# Fully Parallel Associative Memory For Fast Pattern Matching

Yuji Yano, Minoru Honda, Masahiro Mizokami, Tetsushi Koide, and 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, honda, mizokami, koide, hjm}@sxsys.hiroshima-u.ac.jp

## 1. Introduction

Finding the nearest-match between an input-data word of W bit length and a number R of reference-data words is a basic operation for pattern recognition [1] as well as data compression [2]. The nearest-match or winner is defined by the minimum with respect to a distance measure. Practically important are the Hamming (data strings, voice patterns, black/white pictures) and the Manhattan (gray-scale or color pictures) distance. Previously, we demonstrated an efficient architecture for fully-parallel minimum Hamming-distance search [3]. We also proposed an encoding approach to exploit Hamming-distance-search hardware for winner-search according to the Manhattan-distance [4], which has, however, advantages only up to 4bit binaries. Conventional partially parallel Manhattan-distance-search hardware based on multiple SRAMs and external distance calculation plus winner-take-all circuity (WTA) [5] has drawbacks with respect to integration density and short nearest-match times. To overcome these drawbacks, we propose a dedicated mixed analog-digital fully parallel associative-memory architecture for nearest Manhattan-distance search with > 4bit binaries.

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

Figure 1 shows a block diagram of the compact associative memory with fast fully-parallel match capability according to the Manhattan distance. 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 circuit can be 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 structure diagram of the winner-line-up amplifier (WLA) is shown in Fig. 3. We improved the WLA circuit proposed previously in [3] so as to obtain a large regulation range for feedback stabilization and relatively low power dissipation. The new WLA achieves this larger 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 a large 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.

Fig. 4 shows a circuit, which implements the new WLA according to Fig. 3 in CMOS technology. The transistor-count is only 6 transistors per row. A modified version of the fast minimum circuit proposed by Opris et al. [6] 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 $LA_i$. The voltage follower in Fig. 4 is equivalent to the signal followers in Fig. 3 and provides a sufficient driving current for scalability to large row numbers, which are necessary especially for codebook-based data-compression systems.

The WTA-circuit implemented in the test chip is depicted in Fig. 5. It is of O(R) complexity and needs just 17 transistors per row. In order to reduce the negative effects form fabrication induced miss-match of corresponding transistors in different rows and to improve the reliability, we adopt 5 stages of the common-source WTA-configuration proposed by Lazzaro et al. [7], which amplify winner-loser distances by voltage-current-voltage transformations. The final decision circuit in Fig. 5 consists of inverters with an adjusted switching threshold. It generates a "1" for the winner row and a "0" for each loser row.

## 3. Simulated Performance and Test-Chip Design

The test chip was designed in a 0.35µm CMOS technology with 3 metal layers and contains 128 reference words with 16 binaries each 5-bit long [8]. The pattern length is estimated sufficient up to the rather high-end applications of full-color video-signal compression with 4x4 pixel blocks, after rounding of each color to a 5 bit representation. Due to a chip-area limitation, it was only possible to integrate 128 rows, which is, however, a factor 4 larger than previously reported [3,4]. Fig. 6 shows photomicrograph of the designed associative-memory. The data of the test-chip for the associative-memory is summarized in Table 1.

Figure 7 shows the simulated nearest-match times as a function of the distance between winner and input-data word. The data for winner to nearest-loser distances of 1, 2, 5 and 10 bit is plotted. The individual current-regulation (see Fig. 4) in the VDD power supply connection of each word comparator (WC) limits the maximum WC-currents.

For the application to e. g. full-color video-signal compression a codebook size of about 2048 reference patterns and an operation frequency of 1-2 MHz is necessary [5]. Considering preprocessing of rotation, flipping, and inverting of the input data, the codebook size can be reduced to 256 reference patterns. 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 130mW (at 1 MHz) for nearest a 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 47.3mW (at 1 MHz) for the associative-memory core of a full-color real-time motion-picture compression system.

## 4. Conclusion

Associative-memory architecture for fully-parallel minimum-distance search according to the Manhattan-distance is proposed and successfully verified by a chip design in 0.35µm CMOS. The 8.60mm$^2$ test-circuit with 128 reference patterns and 496 equivalent bit per pattern, has high performance, equivalent to a 32bit computer with 20 GOPS/mm$^2$, at low power dissipation of just 30.2mW/mm$^2$.

### Acknowledgment

The test-chip in this study has been fabricated in the chip fabrication program of VLSI Design and Education Center (VDEC), the University of Tokyo with the collaboration by Rohm Corporation and Toppan Printing Corporation.

### References

[1] D. R. Tveter, The Pattern Recognition Basis of Artificial Intelligence, Los Alamitos, CA: IEEE Computer Society, 1998.
[2] A. Gersho and R. M. Gray, Vector Quantization and Signal Compression, Boston, MA: Kluwer Academic, 1992.
[3] H.J. Mattausch et al., IEEE J. Solid-State Circuits, vol. 37, pp.218-227, 2002.
[4] H.J. Mattausch et al., Symposium on VLSI Circuits, pp. 252-255, 2002.
[5] T. Nozawa et al., IEEE J. Solid-State Circuits, vol. 35, pp.1744-1751, 2000.
[6] I. E. Opris, IEEE Trans. Circuits and Systems II, vol. 45, pp. 137-141, 1998.
[7] J. Lazzaro et al., in Advances in Neural Information Processing Systems, I.D.S. Touretzky Ed., San Mateo, CA: Morgan Kaufmann, 1989
[8] Y. Yano, et al., Extend. Abst. of the 2002 international Conf. on Solid State Devices and Materials (SSDM2002), pp. 254-255, 2002.

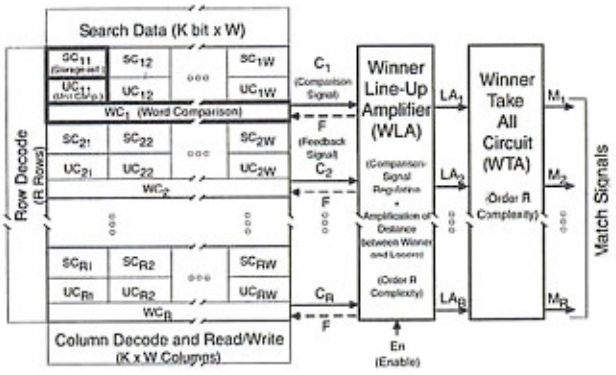



Fig. 1: Block diagram of the compact associative memory with fast fully-parallel match capability according to the Manhattan distance.

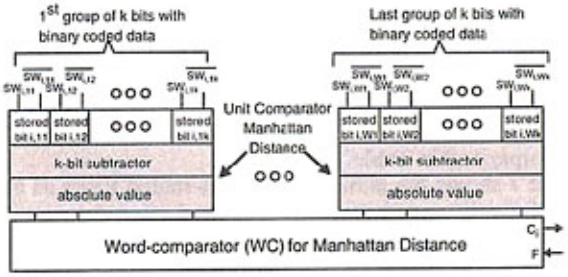



Fig. 2: Architecture of the memory-field of a mixed digital/analog associative-memory for Manhattan-distance search.

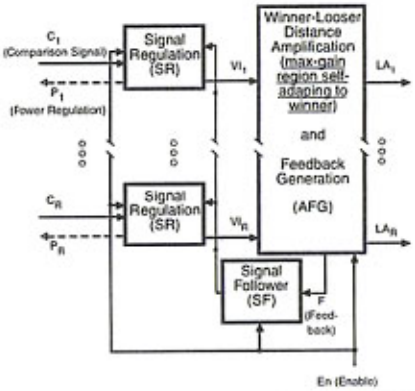



Fig. 3: Structure diagram of the winner line-up amplifier (WLA).

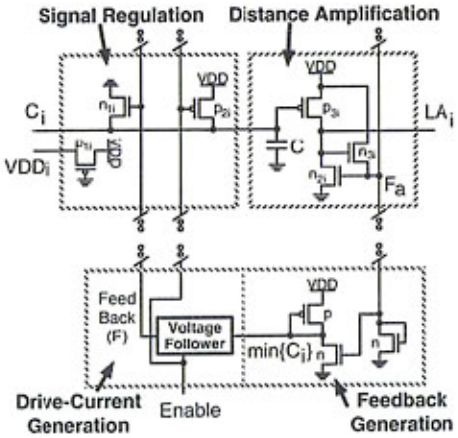



Fig. 4: Circuitry of the winner-line-up amplifier (WLA) with self-adapting maximum-gain region, following automatically the winner-row output $C_{win}$ and thus eliminating the inefficient possibilities of under- or over-regulation.

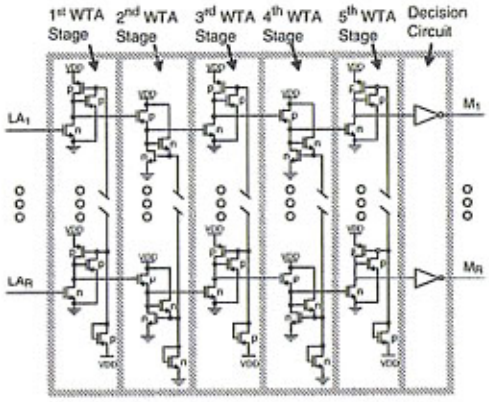



Fig. 5: Winner-take-all (WTA) circuit with 17 transistors per row of the associative memory as used for the test chip.

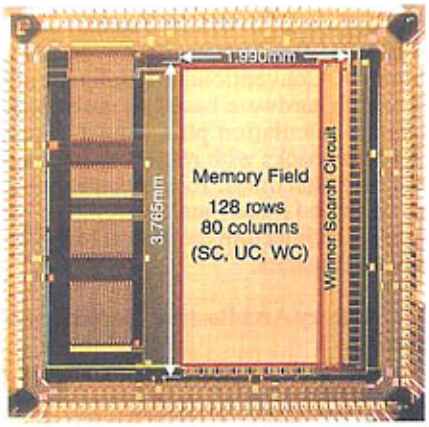



Fig. 6: Photomicrograph of the associative-memory for the Manhattan-distance, designed in 0.35µm CMOS technology.

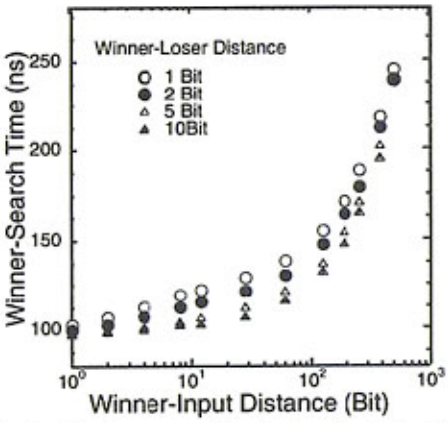



Fig. 7: Simulated performance of the designed test chip. Winner-search times as a function of winner distance to input and to nearest loser are plotted.

Table 1: Characteristics of the designed associative memory.

| | |
|---|---|
| Distance Measure | Manhattan-Distance |
| Reference-Patterns | 128 Pattern (16 binary each 5-bit long) |
| Design Area | 8.60 mm² (2.32mm*3.71mm) |
| Nearest-Match-Unit Area | 0.99 mm² = 11.5% of Design Area (4.45% for WLA, 7.05% for WTA) |
| Nearest-Match Range | 0 – 496 ($(2^5-1)$*16) Equivalent Bit Input-Winner Distance |
| Nearest-Match Times | < 240ns |
| Power Dissipation | < 260mW at 4MHz (< 2.03mW per Ref. Pattern) |
| Technology | 0.35µm, 2-Poly, 3-Metal CMOS |
| Supply Voltage | 3.3V |

# Fully Parallel Associative Memory For Fast Pattern Matching

Y. Yano, M. Honda, M. Mizokami, T. Koide, and H. J. Mattausch

Research Center for Nanodevices and Systems (RCNS)
System Design and Architecture Research Division

Hiroshima University



## Associative-Memory Functionality

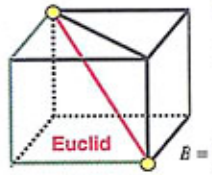
- Find the nearest match for a W-dimension data word among R-reference words.
- Nearest is defined as the minimum of a distance measure, e.g. the Manhattan distance is given by the sum over the absolute differences of the two data-word components.

$$D_{Manhattan} = \sum_{j=1}^{W} |A_j - B_j|$$

Hamming Distance : $D_{Hamming} = \sum_{j=1}^{W} (A_j \oplus B_j)$

**Example : 3-dimension**

$A = (A_x, A_y, A_z)$

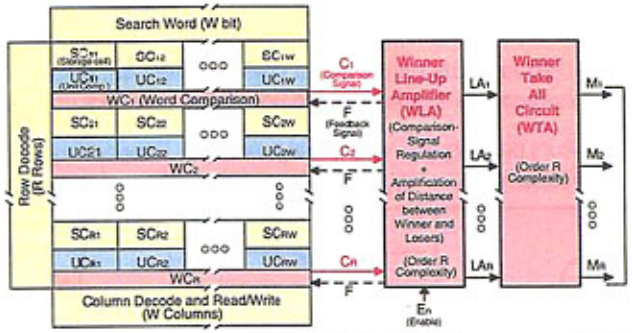



$B = (B_x, B_y, B_z)$

$$D_{Manhattan} = |A_x - B_x| + |A_y - B_y| + |A_z - B_z|$$



## Basic Architecture

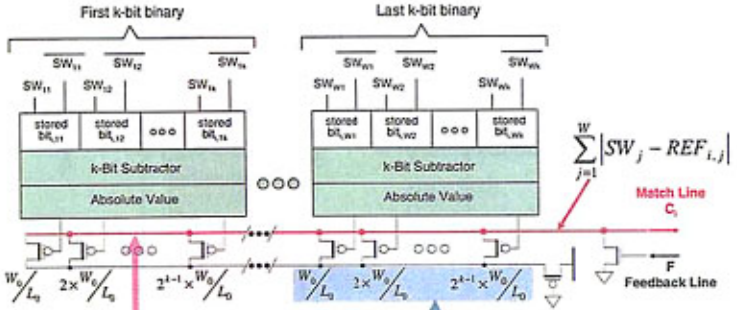



- Unit and word comparators ( UC, WC ) within the memory field.
- A fast and static analog-current-encoding of the word-comparison results.
- High reliability for most practical applications.
- Self-adapting regulation of the word-comparison signals to the point of the largest winner-loser distance amplification for all search cases.
- Winner search circuit with only O(R) complexity.



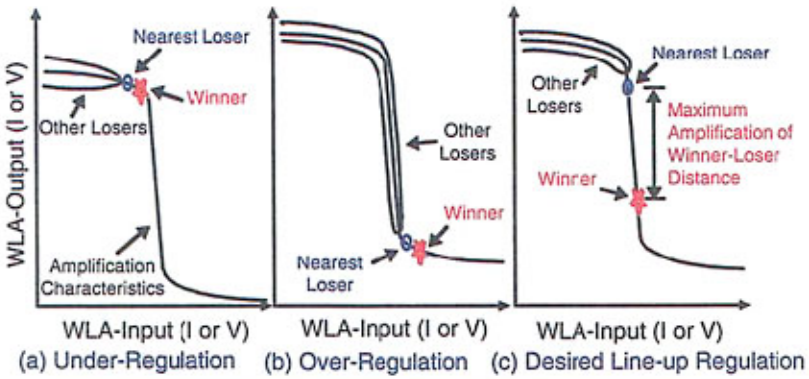
## Structure of Unit/Word Comparator

- K-bit subtractor and absolute-value calculator are needed to realize the Manhattan-distance-search.
- All unit comparators and all word comparators calculate the distance between search word and stored word in parallel.




$$\sum_{j=1}^{W} |SW_j - REF_{i,j}|$$

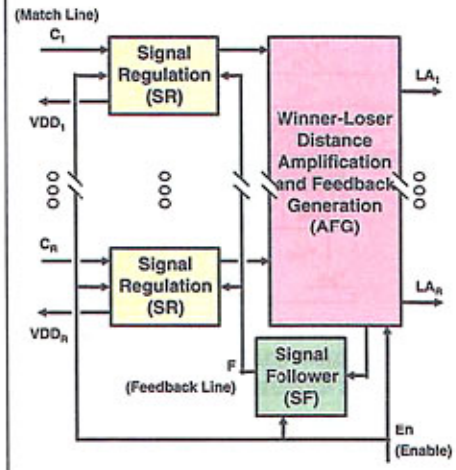
# Winner Line-Up Principle




- Perform a line-up regulation of the effective WC outputs with respect to the narrow region of maximum amplifier gain.
- Maximum winner-loser-distance amplification can be achieved for all search cases.




# Structure of Winner Line-Up Amplifier




- Signal regulation circuits (SR) convert analog input currents to analog output voltages.
- Amplification and feedback generation circuit (AFG) amplifies the difference of analog voltage in each row, and generates feedback signal.
- Signal follower circuit (SF) generates drive-currents for signal regulation circuits.



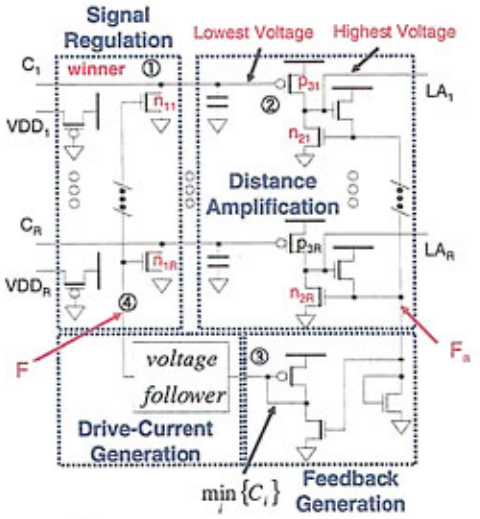

# Circuit of Winner Line-Up Amplifier

① The current on C. is the smallest and the voltage of $C_1$ is the lowest.

② The current-source capability of $p_{31}$ is the largest and the voltage of $LA_1$ is the highest.

③ The voltage of F is the same as that of $C_1$, because $F_a$ controlled by the winner row.

④ For larger winner-input distance the voltage F becomes higher and $n_{li}$ have larger current-sink capability.

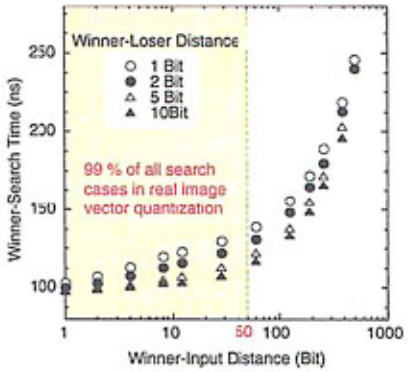






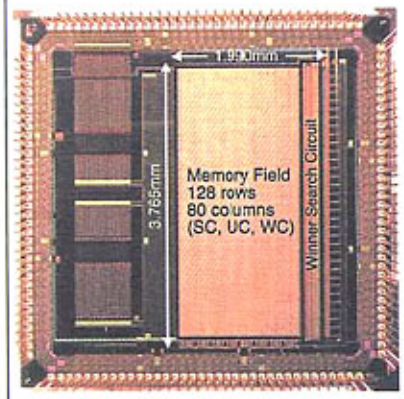
# Simulated Nearest-Match Time

| | |
|---|---|
| Distance Measure | Manhattan Distance, 16 5-bit Binaries |
| Reference Pattern | 128 Pattern of 496 Equivalent Bit |
| Technology | 0.35 μm CMOS, 3-metal |
| Supply voltage | 3.3 Volt |




- For larger winner-input distance, the number of active p-MOSFETs on the match line is increasing.
- Largest search time is about 240 nsec, and shortest search time is about 70 nsec.
- In practice, winner patterns with large winner-input distance are seldom ( < 150nsec ).

## Design-Data of Test Chip



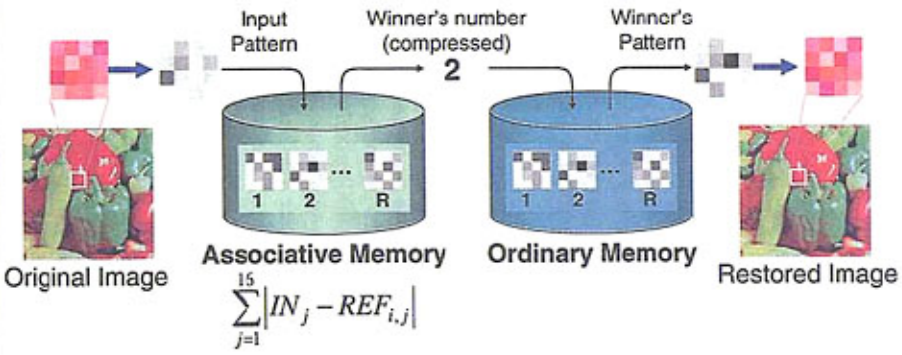

| | |
|---|---|
| Distance Measure | Manhattan Distance, 5 bit binaries |
| Reference Pattern | 128 Pattern of 496 Bit (total 62 Kbit) |
| Design Area | 7.49 mm² (3.765 mm x 1.990 mm) |
| Nearest-Match-Unit Area | 1.02 mm² = 13.6% |
| Nearest-Match Range | 0 - 496 Equivalent Bit Input-Winner Distance |
| Nearest-Match Time | < 240 ns |
| Power Dissipation | < 260 mW at 4MHz ( 30.2mW / mm² ) |
| Performance † | 20 GOPS / mm² |
| Technology | 0.35 μm, 3-metal CMOS |
| Supply Voltage | 3.3 Volt |

† The number of operations when a 32-bit computer runs the same workload as our chip



## Application Example of VQ

VQ : Vector input data is mapped to the identification number (scalar data) of the nearest match pattern.




$$\sum_{j=1}^{16} \left| IN_j - REF_{i,j} \right|$$



## Conclusion

- Associative-memory architecture for fully-parallel Manhattan-distance-search ( MDS ) is proposed.
- Test chip of MDS memory with 128 reference patterns and 496 equivalent bit per pattern has a high performance of 20 GOPS/mm² at the low power dissipation of 30.2 mW/mm².
- Suitable in handheld devices including motion picture communication.