4-6

Circuit Design of a 9ns-HIT-delay 32K Byte Cache Macro

Kazutaka NOGAMI, Takayasu SAKURAI, Kazuhiro SAWADA, Kenji SAKAUE*,
Yuichi MIYAZAWA, Shigeru TANAKA, Yoichi HIRUTA, Katsuto KATOH,
Toshinari TAKAYANAGI, Tsukasa SHIROTORI*, Yukiko ITOH, Masanori UCHIDA*,
and Tetsuya IIZUKA

Semiconductor Device Engineering Laboratory, Toshiba Corp.,
* Toshiba Microelectronics Corp.,
1 Komukai Toshiba-cho, Saiwai-ku, Kawasaki 210, Japan

Introduction

After a Reduced Instruction Set Computer (RISC) was shown to be effective in increasing CPU performance[1], several attempts have been made to further improve the CPU performance by including cache memory on the same chip[2]. However, the formerly reported cache size is limited up to 2K bytes, which is not sufficient to obtain more than 95% hit rate.

This paper describes a 32K byte cache macro with an experimental RISC implemented on the same chip.

Cache Architecture

Figure 1 shows the cache architecture. The cache macro is made of SRAM core but is tuned exclusively to cache design so that it optimizes pipeline data stream between Integer Unit (IU) and Cache resulting in faster machine cycle time than a conventional SRAM and glue logics based cache.

The power consumption is one of the most serious limitations in designing a cache, since a large number of bits, 89 bits in this device, are read out at a time. The macro employs double word line architecture[3,4,5] to minimize both power and silicon area penalty associated with cache entry partitioning.

To best balance the address to HIT signal flow from IU to Cache with respect to pipeline clock, the cache macro utilizes the word line latch which is one of the most unique features. Every WL has a slave latch. Master latch is in the address buffer. These latches act as pipeline latches so that a portion of address decoding time is merged into previous pipeline cycle. Since in the fast SRAMs the delay from address input to Word Line(WL) is about a half of the total delay. This WL latch scheme speeds up the address strobe to HIT delay, which includes read out data comparison, by a factor of 30%. The cache access is a critical path in CPU systems, thus this delay improvement directly enhances the system performance.

The buffer circuit (WL Buffer) placed between a TAG WL and a DATA WL isolates the speed of the TAG part, since DATA part is allowed to be a little slower than the TAG part.

Cache features are listed in Table I. The cache macro is organized a 32Kbyte direct mapped cache, which provides more than 95% hit rate. 64bit data bus achieves more than 400Mbyte/sec bandwidth between IU and Cache.

Design of Memory Core

Figure 2 shows a memory core circuitry of TAG and DATA parts. A new section WL selector is another new technique for high speed operation. This circuit reduces the capacitance of the Section Select(SS) line and Main Word Line(MWL) to 25% and 40% respectively compared to the conventional NOR selector[3]. This contributes to 10% speed up in Address Strobe to HIT delay.

The cache macro also contains a new bump-down delay free circuitry. The write operation is done through NMOS transfer gate at the top of BL and read operation is done through PMOS transfer gate at the bottom of the BL, shown in figure 2. This scheme gives a solution to delay variation problem caused by supply voltage bumpdown[6] without any power penalty.

Figure 3 is a selectively clearable VALID bit scheme using content addressable memory (CAM) and dual port (DP) cells. When the Process ID (PID) is switched, the VALID bits corresponding to a certain PID can be selectively cleared to '0' in a virtual cache system. The poly-Si load CAM cell is 40% smaller than the pure CMOS CAM cell and shows wide power supply voltage margin compared with a formerly reported poly-Si CAM cell. The DP VALID cells shows faster access time because no control circuit is inserted in the Section WL (SWL).

Performance

Features are listed in Table II. The device was fabricated by 1.0μm double Al/double poly twin-well CMOS process of 0.8μm NMOSFET. Whole layout was done under the unified design rule[7], so that this cache macro is available in various device generations up to 0.8μm rule. The chip microphotograph is shown in Fig.4. Figure 5 is the measured waveforms of Address Strobe and HIT with a Address Strobe to HIT delay of 9ns. HIT signal activates DATA output buffer with a HIT to DATA delay of 3ns, so that the Address Strobe to DATA delay is 12ns. Thus, it demonstrates that the cache macro can operate at 80MHz with on-chip IU. The RISC executes most of all the instructions within a single cycle. Therefore the on-chip cache macro is feasible for 80MIPS single chip RISC device based on 0.8μm CMOS technology.

Acknowledgement

The authors would like to express their deep appreciation to H. Yamada and Y.Unno for encouragement throughout the work.

References

[1] D.Patterson, and C.Sequin, "A VLSI RISC," Computer, pp.8-21, Sept. 1982.

[2] M.Horowitz, J.Hennessy, P.Chow, P.Gulak, J.Acken, A.Agarwal, C.Chu, S.McFarling, S.Przybylski, S.Richardson, A.Salz, R. Simoni, D.Stark, P.Steenkiste, S.Tjiang, and M.Wing, "A 32b Microprocessor with On-chip 2KByte Instruction Cache," ISSCC, Dig. of Tech. Papers, pp.30-31, Feb. 1987.

[3] T.Sakurai, J.Matsunaga, M.Isobe, T.Ohtani, K.Sawada, A.Aono, H.Nozawa, T.Iizuka, and S.Kohyama, "A Low Power 46ns 256Kbit CMOS SRAM with Dynamic Double Word Line," IEEE J.Solid State Circ., SC-19, No.5, pp.578-585, Oct. 1984.

[4] T.Sakurai, K.Nogami, K.Sawada, T.Shirotori, T.Takayanagi, T. Iizuka, T.Maeda, J.Matsunaga, H.Fuji, K.Maeguchi, K.Kobayashi, T. Ando, Y.Hayakashi, A.Miyoshi, and K.Sato, "A Circuit Design of 32KByte Integrated Cache Memory," Symp. on VLSI Circ., Tokyo, pp.45-46, Aug. 1988.

[5] K.Nogami, T.Sakurai, K.Sawada, T.Shirotori, T.Takayanagi, T. Iizuka, T.Maeda, J.Matsunaga, H.Fuji, K.Maeguchi, K.Kobayashi, T. Ando, Y.Hayakashi, A.Miyoshi, and K.Sato, "Architecture and Design Methodology of 32KByte Integrated Cache Memory," Euro-

pean Solid State CIRc. Conf., Dig. of Tech. paper, pp.98-101, Sept. 1988.

[6] H.Shimada, Y.Tange, K.Tanimoto, and M.Shiraishi, "An 18ns 1Mb CMOS SRAM," ISSCC, Dig. of Tech. Paper, pp.176-177, Feb. 1988.

Table I Cache Features

Cache size	32Kbyte (data/instruction unified)
Configuration	Direct mapped
Operation	Synchronous
Address	32bit
Process ID	4bit (CAM cell)
Valid bit	1bit/word (selectively clearable)
Line size	16byte
Data bus	64bit

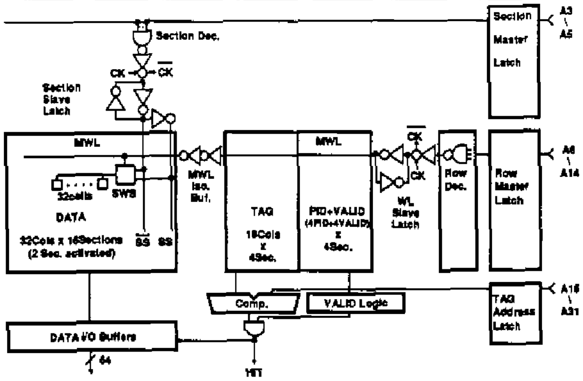

Fig.1. Cache macro architecture

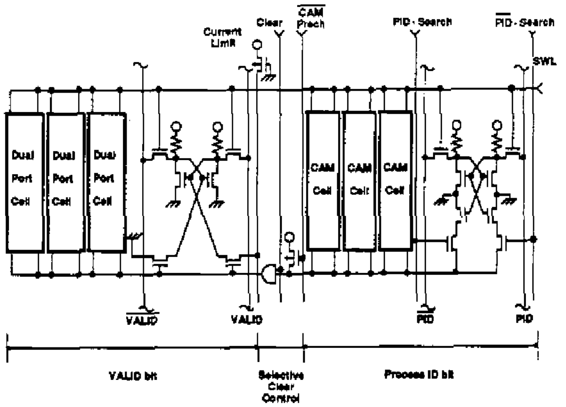

Fig.3. Selective clear scheme using CAM and dual-port cell

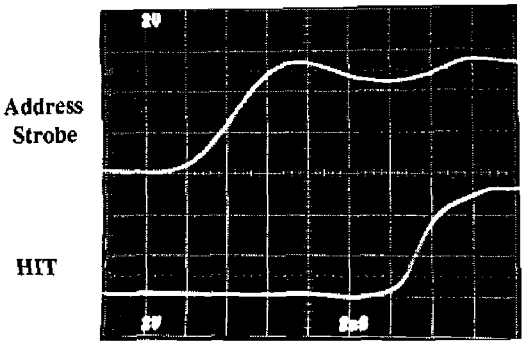

Fig.5. Measured waveforms

[7] T.Kuroda, H.Suzuki, H.Akiba, T.Aoki, T.Shigematsu, and K. Kawagai, "Unified Design Methodology and Device Architecture for Multi-Generation ASIC Application," CICC Technical Digest, 25.7, 1988.

Table II Device Features

Technology	Double Al & double poly twin-well CMOS
Design rule	1.0μm basic rule
MOSFETs	0.8μm gate length
Cell size	
SRAM cell	9.2μm × 13.8μm
Dual-port cell	20.0μm × 13.8μm
CAM cell	25.1μm × 13.8μm
Chip size	14.5mm × 10.8mm
Strobe to HIT	9ns (typical)
Strobe to DATA	12ns(typical)
Cycle freq.	80MHz (typical)
Package	144pin PGA

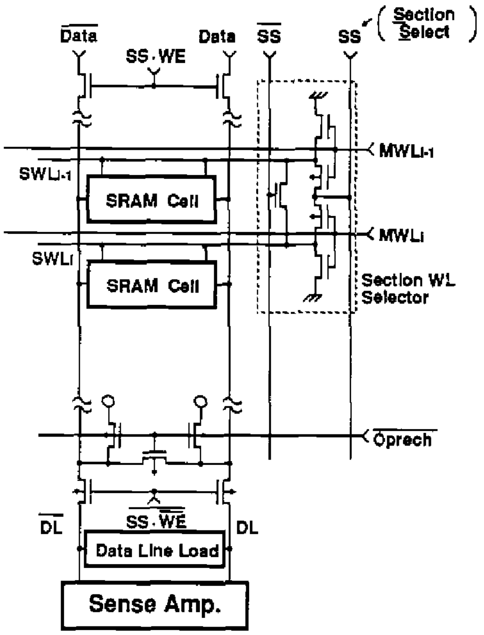

Fig.2. New section WL selector

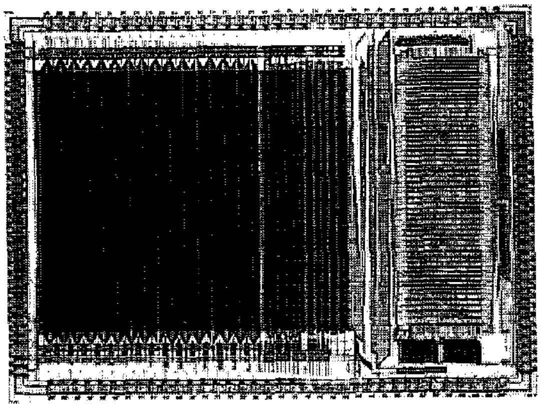

Fig.4. Chip microphotograph