Pipelined, Time-Sharing Access Technique for an Integrated Multiport Memory

Ken-ichi Endo, Tsuneo Matsumura, *Member, IEEE*, and Junzo Yamada, *Member, IEEE*

Abstract—**This paper describes a pipelined, time-sharing access (PTA) technique that realizes an integrated multiport memory for high-speed signal processing. The main features of this technique are as follows. a) $N/2$-port memory cells, with less area and a wider operating margin, are applied for the N-port memory function. b) Memory cell access for multiple ports is performed serially within one cycle instead of as an ordinary parallel operation. c) Memory operation is divided into three pipeline cycles—address selection, memory cell access, and data I/O operation—to reduce the cycle time. A 64-kb four-port memory is fabricated with conventional two-port memory cells to verify the effectiveness of this technique. A 16-ns memory operation with a wide margin is observed under a 3-V supply voltage.**

I. Introduction

MULTIPORT memories, especially two-port (dual port) RAM's, have been used as data RAM's to achieve real-time digital signal processing [1]–[5]. This is because multiport memories consist of static RAM cells with multiple random access ports that enable read/write operations individually. A dual-port RAM using single-port memory cells was reported previously [4], [5]. In this memory, internal latches were only arranged adjacent to the address and data I/O buffers to realize a dual-port function with a time-sharing and pipeline scheme. The operating speed, which was restricted to a serial memory cell access in a cycle, would thus fall off markedly with increased integration. In order to improve the processing system performance, multiport memories must satisfy three key requirements: a high level of integration, a large number of ports, and a high operating speed.

For real-time FFT processing, a two-memory-bank architecture consisting of two 2-port static RAM's has been used, for example, as a four-port data RAM [3]. In this case, a single high-density, high-speed four-port static RAM should result in much higher data transmission efficiency between processing units. However, two serious problems must be overcome to accomplish a four-port memory function with a four-port memory cell: the chip size must be increased, and the operating margin must be rapidly reduced. Adopting these measures, a four-port static RAM using four-port memory cells has been developed up to 16-kb integration [6]. Recently, another multiport memory approach was reported that realizes the multiport function by physically connecting one-port memory cells to each other [7]. Adopting this architecture, the operating margin of the multiport memory cell widens because the margin is equivalent to that of a one-port memory cell. However, the size of the memory cell increases in proportion to the number of ports. On the other hand, a compact full CMOS dual-port memory cell was proposed but a memory cell with three or more ports was not discussed [8].

This paper describes a novel multiport memory technique called pipelined, time-sharing access (PTA) to obtain both high integration independent of memory cell type and high-speed operating with a wide operating margin. Key aspects of the technique are a three-stage pipeline operation that accesses two-port memory cells consecutively in one pipeline cycle, and time-sharing access to two-port memory cells to realize a fast four-port memory function [9]. In Section II, the concept of the PTA technique is discussed, including an appraisal of the technique. In Section III, the PTA technique is applied to the design of a 64-kb four-port memory with synchronous operation from each of the four ports. A circuit design for a parallel-to-serial converter and a dual-sense-amplifier configuration is also discussed. These circuits yield a cycle time of 16 ns. Finally, some experimental results are presented demonstrating stable four-port memory operation over a wide range of supply voltages.

II. Pipelined, Time-Sharing Access Technique

A. Concept

A conceptual diagram of the PTA technique implemented as a four-port memory is shown in Fig. 1. The address and data sequence of a four-port memory operation are indicated in the figure. To achieve a high-speed multiport memory function, the PTA technique offers significant advantages.

1) The same number of peripheral circuits—the address select and data I/O circuits—as the address and data ports is operated in parallel. Memory cells with half the number of ports are operated serially using a time-sharing access scheme. Parallel-to-serial (P/S) conversion is performed by latches and converters, which are arranged at the interface between the memory cell array and the peripheral circuits.

2) Memory operation is divided into three stages: address selection, memory cell access, and data I/O operation. Then, a three-stage pipeline operation is applied to the memory. In particular, accessing the memory cells consecutively in one pipeline cycle improves the operating speed performance.

At first, let's consider the time-sharing access scheme shown in Fig. 1. In the figure, A_0 to A_3 indicate addresses from different address ports and D_0 to D_3 describe data

Manuscript received September 7, 1990; revised December 10, 1990.
The authors are with NTT LSI Laboratories, 3-1, Morinosato Wakamiya, Atsugi, Kanagawa, 243-01, Japan.
IEEE Log Number 9042536.

0018-9200/91/0400-0549$01.00 ©1991 IEEE

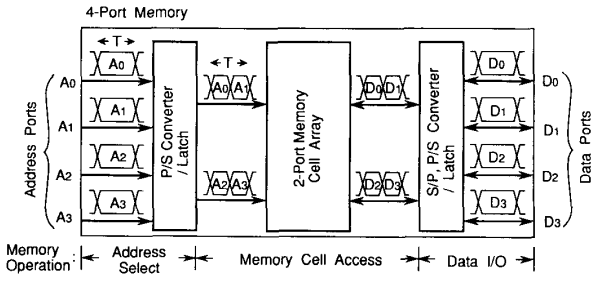


Fig. 1. Conceptual diagram of the pipelined, time-sharing access technique.

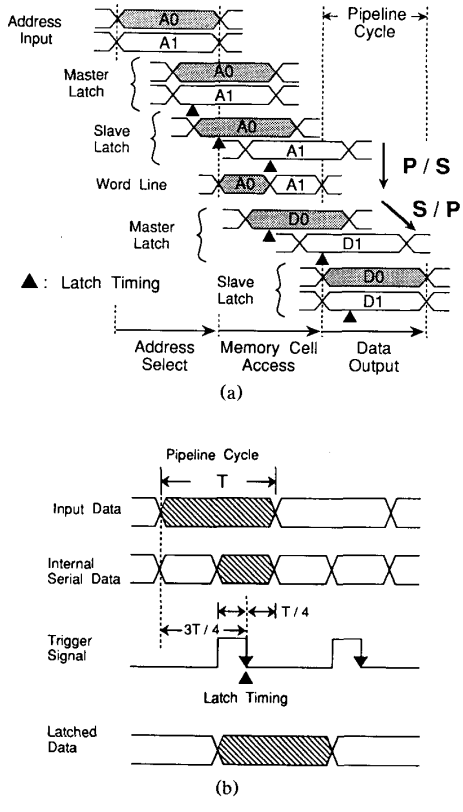


Fig. 2. Pipeline operation. (a) Address and data sequence. (b) Latch timing margin.

from different data I/O ports. All addresses for four ports, latched in the address select circuit simultaneously, are divided into two groups (A_0, A_1 and A_2, A_3). All data for four ports are the same as mentioned above (D_0, D_1 and D_2, D_3). These time-sharing addresses and data are applied to a continuous word-line selection and a data I/O operation that are handled in just one cycle. With this scheme implemented at the memory cell access stage, each two-port memory cell can be selected twice during a cycle T to accomplish a four-port memory function. For example, in the first half of the cycle, two word lines are activated at the same time by addresses A_0 and A_2. Then, data D_0 and D_2 are read out from or written to the selected memory cells in parallel. In the second half of the cycle, addresses A_1 and A_3 and data D_1 and D_3 behave the same way. Of course, it is possible for the same two-port memory cell to be accessed if A_0 to A_3 are the same. Consequently, this scheme allows four-port read access and continuous read/write or write/read access for each serial port to be handled by a two-port memory cell. Turning to the address select circuit and data I/O circuit, all addresses and data are input or output simultaneously within one cycle. In this case, serial-to-parallel (S/P) conversion is performed for the two sets of readout data (D_0, D_1 and D_2, D_3). The latch and converter control the switching of the address and data transmission paths. In this scheme, the cycle time is regulated by the serial accessing of the memory cells, especially the serial write and read operation to the memory cells connected to the same bit-line pair. Therefore, to shorten this cycle time, the three-stage pipeline operation mentioned above as advantage 2) has been introduced.

The pipelined address and data flow diagram for a read operation is shown in Fig. 2(a). The waveforms in this diagram represent valid address and data periods for the two address and data ports. The triangles under the waveforms show the latch timing for addresses and data. This indicates a synchronous memory operation for all ports. The address flow can be divided into two pipeline cycles because latches and converters separate the address select circuit, which operates in parallel, from the memory cell array, which operates serially. The cycle time can be shortened by assigning the memory cell access stage to one pipeline cycle, which is equal to a conventional memory cycle. If the master latch inputs two addresses at the same time, the slave latch outputs the addresses at different times. Therefore, the simultaneous input addresses A_0 and A_1 can be used for serial memory cell access by applying this master and slave latch control to P/S conversion. The sequential data D_0 and D_1 are transmitted to the parallel operating data output buffers through the slave latches with S/P conversion. These latches hold the output data during one pipeline cycle. Consequently, there is a two-pipeline cycle delay between the selected address inputs and the readout data latched in the slave latches.

The timing diagram of the input data latch and the internal serial data latch generated by the PTA technique is shown in Fig. 2(b). All the internal control clocks are generated by the clock generator. Therefore, nothing need be done to correct the skew among external control clocks, such as write enable and chip select. In addition, trigger signals, whose activation period is a quarter of the pipeline cycle T, are utilized for transmission control of the latched data. The data are latched at the timing of a quarter or three quarters of the data activation period using these trigger signals. As a result, a timing margin of a quarter of the pipeline cycle is always maintained between the edge of the data and trigger signals.

A comparison of the pipelined address and data flow between the write operation and the read operation is shown in Fig. 3. The write operation finishes with two pipeline cycles instead of three pipeline cycles for the read operation. In the first cycle of the write operation, the address selection and the parallel input data transmission to the write buffer through the latch are performed at the same time. In the

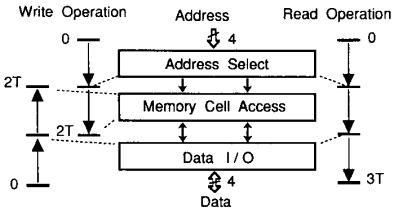


Fig. 3. Pipelined write and read operations.

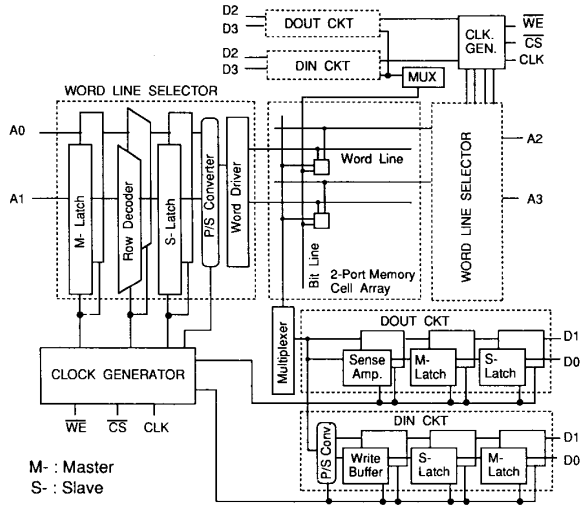


Fig. 4. Block diagram of the PTA four-port memory chip.

second cycle, these data are written to the selected memory cells serially.

By thus combining the time-sharing access and the pipeline operation, a multiport memory capable of stable and high-speed operation is achieved.

B. Configuration

A block diagram of the four-port memory applying the PTA technique is shown in Fig. 4. A memory cell array consists of two-port memory cells selected by two parallel accessing address ports ($A_0(A_1)$ and $A_2(A_3)$). Word-line selectors consist of row decoders and master–slave latches for four address ports, and P/S converters and word drivers for two serial accessing address ports ($A_0(A_2)$ and $A_1(A_3)$). Data input circuits consist of write buffers and master–slave latches for four data ports, and P/S converters for two serial accessing data ports ($D_0(D_2)$ and $D_1(D_3)$). Data output circuits consist of sense amplifiers and master–slave latches with S/P conversion for four data ports. A clock generator is arranged for each port to control the word-line selector and data I/O circuit. This generator is also controlled by the P/S converter and master–slave latch. This arrangement permits many kinds of control clocks to be implemented to provide appropriate timing at each stage of the pipeline operation.

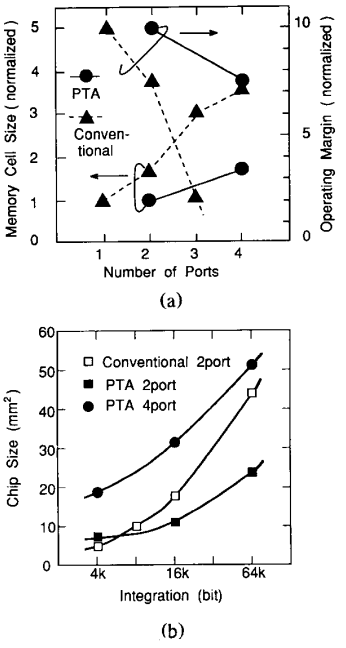


Fig. 5. Performance comparison between the PTA and conventional techniques. (a) Memory cell size and operating margin. (b) Chip size.

C. Effectiveness

A comparison of the memory cell size and operating margin for the PTA technique and a conventional multiport technique is shown in Fig. 5(a). The operating margin indicates the noise tolerance of a flip-flop in the static RAM cell in relation to the amount of dc disturbance, such as voltage offset during a read operation. The solid lines are for the PTA, and the broken ones are for the conventional technique. Clearly, the PTA dramatically improves the performance as the number of ports increases. This is because the PTA uses memory cells with only half the number of ports of the conventional memory cells. In the conventional memory cells with three or more ports, the operating margin becomes narrower because the drivability of the driver transistor of the flip-flop in the static RAM cell becomes lower. In addition, the memory cell size is larger due to the requirements of memory cell layout design.

A comparison of the PTA technique and a conventional multiport technique in terms of chip size is shown in Fig. 5(b). In case of the PTA, the two-port memory consists of one-port memory cells half the size of the two-port memory cell. Moreover, the ratio of the additional circuit (P/S converter, etc.) area to the chip area decreases as the integration increases. Therefore, the PTA technique allows smaller chip size than the conventional technique at integration of 10 kb or more.

III. 64-kb Four-Port Memory Circuit Design

A 64-kb (8K word×8 b) four-port memory was designed to verify the effectiveness of the PTA technique described in the previous section.

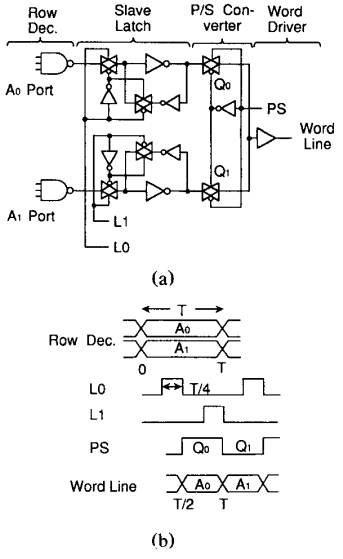


Fig. 6. Word-line selector with latch and parallel/serial converter. (a) Configuration. (b) Clock waveforms.

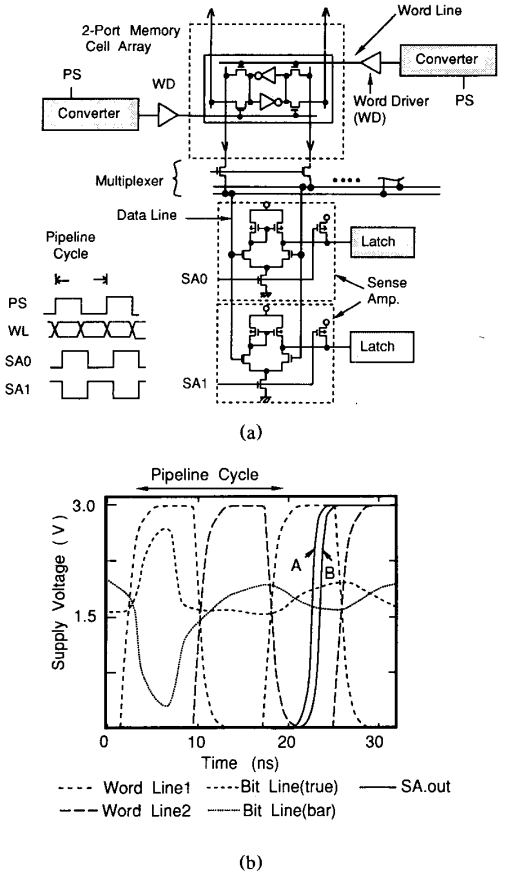


Fig. 7. Dual-sense-amplifier configuration. (a) Configuration. (b) Simulated waveforms.

A. Circuit Design

Peripheral circuits have special design requirements for high data transmission efficiency. Specifically, the circuits are designed to prevent a transmission time loss during the data conversion, and to enable fast memory cell access. To meet these requirements, a P/S converter and a dual-sense-amplifier configuration are discussed in this section.

1) P/S Converter: A P/S converter arranged in a word-line selector is shown in Fig. 6(a). Row decoders and slave latches are provided for two serial accessing address ports (A_0 and A_1). A P/S converter, arranged for each port of a two-port memory cell, is a switching circuit for two data transmission routes. A clock timing diagram for the latch and P/S converter is shown in Fig. 6(b). Latch control clocks L_0 and L_1 determine the A_0 and A_1 address latch timing. The clocks' phase shifts of a half cycle delay for L_0 and L_1 determine the valid address periods for word-line activation. The converter control clock PS has a cycle time T and a time duration of 50%. The clock PS alternately activates transmission gate Q_0 and Q_1 in the P/S converter. The use of the clock PS eliminates the need for a timing margin between two word lines. This indicates that continuous word-line activation can be carried out with no delay.

2) Dual-Sense-Amplifier Configuration: The sense amplifier must operate continuously because of the consecutive word-line activation. If a single sense amplifier is provided for each data-line pair, it is necessary to insert an additional initialization period before the sensing operation. This makes the cycle time longer. Therefore, to shorten this cycle time, the dual sense-amplifier configuration is provided as shown in Fig. 7(a). Complementary clocks SA_0 and SA_1 always make the output nodes of the sense amplifiers initialize to the supply voltage level within one half of a pipeline cycle, which is assigned to an inactive period for one sense amplifier while the other sense amplifier activates. These clocks are synchronous to the converter control clock PS, which controls the word-line activation timing. As a result, with this dual-type configuration, continuous sensing operation of the alternate inverted memory cell data takes less time than with the single-type configuration because the output node of the sense amplifier discharges faster than it precharges.

Simulated waveforms of the four-port read operation after write operation using the dual- and single-sense-amplifier configurations are illustrated in Fig. 7(b). Output waveform A is for the dual-type configuration and B is for the single-type one. The time difference between A and B is due to the extra initialization period. This difference doubles in a pipeline cycle because of continuous activation of the sense amplifier in a cycle. In our example of the 64-kb four-port memory, since the time difference between A and B is estimated at 2 ns, this difference becomes 4 ns during a pipeline cycle.

B. Chip Design

A block diagram of a 64-kb four-port memory composed of two-port memory cells is shown in Fig. 8. The memory cell array is not divided into subarrays. The number of word lines is 256, which is the same as bit-line pairs. The peripheral

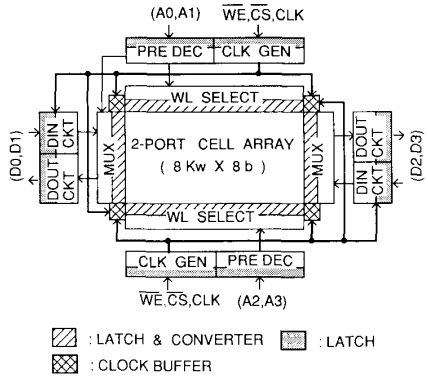


Fig. 8. Block diagram of the 64-kb four-port memory.

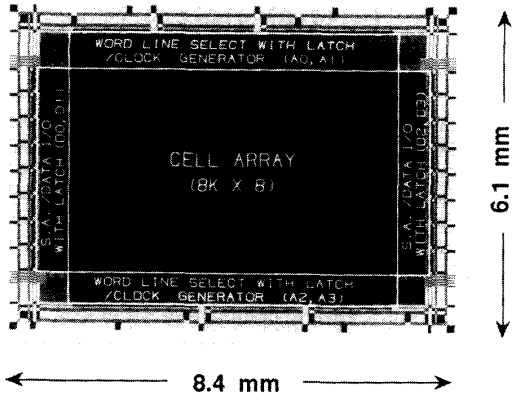


Fig. 9. Photomicrograph of the 64-kb four-port memory chip using the PTA technique.

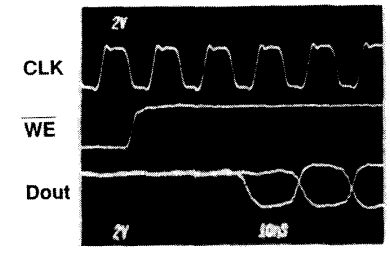


Fig. 10. Typical waveforms for pipeline operation at 3 V and 60-MHz clock rate.

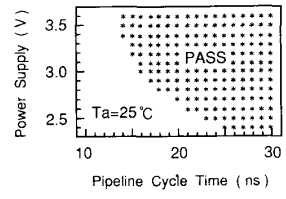


Fig. 11. Schmoo plot.

TABLE I
FEATURES OF THE 64-kb FOUR-PORT MEMORY

Organization	8K words×8 b
Process	0.8-μm n-well CMOS
Memory Cell	Conventional two-port full-CMOS type (13.4×24 μm^2)
Chip Size	6.1×8.4 mm^2
Cycle Time	16 ns (at $V_{cc}=3$ V, $T_a=25$°C)
Supply Voltage	3 V
Operating Current	180 mA at 60 MHz

circuits for parallel accessing—circuit blocks of the A_0, A_1, D_0, and D_1 ports and those of the A_2, A_3, D_2, and D_3 ports—are arranged symmetrically on the chip. Dual-clock-transmission wirings are utilized to reduce large capacitance loads in the word-line selector (WL SELECT) and the multiplexer (MUX). Moreover, the wirings are cut in the middle of the circuit blocks. These wiring techniques reduce the capacitance of a latch control clock line to about one quarter, for example, from 12 to 3 pF. In addition, clock buffers are arranged at the four corners of the memory cell array to provide powerful drivability, as shown by the cross-hatched areas in Fig. 8. These circuit and layout techniques equalize the time for clock distribution and data transmission among all address and data ports. The areas marked by diagonal lines indicate latches and converters, and shaded areas show latches. They occupy only 7% of the chip area. All wirings and circuit block arrangements were derived by a routing program [10]. A conventional two-port full-CMOS static RAM cell was adopted because it has a higher operating margin over a wide range of supply voltages than a polysilicon load memory cell.

A photomicrograph of a 64-kb four-port memory chip is shown in Fig. 9. This chip was designed using the PTA technique and fabricated with 0.8-μm n-well CMOS technology. The chip size is 8.4×6.1 mm^2, which is only 1.2 times larger than a 64-kb two-port memory.

C. Experimental Results

To confirm the multiport memory functions and performances of this chip, we carried out interference tests for simultaneous parallel or serial two-port operation and functional tests for simultaneous three- or more port operation, in addition to the usual one-port testing [11]. Marching and Checkerboard test patterns were utilized for all these tests.

Typical four-port read cycle waveforms for the pipeline operation at a power supply of 3 V and a frequency of 60 MHz are shown in Fig. 10. The main clock CLK controls almost all the internal clocks. D_{OUT} is a data-output waveform with a 50-pF load capacitance. The falling edge of CLK regulates the internal memory operation. Thus, at the first falling edge, $\overline{WE}$ is activated and then the read cycle starts. After activation of $\overline{WE}$, a continuous four-port read operation with a two-pipeline cycle delay is obtained.

A Schmoo plot of the pipeline cycle time versus the power supply voltage for the Marching test pattern is shown in Fig. 11. A stable four-port memory operation over a wide range of supply voltages was achieved, and the minimum pipeline

cycle time was 16 ns. The main features of the memory chip are summarized in Table I.

IV. Conclusions

To achieve an integrated multiport memory that enables simultaneous random read/write access and fast operation, a new pipelined, time-sharing access (PTA) technique has been developed. The time-sharing access scheme permits fabrication of high-density multiport memories with a large number of ports. Four-port memory operation can be obtained by selecting a two-port memory cell twice during a cycle. This thus results in a smaller chip size and a wider operating margin. The pipeline operation shortens the cycle time. Accessing memory cells consecutively in one pipeline cycle improves the operating speed performance. Master–slave latches and P/S, S/P converters are added to a conventional memory to apply the PTA technique.

A 64-kb four-port memory was designed using the PTA technique and fabricated with 0.8-μm CMOS technology. The device had a cycle time of 16 ns at 3 V and room temperature. The additional area of the latches and converters take up several percent of the chip area. The chip size was only 1.2 times larger than a 64-kb two-port memory.

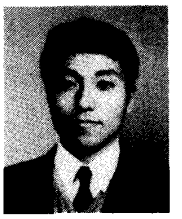
The PTA is a key technique in the development of a high-density, high-speed mutliport memory suitable for data RAM in real-time digital signal processing systems.

Acknowledgment

The authors wish to thank N. Ieda, T. Mano, H. Sawada, and H. Miyanaga for their encouragement and useful discussions.

References

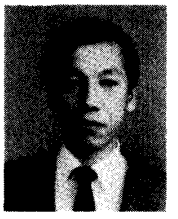
[1] Y. Shimazu *et al.*, "A 50 MHz 24 b floating-point DSP," in *ISSCC Dig. Tech. Papers*, Feb. 1989, pp. 44–45.
[2] N. Jagadish *et al.*, "An efficient scheme for interprocessor communication using dual-ported RAMs," *IEEE Micro*, vol. 9, no. 5, pp. 10–19, 1989.
[3] H. Miyanaga, "A systematic approach to VLSI implementation for real-time digital signal processing," Master's thesis, Mass. Inst. Technol., Cambridge, Jan. 1989.
[4] F. E. Barber *et al.*, "A 200 ns 512×10 bit dual port RAM," in *Proc. Electron. Conf.*, vol. 36, Oct. 1982, pp. 380–382.
[5] F. E. Barber *et al.*, "A 2K×9 bit dual port memory," in *ISSCC Dig. Tech. Papers*, Feb. 1985, pp. 44–45.
[6] "High-speed 1K×8 and 2K×8 four-port static RAMs," Integrated Device Technology, Inc., Santa Clara, CA, Product Data Sheets, Jan. 1989.
[7] K. Sawada *et al.*, "A 5 ns 369 kb port-configurable embedded SRAM with 0.5 μm CMOS gate array," in *ISSCC Dig. Tech. Papers*, Feb. 1990, pp. 226–227.
[8] K. J. O'Connor, "The twin-port memory cell," *IEEE J. Solid-State Circuits*, vol. SC-22, no. 5, pp. 712–720, Oct. 1987.
[9] T. Matsumura *et al.*, "Pipelined, time-sharing access technique for a highly integrated multi-port memory," in *Symp. VLSI Circuits Dig. Tech. Papers*, June 1990, pp. 107–108.
[10] K. Takeya *et al.*, "A generator for high-density macrocells with hierarchical structure," in *Proc. CICC*, 1989, pp. 22.5.1–4.
[11] M. J. Raposa, "Dual port static RAM testing," in *Proc. ITC*, 1988, pp. 362–368.



Ken-ichi Endo was born in Yamanashi, Japan, on June 11, 1961. He received the B.S. and M.S. degrees in electronic engineering from Yamanashi University, Yamanashi, Japan, in 1984 and 1986, respectively.

He joined the Atsugi Electrical Communication Laboratory, Nippon Telegraph and Telephone (NTT) Corporation, Kanagawa, Japan, in 1986, and worked on the circuit design of the MOS integrated circuits. He is presently a Research Engineer in the Integrated Circuit Technology Laboratory, NTT LSI Laboratories, Kanagawa, Japan. His recent work involves the research of memory circuit technologies including ASIC memory.

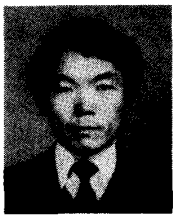
Mr. Endo is a member of the Institute of Electronics, Information and Communication Engineers of Japan.



Tsuneo Matsumura (M'90) was born in Yokohama, Japan, on November 2, 1954. He received the B.S. and M.S. degrees in electrical engineering from Keio University, Tokyo, Japan, in 1978 and 1980, respectively.

He joined the Musashino Electrical Communication Laboratory, Nippon Telegraph and Telephone (NTT) Public Corporation, Tokyo, Japan, in 1980, and worked on high-density DRAM design. He is presently a Senior Research Engineer in the Integrated Circuit Technology Laboratory, NTT LSI Laboratories, Kanagawa, Japan. His recent work involves the research of memory circuit technologies including ASIC memory.

Mr. Matsumura is a member of the Institute of Electronics, Information and Communication Engineers of Japan.



Junzo Yamada (M'86) was born in Nagoya, Japan, on April 3, 1951. He received the B.S. and M.S. degrees in electronic engineering and the Ph.D. degree in computer science from Tokyo Institute of Technology, Tokyo, Japan, in 1974, 1976, and 1990, respectively.

In 1976 he joined the Musashino Electrical Communication Laboratory, Nippon Telegraph and Telephone (NTT) Public Corporation, Tokyo, Japan, where he worked on the design and testing of fault-tolerant MOS memory. He is currently a Memory Circuit Research Group Leader in the Integrated Circuit Technology Laboratory, NTT LSI Laboratories, Kanagawa, Japan. He has been engaged in research on the circuit design of ASIC memory.

Dr. Yamada is a member of the Institute of Electronics, Information and Communication Engineers of Japan.