Associative Memory for High-Speed Nearest Hamming/Manhattan Distance Search with Large Reference Pattern Number

Yuji Yano, Tetsushi Koide, Hans Jürgen Mattausch

Research Center for Nanodevices and Systems, Hiroshima University, 1-4-2 Kagamiyama, Higashi-Hiroshima 739-8527, Japan

Phone:+81-824-24-6265, Fax:+81-824-22-7185, e-mail:{yano, koide, hjm}@sxsys.hiroshima-u.ac.jp

Abstract

Fully-parallel associative memories for minimum Hamming/Manhattan distance search have been designed in 0.6μm/0.35μm CMOS with 3-metal layers. These are suitable for all associative-memory applications which need real-time processing, compact hardware, and low-power dissipation. The performances of designed test chips are about equivalent to a 32 bit computer with 1.34TOPS (Hamming) and 160GOPS (Manhattan). Furthermore, a bank-type associative memory architecture for minimum distance search with large reference number is proposed and test chips have been designed in 0.35μm CMOS technology.

1. Introduction

The pattern-matching function, which finds the nearest-match between an input-data word of W bit length and a number R of reference-data words, is important for realizing recognition, routing calculation at network routers, as well as data/image compression by vector-quantization. The nearest-match or winner is defined by the minimum with respect to a distance measure such as the Hamming (data strings, voice patterns, black/white pictures) and the Manhattan (gray-scale or color pictures) distance. Conventional partially-parallel minimum distance-search hardware based on multiple SRAMs and external distance calculation plus winner-take-all circuitry (WTA) [1] has drawbacks with respect to integration density and short nearest-match times. To overcome these drawbacks, we have proposed a dedicated mixed analog-digital fully-parallel associative-memory architecture for nearest Hamming/Manhattan-distance search [2,3]. Designed minimum Hamming/Manhattan distance search associative memories have high-performance at low-power dissipation. A bank-based architecture is also proposed for enabling minimum distance search with large reference pattern number.

2. Associative Memory Architecture

Figure 1 shows a block diagram of the compact associative memory with fast fully-parallel match capability according to the Hamming/Manhattan distances. The concept for the memory-field is illustrated in Fig. 2. Digital k-bit subtraction and absolute-value calculation units (UC) compare the W binaries, each with k-bit, in all rows of the memory field in parallel with the reference data. The k-bit subtraction and absolute-value-calculation circuit is realized on the basis of a ripple carry adder circuit. In the test chip design, we use a newly devised compact circuit to minimize its design area. The circuit diagram of the winner line-up amplifier (WLA) is shown in Fig. 3. The WLA achieves a large regulation range for feedback stabilization and eliminates the inefficient possibilities of under- or over-regulation by a maximum-gain region which self-adapts to the winner input C_{win}. The signal follower provides the necessary high driving current for scaling to an increased number of reference patterns R. Low power dissipation is achieved by an individual power regulation from the signal-regulation units for each input-signal source. The transistor-count is only 6 per row. A modified version of the fast minimum circuit proposed by Opris et al [4] is applied for combined feedback generation and distance amplification. The minimum function is used in the feedback loop and an intermediate node in each row circuitry is used for the distance-amplified WLA-output LA_i.

Distance amplification and self-adaptation of the maximum-gain region work as follows: Since the winner-row's WC-output C_{win} is lowest, transistor p_{3win} has the largest current-source capability, which must be balanced by the current-sink capability of transistor n_{2win}. Thus the gate voltage F_a of n_{2win}, which is common to all rows, has to rise appropriately and is controlled by the winner row. This in turn is only possible, if the gate voltage of the source follower n_{3win}, being also the output voltage LA_{win}, rises highest. The mechanism works independent of the absolute value of C_{win} and provides the self-adaptability of the maximum-gain region. A gain of about 20-50 over a wide range of absolute C_{win} input voltages is thus achieved. The WTA-circuit implemented in the test-chip is of O(R) complexity and needs just 17 transistors per row. We adopt 5 stages of the common-source WTA-configuration proposed by Lazzaro et al. [5]. This 5-stage WTA amplifies winner-loser distances by voltage-current-voltage transformations and provides a further strong amplification of the winner-loser differences. The final decision circuit consists of inverters with an adjusted switching threshold. It generates a 1 for the winner row and 0 for all other rows.

A newly proposed bank associative memory architecture is shown for the case of 4 banks in Fig. 4. This system has 4 local-winner-search units and a global minimum-distance-winner selection circuit. A local-winner is decided by fully-parallel minimum distance search in each bank in parallel. Each bank consists the associative memory unit, a priority encoder (PE), a circuit for digital-distance calculation of the local winner. The minimum-distance-winner selection circuit determines the global winner among the local-winners and outputs the global winner's bank number as well as bank-internal address.

3. Chip-fabrication and Measurement Results

The Hamming-distance test chip is designed in 0.6μm CMOS with 3-metals and contains 32 reference words with 768 bit binaries (Fig. 5). Design area is 9.75mm^2 and a high performance of < 70nsec minimum distance search at low-power dissipation of 43mW are achieved. The Manhattan-distance test chip was designed in 0.35μm CMOS with 3-metal layers and contains 128 reference words with 16 binaries each 5-bit long. Fig. 6 (a) shows the photomicrograph of the fabricated Manhattan-distance test chip. Fig. 6 (b) depicts the measured average nearest-match times of this chip as a function of the distance between winner and input-data word. The data for winner to nearest-loser distances of 1 and 10 bit are plotted. Some of the chosen row combinations of winner and nearest loser delivered unreliable match results for large winner-input distance. However, this causes no practical problem because vector-quantization (VQ) simulations of real images confirmed, that almost all winner-input distances are less than 50bit. In the practical case with optimized codebook winner patterns with larger winner-input distance are in general expected to be very seldom. Therefore, the measured performance of the designed test-chip is already sufficient for VQ application with a nearest match time < 140nsec. Taking into account that the area for the input-pattern circuit remains the same, we extrapolate an area of about 17.2mm^2 and a power-dissipation of about 180mW for a nearest Manhattan-distance-search memory with 256 reference patterns in 0.35μm CMOS technology. If we furthermore extrapolate the test-chip data to a state-of-the-art 0.13μm CMOS technology with 1.2V power-supply, we expect an integration area of about 6.4 mm^2 and a power dissipation of about 71.7mW. Table 1 shows the data of fabricated test chips for minimum Hamming/Manhattan distance search.

Bank-based associative memories have been designed in 0.35μm CMOS technology and the layout of a 4-bank associative memory for minimum Manhattan distance search with 256

reference pattern number is shown as an example in Fig. 7 (a). Table 2 shows performance data of designed 2/4 bank associative memories.

4. Conclusion

Associative memory architecture for fully-parallel minimum distance search is proposed and test chips are designed in 0.6μm (Hamming) and in 0.35μm (Manhattan) CMOS technologies. The 9.75mm^2 Hamming test-chip with 32 reference patterns and 768 equivalent bit per pattern, has a performance of < 70nsec nearest-match time, equivalent to a 32bit computer with 150GOPS/mm^2, at a power dissipation of 43mW. The 8.6mm^2 Manhattan test-chip with 128 reference patterns and 496 equivalent bit per pattern, has a performance of < 190nsec nearest-match time, equivalent to a 32bit computer with 20GOPS/mm^2,

at a power dissipation of 91mW. These data are sufficient for application in high-performance mobile real-time systems such as systems for image compression by vector-quantization.

References

[1] T. Nozawa et al., IEEE JSSC, vol.35, pp.1744-1751 (2000).
[2] H. J. Mattausch et al., Symp. on VLSI Cir., pp.252-255 (2002).
[3] Y. Yano, et al., Extend. Abst. of SSDM2002, pp.254-255, (2002).
[4] I. E. Opris, IEEE Trans. Cir. and Sys. II, vol.45, pp.137-141 (1998).
[5] J. Lazzaro et al., in Adv. in Neural Info. Proc. Sys., I. D. S. Touretzky Ed., San Mateo, CA: Morgan Kaufmann, (1989).
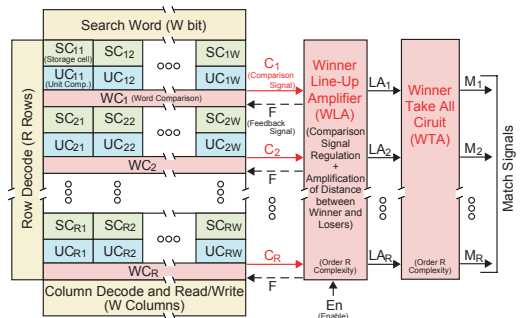
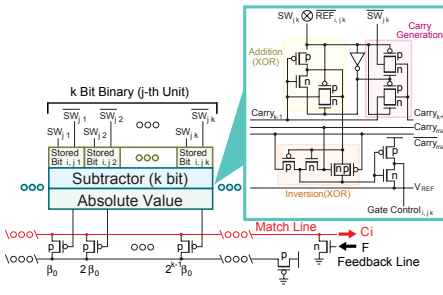Fig. 1: Associative memory architecture.

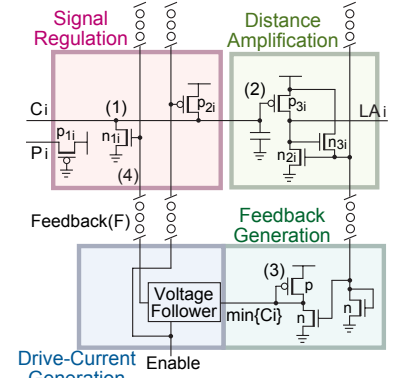
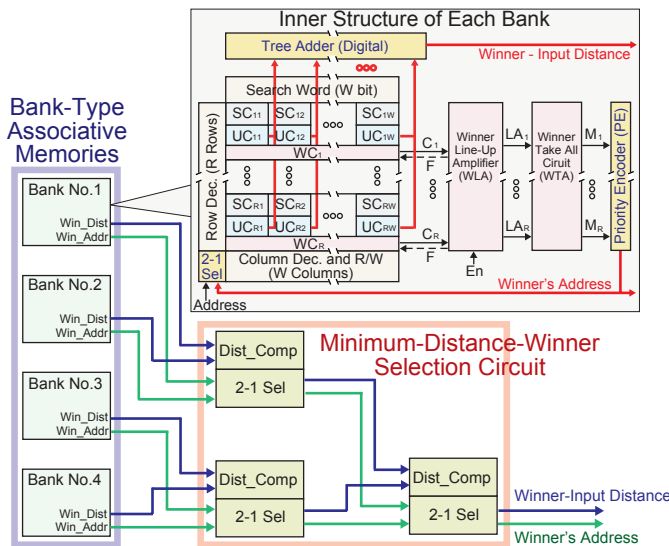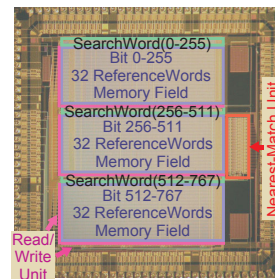Fig. 2: Memory field for minimum Manhattan distance search.
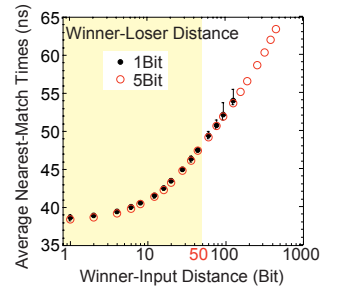

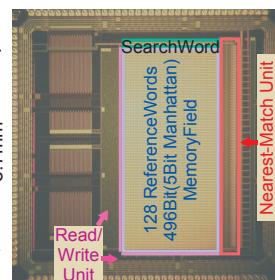Fig. 3: WLA circuit.

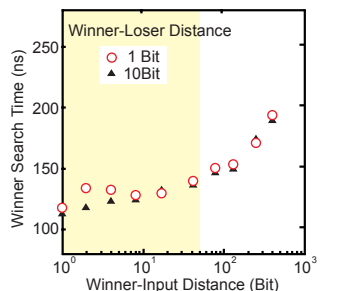Fig. 4: Bank-type associative memory architecture.

(a) chip photo   (b) measured average search times

Fig. 5: Minimun Hamming distance search associative memory.

(a) chip photo   (b) measured average search times

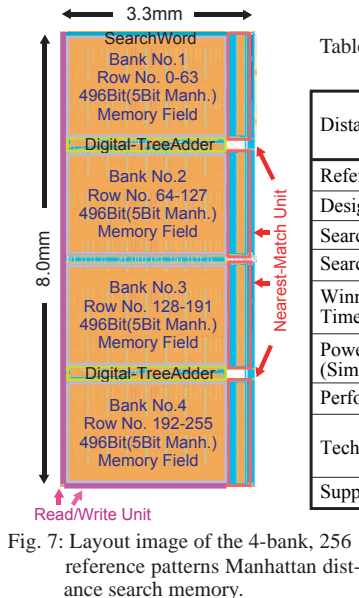Fig. 6: Minimun Manhattan distance search associative memory.

Fig. 7: Layout image of the 4-bank, 256 reference patterns Manhattan distance search memory.

Table 2: Characteristics of the bank-based associative memories.

Distance Measure	Manhattan (5 bit)	
	2-Bank	4-Bank
Reference Number	128 (64 x 2)	256 (64 x 4)
Design Area	11.8mm^2	26.5mm^2
Search Unit Area	0.99mm^2	1.97mm^2
Search Range	0 - 496bit	0 - 496bit
Winner-Search Time (Simulation)	< 260nsec	< 280nsec
Power Dissipation (Simulation)	< 330mW	< 640mW
Performance	128 GOPS	229 GOPS
Technology	0.35μm 3-metal CMOS	0.35μm 3-metal CMOS
Supply Voltage	3.3V	3.3V

Table 1: Performance data of designed associative memory test chips.

Distance Measure	Hamming	Manhattan (5 bit)
Memory Field	32 x 768	128 x 80
Technology	0.6μm CMOS	0.35μm CMOS
Area	9.11 mm^2	8.6 mm^2
Search Range	0 - 400 bit	0 - 480 bit
Winner-Search Time (Measured)	< 70 nsec	< 190 nsec
Performance	1.34 TOPS	160 GOPS
Power Dissipation	43 mW	91 mW
Supply Voltage	3.3V	3.3V