A Nearest-Hamming-Distance Search Memory With Fully Parallel Mixed Digital-Analog Match Circuitry

Tetsushi Koide, Hans Jürgen Mattausch, Yuji Yano, Takayuki Gyohten, and Yoshihiro Soda

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: koide@sxsys.hiroshima-u.ac.jp, http://www.rcns.hiroshima-u.ac.jp/

Abstract- A fully-parallel minimum Hamming-distance search memory has been designed, which uses digital circuitry for bit-comparison and fast analog circuitry for word-comparison as well as winner-take-all (WTA) functionality. The approach allows compact and high-performance integration in conventional CMOS technology. The 0.6μm, 2-poly, 3-metal CMOS design with 32 rows and 128 columns achieves <100ns search times at <260 mW power dissipation. The approach is extendable to increased pattern length and larger row numbers and enables high efficiency for pattern-matching applications such as network routers, code-book-based data compression, or object recognition. As for conventional memories high yield can be achieved with a redundancy concept for rows (including WTA function) and columns.

I. Introduction

The associative-memory functionality is to find the nearest-match between an input-data word of W bit length and a number R of reference-data words. This functionality is a basic operation for such applications as image or voice pattern recognition [1], codebook-base data compression [2,8], routing-table-lookup for network routers, and authentication parts in security system. The nearest-match or winner is defined by the minimum with respect to a distance measure. Practically important distance measures are the Hamming (data strings, voice patterns, black/white pictures) and the Manhattan (gray-scale or color pictures) distance. Previous methods for winner search have been based on: (a) analog neural networks [3], (b) SRAMs and a separate digital winner-take-all (WTA) circuit [4], (c) an analog WTA based on MOSFETs in source follower configuration [5] or a time-domain concept [6]. Problems of these solutions are: Large area-consumption [3, 4, 6] because the search circuits are of order R^2 ($O(R^2)$) or $O(R*W)$ complexity. Long search-times of about 1μs or more [3, 4]. Restricted applicability to small W [5].

II. Associative Memory Architecture

The five main concepts and guidelines for the associative-memory architecture [7] are: (a) Bit- and word-comparison within the memory field for a fully-parallel nearest-match search as well as small size of the comparison circuits. (b) A fast and static analog encoding of the word-comparison results. One possibility, used in the test chip, is a static current-sink capability. (c) Large margins for "good" matches, defined by small winner-input or large winner-loser distances. Smaller margins for "bad" matches, defined by large winner-input plus small winner-loser distances. In this way high reliability for "good" matches is ensured, which is important for most practical applications. (d) The key concept for the performance is a self-adapting regulation of the word-comparison signals to the point of largest winner-loser-distance amplification for all search cases. We call this the winner-line-up principle. (e) Last but not least the complexity of the winner-search circuit should be only proportional to the number of reference words R. This means that the winner-search circuit should have only O(R) complexity.

The structure diagram of the associative-memory architecture for Hamming-distance search is shown Fig.1. The memory part on the left side consists of conventional read/write periphery for storing the reference-data words and for reading out the nearest-match data. The search word is supplied from above, preferably on the bit-lines of the memory field. Each row of the memory field contains storage units (SC) for W bits plus the circuitry for bit (BC) - and word (WC) - comparison. The WC results C_i are transferred to the winner-search circuit on the right side consisting of the winner-line-up amplifier (WLA) and a winner-take-all circuit (WTA). Important is the closely coupled interaction of WLA and WCs, by which the desired maximum amplification of winner-loser distances for all search cases is achieved. The output signals LA_i of the WLA are finally evaluated by the WTA to decide on the row, which contains the winner data.

Figure 2 explains the important interaction between WLA and WCs schematically. Purpose is a lineup regulation of the effective WC-outputs with respect to the narrow region of maximum distance-amplifier gain as explained in part 2(c) and to avoid the inefficient possibilities of under-regulation shown in part 2(a) or over-regulation shown in part 2(b). This results in maximum winner-loser-distance amplification for all search configurations. Without the WLA-regulation such a performance would be impossible.

The basic structure of the WLA consists consequently of signal-regulation (SR) units for each row and a common distance-amplification/feedback-generation (AFG) unit as shown in Fig. 3a. The feedback (F) controls the SR-units in such a way, that the generated, intermediate signal VI_{WIN} of the winner row is just within the narrow maximum-gain region of the distance amplifier in the AFG-unit. The simple WLA-circuit implementation for the test chip is shown in Fig. 3b. It uses only 7 transistors plus a compensation-capacitor per row. The SR-unit consists of two n-MOSFETs. A pull-up n_{1i} in source-follower configuration, with its gate connected to the feedback, transforms the WC-current-sink capability into the intermediate voltage VI_i. A pass-transistor n_{2i} serves to enable or disable the WLA-regulation and to limit the current. The AFG-unit uses 3 n-MOSFETs, 2 p-MOSFETs and a compensation capacitor in each row. Amplification of the winner-loser distance is performed with push-pull amplifiers. Two additional n-MOSFETs n_{3i}, n_{4i} generate 0V input for the push-pull amplifiers and disconnect the SR-units and the AFG-unit when the WLA-regulation is disabled. The compensation capacitor provides sufficient phase-margin for stable WLA operation. The feedback voltage F is generated by pull-down p-MOSFETs p_{1i} in source-follower configuration for each row and by a common pull-up p-MOSFET p_2 for all rows.

Since the winner row has the smallest current-sink capability, it has also the highest intermediate voltage VI_{WIN} and the lowest push-pull amplifier output LA_{WIN}. Thus, the winner row determines the feedback voltage $F \approx LA_{WIN}+V_{th,p}$ and therefore also the current-source capability of all SR-units. A stable line-up state is reached, when the current-source capability of the SR-units balances the current-sink capability of the winner row. We end-up with a situation corresponding to Fig. 2(c). The winner-row output LA_{WIN} will be lowest and the winner-loser distance will be amplified with the maximum push-pull-amplifier gain. The WLA-circuit performs this self-adapting regulation for all possible search cases within its regulation range. An important design task for the WLA is therefore a large regulation range, even with worst-case transistor parameters.

The WTA-circuit implemented in the test chip is depicted in Fig. 4. It is of O(R) complexity and needs just 10 transistors per row. At the core are 2 stages of the common-source WTA-configuration proposed by Lazzaro et al. [9].

III. Chip-fabrication and Measurement Results

The photomicrograph in Fig. 5 shows the fabricated associative memory, which measures just 1.57mm*1mm, even in the rather conventional 0.6μm CMOS technology[1]. 1 bit of the memory field has an area of 226 μm^2, 45% of which are used for BC and WC. The nearest-match unit on right side consumes only 14.3% of the area of the complete associative memory. 5.8% and 8.5% of this area are used for the WLA- and the WTA-circuit, respectively. The WLA was designed for a regulation-range up to 32 bit winner-input distance and large gate length/width was adopted for analog circuits to reduce the effect of process induced variation. Design and fabrication data of the CMOS test chip are summarized in Table 1.

The worst-case performance was determined from the worst-case combination for the physical locations of winner and nearest-loser row within the chip. In Fig. 6(a) the measured worst-case characterization is depicted for one of the functional chips at nominal supply voltage

1 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.

VDD = 3.3V. Region 1 comprises all search configurations, for which the winner is successfully determined. Region 2 contains all search configurations, where both winner- and nearest-loser row are identified as winners but are successfully separated from the other loser rows. This region may be also useful in practical applications, because it corresponds to "bad" matches where the difference between winner and nearest loser is small. Region 3 corresponds to all other unsuccessful search configurations. In Fig. 6(b) the measured average winner-search times are plotted as a function of winner-input distance for 1bit and 5bit distance between the winner and the nearest loser. The average is taken by choosing randomly 10 different combinations of the physical location of the rows, which hold the winner- and nearest-loser words. In the most critical case of 1bit winner-to-nearest-loser distance, the scattering of the winner-search times is plotted in addition. These measured winner-search times are <100ns. The current limitation by the SR-unit leads to an unproblematic power consumption of 200mW at 10 MHz in the test chip. Table 2 shows the simulated results of each winner-search unit for min. and max. power dissipation cases, respectively. Controlling the size of SR-unit transistors, the static current of WC- and WLA-unit and the maximum power dissipation can be optimized. Moreover, scaling to large number and length of stored patterns is facilitated by the low operation frequency of 1-2 MHz sufficient for real applications. Consequently, the possibility of limiting the static current level of WC- and WLA-unit to 10-100µA per row is expected, which solves the static-power-dissipation issue of these current-mode circuits even at large associative-memory size.

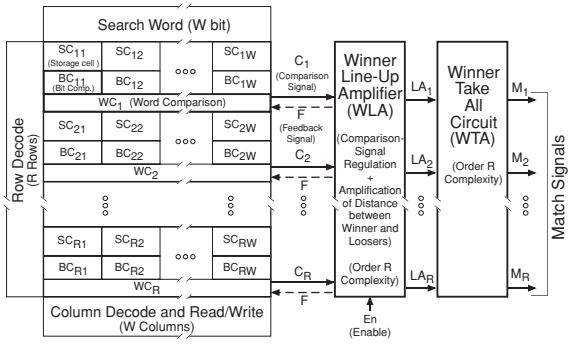

Fig. 1: Block diagram of the proposed architecture for compact associative memories with fast fully-parallel search capability for the smallest Hamming distance.

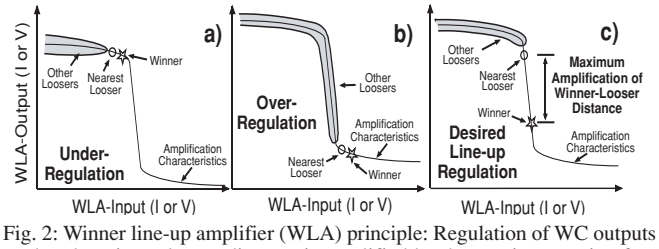

Fig. 2: Winner line-up amplifier (WLA) principle: Regulation of WC outputs so that the winner-looser distance is amplified by the maximum gain of an amplifier for all possible search cases.

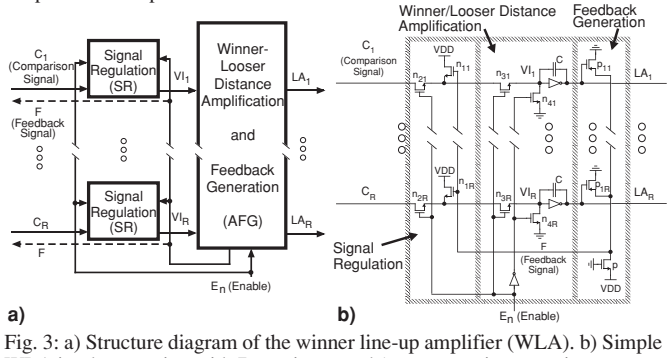

Fig. 3: a) Structure diagram of the winner line-up amplifier (WLA). b) Simple WLA implementation with 7 transistors and 1 compensation capacitor per row of the associative memory. The feedback to the WCs is avoided to achieve a minimum-sized memory field.

IV. Conclusion

Associative-memory architecture for fully-parallel minimum Hamming distance search is proposed and successfully verified by a chip design in 0.6µm CMOS. The 1.57mm^2 test-circuit with 32 reference patterns of 128-bit length, has high performance, equivalent to a 32bit digital computer with 162 GOPS, and at the same time low power dissipation.

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. P. Graf et al., "Analog electronic neural network circuits," IEEE Circuits and Device Mag., vol. 5, pp.44-55, 1989.

[4] A. Nakada et al., "A fully parallel vector-quantization processor for real-time motion-picture compression," IEEE Journ. Solid-State Circuits, vol. 34, pp. 822-830, 1999.

[5] S. M. S. Jalaleddine et al., "Associative IC memories with relational search and nearest-match capabilities," IEEE Journ. Solid-State Circuits, vol. 27, pp. 892-900, 1992.

[6] M. Ikeda et al., "Time-domain minimum-distance detector and its application to low-power coding scheme on chip-interface, Proc. of ESSCIRC'97, pp. 464-467, 1998.

[7] H.J. Mattausch et al., "Compact associative-memory architecture with fully parallel search capability for the minimum Hamming distance," IEEE J. Solid-State Circuits, vol. 37, pp.218-227, 2002.

[8] H.J. Mattausch et al., "Fully-parallel pattern-matching engine with dynamic adaptability to Hamming or Manhattan distance," Symp. on VLSI Circuits, pp.252-255, 2002.

[9] J. Lazzaro et al., "Winner-take-all networks of O(N) complexity," in Advances in Neural Information Processing Systems, I.D.S. Touretzky Ed., San Mateo, CA: Morgan Kaufmann, 1989.

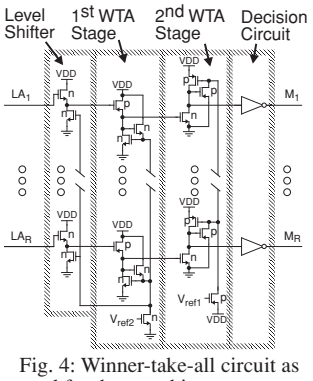

Fig. 4: Winner-take-all circuit as used for the test chip.

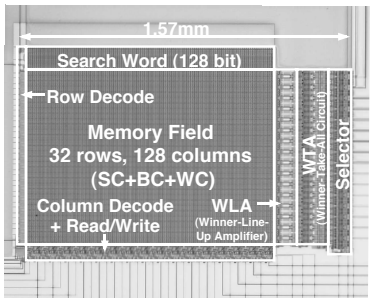

Fig. 5: Fabricated associative-memory test chip in 0.6µm, 2-poly, 3-metal CMOS.

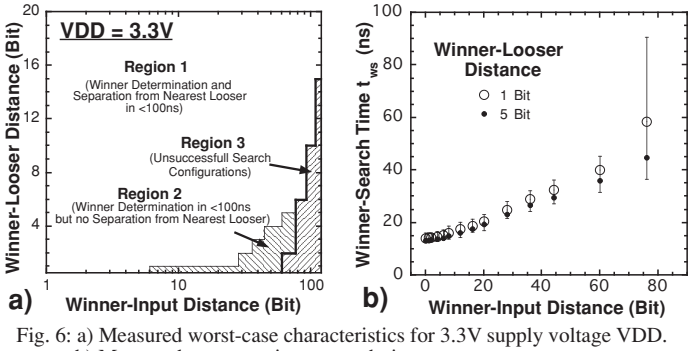

Fig. 6: a) Measured worst-case characteristics for 3.3V supply voltage VDD. b) Measured average winner-search times.

Table 1: Characteristics of the designed associative memory.

Technology	0.6µm, 2-Poly, 3-Metal CMOS
Dimensions	1.57mm*1.00mm=1.57mm^2
Organization	32rows*128columns=4Kbit
Memory-Field Unit (1Bit) (SC+BC+WC)	226 mm^2 (55% for SC, 45% for BC+WC)
Nearest-Match Search Unit	
Distance Measure	Hamming Distance
Area	0.224 mm^2=14.3% of Associative Memory (5.8% for WLA, 8.5% for WTA)
Designed Search Range	0 - 32 Bit Winner-Input Distance
Supply Voltage	3.3V
Power Dissipation	< 200±60mW (Measured)
Fabricated Chips	22 (19 Functional)

Table 2: Simulated power dissipation of WC-, WLA-, and WTA-unit.

(a) winner-input-distance=0, loser-input-distance=1 (min. case)

Circuit	Power dissipation (Simulation)	Ratio
Total	32.00mW	100%
WC	0.93mW	2.9%
WLA	17.54mW	54.8%
WTA	13.53mW	42.3%

(b) winner-input-distance=127, loser-input-distance=128 (max. case)

Circuit	Power dissipation (Simulation)	Ratio
Total	363.00mW	100%
WC	157.34mW	43.3%
WLA	21.23mW	5.8%
WTA	184.43mW	50.9%