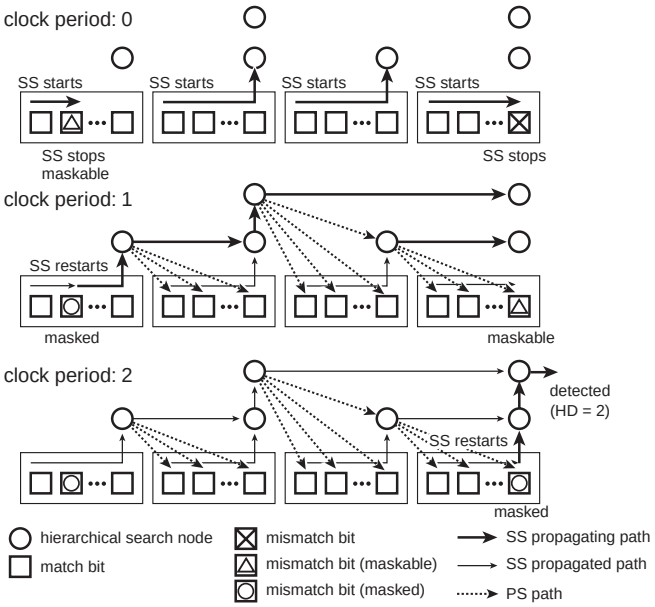

Fig. 2. Hierarchical chained search architecture and operation diagram in a case of HD = 2.

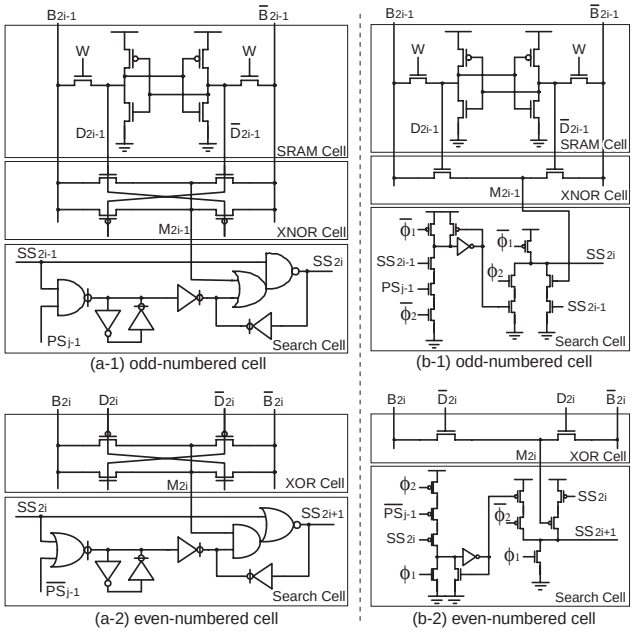

Fig. 3. Circuit configurations of the associative memory cell: (a-1) odd-numbered cell, (a-2) even-numbered cell of static circuit implementation; (b-1) odd-numbered cell, (b-2) even-numbered cell of dynamic circuit implementation.

The search signal passes through match bits via the search circuits. Some propagations are interrupted at the first-encountered mismatch bit. The others pass to the hierarchical nodes and update the permission signals for the next block and hierarchical node as shown by the clock period 0 in Fig.2. In this period, the data with HD = 0 are detected since the search signal is provided from the last hierarchical node without any interruption. Only one mismatch bit, which interrupts the search signal propagation and receives a permission signal from the previous hierarchical node, becomes maskable in each word. During the next clock period, the search signal restarts from the masked bit and updates the permission signals again. It must be noted that the operated clock cycles represent the

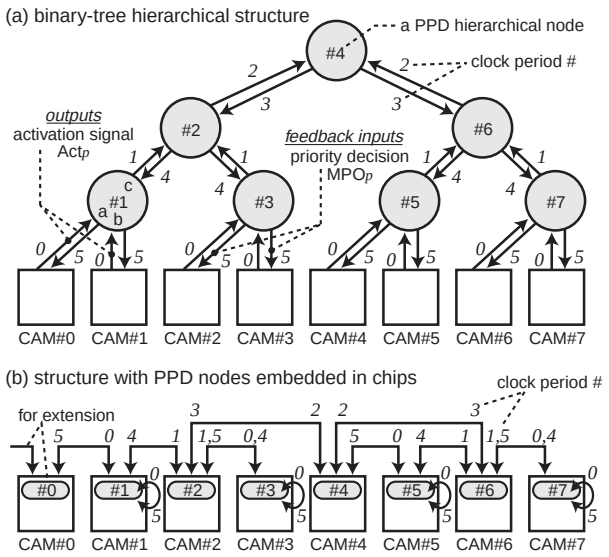

Fig. 4. Hierarchical multi-chip structure using pipelined binary-tree priority decision circuits.

Hamming distance of the detected data. For example, the data with HD = 2 are detected in the clock period 2 as shown in Fig.2. In this manner, the data with HD = *n* are detected in the *n*-th clock period. Thus, the search operation can detect not only the nearest-match data but also all data in the sorted order of Hamming distance in synchronization with the clock cycle.

Fig.3 shows circuit configurations of the associative memory cell. The memory cell is composed of a SRAM cell, an XOR/XNOR circuit for data comparison, and a search circuit. Even-numbered and odd-numbered search circuits are designed in a complementary fashion. SRAM part of even-numbered cell is omitted in Fig.3. In a matched bit, the search signal (SS) always passes to the next bit since the result (M) of comparison is true. In a mismatched bit, the SS stops and waits for the next clock period. In the next clock period, the false M is masked and the SS restarts at the cell where both the search signal (SS) and the permission signal (PS) are true. Therefore, only one mismatched bit is masked in word parallel and all data can be detected in order of Hamming distance. Fig.3 (a) shows a static circuit implementation. It realizes a high tolerance for device fluctuation and a low-voltage operation. Fig.3 (b) shows a compact circuit implementation using dynamic circuits. It saves a search circuit area for large capacity.

Scalable Multi-Chip Architecture

Fig.4 shows a hierarchical multi-chip structure using a binary-tree pipelined priority decision (PPD) circuit. All CAM chips are hierarchically connected via PPD nodes as shown in Fig.4 (a). A CAM chip that detects data of HD = *D* during the *D*-th clock period provides an activation signal (Act_p) to a PPD node. The activation signal is generated by an intra-chip completion signal. The hierarchical PPD nodes transfer the activation signals to the next stage while it determines which one is a priority result. Finally, they return the priority decision results (MPO_p) to the CAM chips. The additional latency for the multi-chip structure is a logarithmic order of the number of chips. A PPD node is embedded in a CAM chip as shown in Fig.4 (b). Therefore, all CAM chips are implemented by the

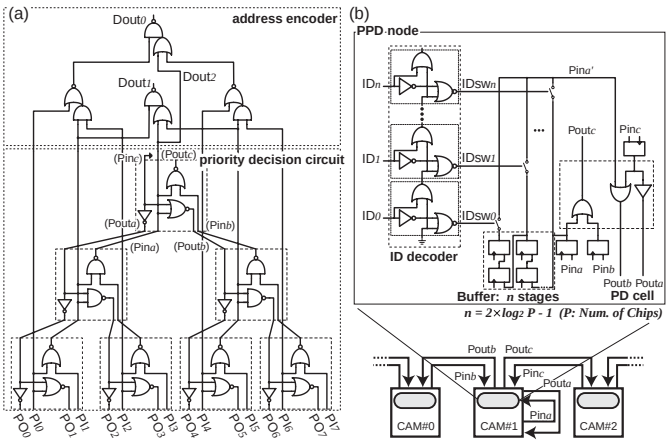

Fig. 5. Simplified binary-tree priority decision circuit: (a) intra-chip priority decision circuit and address encoder, (b) inter-chip pipelined priority decision circuit.

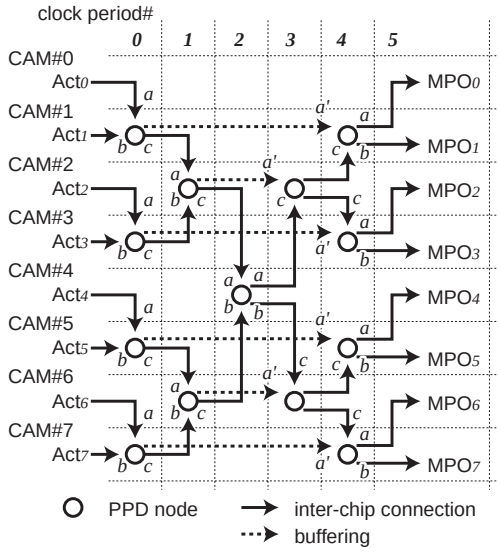

Fig. 6. Timing diagram of PPD circuit in case of 8 chips.

same circuit configuration. This feature enables a multi-chip structure without any additional processor chip.

The intra-chip priority decision is carried out by a binary-tree priority address encoder as shown in Fig.5 (a). An inter-chip PPD circuit is designed based on the binary-tree priority decision circuit. A PPD node consists of a priority decision cell, an ID decoder, and register buffers as shown in Fig.5 (b). A priority decision cell has three inputs (*Pin*) and three outputs (*Pout*) in a similar configuration to the intra-chip priority decision circuit. In the intra-chip priority decision circuit, an input of Pin_a is also used for a return path from the upper hierarchical level. On the other hand, the inter-chip priority decision circuit loses the original input of Pin_a since the operations are pipelined. Therefore, an input of Pin_a is buffered by shift registers in each PPD node. The shift registers are prepared according to the maximum number of chips. The number of buffer stages is set by the chip ID since the return path length is different for the hierarchical levels. Fig.6 shows a timing diagram of the inter-chip PPD circuit. The number of buffer stages can be determined by the least true bit of a chip ID because of a binary-tree structure. An inter-chip completion signal MCO_p is acquired by $Pout_c$ of the top node, for example, $Pout_c$ of CAM#4 in a

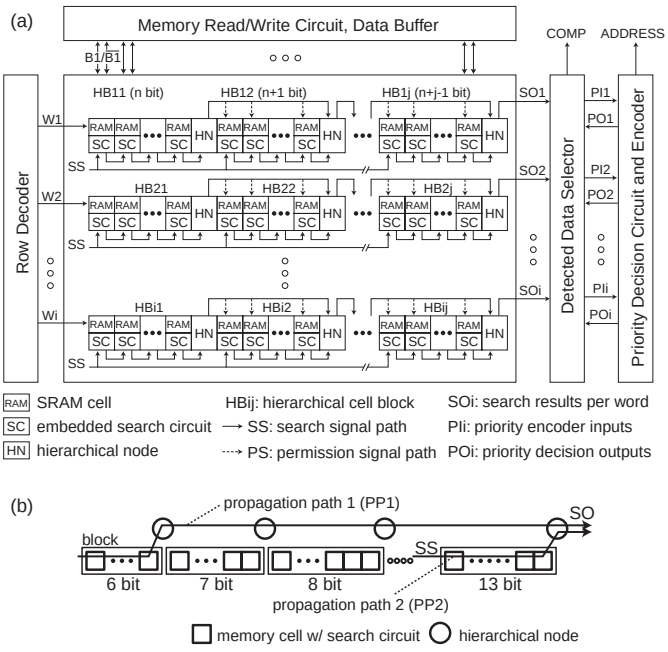

Fig. 7. Block diagram: (a) associative memory array, (b) word structure.

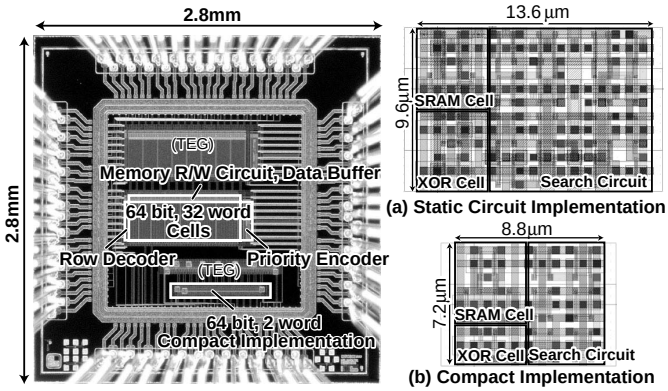

Fig. 8. Chip microphotograph and memory cell layouts: (a) static circuit implementation, (b) dynamic circuit implementation.

multi-chip structure with eight chips. The completion signal is provided to each chip along a return path.

Chip Implementation

We have designed 64-bit 32-word associative memories with the static circuit implementation using a 1P5M 0.18 μm CMOS process[1]. Fig.7 (a) illustrates a block diagram of the associative memories. Fig.8 shows the chip microphotograph and the cell layouts. The associative co-processor is composed of a 64-bit 32-word associative memory array, a memory read/write circuit with data buffers, a word address decoder, and a 32-input priority encoder with detected data selectors. A two-stage hierarchical structure is implemented as shown in Fig.7 (b). A hierarchical node is achieved by a 2-input AND gate. In the 2-stage hierarchical structure, the number of hierarchical nodes on each propagation path is different. Therefore, the number of blocks and each bit length need to be optimized for the min-

[1]The VLSI chip in this study has been fabricated through VLSI Design and Education Center (VDEC), University of Tokyo in collaboration with Hitachi Ltd. and Dai Nippon Printing Co.

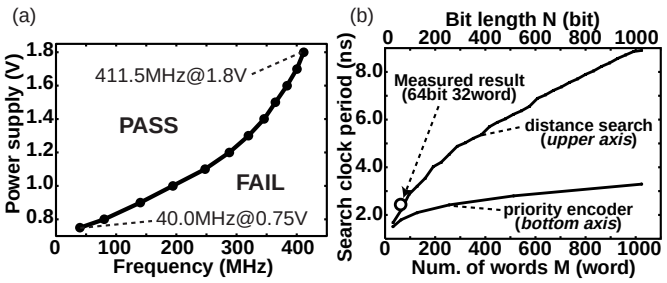

Fig. 9. Measurement results: (a) operation frequency vs. power supply voltage, (b) data capacity vs. search clock period.

TABLE I Estimated area of the associative memories.

Database capacity	Area (Module size)
4K (64 b × 64)	0.98 mm^2 (0.79 × 1.24)
16K (128 b × 128)	3.02 mm^2 (1.40 × 2.16)
64K (256 b × 256)	11.05 mm^2 (2.63 × 4.20)
256K (512 b × 512)	38.34 mm^2 (5.08 × 7.55)
1M (1024 b × 1024)	146.40 mm^2 (10.00 × 14.64)

imum critical path. We have also designed a 64-bit 2-word associative memory using dynamic circuit implementation for feasibility and performance evaluation. The memory cell area is 13.6 × 9.6 μm^2 and 8.8 × 7.2 μm^2 in the static and dynamic circuit implementations, respectively.

Measurement Results and Discussions

The measurement results show that the operation speed is 411.5 MHz and 40.0 MHz at a supply voltage of 1.8 V and 0.75 V, respectively. Fig.9 (a) shows the operation speed as a function of a supply voltage from 0.75 V to 1.8 V. The total search time increases in proportion to the distance of detected data. For example, the nearest-match detection is completed in 17 clock periods (i.e. 41.3 ns) when the nearest-match data has a 16-bit distance from an input. The worst-case operation of nearest-match detection or data sorting requires 65 clock periods, thus, it takes 158.0 ns. Fig.9 (b) shows the relation between the search clock period and the data capacity. The hierarchical search architecture maintains a high-speed search operation in a large data capacity.

The 64-bit 32-word associative memory module occupies 475 μm × 1160 μm (0.55 mm^2). Table I shows estimated core area of the associative memories in various data capacities. The number of transistors in the proposed memory cell is larger than the conventional analog approaches [3]–[6]. The analog approaches are, however, difficult to follow device scaling especially in a DSM process with keeping their performance and marginal capacity. The present approach can follow device scaling and operate in a low voltage supply because of the synchronous digital search logics embedded in memories. Besides, it has no limitation of the capacity and the search distance. Therefore, the associative memories have more potential for a practical use and a large capacity than the conventional designs.

The power dissipation of the associative co-processor is < 51.3 mW at a supply voltage of 1.8V and an operation frequency of 400 MHz. In a low-voltage operation, it is 1.18 mW at a supply voltage of 0.75V and an operation frequency of 40 MHz. Search accuracy of the conventional analog approach is unstable in a low-voltage operation. The present search results are strictly exact regardless of the power supply voltage. This feature contributes to not only a low power operation but also the suitability to a system-on-a-chip application. Table II summarizes the chip specifications.

TABLE II Prototype specifications.

Process	1P5M 0.18 μm CMOS process
Power Voltage Supply	0.7 V – 1.8 V
Organization	64 bit × 32 word memory cells
Functions	Nearest-match detection
	Data sorting
Module Size	475 μm × 1160 μm (0.55 mm^2)
Num. of Transistors	88.5k transistors
Memory Cell Size	9.6 μm × 13.6 μm (130.56 μm^2)
Operation Speed	411.5 MHz (@ 1.8V, measured)
	40.0 MHz (@ 0.75V, measured)
Worst-Case Search Time	158.0 ns (0-bit to 64-bit distance)
Power Dissipation	51.3 mW (@ 1.8V, 400MHz)
	1.18 mW (@ 0.75V, 40MHz)

Conclusions

We have proposed a new concept and circuit implementation of high-speed and low-voltage associative memories with exact Hamming distance computation. A hierarchical search architecture attains a high-speed search operation in a large database. The digital circuit implementation enables a high tolerance for device fluctuation and a low-voltage operation under 1.0 V. Furthermore, it is capable of a continuous search operation for data sorting in addition to the traditional nearest-match detection. The proposed hierarchical multi-chip architecture realizes high capacity scalability by taking advantage of the digital associative memories. We have designed 64-bit 32-word associative memories using a 0.18 μm CMOS process. It achieves 411.5 MHz and 40.0 MHz operations at a supply voltage of 1.8 V and 0.75 V, respectively.

References

[1] T. Ogura, J. Yamada, S. Yamada, and M. Tanno, "A 20-kbit Associative Memory LSI for Artificial Intelligence Machines," *IEEE J. Solid-State Circuits*, vol. 24, no. 4, pp. 1014 – 1020, Aug. 1989.

[2] H. Miyatake, M. Tanaka, and Y. Mori, "A Design for High-Speed Low-Power CMOS Fully Parallel Content-Addressable Memory Macros," *IEEE J. Solid-State Circuits*, vol. 36, no. 6, pp. 956 – 968, Jun. 2001.

[3] T. Yamashita, T. Shibata, and T. Ohmi, "Neuron MOS Winner-Take-All Circuit and Its Application to Associative Memory," *IEEE ISSCC Dig. Tech. Papers*, pp. 236 – 237, Feb. 1993.

[4] M. Nagata, T. Yoneda, D. Nomasaki, M. Sato, and A. Iwata, "A Minimum-Distance Search Circuit using Dual-Line PWM Signal Processing and Charge-Packet Counting Techniques," *IEEE ISSCC Dig. Tech. Papers*, pp. 42 – 43, Feb. 1997.

[5] M. Ikeda and K. Asada, "Time-Domain Minimum-Distance Detector and Its Application to Low-Power Coding Scheme on Chip-Interface," *Proc. of Eur. Solid-State Circuits Conf. (ESSCIRC)*, pp. 464 – 467, 1998.

[6] H. J. Mattausch, N. Omori, S. Fukae, T. Koide, and T. Gyohten, "Fully-Parallel Pattern-Matching Engine with Dynamic Adaptability to Hamming or Manhattan Distance," *IEEE Symp. on VLSI Circuits Dig. Tech. Papers*, pp. 252 – 255, 2002.

[7] Y. Oike, M. Ikeda, and K. Asada, "A High-Speed and Low-Voltage Associative Co-Processor With Exact Hamming/Manhattan-Distance Estimation Using Word-Parallel and Hierarchical Search Architecture," *IEEE J. Solid-State Circuits*, vol. 39, no. 8, pp. 1383 – 1387, Aug. 2004.