D-1-3

Optimizing Associative Processor Architecture for Intelligent Internet Search Applications

Huaiyu Xu, Yoshio Mita and Tadashi shibata

Department of Electronic engineering, The University of Tokyo
*Department of Frontier Informatics, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656 Japan
Phone: +81–3–5481–8567 Fax: +81–3–5841–8567 e–mail: jo@if.t.u–tokyo.ac.jp, {mita, shibata}@ee.t.u-tokyo.ac.jp

1. Introduction

World-Wide-Web (WWW) based information searching plays an important role in our daily life [1]. However, what is most intolerant is that a search query often results in " No matches found", although the database contains very useful information that is in need. This is because most of the search engines employ a very simple word-matching technique. It is therefore essential to develop an intelligent search engine that is capable of finding the information in need more flexibly taking the individual's preference, specific interest etc. into account. The similarity-measure based search [2,3] can be a good candidate for this. However, due to the very time-consuming computation required for similarity-measure evaluation, this is not in practical use at present in Internet search applications.

The purpose of this paper is to develop an optimized hardware organization for intelligent Internet search engines employing the associative processor (AP) architecture [4]. Due to the parallel processing on the chip, more than 10^4 times faster search has been demonstrated for 40,000 items. Optimization has been carried out by implementing typical architectures either in a VLSI chip or in FPGA's. An interactive search engine for an E-commerce real-estate agent system has been developed and demonstrated on the FPGA-based implementation.

2. Associative Processor (AP) Architectures

The AP stores the information by means of template vectors, calculates the similarity (= Manhattan distance: MD) between an input vector (query) and template vectors, and returns the address of the maximal likelihood vector (Top 1). The second, third...etc. most similar vectors are also retrieved as Top 2,3...M. To be applied in every intelligent Internet search engines, in which the dimension of the vector is unpredictable, we first designed a very flexible VLSI AP. It allows to use an arbitrary number of elements in a vector and various template grouping structures. In spite of such versatile features, however, the chip can handle only 64 vectors (128 dimension) using external memories for templates (Fig.1).

Therefore more area efficient organization that allows a larger number of template integration on a chip has been studied. Three different organizations are compared in Fig.2 for each Manhattan-distance (MD) calculation module.

Type I represents the organization similar to that in Ref. [4] where the template memory (TM), MD module, and winner take all (WTA) are provided for each template vector. In order to weight each element according to personal preference in Internet search, 7-bit multiplier was added in the MD calculation module, thus substantially increasing the area of an MD module. The area ratio estimated based on the chip implementation [4] and gate counts obtained from FPGA implementation is shown in the figure.

In order to increase the number of template vectors on a

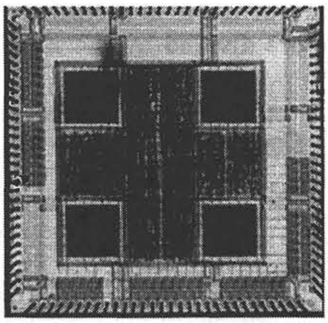

Fig.1 Photomicrograph of the AP test chip featuring variable vector dimension and grouping structure.
(0.6-μm triple-metal CMOS)

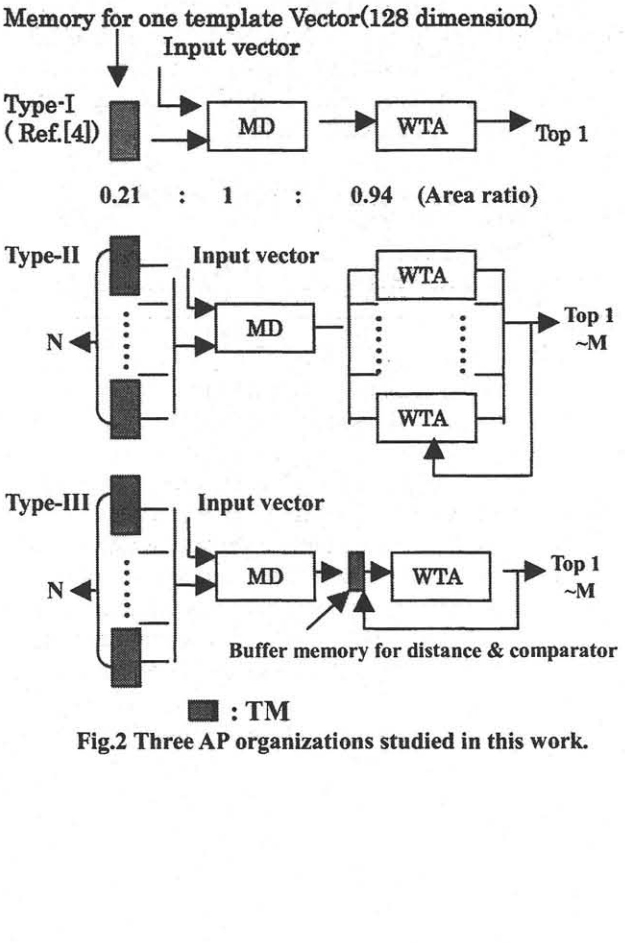
Memory for one template Vector(128 dimension)
Input vector
Type-I
(Ref.[4])
MD
WTA
Top 1
0.21 : 1 : 0.94 (Area ratio)
Type-II
Input vector
WTA
N
MD
Top 1
~M
WTA
Type-III
Input vector
N
MD
WTA
Top 1
~M
Buffer memory for distance & comparator
■ : TM

Fig.2 Three AP organizations studied in this work.

chip, the Type II and Type III were examined in this work. A single MD module is provided for N TM's and N WTA's in Type II, while one MD module and one WTA for N TM's in Type III. In Type III, a buffer memory storing distance values and a comparator are added for sequential winner search, while with a slight area penalty.

Five pieces of architecture are considered. Architecture A has the configuration of Type I. The Type II configurations with N=10 and N=100 give architectures B and C, respectively. The Type III configurations with N=10 and N=100 represent architecture D and E, respectively.

The number of storable template vectors, which is the most important characteristic value, is roughly estimated, assuming the die size of 1cm×1cm and 0.18-μm design rules. The area data obtained in designing the chip of Ref.[4] were simply scaled down from 0.6μm to 0.18μm. The performance comparison among various architectures is summarized in Table I, where figure of merit is defined as the number of template vectors divided by the number of clock cycles.

Table I Comparison of various AP architectures

	A	B	C	D	E
# of Temp. vectors /chip	490	842	908	2519	4640
# of clock cycle	1916	2060	3500	3060	13500
F. of merits	0.26	0.40	0.26	0.8	0.34

3. Example of Intelligent Internet Search Application

Experiments of intelligent search were carried out using FPGA prototype. Fig.3 shows a screenshot of an E-commerce house-search application. The searching agent receives query via WWW server, and returns the house best matching to the customer's need in real time. The operation was demonstrated using a FPGA with the architecture B having 100 template vectors (Fig.4). The same searching function is implemented using only C++ language and the performances are compared in Table II.

Table II Comparison of computation time in seconds.

#of Temps. Type	100	2519	40000
Software	1.6×10^{-4}	4.5×10^{-2}	4.2
B-type AP	6.18×10^{-5}	6.18×10^{-5}	6.18×10^{-5}
D-type AP	9.18×10^{-5}	9.18×10^{-5}	9.18×10^{-5}

Top 100 vectors were searched
Software: Running on 600MHz, IBM PC.
B-type AP, D-type AP: Running at 30MHz.

With 2519 template vectors, which can be stored in one chip with D architecture AP, the AP is 490 times faster than the software. With 40,000 template vectors, which would be necessary for search in a sufficient number of houses in a city, the search time by software becomes intolerant. On the other hand, the D-type AP returns the result without any appreciable delay. The AP is 45000 times faster than the software at a clock frequency of only 30MHz. (However, we need 16 D-type AP chips on a board.) Therefore similarity based intelligent search becomes feasible by employing the AP chips.

4. Conclusion

Associative processor architecture has been optimized for intelligent Internet search applications. The search engine for E-commerce application has been developed and successfully implemented in the AP architecture using the FPGA prototype. More than 10^4 faster search would be possible for template vectors over 40,000 at a clock frequency of 30MHz, allowing high-performance and low power implementation.

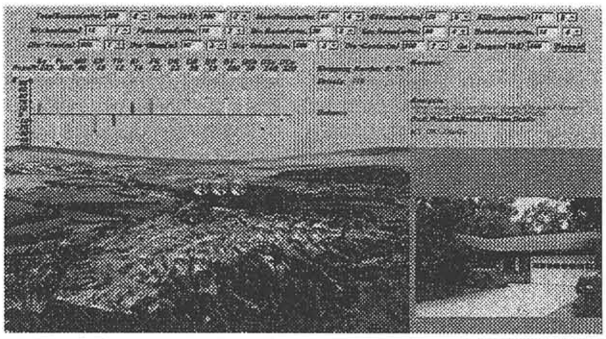

Fig.3 Demonstration of typical intelligent Internet search engine

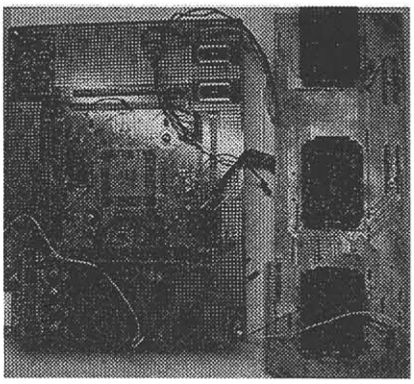

Fig.4 Prototype Board

Acknowledgements

The VLSI 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.

References:

[1] "An agent marketplace for buying and selling goods," In Proceedings of the First International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology. London, U.K

[2] W. Wilke, " CBR and electronic commerce on the WWW," Invited talk on the International conference on Case-Based Reasoning (ICCBR-97), Providence, Rhode Island (1997).

[3] A.Aamodt and E. Plaza, "Case-based reasoning: foundationl issues, methodological variations, and system approaches," AI communications, 7(1) 39-59 (1994)

[4] A. Nakata, T. Shibata, M.Konda, T. Morimoto, and T. Ohmi, "A fully-parallel vector quantization processor for real-time motion picture compression," IEEE Journal of Solid-State Circuits, Vol.34, No. 6, pp. 822-830 (1999).