IV-3

A Circuit Design of 32KByte Integrated Cache Memory

Takayasu SAKURAI, Kazutaka NOGAMI, Kazuhiro SAWADA, Tsukasa SHIROTORI*,
Toshinari TAKAYANAGI, Tetsuya IIZUKA, Takeo MAEDA, Junichi MATSUNAGA,
Hiromichi FUJI, Kenji MAEGUCHI, Kiyoshi KOBAYASHI, Tomoyuki ANDO,
Yuki HAYAKASHI, Teruo MIYOSHI, and Kazuyuki SATO

TOSHIBA Corporation, * TOSHIBA Microcomputer Eng. Corp.,
1, Komukai-Toshiba-cho, Saiwai-ku, Kawasaki, 210 Japan

Introduction

A cache memory is effective in enhancing CPU system throughput[1]. A circuit design aspect of a newly developed integrated cache memory which includes 32Kbyte DATA memory, with a typical HIT delay of 18ns is described. The present memory achieves four times larger DATA memory size together with much faster operation speed compared with the recently reported integrated cache memory[2].

The memory includes 32Kbyte DATA (INSTRUCTION) memory, 34Kbit TAG memory, 8Kbit VALID flag, 2Kbit LRU flag, comparator, and CPU interface logic circuits. The inclusion of DATA memory is important in improving system cycle time as shown in Fig.1. It is also important for reducing board area and cost, because it replaces about ten LSI's.

Device Outline

The features of the device are listed on TABLE I. Newly proposed way-slice architecture is adopted. On-chip CPU interface circuit is confined to critical path logics in order to increase system design flexibility. The interface is programmed with Al masterslice for specific CPUs. The device also has a 32bit x 8Kword SRAM mode which is suitable for any 32bit CPU system. As for a process, 1.2μm design rule with 1.0μm NMOS gate poly length is used to optimize yield and speed tradeoff.

Double Word Line for Cache Memory

As a core architecture, double word line structure[3] is adopted as shown in Fig.2. Different from the double word line structure for a usual SRAM where only one section is activated at a time, three section are activated, namely a DATA section, a TAG+LRU section and a VALID section. This maximizes silicon efficiency and minimizes power consumption since only required bits are activated. Power consumption is the most severe design bottleneck in the integrated cache memory where at least 54bit should be read out at very high speed. The device utilizes an address transition detector and bit line equalibration signals are distributed hierarchically in accordance with the divided section scheme to increase speed.

Flush Clear Function

4T SRAM cell is suitable for VALID bit flush clear, being different from 6T SRAM cell. By inserting NAND gate to every word line as is shown in Fig.2, flush clear is achieved with minimum current peak caused by activation of all word lines. Without the NAND gates, the total word line capacitance of 1000pF are to be charged and discharged compared with 50pF with the NAND gates. The clear function is measured to take less than 50ns typically without any noise problems.

Sense Amplifier and Comparator

The design of sense amplifiers also effects the power and the speed. Latch-type sense amplifiers are rather slow as in DRAM designs, because the latch timing should wait with margins until all bit line signals are definite. Current-mirror type sense amplifies in Fig.3 are employed in this device, which is fast because sense operation is self-aligned and not timing dependent.

The problem with the current-mirror sense amplifier is its rather high power consumption. Figure 3 also shows the relationship between delay and power when tuning the MOSFET size. As seen from Fig.3, delay is almost saturated at 0.4mA, so that this value is adopted for a TAG sense amplifier. On the other hand, DATA sense amplifier is not required to be so fast as a TAG memory, because DATA is needed just after TAG read-out data is compared with TAG address. In order to minimize the power consumption without degrading the access speed, 0.2mA current dissipation is assigned for a DATA sense amplifier.

The operation speed of the current-mirror sense amplifier also depends linearly on the output capacitance which is designed to be as small as possible at both a circuit level and a layout level.

Figure 4 shows a new comparator circuit. The output node of the sense amplifier drives the C^2MOS gate which is the first stage of the comparator. This comparator does not consume static power, being different from the comparators formerly reported[2][4].

Output Buffer and Noise

Noise caused by inductance puts another constraint in fast multi-bit memory design. In order to achieve both high speed and low noise, DATA output buffer is placed just after the sense amplifier in this device, as shown in Fig.5. Then, the DATA output buffer directly drives the output transistors rather slowly through a long data bus. This small slew rate drive of the output transistor gate is good in reducing the L x di/dt noise at the output.

On the other hand, in the conventional SRAM, the sense amplifier drives a long bus line and after the long bus line, an output buffer is placed just before the output transistors. Since the output transistor should be driven slowly, the conventional design has two slow nodes, the long data bus line and the gate of output transistor. The present design merges these two delays in one and achieves faster operation still maintaining low inductive noise.

The present RAM has six V_{DD} and six V_{SS} pins. Separate V_{DD}–V_{SS} pairs are assigned for DATA output pins, a memory core and peripheral control circuits. This also eliminates malfunctions caused by the inductive noise.

Test Mode and Other Scheme

The device includes a test mode where all of the memory cells can be directly written and read out from outside. This mode is beneficial to search for the sparable cells and rows. The RAM includes 4 spare rows, where disabling of bad rows is physically achieved by blowing the word line fuse as is shown in Fig.2, without degrading the access time.

The cache memory does not have a low voltage data retention mode where the soft error rate is high. Since parity circuit is estimated to increase 15% in silicon area and 10% in access time, parity is not included.

Performance

Figure 6 shows a chip microphotograph of the integrated cache memory whose chip size is 8.5 x 9.9mm^2, and memory cell size is 7.9 x 13.9μm^2. Figure 7 shows the measured address to HIT and address to DATA shmoos. Typical address to HIT delay was 18ns and address to DATA delay was 23ns. Power consumption is measured typically 0.5W.

Acknowledgement

The authors would like to thank K.Suzuki, Y.Unno, Nishibe, N.Yamada, O.Ozawa, Y.Ito for encouragement throughout the work.

References

[1] A.J.Smith, "Cache Memories," Computing Surveys, 14, 3, pp.473-530, 1982.

[2] T.Watanabe et al, "An 8Kbyte Intelligent Cache Memory," ISSCC, pp.266-267, Feb.1987.

[3] T.Sakurai et al, "A Low Power 46ns 256kbit CMOS Static RAM with Dynamic Double Word Line," J.SSC, SC-19, pp.578-584, 1984.

[4] A.Suzuki et al, "A 19ns Memory", ISSCC, pp.134-135, Feb.1987.

TABLE I Features

Cache size	32Kbyte per chip Expandable to 128Kbyte (4 chips)
Partitioning	Data/Instruction unified
Line size	16byte
Valid bit	4bit per entry (Flush clear available)
Bus size	4byte (Byte Enable supported)
Entry	2048 sets
Tag size	17bit
Placement	Direct mapping Expandable to 2-way set-associative (2-chips or 4-chips) with LRU algorithm and to n-way set-associative (n-chips) with random replacement
Address to HIT	18ns
Address to DATA	23ns
Supply voltage	5V ± 5%
Power	0.5W
Package	100pin PGA & plastic flat package
Technology	N-sub Twin-well CMOS (1μm NMOS) double poly-Si & double Al
Memory cell	High resistive poly load 4T SRAM cell

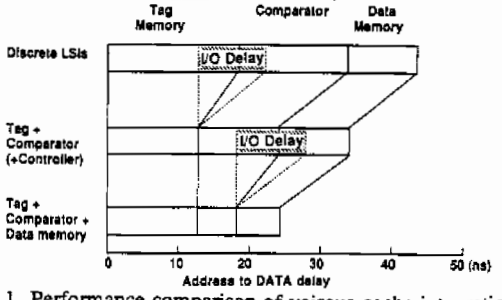

Fig.1 Performance comparison of vairous cache integration

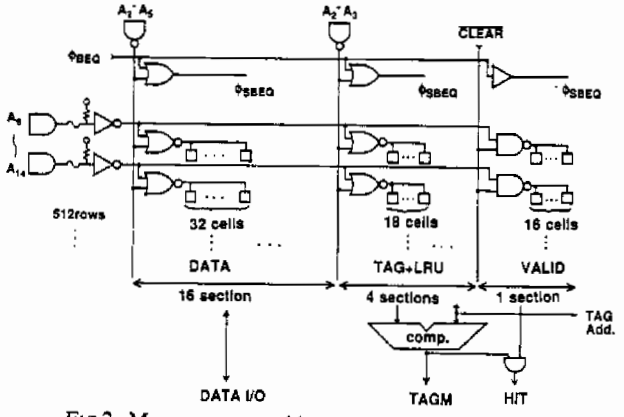

Fig.2 Memory core architecture (double word line)

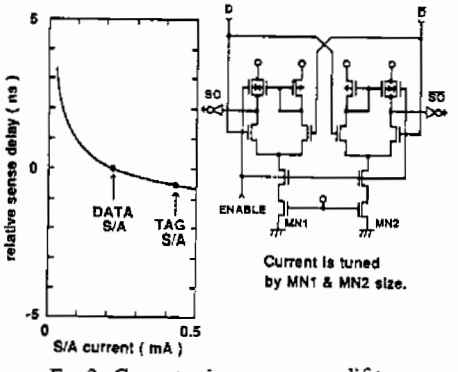

Fig.3 Current-mirror sense amplifier

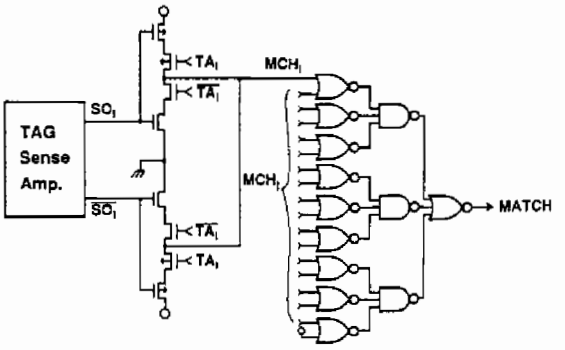

Fig.4 Static comparator

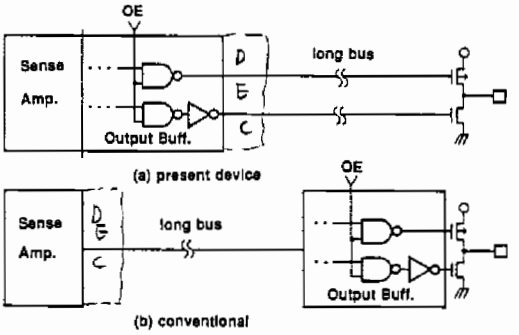

Fig.5 Output buffer near to sense amplifier

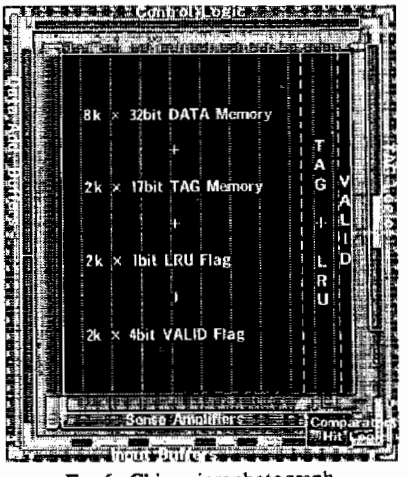

Fig.6 Chip microphotograph

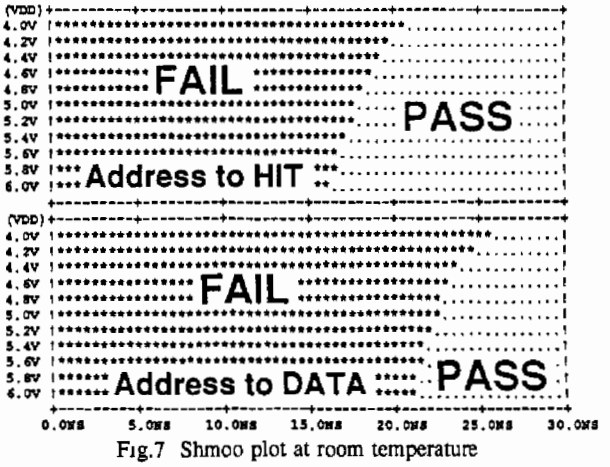

Fig.7 Shmoo plot at room temperature