[See page 423 for Figure 4.]

SESSION XX: SPECIAL PURPOSE ACCELERATORS

FAM 20.2: A Character String Search Processor

Hachiro Yamada, Masaki Hirata, Hajime Nagai, Kousuke Takahashi

NEC Corporation

Kawasaki, Japan

SPECIAL DEDICATED CHIPS for searching specific strings in the input text have been developed at various organizations[1]. Although the main goal of these activities has been to improve the searching speed, some sophisticated search operations, such as approximate comparison for variable-length strings in the non-anchor mode, still require a relatively long time[2].

This paper will describe a character string search processor (SSP) for text retrieval, which uses an architecture to compare 80M strings/sec, several hundred times higher speeds than reported previously[3].

The SSP consists of a 8Kb content-addressable memory (CAM), 20K-gate finite-state automation (FSA) logic, a priority encoder and other control logic (Figure 1). The input text stream in the form of character codes enters the CAM through data decoders, and then is concurrently compared with the contents of the stored characters in the CAM. String comparison results, produced by the FSA logic, are transformed into coincident pattern string address codes through the priority encoder.

The total CAM capacity, 528 characters (16 bit codes) includes 512 characters for 64 pattern strings, 8 mask characters for the fixed-length don't care (FLDC) match mode, and 8 delimiter characters for the anchor match mode. Parallel operation of this SSP chip would further expand the word capacity and character length.

The CAM consists of 8 x 528 PCAM (pair-bit-CAM) cells and 8 two-bit data decoders. As shown in Figure 2, each PCAM cell is made up of four conventional static RAM cells and a read-write circuit. The two-bit data decoder energizes one of four data lines to select one SRAM cell from each PCAM cell. A write operation to a PCAM cell first clears all the SRAM cells and then writes *1* in the cell selected by a 2b write data. When the PCAM is accessed later by the same 2b data as used for writing, it will turn off a transistor and maintain the response line at the high level which shows the data is matched. Access by a different data will discharge the response line and show a mismatch. Use of eight PCAM cells connected together at the drains of each transistor to form a wired-AND logic can detect the match of a 16b coded character.

The CAM is suitable for implementation by VLSI technology, because the logic portion, such as the data decoder, can be partitioned from the regularized storage portion. The minimized stray capacity on the bit line enables the PCAM cell to operate at a high speed. A single word PCAM can store two characters having 1b differences from each other, although the cell needs a few more transistors than a conventional CAM cell[4]. This principle is useful for storing a capital and a small letter in the alphabet in the same CAM word.

The FSA logic is designed to make it

The FSA logic is designed to make it possible to compare the stored strings with those input text data having errors, such as omission, insertion or substitution of a single character. It contains 512 logic cells (Figure 3) to emulate the finite state transition diagram. Flip-flops FFe and FFa hold exact and approximate match pointers, respectively, each of which indicates the comparison result for the stored pattern string in the CAM. Delimiter flip-flop FFd stores the endmark for the variable-length pattern string, if necessary.

A character-match signal transfers the pointers in FFe and FFa to the next logic cell through gates G0, G4 and G5. A character mismatch signal transfers the pointer in FFe to both FFa's in the same and next logic cells. The FFe (or FFa) for a particular cell, in which the FFd has the end mark, outputs the exact (or approximate) match signal to the priority encoder.

The pointer transfer operations, described, are carried out for every character input in the non-anchor mode. In the anchor mode, the logic cells remain inactive until any delimiter characters are detected.

In the multiple variable-length don't care (VLDC) mode, the SSP can search the string, which consists of VLDC character strings and specified sub-strings, registered in the CAM.

The geometric design of the CAM portion was full-custom to minimize the cell size and delay times, while the logic portion, including 22K gates, was designed using standard cells. The SSP chip has 217,600 transistors in an 8.62 x 12.76mm die area. The technology used was a double-metal 1.6μm N-well CMOS process. Figure 4 shows a photomicrograph of the chip. Table 1 summarizes the specifications.

Typical waveforms in case of matching are shown in Figure 5. Response time is less than 50ns. The character input rate of greater than 10M characters/sec is observed.

The string search processor (SSP) is able to receive the document data of 10M characters/sec directly from communication networks or disks. Because the SSP chip can handle 64 strings in parallel, it facilitates a high-speed text retrieval system that can compare strings at an 80M strings/sec rate.

Acknowledgments

The authors would like to thank N. Yoshida, K. Yoshimi and H. Sakuma for direction and encouragement. They also wish to thank Y. Kitamura and K. Matsumoto for their technical suggestions.

[1] **Yianilos, N.P., "Dedicated Comparator Matches Symbol Strings Fast and Intelligently",** *Electronics*, **p. 113-117; Dec., 1983.**

[2] **Faloutsos, C., "Access Methods for Text",** *Computing Surveys*, **Vol. 17, No. 1, p. 49-74; Mar., 1985.**

[3] **Takahashi, K., et. al., "A New String Search Hardware Architecture for VLSI",** *Proc. of 13th ISCA*, **p. 20-27; 1986.**

[4] **Kadota, H. et. al., "An 8Kb Content-Addressable and Reentrant Memory",** *ISSCC DIGEST OF TECHNICAL PAPERS* **p. 42-43; Feb., 1985.**

0193-6530/87/0000-0272$01.00 © 1987 IEEE

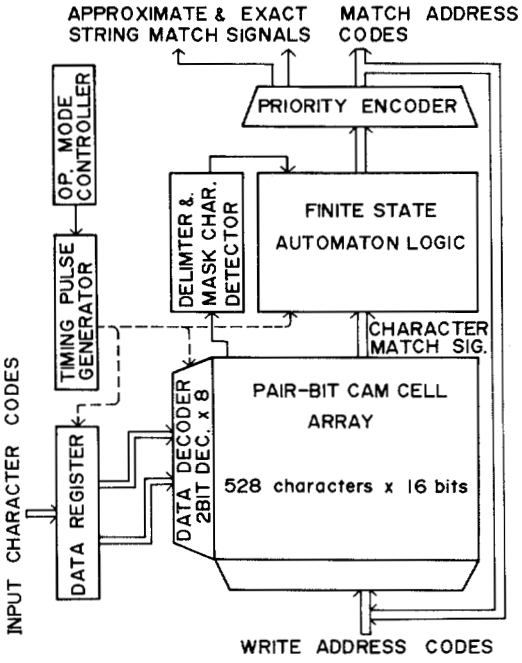

FIGURE 1—String search processor (SSP) block diagram.

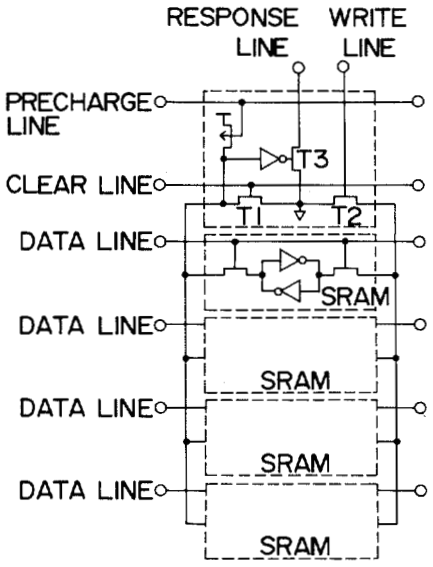

FIGURE 2—Circuit diagram for the Pair-bit Content Addressable Memory (PCAM) cell.

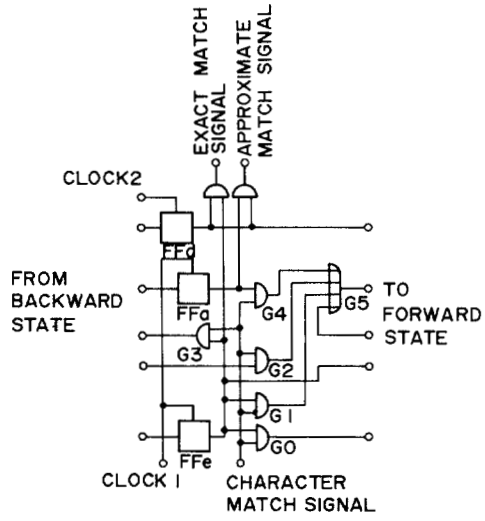

FIGURE 3—Finite State Automation (FSA) logic cell circuit.

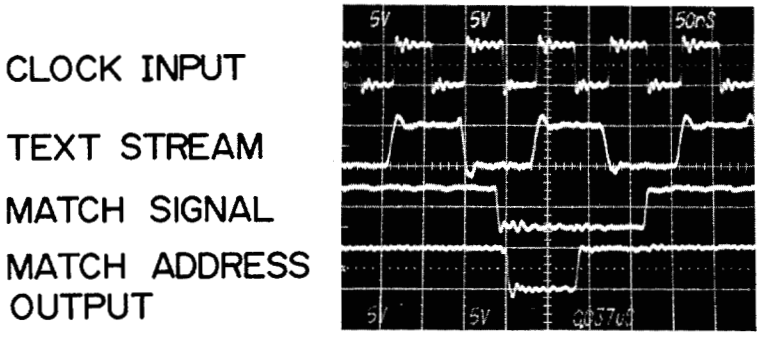

FIGURE 5—Operating waveforms in the search operation.

Operation mode	Exact/Approximate match
	Anchor/Non-anchor
	FLDC/VLDC
Pattern string	Max. 64 strings
	Max. 512 character length
Character input rate	10M characters/sec
Power dissipation	500mW (10MHz)
Number of devices	217,600
Die size	8.62 x 12.76 mm
Process technogy	2 metal 1.6 μm CMOS

TABLE 1—SSP specifications.

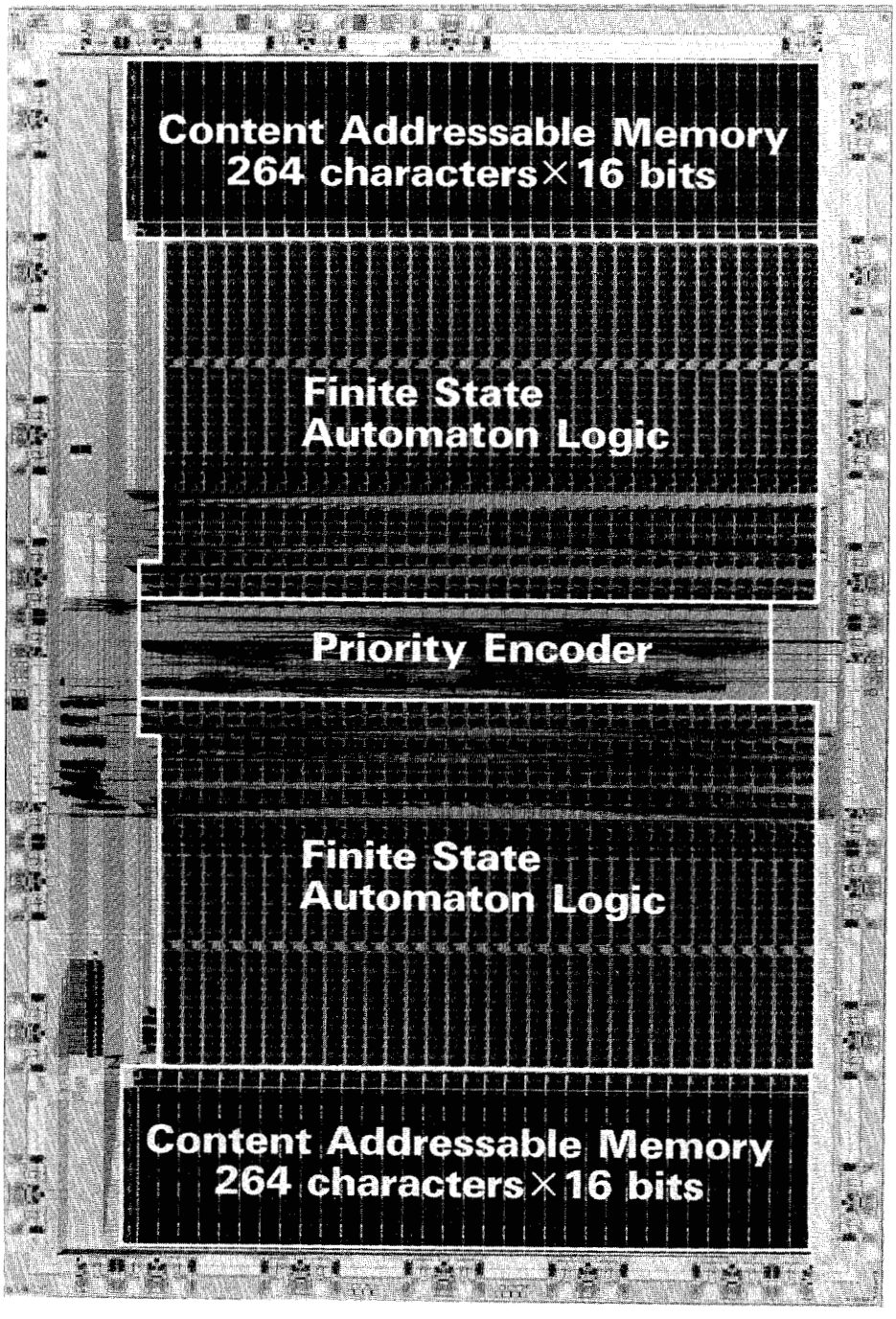

FIGURE 4–String search processor (SSP) photomicrograph.