# Bank-Type Associative Memory for High-Speed Nearest Manhattan Distance Search in Large Reference-Pattern Space 

Tetsushi Koide, Yuji Yano, and Hans Jürgen Mattausch<br>Research Center for Nanodevices and Systems, Hiroshima University, 1-4-2 Kagamiyama, Higashi-Hiroshima 739-8527, Japan<br>Phone: +81-82-424-6265 Fax: +81-82-422-7185 e-mail: \{koide, hjm\}@sxsys.hiroshima-u.ac.jp

## 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 refer-ence-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 Manhat$\tan$ (gray-scale or color pictures) distance. Conventional partiallyparallel 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 asso-ciative-memory architecture for nearest Hamming/Manhattan-distance search $[2,3]$. However, the reliable handling of a large number of reference-patterns remained as an open issue. In this paper, we present a bank-type associative-memory architecture as a reliable solution to the remaining issue of nearest-match search in a large reference-pattern space. Additionally, we disclose an efficient circuit for implementing the Manhattan-distance-search function within the memory field. Designed minimum Manhattan-distance-search associative memories have high-performance at low-power dissipation. The performances of designed test chips in $0.35 \mu \mathrm{~m}$ CMOS, as compared to a 32 bit computer, ranges from 128GOPS to 230GOPS.

## 2. Associative-Memory Architecture for Manhattan-Distance Search

Figure 1 shows the block diagram of our compact associativememory architecture with fast fully-parallel match capability according to the Manhattan distances [2]. 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 kbit subtraction and absolute-value-calculation circuit is usually 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. Fig. 3 shows the circuit diagram for our k-bit subtraction and absolute-value calculation unit. The transistor number of the this optimized circuit is only $20 \mathrm{k}-2$. On the other hand, that of the conventional implementation needs 50 k - 18 transistors. So nearly $60 \%$ reduction of the transistor number can be achieved by using our newly developed circuit.
Fig. 4 shows the circuit diagram of the winner line-up amplifier (WLA) [2]. 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 $\mathrm{C}_{\text {win }}$. 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 in the signal-regulation units for each input-signal source. The transistor-count is only 6 per row. The circuitry is based on a modified version of the fast minimum circuit proposed by Opris et al [4] and 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 dis-tance-amplified WLA-output LAi. A gain of about 20-50 over a wide range of absolute $\mathrm{C}_{\text {win }}$ input voltages is thus achieved. The final Win-ner-Take-All (WTA) ${ }^{\text {win }}$ circuit implemented in the test-chips is of $\mathrm{O}(\mathrm{R})$ complexity and needs just 17 transistors per row and provides further strong amplification by voltage-current voltage transformation [2]. It generates a 1 for the winner row and 0 for all other rows.
The proposed bank-type associative memory architecture is shown
for the case of 4 banks in Fig. 5. This system has 4 local winnersearch units, implemented as banks which apply the above described fully-parallel architecture, and a global digital minimum-distancewinner selection circuit. Each local-winner is decided by fully-parallel minimum distance search in the banks in parallel. In addition to the associative memory core, each bank has a priority encoder (PE) and a circuit for digital-distance calculation of the local winner. The minimum-distance-winner selection circuit determines the global winner among the local-winners by digital calculation in a tournamentselection way and outputs the global winner's bank number as well as bank-internal address. With this bank-type approach our fully-parallel associative memory architecture can be applied to search problems in an in principle infinite space of reference patterns. For example, in the application case of code-book based vector quantization for video compression a typical code-book size is 1024 reference patterns. This could be realized with an 8 -bank structure and 128 reference patterns per bank. Local winner search in each bank and global winner selection circuit could be pipelined, so that the search throughput would be determined only by the local search time of a single bank.

## 3. Chip-fabrication and Measurement Results

A single-bank Manhattan-distance test chip was designed in $0.35 \mu \mathrm{~m}$ CMOS with 3-metal layers and contains 128 reference words. A reference word consists of 16 binaries each 5-bit long. Fig. 6 (a) shows the photomicrograph of the fabricated single-bank 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 pattern. The results 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 (distance error $<1 \%$ ) for large win-ner-input distances. However, this causes no practical problem in the video-compression application, because vector-quantization (VQ) simulations of real images confirmed, that almost all winner-input distances are less than 50bit. Table 1 summarizes the data of the fabricated single-bank test chip for minimum Manhattan distance search.
Multi-bank associative memories have been also designed in $0.35 \mu \mathrm{~m}$ CMOS technology. Fig. 7 (a) shows the layout of a 2-bank associative memory for minimum Manhattan-distance search with 128 reference pattern number. Fig. 7 (b) depicts the simulated average near-est-match times of this layout, indicating that high-speed search time can be expected. The layout of a 4-bank associative memory with 256 reference patterns is shown in Fig. 8 and the photomicrograph of the fabricated chip is shown in Fig. 9. Table 2 summarizes performance data of designed 2 and 4 bank associative memories. In addition to the scalability of the bank number N , the introduction of a bank-selective activation methodology can reduce the power dissipation of the whole chip. This is in particular useful if the referencepattern space can be categorized, so that the bank (or banks) which must contain the winner can be identified in a preprocessing step. Consequently, only 1 bank has to be activated for local winner search in the best case, reducing the power dissipation by approximately a factor $1 / \mathrm{N}$.

## 4. Conclusion

A bank-type associative memory architecture for fully-parallel minimum distance search is presented and verified by test chips in $0.35 \mu \mathrm{~m}$ CMOS technology. The proposed architecture extends the possibility of fully-parallel nearest-match search to an in principle infinite space of reference patterns. Designed 1-, 2- and 4-bank test chips in $0.35 \mu \mathrm{~m}$ CMOS verify nearest-match times below 280 nsec , a performance between 120GOPS and 230GOPS, and a power dissipation
between 90 mW and 150 mW per bank. For search problems with categorizable reference-data space the power dissipation can be reduced to the value for one bank in the best case. These data are sufficient for application in high-performance mobile real-time systems such as image-compression systems by vector-quantization.

## Acknowledgment

The test-chips have been fabricated in the chip fabrication program of VDEC, the Univ. of Tokyo with the collaboration by Rohm Co. and Toppan Printing

Co. Part of this work was supported by the 21st Century COE program and Ministry of Education, Sciences and Culture, Japanese Government and a CASIO Foundation's Research Grant.

## 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).


Fig. 8: Layout image of the 4-bank, 256 reference-pattern Manhattan-distancesearch memory.


Fig. 9: Chip photo of the 4-bank associa-tive-memory ( $0.35 \mu \mathrm{~m}$ CMOS).

Table 2: Characteristics of the bank-type Manhattan-distance associative memories.

| Distance Measure | Manhattan (5 bit) |  |
| :--- | :--- | :--- |
|  | 2-Bank | 4-Bank |
| Reference Number | $128(64 \times 2)$ | $256(64 \times 4)$ |
| Design Area | $11.8 \mathrm{~mm}^{2}$ | $26.5 \mathrm{~mm}^{2}$ |
| Search Unit Area | $0.99 \mathrm{~mm}^{2}$ | $1.97 \mathrm{~mm}^{2}$ |
| Search Range | $0-496 \mathrm{bit}$ | $0-496 \mathrm{bit}$ |
| Winner-Search <br> Time (Simulation) | $<260 \mathrm{nsec}$ | $<280 \mathrm{nsec}$ |
| Power Dissipation <br> (Simulation) | $<330 \mathrm{~mW}$ | $<640 \mathrm{~mW}$ |
| Performance | 128 GOPS | 229 GOPS |
| Technology | $0.35 \mu \mathrm{~m}$ <br> $3-\mathrm{metal}$ <br> CMOS | $0.35 \mu \mathrm{~m}$ <br> $3-\mathrm{metal}$ <br> CMOS |
| Supply Voltage | 3.3 V | 3.3 V |



Generation
Fig. 4: Circuitry of the winner-line-up amplifier (WLA) with self-adapting maxi-mum-gain region.

Table 1: Summary of characteristic data of the fabricated single-bank test chip for minimum Manhattan distance search.

| Distance Measure | Manhattan (5 bit) |
| :--- | :--- |
| Memory Field | $128 \times 80$ |
| Technology | $0.35 \mu \mathrm{~m} \mathrm{CMOS}$ |
| Area | $8.6 \mathrm{~mm}^{2}$ |
| Search Range | $0-480 \mathrm{bit}$ |
| Winner-Search <br> Time (Measured) | $<190 \mathrm{nsec}$ |
| Performance | 160 GOPS |
| Power Dissipation | 91 mW |
| Supply Voltage | 3.3 V |

Fig. 5: Bank-type associative memory architecture with fully parallel search in each bank and global winner determination by tournament search.


Fig. 7: 2-Bank minimun Manhattan-distance search associative memory. (a) chip layout (b) simulated average search times

