the replaceable unit: If, e.g., the reconfiguration is to be performed in units of half chip capacity, this can be implemented very easily by doubling the number of FST entries, i.e. by assigning each entry to half a block instead to an entire one. This strategy may be generalized to smaller binary fractions $1/2^g$ (g=1,2,...) of the chip capacity. The management of the size of reconfiguration is performed completely outside the main memory itself, thus causing no extra switching hardware within the main memory array.

This scheme, due to its inherent flexibility, appears to be extremely well suited to provide for fault-tolerance in memories which consist of chips with very large bit capacity: Especially in such memory chips internal faults- either originating from fabrication or stochastically occuring during memory operation- have to be taken into account. On the one hand it is hardly possible to care for such internal hard faults by providing extra switching mechanisms inside the chip due to the very large number of different potential fault patterns that had to be cared for and due to the resulting large additional hardware effort. On the other hand, it would be extremely wasteful to deactivate an entire chip only because of a few faulty bit cells. This especially holds for the cases where, e.g., as may be expected for the future, entire bit slices of a large capacity memory are realized by single chips or an entire memory is implemented on a single wafer. For such memories it is decisively important to be able to easily deactivate arbitrary parts of the bit cell array on the wafer. Especially in these cases, our method to define the units of reconfiguration at a quasi-logical level and to perform all necessary reconfiguration operations outside the memory bit cell array appears to be a very elegant solution to the problems of limited reconfigurability of faulty regions in VLSI/WSI chips. Our scheme thereby provides in a uniform way means to deal with
- faults caused by imperfect production, so that partially defective chips nevertheless may contribute to the production yield
- faults occuring during normal memory operation.

6. CONCLUSION

In this paper the realization of a fault-tolerant stand-by reconfiguration scheme by a VLSI chip was discussed in some details. The basic idea of the scheme is that the reconfiguration of the memory in case of faults is totally carried by a network at the memory interface. This network is dynamically controlled by a table storing the fault status of the memory. Compared to conventional stand-by techniques, the introduced scheme has several considerable advantages, especially concerning the resulting memory reliability and the flexibility to adapt the scheme to an already existing memory. Due to the regular structure of the reconfiguration hardware, it was implemented by one VLSI chip that can be modularly added to the memory interface. The architecture of the chip was described in details. Schemes to operate the chip in arbitrary memory systems were discussed. The advantages of the implemented scheme for the reconfigurability especially of VLSI/WSI memoris were demonstrated.

REFERENCES

[1] Aichelmann, F.J., Fault-tolerant design techniques for semiconductor memories, IBM Journal Res. Devel. Vol. 18, March 1984, p. 177-183
[2] Borgerson, B.R., Freitas, R.F., A reliability model for gracefully degrading and stand-by sparing systems, IEEE Transact. on Computers, C-24, p. 517-525
[3] Chen, C.L., Hsiao, M.Y., Error-correcting codes for semiconductor memory applications, IBM Journal Res. Develop. Vol.18, March 1984, p. 124-134
[4] Goldberg J. et al., An organization for a highly survivable memory, IEEE Transact. on Computers, C-23, p. 691-705
[5] Grosspietsch, K.E., Kaiser, J., Nett, E., A dynamic stand-by system for random access memories Proc. Int. Symp. Fault-Tolerant Computing FTCS-11, Portland 1981, p. 268-270
[6] Grosspietsch, K.E., Kaiser, J., Nett, E., A dynamic reconfiguration scheme for stand-by memory systems Digital Processes, 6(1980), p. 257-270
[7] Losq, J., Influence of fault-detection and switching mechanisms on the reliability of stand-by systems, Proc. Int. Symp. Fault-Tolerant Computing FTC-5, Paris 1975, p. 81-86
[8] Peterson, W.W., Weldon, E.J., Error correcting codes, MIT Press Cambridge, Mass. 1972
[9] Sarrazin, D.B., Malek, M., Fault-tolerant semiconductor memories (Survey paper), Computer, August 1984, p. 49-56
[10] Siewiorek, D.P., Swarz, R.S., The theory and practice of reliable system design, Digital Press 1982
[11] Stiffler, J.J., Coding for random access memories IEEE Transact. on Computers, C-27, p. 256-531
[12] Tanner, R.M., Fault-tolerant 256k memory designs IEEE Transact. on Computers, C-33, p. 314-322

5.3

ORTHOGONAL MEMORY - A STEP TOWARD REALIZATION OF LARGE CAPACITY ASSOCIATIVE MEMORY

Akio Kokubu, Minoru Kuroda* and Tatsumi Furuya
Electrotechnical Laboratory
Sakura-mura, Niihari-gun, Ibaraki 305, Japan
* Matsushita Electric Works, LTD., Kadoma, Osaka, Japan

A one kilo-bit memory with orthogonal access control is designed and fabricated using the nMOS process to show that large capacity associative memories can be implemented easily with available silicon technology. The orthogonal memory is integrated on a single chip by adding further access control gates to the six-transistor static RAM cell, and is realized with 11,300 transistors on a square of 4.8 mm by 4.8 mm by using the 5 micron design rule. An associative processor is proposed as an application of orthogonal memories. To realize this, an LSI switching element is designed and fabricated.

1. INTRODUCTION

Realization of large capacity associative, or content-addressed, memories on the scale of the main memory has been a dream of present-day computer architects. This is because a revolutionary change in computer architecture and a drastic improvement of system performance can be expected with associative memories, especially in such applications as database processing, image processing, and computer aided design.

Progress in VLSI technology has been making the implementation of large capacity associative memories on silicon chips possible. Recently, several associative memories whose bit capacity is a few kilo-bit have been fabricated and published in the literatures [1,2] to let computer architects reconsider their applications.

In applications of associative memories, various logical and arithmetic functions ranging from simple matching to complex functions such as 'greater than', 'less than', 'range', 'maximum', 'minimum', and 'average' are required. Integration of such complex functions will consume area on a VLSI chip in a certain degree and result in a poor bit density. This is why associative memories were not appreciated by semiconductor industry people in the past.

From the chip designer's point of view, it is considered that there are different approaches to realize associative memories on a VLSI chip depending on the complexity of integrated functions and the capacity of integrated bits, because the chip area is specified as a constant by such factors as cost and yield.

One approach is to integrate as many complex functions as possible on a chip (rather than bits). The compare and arithmetic functions are usually included in each cell in this approach. Integration of many complex functions results in a small number of bits. This can be found in the fully parallel architecture [3] where complex functions unloaded from CPU to an associative memory are performed in a highly parallel manner. This is called a complex associative chip approach.

The other approach is to integrate as many bits as possible on a chip (rather than functions). Integration of many bits can result in poor on-chip functions. In this approach, the compare and arithmetic functions are usually not included in each cell, and most associative functions are performed off-chip. This makes the circuit simple and allows great operational flexibility. This can be found in the word parallel and bit serial architecture [4] where complex functions unloaded from CPU to a dedicated associative processor (and an associative memory) are performed in a partially parallel or serial manner. This is called a simple associative chip approach.

In general, associative memories have been treated as a special dedicated hardware further attached to a main memory. It has been required that software must use both the main and the associative memories efficiently. From the system designer's point of view, there are some trade-offs between the improvement of system performance and the difficulty of software development when such special hardware as associative memory is used in computer systems. Highly parallel execution of complex functions in associative memories may lead to high system performance, depending on how many parallel operations are executed in programs. However, efficient software which utilizes such special hardware is hardly developed, and inefficient software may lead to low system performance.

In the case of the simple associative chip approach, we could use the memory as a main memory as well as an associative memory (when the memory capacity is large). The off-chip operations performed by CPU (or a dedicated associative processor) and the operational flexibility of the simple associative chip make software development relatively easy.

We adopt the simple associative chip approach in this paper because software is the most important factor in modern computer design, and it has a great effect on system performance in the user's environments. Therefore, the orthogonal memory proposed here is a simple associative chip and a word parallel bit serial type memory.

An orthogonal access concept was proposed for the first time in [5]. This concept was unveiled on the market in a new class of computer called STARAN in early 1970 [6]. The STARAN is classified as an associative processor and was expected to be used as a dedicated computer in the field of database processing and image processing to increase throughput rates drastically. The STARAN architecture provides both word slice and bit slice accesses to memory array. This memory array, an associative array module, consists of conventional memory chips and a switching network. In STARAN, data is logically stored on a memory array in a normal matrix form, and data is physically stored on the memory chips in a skewed matrix form. Normal to skew transformation is performed with the switching network. The orthogonal memory is a memory which provides both word slice and bit slice accesses without using such transformation.

2. DESIGN OF THE ORTHOGONAL MEMORY

2.1. Orthogonal Access

The orthogonal memory has two access modes which are orthogonally related as shown in Figure 1. One is the word slice access mode which is the same as that used in the conventional RAM. The other is the bit slice access mode by which the word parallel bit serial operation can be performed.

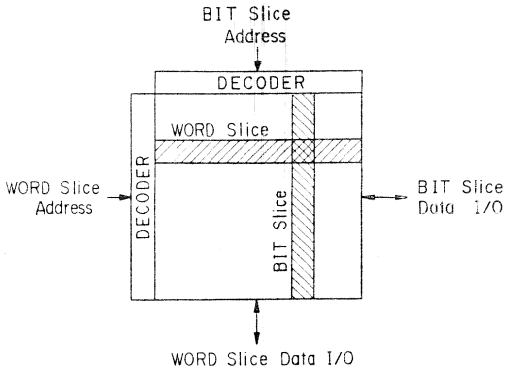


Figure 1 Orthogonal Access

In word slice access, the word address is given to the decoder and the selected word lines are activated to high. Then the word data can be read from or written to the bit lines. On the other hand, bit slice access decodes the address to the bit lines. When the bit address is given to the decoder, the selected bit lines are activated to high. This makes the bit data to be read from or written to the word lines.

Figure 2 shows a conceptual block diagram of a memory chip with orthogonal control. Organization of the orthogonal memory is 32 words by 32 bits because the bit capacity is specified as one kilo-bit by both the design rule available and the chip area given.

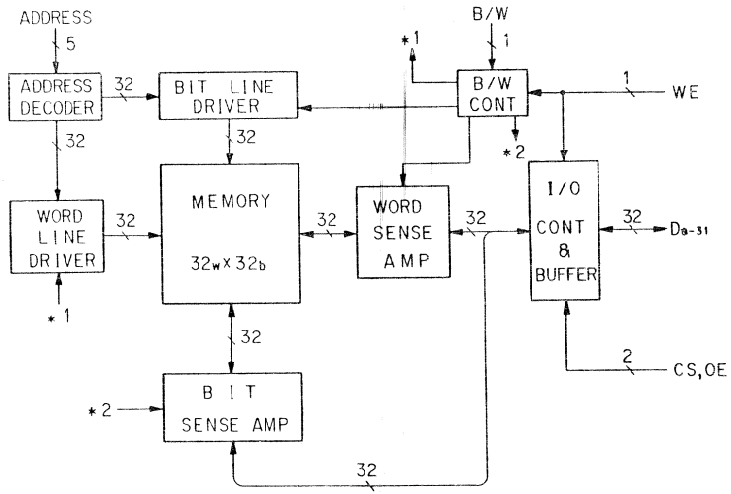


Figure 2 Conceptual Block Diagram

Five address lines enter the address decoder which drives the decode lines. Each decode line is connected to the word line driver and the bit line driver. If the mode is word slice access, bit line drivers are disabled by the control signals generated at the bit/word control circuit (B/W CONT) when the decode lines are driven. Similarly, the word line drivers are disabled if the mode is bit slice access.

Read/write operations are performed using the sense amplifiers and the I/O control buffers (I/O CONT and BUFFER). A sense amplifier is provided for each word line and bit line. Outputs of word slice and bit slice accesses are controlled by signals generated at B/W CONT. Chip select (CS) and output enable (OE) can be used to expand word and bit sizes.

2.2. Orthogonal Memory Cell

Figure 3 shows the orthogonal memory cell which consists of a flip-flop circuit with four transfer gates. Compared with the conventional static RAM cell, two transfer gates (Q6 and Q8) are added for bit slice access to drive the word lines. They are paired with transfer gates Q5 and Q7, and are connected to the flip-flop circuit. It is noted that the gate of Q7 is connected to the further word line (0) paired with the word line (1). This makes orthogonal control possible.

In bit slice access, bit lines (0) and (1) are activated to high, and transfer gates Q6 and Q8 are turned on. When the cell state is held as A = high and B = low, the word line (1) is activated to high via transfer gate Q6. Transfer gate Q5 is also turned on by activation of the word line (1). The cell state, nevertheless, is not affected because A is already high. The word line (0) is not activated to high via transfer gate Q8 because B is low. Therefore, transfer gate Q7 is not turned on and the cell state is not affected.

In word slice access, word lines (0) and (1) are activated to high, and transfer gates Q5 and Q7 are turned on. Operation of the memory cell proceeds in the same way as bit slice access.

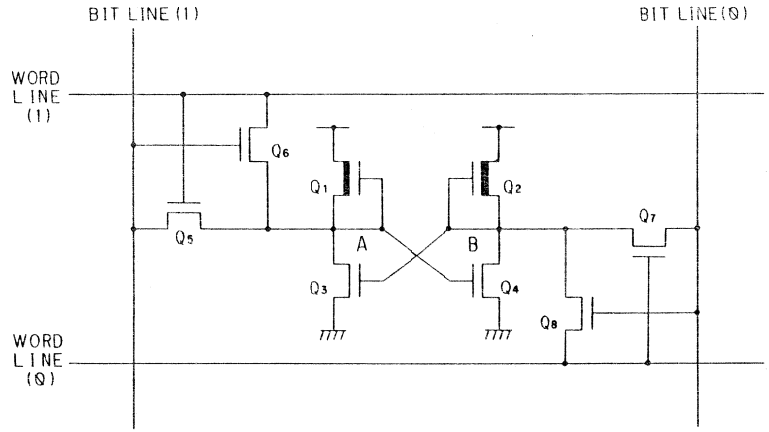


Figure 3 Orthogonal Memory Cell

2.3. Orthogonal Control

The circuit components distributed around the one bit memory cell are shown in Figure 4. In the chip, 32-by-32 cells are arranged as an array. Therefore, there are 32 combinational rows of the word line driver, the bit sense amplifier and buffer, and the bit write driver. Similarly, there are 32 combinational columns of the bit line driver, the word sense amplifier and buffer, and the word write driver.

The word line driver activates two word lines through the tri-state super buffers. The word lines run horizontally along a cell row to the bit sense amplifier and buffer. Similarly, the bit line driver activates two bit lines through the tri-state super buffers. The bit lines run vertically along a cell column to the word sense amplifier and buffer. A read/write line in the row is connected to a corresponding read/write line in the column to form a common data I/O line. This makes the number of pads in the chip small.

In bit slice access, the cell column is selected by activation of the bit lines, and a set of cell states are read from or written to the word lines. Output of the word line drivers is set to high impedance state by disabling the word slice enable signal. For read operation, the word line signal from the selected cell column is amplified by the differential bit sense amplifier and read through the tri-state super buffer enabled by the bit read enable signal. The tri-state super buffer of the word sense amplifier is disabled simultaneously because the data I/O line is common. For write operation, the word line signal enabled by the bit write driver is written to the selected cell column. The word write driver is disabled simultaneously because the data I/O line is common.

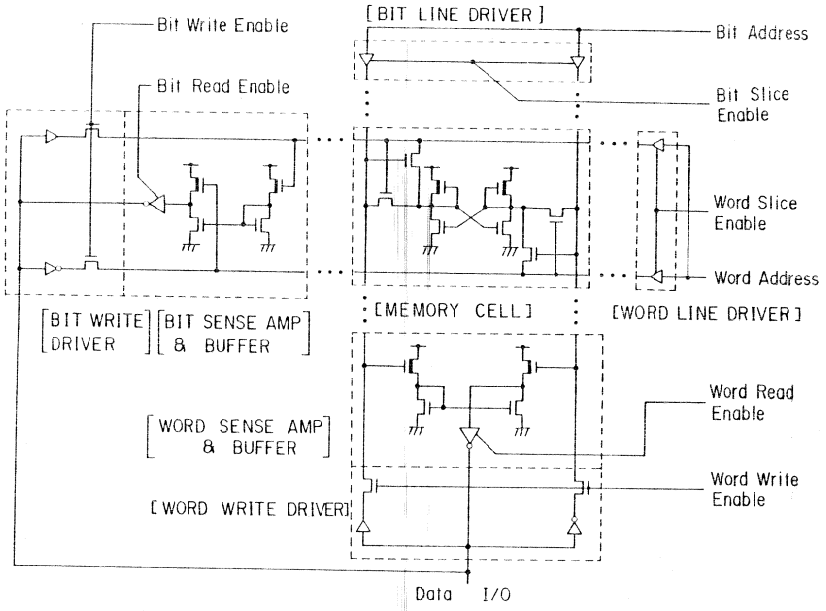


Figure 4 Orthogonal Control

2.4. Layout

Our floor plan is division of the memory into four segments and arrangement of the address decode lines inside the array. These segments are the higher word and the lower word parts in word direction, and the higher bit and the lower bit parts in bit direction. The layout pattern of the chip according to the floor plan is shown in Figure 5.

If the memory is not divided into segments, the following will result:
(1) Five bit address lines will be decoded; two groups of 32 decode lines which are connected to both word line drivers and bit line drivers will occupy a large area in the chip.
(2) Outputs of sense amplifiers will be 32 bits each in bit and word directions; two groups of 32 output lines which are connected to the 32 common tri-state pads occupy a large area in the chip.

Figure 6 shows a block diagram of the divided memory organization. Two groups of word drivers are provided for the higher word and the lower word parts, respectively. Two groups of bit drivers are also provided for the higher bit and the lower bit parts, respectively. Four address lines (A0-A3) are connected to the address decoder which activates one of 16 decode lines. The fifth address line (A4) is used to select a driver group.

Bit sense amplifiers are connected to the higher word and the lower word parts. Word sense amplifiers are connected to the higher bit and the lower bit parts. A common I/O line connected to the outputs of row and column is led to a pad. The tri-state super buffer which is controlled by an external signal is used with the pad.

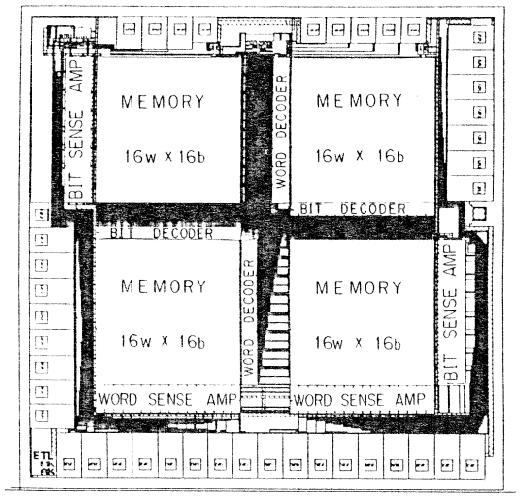


Figure 5 Layout Pattern

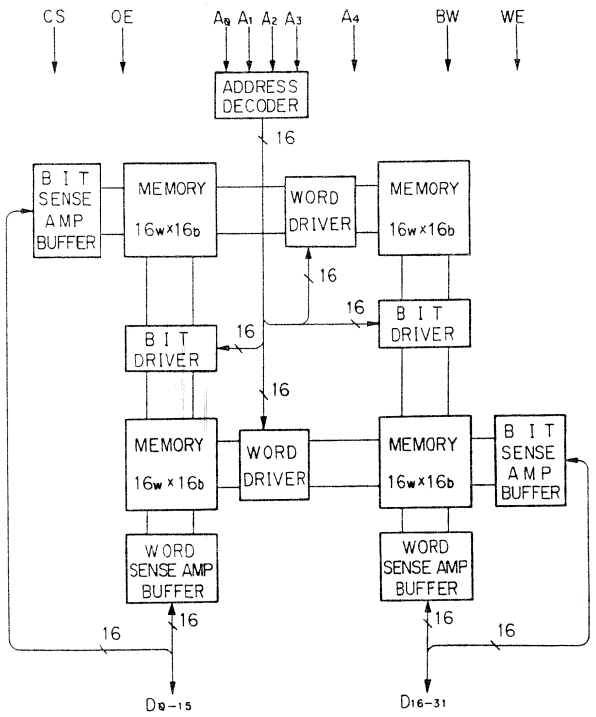


Figure 6

Real Block Diagram

3. EVALUATION OF THE ORTHOGONAL MEMORY

The chip described in previous sections was designed according to the 5 micron rule and fabricated using the nMOS process. The chip size was specified as 4.8mm X 4.8mm by Multi Project Chip fabrication. The total number of pins is 43. The memory cell, the word line driver, and the word sense amplifier and word write driver are 90 micron X 90 micron, 160 micron X 90 micron, and 290 micron X 90 micron in size, respectively. The chip consists of 11,303 transistors in total. The number of transistors in each block and their percentages are shown in Table 1.

Table 1. Number of transistors in each block of the orthogonal chip

Function block	Transistors	Percentages(%)
Memory array	8,192	73
Decoder (word)	352	3
(bit)	352	3
Sense (word)	832	7
(bit)	832	7
Address	208	2
Control	535	5
Total	11,303	100

The orthogonal memory designed can be compared with the conventional static RAM with the same design rule. The number of transistors calculated for the static RAM is shown in Table 2. The percentages in Table 2 are normalized with the total orthogonal memory transistors. We find that the static RAM is designed with 63% of the transistors used in orthogonal memory design, or the orthogonal memory is designed with 158% of the transistors used in static RAM design.

Table 2. Number of transistors in each block of the static RAM

Function block	Transistors	Percentages(%)
Memory array	6,144	54
Sense amplifier	832	7
Address decoder	208	2
Total	7,184	63

4. DISCUSSIONS ON ORGANIZATION AND APPLICATION

4.1. Orthogonal Memory Chips with the Advanced Process

The chip described was fabricated using the rather old 5 micron design rule. A submicron process which will make it possible to increase chip capacity will be surely available in the near future. Possible organizations of VLSI orthogonal memories using the submicron process can be considered as follows:

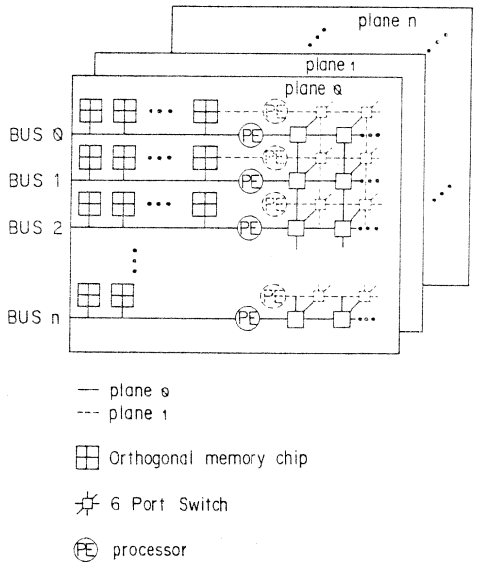


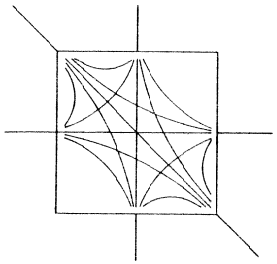
Figure 7 Associative Processor

One organization is a chip with large word and bit sizes. It will be possible to fabricate a 256 x 256 bit (or, even a 512 x 512 bit) orthogonal memory with the most advanced process. This type of memory organization is suitable for storing an image in an easy way and for processing the image in a short time. With present packaging technology, however, the number of pins can not be increased to the number corresponding to such a size. Therefore, decreasing I/O lines to keep the number of pins small is required, and a large bit width processing unit must be considered for integration on the same chip.

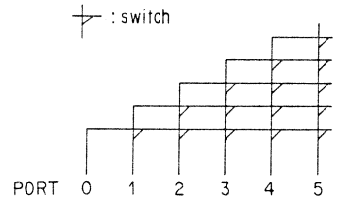
The second type of organization integrates as many relatively small size orthogonal memories (such as those described in previous sections) as possible on a chip. The small size (32 x 32 bit) orthogonal memory is considered as one of memory blocks in this type. A memory cell is accessed by a memory block number and an address within the memory block. This type of organization can be used in such applications as a fast table storage for database systems.

4.2. Partitionable Orthogonal Memory

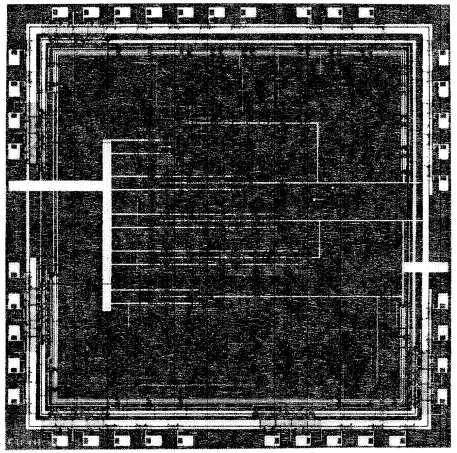
An associative processor system with orthogonal memory is proposed for multi-dimensional information processing such as image processing and computer aided design. Figure 7 shows the block structure of total system. A line of orthogonal memories are connected to a bus, and all of the buses are connected to a switching network. The system operates either in SIMD mode or Multi-SIMD mode. The whole memory is partitioned to many groups (each group consists of multiple buses) by the switching network. Our interest is to process pyramid data structures in image processing using Multi-SIMD execution.



(a) Interconnection



(b) Switching Structure



(c) Photograph

Figure 8 Switching Element

The switching network consists of a large number of switching components. A switching component consists of the 6-port switching elements shown in Figure 8(a). This element enables the connection between any two arbitrary ports, and is fabricated in the Multi Project Chip. Figure 8(b) shows the switching structure of the chip. Figure 8(c) is a photograph of the chip.

A line of orthogonal memories connected to a bus will be integrated on a VLSI chip. The three dimensional circuit technology can be applied to construct the switching network because a number of vertical interconnections are suitable for the three dimensional structure [7].

5. CONCLUSION

A one kilo-bit memory with orthogonal access control was designed and fabricated using the nMOS process to show that large capacity associative memories can be implemented easily with available silicon technology. The orthogonal memory was integrated on a single chip by adding further access control gates to the six-transistor static RAM cell, and was realized with 11,300 transistors on a square of 4.8 mm by 4.8 mm by using the 5 micron design rule. Organization of the orthogonal memory is 32 words by 32 bits. Dividing the cell array into 4 segments made the chip area small, and kept the number of transistors only 1.58 times that of the one kilo-bit static RAM with the same design rule.

An associative processor for multi-dimensional information processing was proposed as an application of orthogonal memories. To realize this, an LSI switching element was designed and fabricated.

We found that a large capacity orthogonal memory chip would be available soon with the advanced fabrication process. Realization of such large scale orthogonal memories will have an enormous impact on the future of computer architecture.

REFERENCES

[1] Kodata, H., Miyake, J., Nishimichi, Y., Kudo, H. and Kagawa, K., An 8Kb Content-Addressable and Reentrant Memory, International Solid-State Circuits Conference, Digest of Technical Papers, pp.42-43, 1985.
[2] Ogura, T., Yamada, S., Tanno, M. and Ishikawa, K., A 4 Kb Associative Memory LSI, IECE Technical Report, SSD83-78, pp.45-52, 1983 (in Japanese).
[3] Ogura, T., Nikaido, T. and Miyahara, N, A 1-kbit Associative Memory LSI, IECE Technical Report, EC80-44, pp.13-21, 1980 (in Japanese).
[4] Foster, C.C., Content Addressable Parallel Processors, Van Nostrand Reinhold, New York, 1976.
[5] Shooman,W., Parallel Computing with Vertical Data, Proc. EJCC, pp.111-115, 1960.
[6] Batcher, K.E., STARAN Parallel Processor System Hardware, AFIPS Conf. Proc. (Proc. NCC), Vol.43, pp.405-410, 1974.
[7] Fernandez, D.S., Three-dimensional Integration, VLSI Design, Vol.V, No.4, pp.82-85, April 1984.

Session 6: Hardware Algorithms