A 110MHz / 1MBIT SYNCHRONOUS TAG RAM

Yasuo Unekawa, Tsuguo Kobayashi, Tsukasa Shirotori*, Yukihiro Fujimoto,
Takayoshi Shimazawa, Kazutaka Nogami,Takehiko Nakao, Kazuhiro Sawada,
Masataka Matsui, Takayasu Sakurai, Man Kit Tang**, and Bill Huffman**

Semiconductor Device Eng. Lab., Toshiba Corp., Toshiba Microelectronics Corp.*,
1 Komukai Toshiba-cho, Saiwai-ku, Kawasaki 210, Japan
Silicon Graphics Inc.**
2011 North Shoreline Blvd. Mountain View, CA, 94039, USA

1. INTRODUCTION

Most of the recent micro-processors include the first level cache on chip. However, a chip size constraint limits the on-chip cache capacity to 16KBytes. Therefore a large off-chip secondary cache is indispensable for high-performance computer systems. The synchronous Tag RAM reported in this paper holds addresses and status bits of cached data and can be used to build a secondary cache system of up to 16MBytes with external commodity synchronous SRAM's. In order to handle the large secondary cache, the present Tag RAM contains 1.189Mbit of 4T SRAM cells, the largest capacity ever reported for a Tag RAM.

Short cycle time and small clock to D_{OUT} (data output) delay of the Tag RAM is crucial for a high-performance cache system. 9ns cycle operation and clock to D_{OUT} of 4.7ns in typical condition are achieved by a use of circuit techniques such as a pipelined decoding scheme, a single PMOS load BiCMOS main decoder, a BiCMOS sense-amplifying comparator, a highly linear Voltage-Controlled Oscillator (VCO) for a Phase Locked Loop (PLL) and doubly placed self-timed write circuits. Since pure CMOS implementation can not achieve the required speed, the device is manufactured with 0.7μm double-polysilicon and double-metal BiCMOS technology.

2. FEATURES

Figure 1 shows a block diagram of the present Tag RAM. It contains 8K entries x 4-ways x 20bits Tag memory, 8K entries x 4-ways x 12 bits memory for State bits and 8K entries x 4-way x 4 bits of Dirty bits. 8 redundancy rows are included. It also contains comparators to compare read-out Tag address with Higher Physical Address. Double WL structure [1] is adopted for reducing memory cell power consumption and WL delay. In the State bits, Virtual Synonym bits are included which are used to resolve the first cache sysnonym problem.

Since the Dirty bit is written with the address given in the preceding cycle, a flip-flop is inserted between a State memory word line (WL) and a Dirty memory WL to hold the WL information for one cycle. Additional advantage of this configuration is the reduction of main WL capacitance and hence accelerates the Tag part operation.

The write control logic for Dirty bits is also integrated on chip and JTAG is supported to increase on-board testability. A P-epi wafer is adopted for a substrate, which is biased with on-chip self-subbias circuit to reduce junction capacitances. Table 1 summarizes the key features of the chip.

3. CIRCUIT DESIGN DETAILS

In order to meet the high-speed requirement together with the large memory capacity requirement, several novel circuit techniques are introduced.

3.1 PIPELINED PARTIAL DECODING SCHEME

Pipelined decoding scheme (see Fig.1) is used to reduce the partial decoding time. Partial decoders are place in between master and slave address latches. Since the slave latch is placed after the partial decoder, the partial decoding can be done in address set-up time. If the master latch is also placed after the partial decoder, the master latch erroneously latches the address of the succeeding cycle due to the internal clock delay and rather short address hold time. This achieves the gain of 2ns over the conventional partial decoding scheme, where the address F/F is placed just at the address input.

3.2 SINGLE PMOS LOAD BICMOS MAIN DECODER

Figure 2 shows a proposed BiCMOS main decoder. A normally-on PMOS load (P1) is adopted to minimize the input capacitance of the main decoder to reduce the driver delay of the partial decoder. The serially connected NMOS structure, however, is slow in nature. The bipolar transistor Q1 is added to enhance the drivability of the serial structure and in consequence the P1 can be designed to have high drivability to realize high-speed pull-up. The present circuit reduces the address decoding time by 0.5ns compared with the conventional full CMOS decoder + BiCMOS buffer scheme.

3.3 SENSE-AMPLIFYING BICMOS COMPARATOR

Conventionally, a comparator for a cache is built with a MOS comparator inserted between a bit line (BL) and a BiCMOS sense amplifier [2]. The newly proposed Sense-Amplifying Comparator (SAC), whose circuit diagram is shown in Fig.3, replaces the conventional MOS comparator with a bipolar comparator (Q3 & Q6) merged into a bipolar sense amplifier (Q1, Q2, Q4, Q5). This configuration eliminates the MOS comparator delay and gains 0.5ns.

Because the Tag data read-out is also required a fast operation in the present Tag RAM, a Tag data sense amplifier (S/A, Q8 ~ Q10) is placed in parallel to the SAC. All circuits from the BL through the HIT signal generator take ECL-based configuration to reduce critical path delay.

3.4 ON-CHIP PLL

A PLL is integrated on chip which cancels internal clock delay. The linearity of the VCO is a key to obtain large lock frequency range. The proposed VCO is shown in Fig.4, together with a linearity comparison with the conventional VCO. Due to the high linearity, the PLL is measured to lock frequencies from 50MHz to 150MHz stably with a 0.4ns jitter. The inclusion of the PLL on a chip can reduce of the cycle time by 1ns.

3.5 DOUBLY PLACED SELF-TIMED WRITE CIRCUITS

In order to minimize the clock to D_{OUT} delay, sense amplifiers should be placed near to the pads. This rules out the possibility of BL partitioning and the placement of S/A's at the center of a memory array. Then the BL gets highly capacitive due to the 1028 memory cells connected to each BL. The RC delay of the BL amounts to 2.5ns and hinders a fast write and write-recovery operation, although it does not cause a problem in a read-out operation because of the inherently small (0.2V) BL swing needed for a BiCMOS S/A.

In order to reduce the BL RC delay, write circuit and BL precharge circuit are placed at both ends of each BL (see Fig.5). This configuration reduces the write operation delay by 1ns and eliminates the case where the write operation determines the cycle time. The write operation is controlled by a self-timed write pulse and after the write pulse, the BL precharge is taken place automatically.

4. RESULTS

Figure 6 shows a total chip layout whose size is 14.8mm x 14.8mm. The minimum clock cycle time is 9ns in typical condition which corresponds to 110MHz clock frequency. If the RAM is designed with a pure CMOS technology without the circuit ideas mentioned in Section 3, the clock cycle is estimated to be 18ns. If the RAM is designed with a BiCMOS technology without the above-mentioned circuit ideas, the clock cycle is estimated to be 13ns. The 4ns improvement of the present design over the conventional BiCMOS design comes from the additive speed gains of the circuit ideas described in chapter 3.1 through 3.4.

Figure 7 and 8 show simulated waveforms and the delay distribution of a Tag look-up cycle operated at 9ns cycle time, respectively.

REFERENCES

[1] T.Sakurai et al., "A Low Power 46ns 256Kbit CMOS Static RAM with Dynamic Double Word Line," JSSC, SC19, pp.578-585, Oct.1984.
[2] H.Hara et al., "0.5μm 3.3V BiCMOS Standard Cells with 32KByte Cache," JSSC, SC27, pp.1579-1584, Nov.1992.

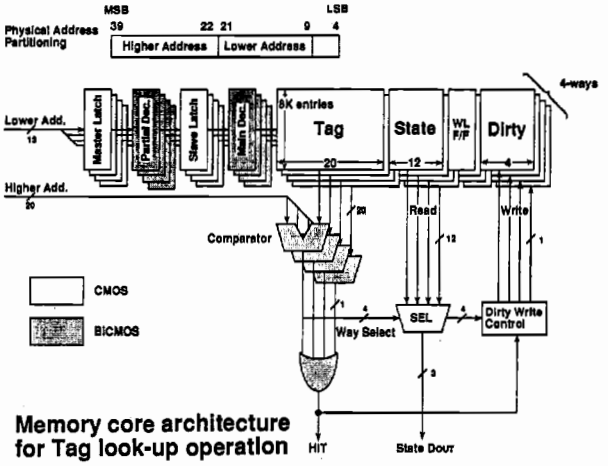

Figure 1 Memory core architecture for Tag look-up operation

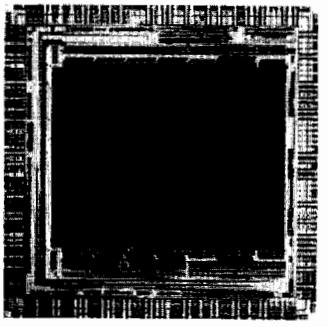

Figure 6 Chip layout

Entry size	8K lines
Line size	512 Bytes
Mapping	4-way set-associative
SRAM cell	Highly resistive poly-Si load 4T cell
SRAM cell size	8.0μm x 4.8μm
Chip size	14.8mm x 14.8mm
Power	3W (@75MHz)
Package	155pin ceramic PGA
Technology	0.7μm double poly-Si, double Al BiCMOS
Other features	Integrated Dirty bit logic Self-timed write On-chip PLL JTAG supported 8 rows redundancy

Table 1 Features

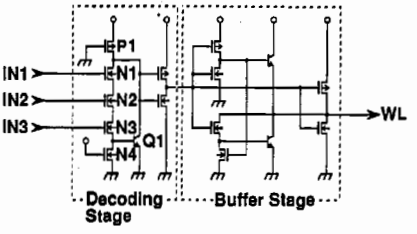

Figure 2 BiCMOS Main Decoder

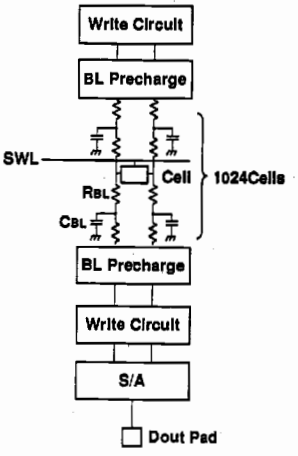

Figure 5 Doubly placed precharge and write control

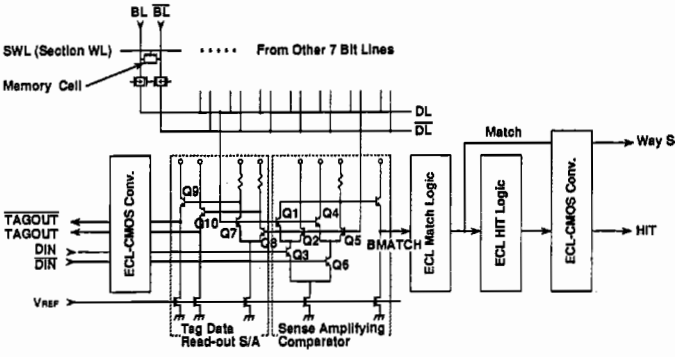

Figure 3 Sense Amplifying Comparator

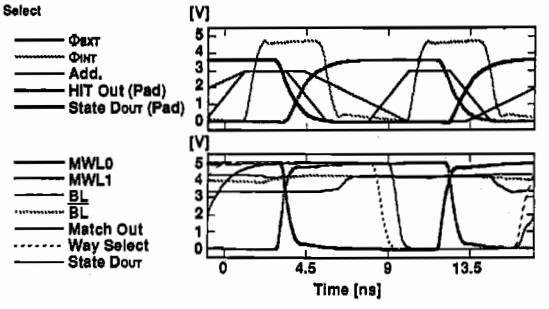

Figure 7 Simulated waveforms of a Tag look-up cycle

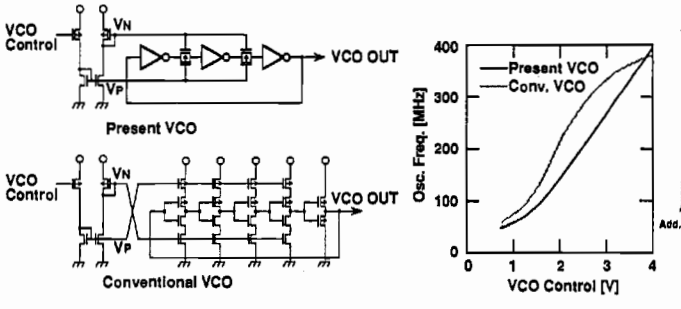

Figure 4 Voltage Controlled Oscillator and measured property

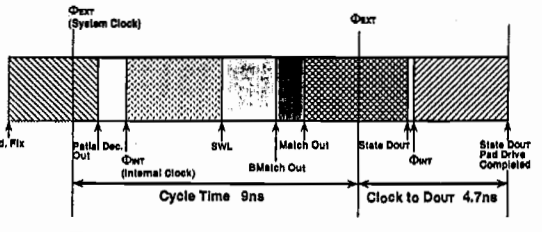

Figure 8 Distribution of delay time in Tag look-up cycle