A 10-ns 54×54-b Parallel Structured Full Array Multiplier with 0.5-μm CMOS Technology

Junji Mori, Masato Nagamatsu, Masashi Hirano, Shigeru Tanaka, Makoto Noda, Yoshiaki Toyoshima, Kazuhiro Hashimoto, *Member, IEEE*, Hiroyuki Hayashida, and Kenji Maeguchi

Abstract—**This paper describes a 54×54-b multiplier fabricated by double-metal 0.5-μm CMOS technology. The 54×54-b full array has been adopted to complete multiplication within one latency. A 10-ns multiplication time has been achieved by optimizing both the propagation time of the part consisting of 4–2 compressors [1] and the propagation time of the final adder part. The n-channel pass-transistor circuit and the p-channel load circuit have been employed at the critical blocks to improve the multiplication speed. This multiplier is intended to be applied to double-precision floating-point data processing based on the IEEE standard up to a clock range of 100 MHz.**

I. Introduction

RECENTLY, the performance of microprocessors has been improving rapidly. This is because the integration density has increased by the advancement of semiconductor technology, and large-scale circuits can be built in one-chip silicon. As is seen in a RISC-type processor, not only the CPU core but also cache memory and FPU are put on a single chip. This technology enables high-speed processing for pipeline and parallel processing.

Along with improved performance of microprocessors, it is essential to improve the performance of the floating-point unit. In particular, the multiplier has an important role to play. High-speed floating-point operation has been required for digital signal processing, circuit simulation, image processing, and so on. In many cases, the system performance depends on the floating-point multiplication time of the mantissa. To enhance the performance of the multiplier, a parallel architecture is desired to reduce the propagation delay.

This paper describes a parallel structured floating-point multiplier macro. This multiplier has adopted a 54×54-b full array, and optimized the speed of the array part and the final adder part. The authors also used pseudo-CMOS circuits for important blocks to shorten the propagation time.

Section II describes the required performance of multipliers to be used in future high-speed processors. The basic structure of this multiplier is given in Section III. Section IV describes the timing analysis (optimizing the array part and the final adder part) to improve the total multiplication speed. In Section V, the pseudo-CMOS circuits employed in this multiplier are explained. Process technology and fabrication are described in Section VI. The evaluation and chip characteristics are described in Section VII. Section VIII is the conclusion.

II. Target Performance

The performance required for the multipliers, which will be used in future processors, is described.

In a RISC architecture, the hardware is simplified compared with the CISC architecture, and the clock cycle has been shortened. In pipeline processing, all stages execute at the same time, thus a multiplier requires a high-speed equivalent to other macros such as a floating-point ALU, cache memory, and control block. One of the targets for the speed of a next-generation processor is a 100-MHz operation clock speed. The authors have established a 10-ns multiplication speed.

The second requirement is a double-precision multiplier. Since a single-precision operation is less accurate than an integer operation, generally many floating-point operations are performed by double precision. It is obvious that floating-point multiplications in many C compliers are performed using double precision to prevent loss of accuracy. If double precision is performed by using a single-precision multiplier, using the multiplier multiple times is unavoidable. This results in slower processing speed. Based on the IEEE standard, the mantissa of double-precision data has 52 b (including the hidden bit). Since this multiplier adopts the Booth's algorithm, the sign bit is needed. Moreover, one guard bit is added, and the total becomes 54 b.

The third requirement is that multiplication is performed in one latency. For many RISC-type processors, mainly with pipeline processing, a multiplication instruction has some latency. Pipeline processing is an effective means to increase the operating frequency without deterioration in visible processing ability. Ordinarily, the multiplier is divided into several stages. However, latency will then occur, which will cause the regular pipelining to break. Furthermore, having such latency, if an operation result is used for a successive operation, the next operation has to wait until the previous operation finishes. Moreover, to prevent register conflicts, a procedure, such as a score boarding, is required. One latency is even more essential for a processor that can process some instructions concurrently. A 54×54-b full array multiplier must operate with one latency for multiplication. A 53×27-b

Manuscript received September 12, 1990; revised December 14, 1990.

The authors are with the Semiconductor Device Engineering Laboratory, Toshiba Corporation, 1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki 210, Japan.

IEEE Log Number 9042540.

0018-9200/91/0400-0600$01.00 ©1991 IEEE

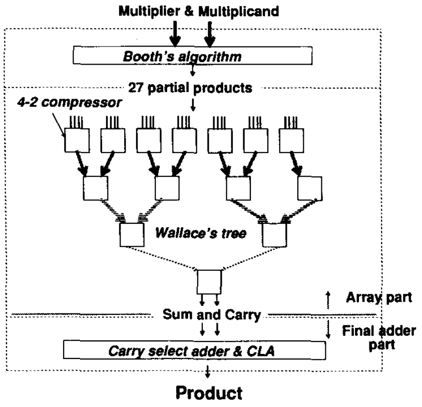


Fig. 1. Structure of this multiplier.

half array has been used in a conventional double-precision multiplier [2]. This array needs two latencies because the product is obtained by flowing through the array twice.

To meet these requirements, the authors propose here a 54×54-b double-precision full array multiplier using a parallel architecture with 0.5-μm CMOS technology.

III. Basic Structure

The structure of this multiplier is shown in Fig. 1. This multiplier uses a parallel structure with Booth's algorithm and Wallace's tree.

By applying the Booth algorithm [3], the number of partial products is halved. Since this multiplier performs 54-b multiplications, 27 partial products are generated.

By adopting the Wallace tree [4] and a 4–2 compressor, only four addition stages are needed in order to add 27 partial products. This addition is performed in the form of a tournament by using the Wallace tree method. The addition of the partial products used the 4–2 compressor, which can sum up four partial products concurrently. The 4–2 compressor can add without propagating the carry to a higher position, and it generates a 108-b sum and carry.

Finally, the 108-b sum and carry is added with the 108-b carry propagation. This adder consists of a 4-b carry look-ahead adder and a 16-b carry select adder to propagate the carry at high speed.

A multiplier like this is divided into two parts. The first part is the partial-product generation and the carry-save tournament addition part. The second part is the carry propagation part which adds the 108-b sum and carry in the last stage. For convenience's sake, let us call the former "the array part" and the latter "the final adder part."

An ordinary multiplier which is equipped in processors normally divides these two parts into two pipeline stages with a pipeline register insertion in the middle. Therefore, the multiplier has at least two latencies. However, this multiplier does not have a pipeline register. These two parts are performed in one cycle to overlap the propagating time of the two parts.

IV. Analysis of Multiplication Time

The authors analyzed the propagation delay time as shown in Fig. 2, designed a multiplier to complete the 54-b multiplication in 10 ns, and estimated the propagation delay for each block.

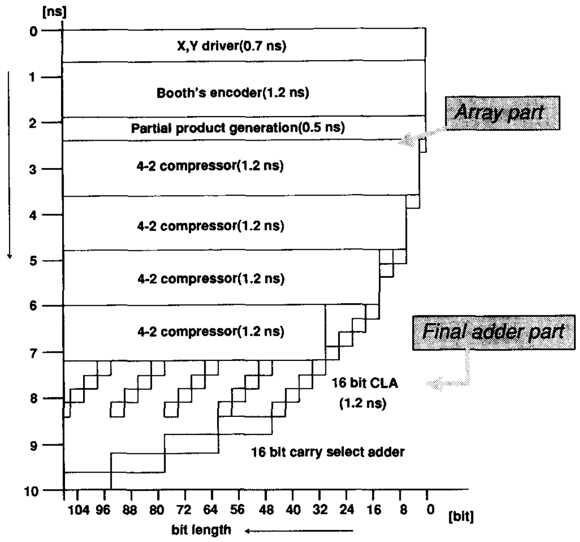


Fig. 2. Analysis of multiplication speed (1).

As is the case for this multiplier, the total operation time is shortened because of the overlapping propagation time for the array part and the final adder part since the array part and the final adder part are not divided by the pipeline register. Note the output of the array. Due to the tournament addition of the Wallace tree, the number of levels of addition varies by the bit position. This is because the partial product shifts by some bits to the left at each level. The output of the array is 108 b, and only more than the thirtieth bit in among these 108 b has to pass through the four levels of the 4–2 compressor. The other bits have a speed capacity equivalent to one level of the 4–2 compressor or more. Therefore the final adder part can begin to operate prior to the finish of the array part operation.

However, there is one cautionary note. The authors used a 16-b carry select adder and a 4-b carry look-ahead adder in the final adder part to perform high-speed carry propagation. Since the propagation delay is different in the array part, time loss occurs if these do not match the bit area.

Therefore the final adder is shifted by 2 b to match the array so that the bit position can be matched with the 4-b CLA. Fig. 3 shows the boundary analysis. The carry propagation time of the 16-b carry select adder is optimized within the delay time of the 4–2 compressor. It is required that the carry propagation time in the least significant 30 b of the product is equalized to the delay time of the 4–2 compressor. By using pseudo-CMOS circuits for these blocks, the propagation delay of each block was optimized and the loss time became 0 ns.

V. Circuit Design

In Fig. 2, it is obvious that the most affected areas are the 4–2 compressor and the final adder part. To increase the speed of these two circuits, the authors decided to apply pseudo-CMOS circuits. To be precise, an n-channel pass transistor was used for the 4–2 compressor. Furthermore, a

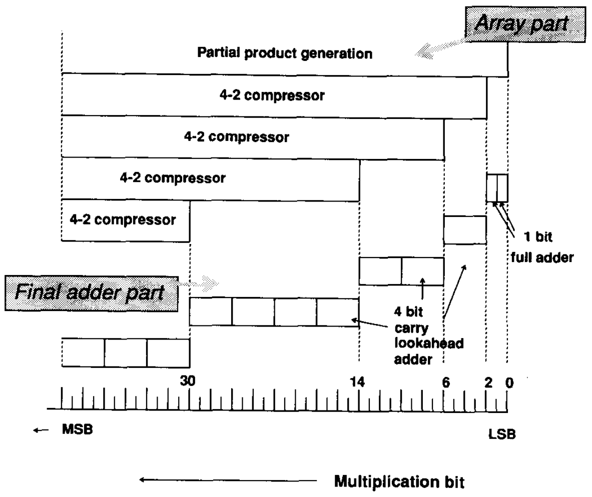


Fig. 3. Analysis of multiplication speed (2).

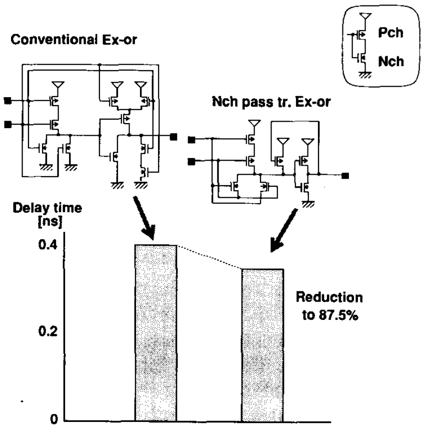


Fig. 4. Effect pass transistors.

p-channel load circuit was used for the CLA to attain higher speed.

An n-channel pass transistor was used for the XOR circuit, as shown in Fig. 4, to shorten the propagation delay of the 4–2 compressor. This XOR consists of only seven transistors. For comparison, the conventional circuit has ten transistors. The propagation delay of this XOR was 0.35 ns by circuit simulation. The XOR takes 0.4 ns when it is designed by using CMOS circuit. The propagation delay has been shortened to 87.5%, and a faster operation speed can be expected.

However, if signals are propagated by the n-channel transistor only, the level will not be sufficient. This case occurs when the inputs are (1,1). The output is raised by using the p-channel to eliminate such a problem. This rise is to prevent a through current, and is kept small so that it does not affect the speed directly.

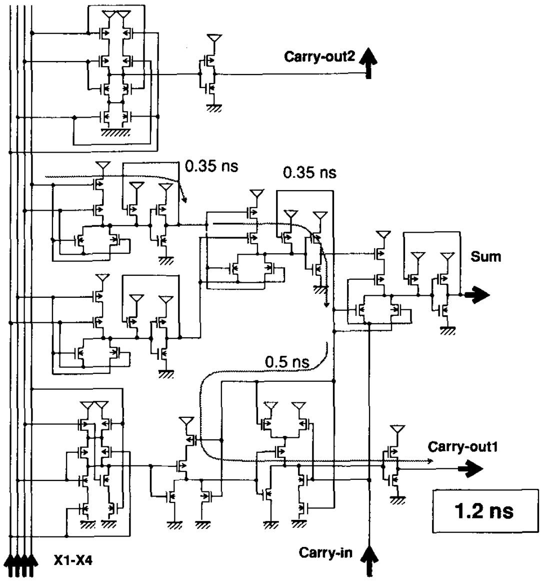


Fig. 5. 4–2 compressor circuit.

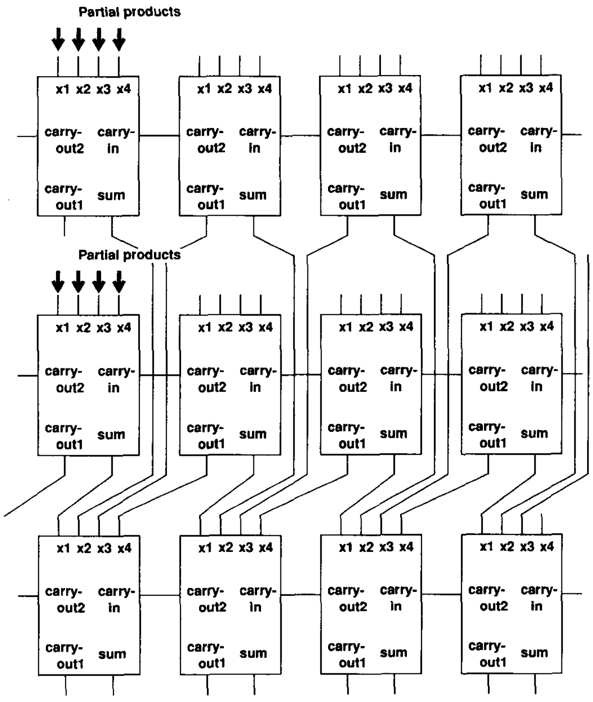


Fig. 6. Interconnection of 4–2 compressors.

The 4–2 compressors are used to speed up the array part during circuit design. Fig. 5 shows the circuit diagram of the 4–2 compressor. Carry-out2 is connected to next 4–2 compressor's carry-in. However, carry-out2 is never generated by carry-in. Fig. 6 is a diagram showing the interconnection between the 4–2 compressors in the array. It also shows that the carry is propagated by 1 b per one level of the 4–2 compressor, and is propagated to the next addition level while the carry is stored. Since the XOR operations for the individual two inputs of four are concurrently performed, four inputs can compress to sum at three XOR gate propagations. It takes four gate propagations when using an ordinary carry-save adder.

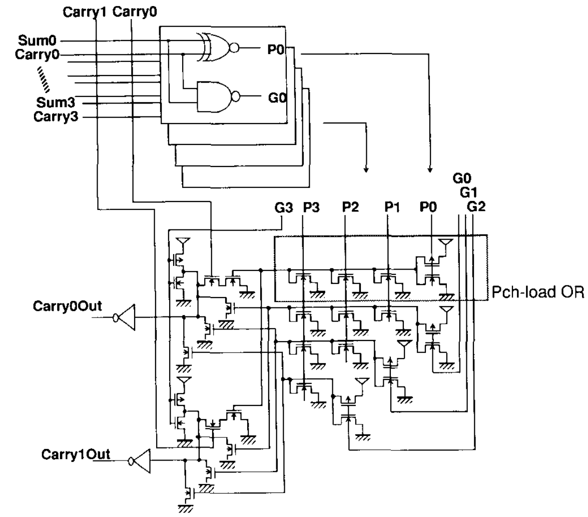


Fig. 7. Four-bit CLA circuit.

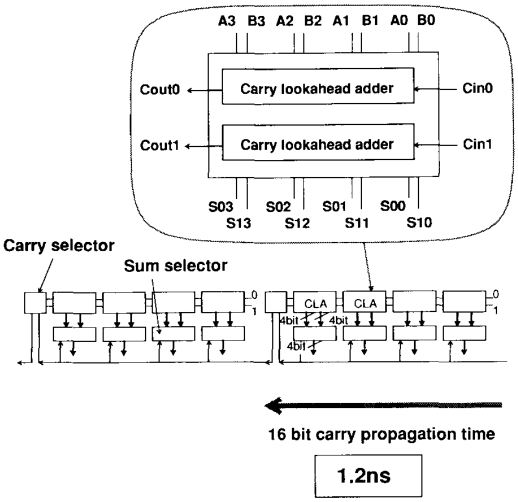


Fig. 8. Sixteen-bit CLA circuit.

The critical path of this 4–2 compressor is shown in Fig. 5 by a gray line. The propagation delay achieved was 1.2 ns.

A p-channel load circuit is used for the carry look-ahead adder (CLA) in the final adder to propagate the carry at high speed. As shown in Fig. 7, a multibit input is required for the OR circuits in the CLA circuit. The speed of this OR circuit greatly influences the propagation of the carry. Thus the p-channel load circuit was adopted in the CLA. The power consumption increases by using this p-channel load circuit, since a through current sometimes flows. However, switching is made only at one n-channel transistor. This avoids totem-pole connected transistors, and a high-speed operation is maintained. The final adder part is constructed from a combination of 4-b unit CLA's. Since the carry propagation passes only one level of the OR circuit, the passing time was 0.2 ns by circuit simulation. As also shown in Fig. 7, the CLA circuit needs both carry propagation and generation signals in order to determine the gates of the CLA circuit. These signals are generated simultaneously on all bits, and it takes 0.4 ns.

When the carry generation and propagation for 16 b are calculated, 0.2 ns×4+0.4 ns* = 1.2 ns (*for generating both carry generation and propagation signals). Fig. 8 shows the structure of a 16-b carry select adder. This propagation time is equivalent to the passing time of the 4–2 compressor. Therefore, the lower 30 b could be overlapped perfectly with the array part, and could eliminate time loss.

Moreover, the authors have adopted the carry-select adder in the final adder. This method is the most efficient way to propagate the carry at high speed while the circuit increases. The carry propagation and sum generation circuits are needed twice in the carry-select adder. However, it is sufficient that the circuit for signals which propagate and generate the carry is one. The carry-select adder of this multiplier is constructed from 16-b units. The carry select time is 0.4 ns with signal buffering.

The authors analyzed the propagation delay as mentioned above, and optimized the array part and the final adder part, as shown in Fig. 2.

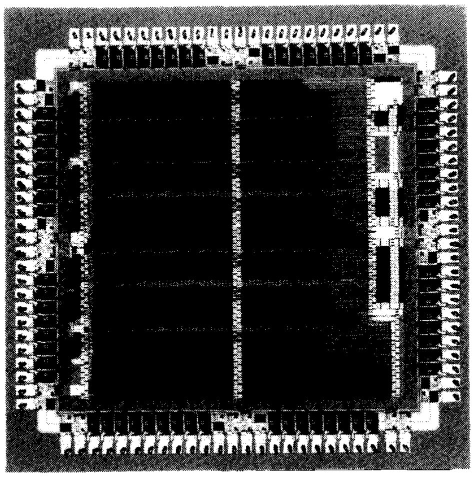


Fig. 9. Photograph of chip.

VI. Fabrication

The double-metalization technology has been used. By adding two feedthrough lines in the cell of the 4–2 compressor and placing the cell in a row, the complex interconnection in the Wallace tree has been constructed effectively in the multiplication array. Fig. 9 shows a photograph of the chip. The active size is 3.62×3.45 mm and 81 600 transistors are integrated in it. The layout is packed to the left. A large load is attached to both the multiplier and the multiplicand since the partial product is generated concurrently and in parallel. The circuit for the partial product is placed vertically straight to reduce this load.

The chip has been fabricated using 0.5-μm CMOS technology. Regarding increased speed, the effect of the 0.5-μm process is an important point in achieving the 10-ns multiplier. Table I shows the device characteristics. The choice of the power supply voltage requires consideration of the trade-off between speed, reliability, and power dissipation [5]. The

TABLE I
DEVICE CHARACTERISTICS

Process	Double metal (Al) 0.5-μm p-sub twin-well CMOS
Gate length	0.5 μm (n-channel) 0.6 μm (p-channel)
Gate oxide	11 nm
1st Al width space	1.0 μm 0.8 μm
2nd Al width space	1.2 μm 1.0 μm
Supply voltage	3.3 V

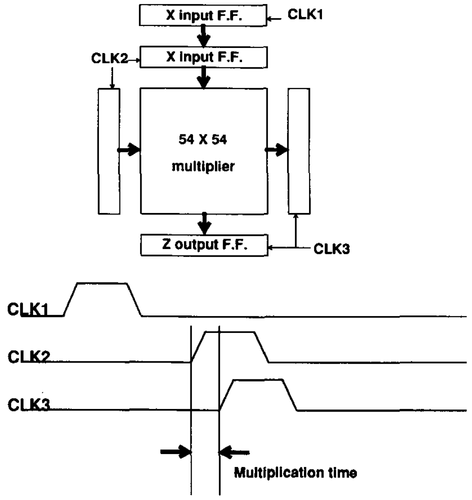


Fig. 10. Evaluation method.

problem with voltage is the restriction caused by the pressure resistance of the oxide film. A lower power supply voltage of 3.3 V has been employed in this chip. This I/O circuit has been made for 3.3 V, and a 3.3-V supply voltage is provided from outside of the chip.

VII. Evaluation

The evaluation method is shown in Fig. 10. Flip-flops were attached to the input and output for peripheral circuits. The operation time was from the edge of the input clock (CLK2) to the edge of the output clock (CLK3). The interval of these edges was gradually shortened to confirm if the result was a failure or not. Fig. 11 shows the schmoo plot at room temperature. As shown in the plot, a 10-ns multiplication time was achieved at 3.3 V.

Three thousand multiplication patterns were input as inputs for this measurement. These patterns included 2900 random patterns and the 100 intentional patterns propagating the carry from 0th bit to 108th bit. Examples of the intentional input patterns (this is worst case) are as below:

Multiplier	Multiplicand (Booth's decoder side)	
(∅x3F FFFF FFFF FFFF)	×(∅x00 0000 0000 0001)	← set all outputs to 1
(∅x00 0000 0000 0000)	×(∅x3F FFFF FFFF FFFF)	← critical pattern
(∅x3F FFFF FFFF FFFF)	×(∅x00 0000 0000 0001)	← set all outputs to 1
(∅x00 0000 0000 0000)	×(∅x3F FFFF FFFF FFFE)	← critical pattern.

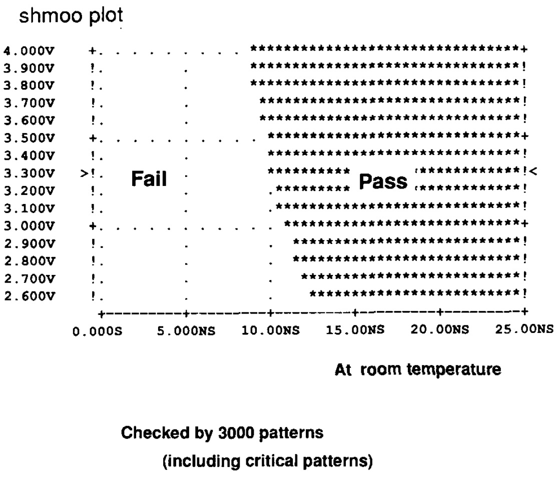


Fig. 11. Schmoo plot.

TABLE II
CHIP CHARACTERISTICS

Multiplier & multiplicand	54 b (with sign)
Product	108 b (with sign)
Multiplication time	10 ns
Power at 100 MHz	870 mW
Chip size	3.62×3.45 mm
Number of transistors	81600

The case in which the Booth's decoder makes only one partial product to be a minus one value has been able to become the critical input pattern. In the situation where only one partial product has an all-one value and the other partial products have a zero value, each of the nodes in the array is active since the array consists of XOR gates, and the input of the final adder is $(-1)+1$. The 108-b carry generation occurs in the final adder part. Random patterns were used to confirm timing analysis.

The power dissipation was 870 mW at 100 MHz. Standby power dissipation existed since a p-channel load circuit has been employed. The power dissipation was 33 mW at standby. The chip characteristics are shown in Table II.

VIII. Conclusion

The authors set a target to create a multiplier with 10 ns, double precision, and one latency. The authors have analyzed the timing of the partial product addition to achieve 10-ns multiplication speed. The speed of the array part and the final adder part have been optimized so that the carry select adder can propagate the carry most effectively. The array part has a circuit with existing 4–2 compressors with an n-channel pass transistor, and the final adder part has a p-channel load circuit. A 10-ns multiplication time has been achieved by using the pseudo-CMOS circuits and optimizing a multiplication speed with 0.5-μm CMOS technology.

Acknowledgment

The authors wish to thank T. Iizuka, H. Nakatsuka, H. Yamada, Y. Miyazawa, and H. Ushikubo for their encouragement and F. Umibe for reviewing and revising the original manuscript.

References

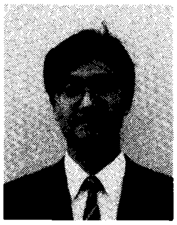
[1] M. Nagamatsu *et al.*, "A 15 ns 32×32-bit CMOS multiplier with an improved parallel structure," in *1989 CICC Dig. Tech. Papers*, 10.3.

[2] L. Korn and S.-W. Fu, "A 1,000,000 transistor microprocessor," in *1989 ISSCC Dig. Tech. Papers*, pp. 54–55.

[3] A. D. Booth, "A signed binary multiplication technique," *Quart. J. Mech. Appl. Math.*, vol. 4, part 2, 1951.

[4] C. S. Wallace, "A suggestion for fast multipliers," *IEEE Trans. Electron. Comput.*, vol. EC-13, pp. 14–17, Feb. 1964.

[5] M. Kakumu *et al.*, "0.5 μm gate 1M SRAM with high performance at 3.3 V," in *1989 VLSI Technology Dig. Tech. Papers*, pp. 63–64.



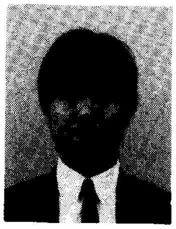
Junji Mori was born in Aichi Prefecture, Japan, on March 9, 1965. He graduated from Higashiyama Engineer High School, Nagoya, Japan, and from Toshiba Computer School, Kawasaki, Japan, in 1983 and 1984, respectively.

In 1984 he worked for Toshiba Computer School as an Instructor, and in 1985 he joined the Semiconductor Device Engineer Laboratory, Toshiba Corporation, Kawasaki, Japan. He has been engaged in research and development of CMOS graphic processors and high-performance logic LSI including microprocessors.

Masato Nagamatsu was born in Oita Prefecture, Japan, on July 29, 1959. He received the B.S. and M.S. degrees in electronic engineering from the University of Tokyo, Tokyo, Japan, in 1982 and 1984, respectively.

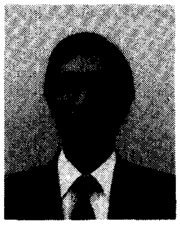
In 1984 he joined the Semiconductor Device Engineering Laboratory, Toshiba Corporation, Kawasaki, Japan. He has been engaged in the design of CMOS logic VLSI.

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



Masashi Hirano was born in Toyama, Japan, on March 2, 1967. He graduated from the Ohsawano Technical High School, Toyama, Japan, in 1985.

In 1985 he joined the Semiconductor Device Engineering Laboratory, Toshiba Corporation, Kawasaki, Japan. He has been engaged in the research and development of CMOS logic LSI's.



Shigeru Tanaka was born in Tokyo, Japan. He received the B.S., M.S., and Ph.D. degrees in physics from Tokyo University, Tokyo, Japan, in 1974, 1976, and 1979, respectively.

In 1980 he joined the Semiconductor Device Engineering Laboratory, Toshiba Corporation, Kawasaki, Japan. He has been engaged in the research and development of IIL devices, CMOS/SOS gate arrays, and CMOS graphic processors. He is now engaged in the research and development of high-performance logic VLSI including microprocessors as a Manager of the Logic Device Design Section.

Dr. Tanaka is a member of the Information Processing Society of Japan and the Institute of Electronics and Communication Engineers of Japan.

Makoto Noda received the B.S. and M.S. degrees in electronic engineering from Yamagata University in 1973 and 1975, respectively.

In 1975 he joined Toshiba Corporation. Since then he has been engaged in the research and development design of microprocessors, digital signal processors, and BiCMOS LSI's. He is now Manager of the Advanced Logic/Memory Technology Department, Semiconductor Device Engineering Laboratory, Kawasaki, Japan.

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

Yoshiaki Toyoshima was born in Abiko, Japan, on May 22, 1958. In 1981 he received the B.E. degree in electronic engineering from Waseda University, Tokyo, Japan.

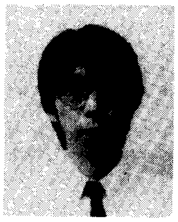
He joined the Semiconductor Device Engineering Laboratory, Toshiba Corporation, Kawasaki, Japan, in 1984. He has been engaged in the research and development of process/device technology for CMOS logic VLSI's. His main concern is MOS device physics, especially CMOS performance and reliability.

Mr. Toyoshima is a member of the Japan Society of Applied Physics.

Kazuhiro Hashimoto (M'90) was born in Tokyo, Japan, on July 21, 1951. He received the B.S. and M.S. degrees in metallurgical engineering from Waseda University, Tokyo, Japan, in 1975 and 1977, respectively.

He joined Toshiba Research and Development Center, Kawasaki, Japan, in 1977, and moved to a newly organized laboratory, the Semiconductor Device Engineering Laboratory, in 1979. In 1985 he worked for Hewlett-Packard Laboratories, Palo Alto, CA, as a Visiting Engineer, where he studied device characteristics of small-geometry SOI devices. He has been engaged in the research and development of device and process technology for MOS LSI memories such as MNOS, DRAM, and CMOS SRAM, and VLSI logic devices. Currently, he is interested in sub-half-micrometer CMOS technology, especially for high-speed and high-density applications.

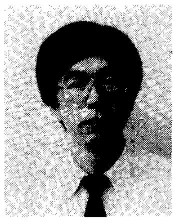
Mr. Hashimoto is a member of the Japan Society of Applied Physics and IEEE Electron Devices Society. He served as a subcommittee member of device technology for the International Electron Device Meetings (IEDM) in 1989 and 1990.



Hiroyuki Hayashida was born in Hiroshima, Japan, on August 29, 1961. He received the B.E. degree in 1985 and the M.E. degree in 1987, both in electronic engineering, from Hosei University, Tokyo, Japan.

He joined the Semiconductor Device Engineering Laboratory, Toshiba Corporation, Kawasaki, Japan, in 1987. He has been engaged in the research and development of process technology for CMOS logic VLSI's.

Mr. Hayashida is a member of the Japan Society of Applied Physics.



Kenji Maeguchi was born in Sapporo, Japan, on March 30, 1948. He received the M.S. degree in precision engineering from Hokkaido University, Sapporo, Japan, in 1973.

In 1973 he joined the Research and Development Center, and in 1979 the Semiconductor Device Engineering Laboratory, Toshiba Corporation, Kawasaki, Japan, where he was engaged in the research and development of MOS/SOS integrated circuits. He is now engaged in the development of CMOS/bulk integrated circuits.

Mr. Maeguchi is a member of the Japan Society of Applied Physics.