Design of a Scalable Systolic Synchronous Memory

Gab-Joong Jeong, Kyoung-Hwan Kwon, and Moon-Key Lee

Abstract

This paper describes a scalable systolic synchronous memory for digital signal processing and packet switching. The systolic synchronous memory consists of the 2-D array of small memory blocks which are fully pipelined and communicate in three directions with adjacent blocks. The maximum delay of a small memory block becomes the operation speed of the chip. The array configuration is scalable for the entire memory size requested by an application. It has the initial latency of N+3 cycles with N×N array configuration. We designed an experimental 200 MHz 4Kb static RAM chip with the 4×4 array configuration of 256b SRAM blocks. It was fabricated in 0.8 ㎛ twin-well single-poly double-metal CMOS technology.

Ⅰ. Introduction

There have been many previous works to improve the access time of memory device. From the efforts, divided word-line technique [1, 2] reduces the RC delay from several components of the access time of memory device and pipelined architecture is introduced in SRAM to support fast access time [3]. Improvement of memory bandwidth can be achieved by dividing internal structure into many blocks and expanding data bus width. Switching speed of memory device can be increased by dividing internal structure into many blocks because an actual data path to specific memory cell is shorten and capacitive load becomes small. Although those technologies reduced the access time of memory device, many applications of memory device require some new features like scalability that cannot be solved with conventional memory architecture. Scalability feature includes a very fast access time regardless of memory size. Especially, in the fields of embedded memory for digital signal processing and packet switching that are currently focused on, requirements for scalability as well as fast access time have grown stronger. We introduced a systolic fashioned three directional data flow in memory device [4-6].

This paper presents a scalable systolic synchronous memory which has scalability using pipelined architecture with three directional communication of adjacent blocks. It raises memory bandwidth by dividing the internal memory structure into two dimensional array configuration of small random access memory block. It supports random address access. Memory access time of the systolic synchronous memory depends on the access time of a small memory block and is independent of entire array configuration. We can enlarge the entire memory capacity by increasing the array configuration. The small memory block is designed for the access time requirement of the system. Although it has the initial latency of N + 3 cycles for N×N array configuration, it handles one of 'read' and 'write' operations at every cycle. The initial latency can be compared to the initial latency of 2N + 3 in the method of SCRAM(Scalable Cellular RAM)[7, 8]. The smaller initial latency can drive performance increase when the array configuration is large.

In Section Ⅱ, proposed entire architecture and its address decoding method are described. Implemented circuits are described for decoders and memory block in Section Ⅲ. Experimental 4Kb chip and its performance are described in Section Ⅳ. Conclusions will be given in Section Ⅴ.

Ⅱ. Architecture

1. Pipelined Architecture

We divided typical asynchronous RAM structure into five blocks for the proposed scalable systolic synchronous memory. The five blocks are primary decoder, row decoder, column decoder, small memory block, and output buffer. Fig. 1 shows the entire architecture of the systolic synchronous memory. Each block is fully pipelined and treats new input data at every cycle. Primary decoder decodes input address and generates three new signals for pipeline control. Row decoder and column decoder decode row and column addresses individually which are generated by the primary decoder. They generate row and column selection signals to validate a diagonal data path. A small memory block

Manuscript received May 29, 1997; accepted July 26, 1997.

G. J. Jeong, K. H. Kwon, and M. K. Lee are with Dept. of Electronic Eng., Yonsei University, Korea.

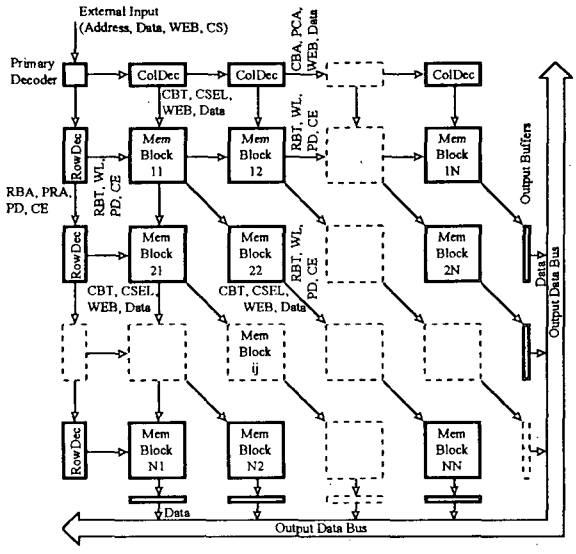


Fig. 1. Entire architecture of a scalable systolic synchronous memory.

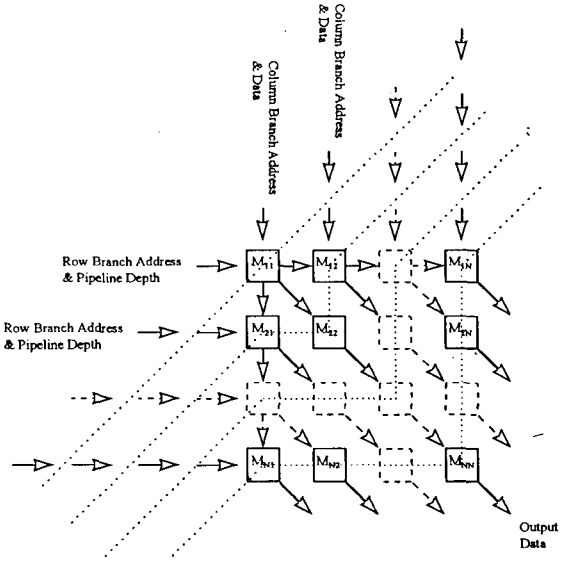


Fig. 2. Systolic data flows of the proposed memory architecture.

is designed with I rows and J columns which are suitable to support the memory access time requirement of an application. The two dimensional array configuration size of the memory blocks is determined by the required entire memory size of the application. The entire memory size of N×N array configuration is $I \times J \times k \times N^2$ where k is the bit size of a column. Each block communicates with adjacent blocks in systolic fashioned data flow. Fig. 2 shows the systolic data flows.

2. *Array Element Selection Algorithm*

Primary decoder in the first pipeline stage generates row branch address(RBA), column branch address(CBA), and pipeline depth (PD) to select an array element. In the case of N×N array configuration, it generates three variables above with $\log_2 N$ bit size using $2\log_2 N$ bits of input address. Fig. 3 shows the primary decoding algorithm. Remained bits of input address are used for the actual access of memory cells in a small memory block. Among them, $\log_2 I$ bits are used for word line(WL) and the others are used for column selection(CSEL) in a memory block. To reduce the decoding data path delay for WL and CSEL, primary decoder decodes those bits partially and transfers them to row and column decoders individually.

In the second pipeline stage, row and column decoders operate in parallel. The row decoder gets RBA, PD, and partially decoded row address(PRA) from the primary decoder or previous row decoder. It generates row branch trigger(RBT) for an adjacent memory block. The RBT is 'high' if the RBA is zero. It transfers PD to an adjacent memory block in horizontal. It generates word line(WL) data using PRA and transfers the WL to the memory block. For the next adjacent row decoder, it transfers RBA and PD which are decreased by 1 and input PRA. The column decoder gets CBA, partially decoded column address (PCA), and input data for 'write' operation from the primary. decoder or previous column decoder. It generates column branch trigger(CBT) 'high' when CBA is zero. It transfers column selection(CSEL) signal to an adjacent memory block in vertical. In 'write' mode, it transfers input data vertically. For the next adjacent column decoder, it transfers CBA which is decreased by 1, data for 'write', and input PCA.

There is the array of small memory blocks in the third pipeline stage. Each memory block decides one of two operations, propagation and internal memory access using RBT, CBT, and PD which are propagated from row and column decoders and previous memory block. Pipeline cycle in this array is determined by the array configuration. Only one diagonal data path is

```
Algorithm: Primary Decoding( RBA, CBA, PD, address )
    x = log2N bits of external address;
    y = another log2N bits of external address;
    Temp = x - y;
    if( Temp >= ZERO ) {
            RBA = | Temp |;
            CBA = 0;
            PD = x;
    } else {
            RBA = 0;
            CBA = | Temp |;
            PD = y;
    }
```

Fig. 3. Primary decoding algorithm.

selected to transfer data to the end of the pipeline. If the PD is zero in 'read' and 'write' modes, actual memory cell access arises in a memory block which is located in the valid diagonal data path. Memory cell access happens only once for one input address. Each memory block transfers PD which is decreased by 1 to the next adjacent memory blocks in three directions. Memory blocks adjacent to the row decoders transfer signals and data in vertical and diagonal directions and memory blocks adjacent to the column decoders transfer signals and data in horizontal and diagonal directions. Other memory blocks transfer signals and data in diagonal direction. Therefore, the traveling latency of input address and data in the array is N cycles for N $\times$N array configuration. Initial latency is N + 3 cycles. One cycle delay occurs at the primary decoder and another one cycle delay occurs at the row and column decoders. At the last pipeline stage, there is another one cycle delay for output buffer. Output buffers are attached at the memory blocks that are located at the boundary of memory block array. Fig. 4 shows an example of the initial latency of 7 cycles for 4$\times$4 configuration.

III. Circuit Implementation

1. RAM Block

Designed memory block has 256 memory cells which consist of 16 rows and 4 columns with 4b data width. It has the memory cell block, a PD decrement block, and a control logic. Fig. 5 shows the schematic diagram of the memory block. The control logic decides an operation from propagation and memory cell access. It uses RBT and CBT to decide the direction of data propagation. CBT controls vertical data propagation and RBT controls horizontal data propagation. If both of them are 'high' level, the data propagation direction is diagonal. In memory cell access, the control logic checks whether PD is zero or not when CBT and RBT are in 'high' level. It decides whether memory cell access is needed or not in its memory block. The decrement block decreases PD by 1 before propagation.

Fig. 6 illustrates one bit column of a memory cell block and its timing diagram. We used conventional 6-Tr full CMOS SRAM cells and simple single ended differential amplifiers. Memory access arises synchronously with external clock. Input data are

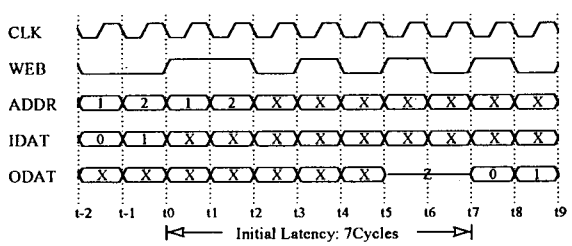


Fig. 4. An example of the initial latency of 7 cycles for 4$\times$4 configuration.

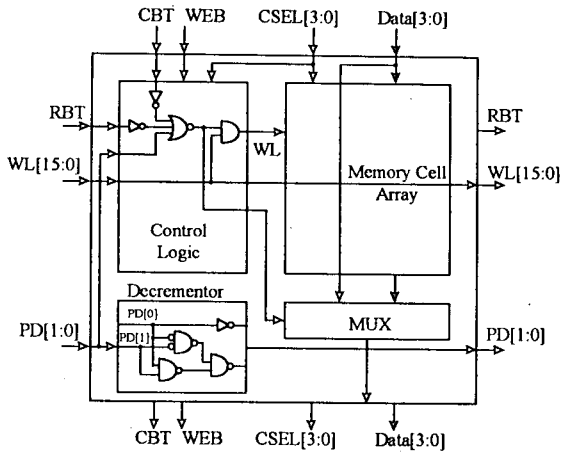


Fig. 5. Schematic diagram of a memory block.

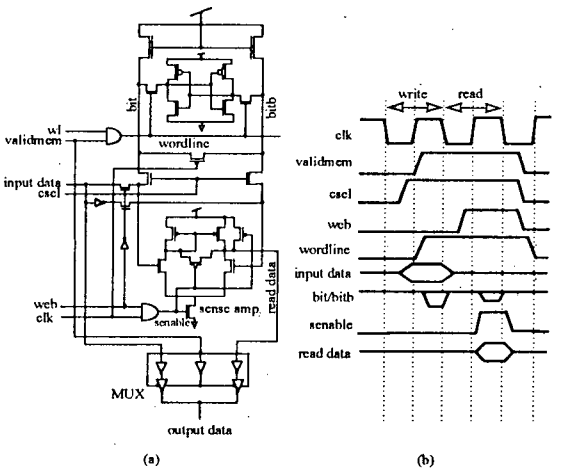


Fig. 6. Circuit and timing diagrams of one bit memory column.

latched at the falling edges of clock. During 'low' level of clock, precharge cicuit precharges bit lines and the loads of sense amplifiers to Vdd. At the rising edge of clock, either sensing or writing occurs. Sensed data for 'read' travel multiplexer for propagation during 'high' level of clock. The 'write' operation finishes at the falling edge of clock. Bit lines and loads of sense amplifiers go back to precharge state during the next 'low' level of clock. When the memory enable signal(validmem) is 'low' level, the memory block propagates data only.

2. Decoders and Output Buffer

Input address decoding is performed by three decoding blocks. Fig. 7 shows the schematic diagram of the primary decoder. It consists of one 2b subtractor, one 2b two's complementor, and signal selection circuits to generate RBA, CBA, and PD. It also has three 2-4 decoders to decode row and column addresses which are used for memory cell access in a memory block. The row decoder has two decrement blocks to subtract 1 from PD and RBA. It has logic circuits for word line data generation.

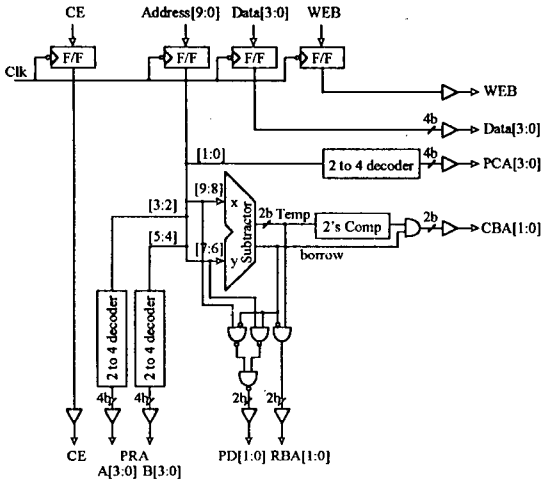


Fig. 7. Schematic diagram for primary decoder.

The column decoder has one decrement block to subtract 1 from CBA. In the column decoder, 4b PCA signal from input data can be used for column selection signal(CSEL) because the designed memory block has only 4 columns. Output buffers are located at the boundary memory blocks. It consists of flip-flops and tri-state buffers. The tri-state buffers are controlled by RBT, CBT, and WEB which are propagated by adjacent memory block. For 'read' operation, RBT and CBT signals are 'high' in only one output buffer when WEB is 'high'. At the view of scalability, the only factor, data bus length, is increased by the increase of array configuration. The propagation delay caused by the pure bus wire length does not affect the scalability of the systolic synchronous memory seriously.

Ⅳ. Experimental Results

1. Characteristics of Systolic Synchronous Memory Chip

Experimental 4Kb systolic synchronous memory chip was fabricated in a 0.8-μm, twin-well, single-poly, double-metal CMOS technology. It was designed by full custom layout. The chip size is 3.7 mm×3.7 mm. It has separated 4b data I/O and 4×4 configuration of 256b SRAM blocks. We used typical 6-Tr SRAM cell. The cell area is 405 μm^2. Maximum operating clock frequency is 200 MHz from simulation. Estimated power consumption is 1173 mW at 200 MHz with 5 V supply. The memory cell area and routing area of systolic synchronous memory chip can be shrunk by layout optimization. The measurement and simulation results are presented in Table 1. Fig. 8 shows the microphotograph of the experimental systolic synchronous memory chip.

2. Performance

Using the scalability of systolic synchronous memory, we simulated the access time of large memory size with various configurations. Fig. 9 shows the access time comparison of systolic synchronous memory and typical asynchronous SRAM [9]. We focused the comparison on the access time of similar process technology of the experimental chip. The increase of power consumption followed by the increase of configuration size can be reduced by disabling inactive blocks with gated clock. Fully active memory blocks are N among N^2 memory blocks because one diagonal data path is fully active for one operation. We can estimate that the reduction rate of on-chip power consumption is 1/N by the gated clock for N×N configuration. General low power and high speed circuit techniques of SRAM can be applied to design the memory block of systolic synchronous memory.

Table 1. Characteristics of the experimental chip.

Process Technology	0.8μm 2-Metal 1-Poly CMOS
Chip Size	3.7×3.7mm^2
Organization	4×4×256b(16×4×4b)
Transistors	44.8K
Initial Latency	7 Clock Cycles
Throughput	800Mbps
Operating Frequency	200MHz
Power Dissipation	1173 mW(at 200MHz)
Power Supply	5 V
I/O	4-bit Separated I/O
Package	24-pin DIP

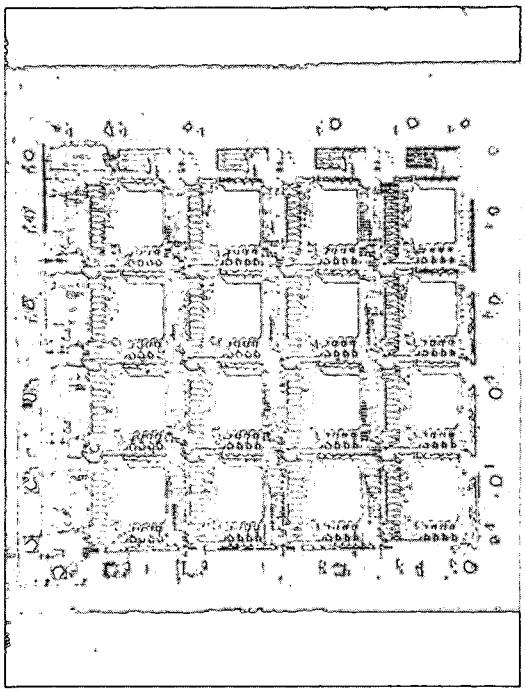


Fig. 8. Microphotograph of the systolic synchronous memory chip.

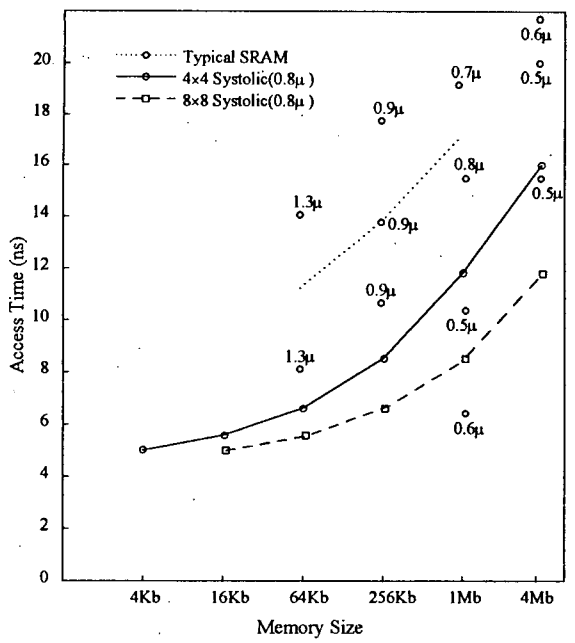


Fig. 9. Access time comparison of systolic synchronous memory and typical asynchronous SRAM.

V. Conclusions

In this paper, we described a scalable systolic synchronous memory. It is a scalable memory architecture that has the initial latency of N + 3 cycles for N×N configuration using three directional data flow shown in Section II. It is suitable for the packet switching and DSP system which need fast access time and large memory size. It provides the scalability with little circuit redesign in memory block. Large memory capacity can be achieved without an access time variation by increasing the configuration size. On-chip power consumption can be reduced by using gated clock for inactive blocks described in Section IV. Our continuing research is layout area optimization and power reduction circuit implementation for the systolic synchronous memory. We are designing an 622Mbps CMOS ATM switch chip embedded the systolic synchronous memory as an application example.

References

[1] M. Yoshimoto, K. Anami, H. Shinohara, T. Yoshihara, H. Takagi, S. Nagao, S. Kayano, and T. Nakano, "A divided word-line structure in the static RAM and its application to a 64k full CMOS RAM," IEEE J. Solid-State Circuits, Vol. SC-18, No. 5, pp. 479-485, Oct. 1983.

[2] T. Hirose, H. Kuriyama, S. Murakami, K. Yuzuriha, T. Mukai, K. Tsutsumi, Y. Nishimura, Y. Kohno, and K. Anami, "A 20-ns 4-Mb CMOS SRAM with hierarchical word decoding architecture," IEEE J. Solid-State Circuits, Vol. 25, No. 5, pp. 1068-1074, Oct. 1990.

[3] D. Schmitt-Landsiedel, B. Hoppe, G. Neuendorf, M. Wurm, and J. Winnerl, "Pipeline architecture for fast CMOS buffer RAM's," IEEE J. Solid-State Circuits, Vol. 25, No. 3, pp. 741-747, June 1990.

[4] M. K. Lee, K. W. Shin, and J. K. Lee, "A VLSI array processor for 16-point FFT," IEEE J. Solid-State Circuits, Vol. 26, No. 9, pp. 1286-1292, Sept. 1991.

[5] G. J. Jeong, K. H. Kwon, M. K. Lee, and S. H. Ahn, "High throughput systolic memory architecture using three directional data flows," in Proc. IEEE ICECS, Rodos, Greece, pp. 667-670, Oct. 1996.

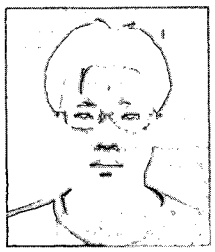
[6] K. H. Kwon, G. J. Jeong, M. K. Lee, and S. H. Ahn, "A scalable memory system design," in Proc. Int. Conf. VLSI Design, Hyderabad, India, pp. 257-260, Jan. 1997.

[7] A. G. Dickinson and C. J. Nicol, "A systolic architecture for high speed pipelined memories," in Proc. IEEE Int. Conf. Computer Design, Cambridge, MA, pp. 406-409, Oct. 1993.

[8] A. G. Dickinson and C. J. Nicol, "A scalable pipelined architecture for fast buffer SRAM's," IEEE J. Solid-State Circuits, Vol. 31, No. 3, pp. 419-429, Mar. 1996.

[9] B. Prince, Semiconductor memories, Wiley, New York, 1991.

Gab-Joong Jeong received the B.S. and M.S. degrees in Electronic Engineering from Kyungpook National University, Taegu, Korea, in 1987 and 1989, respectively. He is currently working toward the Ph.D. degree in the Department of Electronic Engineering at Yonsei University, Seoul, Korea. He worked as a Senior Engineer in R&D Center of LG Semiconductor, Inc. from 1989 to 1994. His current research interests include VLSI design and packet switching.



Kyoung-Hwan Kwon was born in Seoul, Korea on 1972. He received the B.S. and M.S. degrees in Electronic Engineering from Yonsei University, Seoul, Korea, in 1995 and 1997, respectively. He is currently working on mobile technology at LG Semiconductor, Inc. His current research interests include VLSI architecture, Microprocessor-based system design, and portable system design.

Moon-Key Lee received the B.S., M.S., and doctor degrees in electrical engineering from Yonsei University, Seoul, Korea in 1965, 1967, and 1973, respectively. Also, he received the Ph.D. degree in electronic engineering from the University of Oklahoma, Oklahoma, in 1980. He is currently a professor in the Department of Electronic Engineering at Yonsei University, Seoul, Korea and a visiting professor at the University of Illinois at Urbana-Champaign, Illinois. He was Chairman of The Korea Institute of Telematics and Electronics(KITE) at 1995. From 1980 to 1982, he was Director of Semiconductor Design Division at the Korea Institute of Electronic Technology(ETRI), Kumi, Korea. He is a founder of Research Institute of ASIC Design(RIAD), which was established in 1989 and located at Yonsei University, Seoul, Korea. His current research interests include high performance Microprocessor, Digital Signal Processor, VLSI design, and packet switching.