第 25 卷第 4 期 半 导 体 学 报 Vol. 25, No. 4
2004 年 4 月 CHINESE JOURNAL OF SEMICONDUCTORS Apr., 2004

A Low-Power Super-Performance Four-Way Set-Associative CMOS Cache Memory*

Sun Hui, Li Wenhong and Zhang Qianling

(*ASIC and System State Key Laboratory, Fudan University, Shanghai* 200433, *China*)

Abstract: A 1.8-V 64-kb four-way set-associative CMOS cache memory implemented by 0.18μm/1.8V 1P6M logic CMOS technology for a super performance 32-b RISC microprocessor is presented. For comparison, a conventional parallel access cache with the same storage and organization is also designed and simulated using the same technology. Simulation results indicate that by using sequential access, power reduction of 26% on a cache hit and 35% on a cache miss is achieved. High-speed approaches including modified current-mode sense amplifier and split dynamic tag comparators are adopted to achieve fast data access. Simulation results indicate that a typical clock to data access of 2.7ns is achieved.

Key words: cache; set-associative; sequential access; parallel access; current-mode sense amplifier; comparator
EEACC: 2570D; 1265D
CLC number: TN432 **Document code**: A **Article ID**: 0253-4177(2004)04-0366-06

1 Introduction

Growing demand for battery-operated devices such as portable computers and mobile phones has led to extensive study of low-power super-performance microprocessors. In these microprocessors, cache memory usually consumes a significant fraction of overall power consumption. For example, StrongARM-1 consumes about 43% of chip power in I-cache and D-cache together[1], and Pentium Pro consumes about 33% in caches[2].

Set-associative caches are widely adopted in modern microprocessors to achieve low miss rates for typical applications. Until now, however, most set-associative caches access the tag and data array in parallel, then select the data from the matching way on a cache hit or discard all the data on a cache miss[3]. Although parallel access could achieve fast data access, it also wastes dynamic energy dissipation.

Low power is very important for memory design[4,5]. Since low power requirement dominates in the 32-b RISC microprocessor in which the proposed cache memory is to be embedded, we designed a 64-kb four-way set-associative cache memory using sequential access. Simulation results indicate that compared to a conventional parallel access cache memory with the same storage and organization, power reduction of 26% on a cache hit and 35% on a cache miss has been achieved. To achieve fast data access, high-speed circuit modules including high-speed current-mode sense amplifier and split dynamic tag comparators are utilized. Ow-

* Project supported by National High Technology Research and Development Program of China (No. 2002AA1Z1060)
Sun Hui female, was born in 1978, master candidate. Her research interests focus on VLSI design and low power design.
Li Wenhong male, was born in 1967, PhD. His research interests include ASIC design and SOC design.
Zhang Qianling female, was born in 1936, professor. Her research areas include microelectronics and SOC design.
Received 25 July 2003, revised manuscript received 26 November 2003

ing to these high-speed approaches, a typical clock to data access of 2.7ns and clock to hit access of 2.1ns have been achieved.

2 Sequential access set-associative cache

2.1 Cache memory overview

Figure 1 shows the block diagram of the 1.8-V 64-kb four-way set-associative CMOS cache memory. As shown in the Fig. 1, this 64-kb four-way set-associative cache memory is composed of tag arrays, data arrays, sense amplifiers, address decoders, timing controller, and tag comparators. In the proposed cache, the tag and status bits access is the same as conventional cache, but data access is limited to only the desired word in the data array to reduce energy dissipation and critical path delay on the word-lines. The subbank architecture of the data array is shown in Fig. 2.

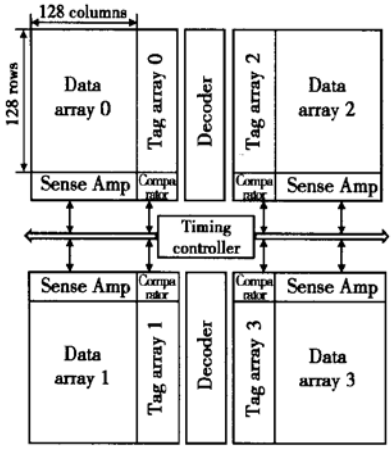

Fig. 1 Block diagram of the 1.8-V 64-kb four-way set-associative CMOS cache memory

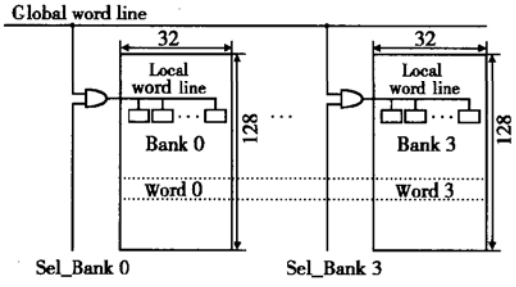

Fig. 2 Subbank architecture of the data array

2.2 Sequential access mode

Figure 3 shows the relevant components and operations of parallel access and sequential access set-associative cache. In a parallel access set-associative cache memory shown in Fig. 3(a), all the data ways and tag ways are accessed in parallel before the matching way is decided by tag comparison. During a cache hit access, data from the matching way is selected and transported to the output data bus. During a cache miss access, however, all the data read out from the data ways are discarded[3]. The energy dissipated by a parallel access N-way set-associative cache can thus be expressed as

$N \times$tag array energy+ $N \times$data array energy

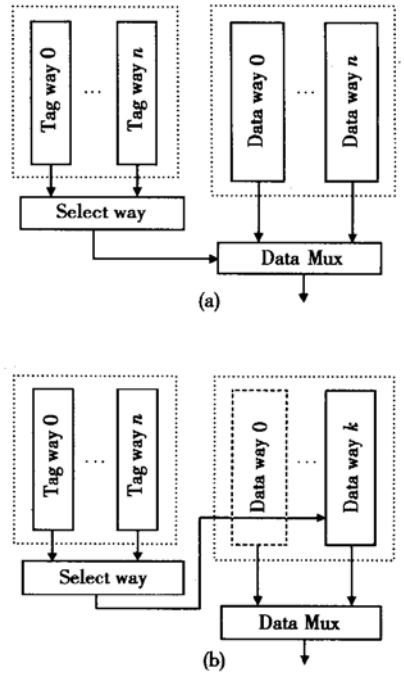

Fig. 3 Access and operation for parallel and sequential access set-associative cache memory (a) Conventional parallel access; (b) Sequential access

In a sequential access set-associative cache memory shown in Fig. 3(b), all the tag ways are probed before the matching way is decided by tag comparison. During a cache hit access, the matching data way is accessed. During a cache miss access, no data way is accessed[6]. Thus the energy dissipation of a sequential access N-way set-asso-

ciative cache can be described as

$N \times$tag array energy+ $1 \times$data array energy

during a cache hit access and

$N \times$tag array energy

during a cache miss access.

A conclusion can then be drawn that for a N-way set-associative cache memory, sequential access can achieve power savings of $(N-1) \times$data array energy on a cache hit and $N \times$data array energy on a cache miss.

3 Design approaches for high-speed

3.1 High-speed current-mode sense amplifier

A fast sense amplifier is very important for reducing read path delay and achieving fast data access in memories. Current-mode sense amplifiers are widely used in memory design today because they can amplify a small difference of current on the bit-lines to a full-rail voltage difference. In the proposed 1.8-V 64-kb four-way set-associative cache memory, a new super performance current-mode sense amplifier is designed to achieve high amplification speed with low power consumption.

Figure 4 shows the schematic of the current-mode sense amplifier. It has been modified on the basis of the circuit in Ref. [7]. Unlike the amplifier in Ref. [7], a current conveyor circuit (MP6~MP10) is used to function as a current transporting and column-select device. The current conveyor presents a virtual short to the bit-lines due to its cross-coupled pair MP6 and MP7[8]. These results in an equal voltage at the bit-lines while at the same time providing a current difference on nodes I and NI. The current conveyor has zero input resistance during sensing. This property makes it insensitive to the bit-line capacitance.

The current-mode sense amplifier shown in Fig. 4 includes a current transporting and column-select structure (MP6~ MP10), a controlled cross-coupled transistor structure (MP1, MP2, MN1, MN2), control transistors (MP3, MP4, MN3, MN4), sense control transistor MN5 and recovery transistor MP5. SAE is the sense amplifier enable signal, WL is the word-line select signal, CS is the column-select signal, and CSN is the complement of CS.

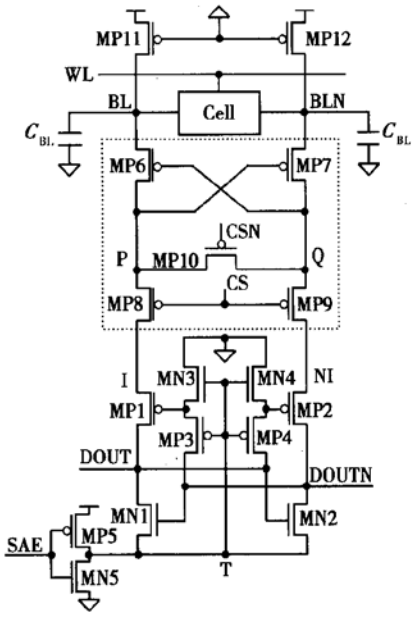

Fig. 4 High-speed current-mode sense ampflifier

There are two operation phases, one is equalizing phase and the other is evaluating phase. In equalizing phase, SAE is low, CS is high, and CSN is low. MP10 is on and the voltages on nodes P and Q are equalized. Node T is pulled to high voltage by MP5. When CS turns low and CSN turns high, MP10 is off and MP8, MP9 are turned on. The current-mode sense amplifier enters the evaluating phase. During this time, WL is asserted thus turning on the access transistors of the memory cell. A current difference will appear on the complementary bit-lines BL and BLN owing to the opposite polarities existing on the memory cell inverters. The current difference is conveyed to I and NI through the current conveyor. During this time, SAE turns high, MN5 is on and the current difference on I and NI is evaluated. Corresponding CMOS voltages will appear on the output nodes DOUT and DOUTN.

Using a 0.18μm/1.8V CMOS technology, the

sensing delay and average power dissipation against different bit-line capacitances are simulated for the proposed current-mode sense amplifier and the recently reported current sense amplifier from Ref. [9]. For comparison, the simulating conditions are set to be same. Fig. 6 shows the simulation results. It can be observed that the proposed current-mode sense amplifier is insensitive to bit-line capacitance and has a faster sensing speed and lower average power dissipation.

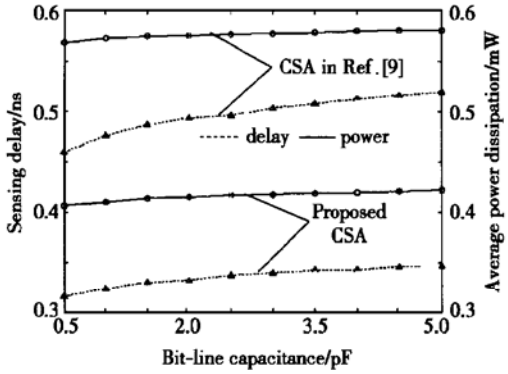

Fig. 5　Sensing delay and average power dissipation against bit-line capacitance

3.2　Split dynamic tag comparator

Tag comparison delay is another key component of the critical path delay of the proposed cache. To reduce tag comparison delay and achieve fast data access, a high-speed split dynamic tag comparator is used. Figure 6 is the schematic of the split dynamic tag comparator used in the proposed cache. The condition for a HIT is when all 21 TAG bits match the 21 incoming tag bits in the address and the corresponding status bit VALID is true (high). As it is shown in Fig. 6, a 22-bit tag comparison is carried out using two super-performance 11-bit dynamic comparators. This procedure reduces the capacitive load on the comparison line and results in a faster speed of comparison[10]. Usually the match output of each 11-bit dynamic comparator is precharged high and remains high if the comparison result is a match. The two match outputs are then combined to generate the HIT signal for a 22-bit tag comparison. Simulation results indicate that splitting one 22-bit tag comparison into two 11-bit comparisons reduces the tag comparison delay by 20% without increasing power consumption.

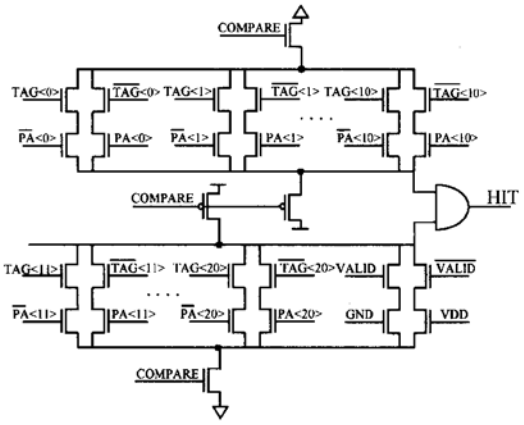

Fig. 6　Split dynamic tag comparator

4　Simulation results and chip implementation

The 64-kb four-way set-associative cache is simulated using a 0.18μm/1.8V 1P6M logic CMOS technology with V_{DD} = 1.8V. Results from the simulation are given in Fig. 7 and Fig. 8.

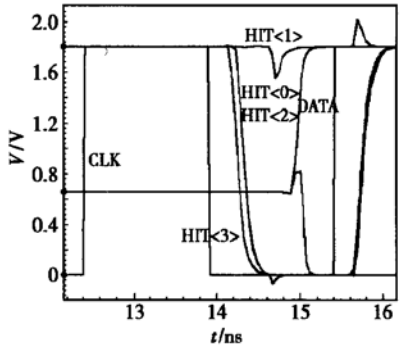

Fig. 7　Simulated transient waveforms of the 1.8-V 64-kb four-way set-associative cache at V_{DD} = 1.8V during hit access

Figure 7 shows one cache hit access cycle. In this figure, CLK is the clock signal, DATA is one of the data output bits. HIT ⟨0⟩, HIT ⟨1⟩, HIT ⟨2⟩, and HIT ⟨3⟩ are the four way select signals of the cache memory. Simulation results indicate that the clock to data access time is 2.7ns and to hit access time is 2.1ns.

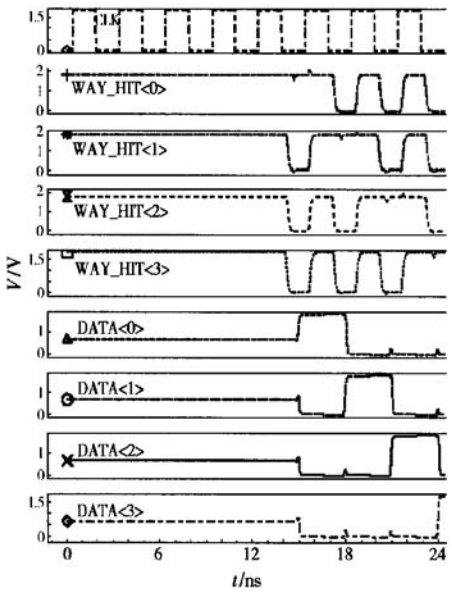

Fig. 8 Simulation waveforms of four consecutive load operations

In Fig. 8, CLK is the clock signal, HIT ⟨0:3⟩ are the hit signals of the four ways. DATA ⟨0:3⟩ are the lowest four bits of output databus DATA ⟨0:31⟩. Data "00000001", "00000002", "00000004" and "00000008" are stored into the four data arrays in the first four clock cycles. Next, four consecutive load operations are executed. In each load operation, a different way is hit. Correspondingly, one of HIT ⟨0⟩, HIT ⟨1⟩, HIT ⟨2⟩, and HIT ⟨3⟩ keeps high in one of the four load cycles and the stored data are read out. It can be seen that the proposed cache memory works well at a clock frequency of 333MHz (clock cycle is 3.3ns).

Power consumption of the proposed cache memory is 210mW during a cache hit access and 185mW during a cache miss access at a supply voltage of 1.8V and clock frequency of 333MHz. For comparison, a conventional parallel access cache with the same storage and organization is also designed and simulated using the same technology. Simulation result indicates that using sequential access, power reduction of 26% on a cache hit and 35% on a cache miss have been achieved.

The proposed 1.8-V 64-kb four-way set-associative cache is fabricated at SMIC 0.18μm/1.8V 1P6M logic CMOS technology. Figure 9 is the photo of the fabricated RISC microprocessor chip. The proposed cache is used as I-Cache and D-Cache in the microprocessor as shown in Fig. 9. Test result of the fabricated cache is not available yet because the package is not finished. The layout area of the cache memory is 1422.16μm×1001.72μm. Table 1 lists the technology and organization of the presented cache memory.

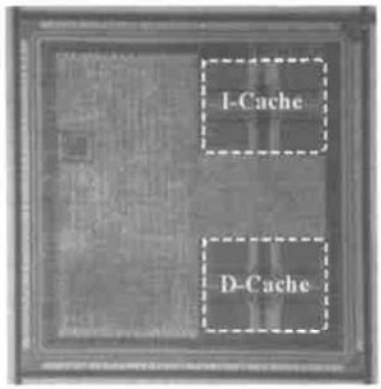

Fig. 9 Photo of the fabricated RISC microprocessor chip

Table 1 Cache technology and organization

Technology	0.18μm/1.8V CMOS technology
Metal system	Six metal layers
Memory cell type	Six-transistor CMOS
Memory cell size	4.20μm×2.74μm
Cache storage	64kbits data+ 11kbits tag
Cache organization	128line×4way×128bit
Total cache transistors	0.5million
Cache block size	1422.16μm×1001.72μm
Cycle time	3ns
Access time	2.7ns
Cache power	210mW/185mW
Clock frequency	333MHz
Power supply	1.8V

5 Conclusion

In this paper, a 1.8-V 64-kb four-way set-associative CMOS cache memory implemented by 0.18μm/1.8V 1P6M logic CMOS technology has been described. Simulation results indicate that power reduction of 26% on a cache hit and 35% on a cache miss have been achieved by using sequential access. High-speed circuit modules including

high-speed current-mode sense amplifier and split dynamic tag comparators are utilized to achieve fast data access. Using a 0.18m/1.8V CMOS technology, simulation results indicate a typical clock to data access of 2.7ns and to hit access of 2.1ns. Power consumption of the proposed cache memory is 210mW during a cache hit access and 185mW during a cache miss access with V_{DD} = 1.8V and clock frequency of 333MHz. These features qualify the proposed cache memory to be integrated into high performance microprocessors.

References

[1] Montanaro J, Witek R T, Anne K, et al. A 160MHz, 32b 0.5W CMOS RISC microprocessor. IEEE J Solid-State Circuits, 1996, 31(11): 1703

[2] Manne S, Klauser A, Grunwald D. Pipeline gating: speculation control for energy reduction. Proceedings of the 25th Annual International Symposium on Low Power Electronics and Design(ISLPED), 1998: 132

[3] Mizuno H, Matsuzaki N, Osada K, et al. A 1V, 100MHz, 10mW cache using a separated bit-line memory hierarchy architecture and domino tag comparators. IEEE J Solid-State Circuits, 1996, 31(11): 1618

[4] Pan Liyang, Zhu Jun, Liu Zhihong, et al. A novel flash memory using band-to-band tunneling induced hot electron injection to program. Chinese Journal of Semiconductors, 2002, 23(7): 690

[5] Pan Liyang, Zhu Jun, Liu Kai, et al. Novel p-channel selected n-channel divided bit-line NOR flash memory using source inducted band-to-band hot-electron injection programming. Chinese Journal of Semiconductors, 2002, 23(10): 1031

[6] Benschneider B J, Black A J, Bowhill W J, et al. A 300MHz 64b quad-issue CMOS RISC microprocessor. IEEE J Solid-State Circuits, 1995, 30(11): 1203

[7] Kristovski G V, Pogrebnoy Y L. New sense amplifier for small-swing CMOS logic circuit. IEEE Trans Circuit and Syst-II: Analogy and Digital Signal Processing, 2000, 47(6): 573

[8] Wang J S, Lee H Y. A new current-mode sense amplifier for low-voltage low-power SRAM design. Eleventh Annual IEEE International ASIC Conference, 1998: 163

[9] Izumikawa M, Igura H, Furuta K, et al. A 0.25μm CMOS 0.9V 100MHz DSP core. IEEE J Solid-State Circuits, 1997, 32(1): 52

[10] Reinman G, Jouppi N P. CACTI 2.0: an integrated cache timing and power model. WRL Research Report, 2000

一种低功耗的高性能四路组相联 CMOS 高速缓冲存储器*

孙　慧　李文宏　章倩苓

(复旦大学专用集成电路与系统国家重点实验室，上海　200433)

摘要: 采用 0.18μm/1.8V 1P6M 数字 CMOS 工艺设计并实现了一种用于高性能 32 位 RISC 微处理器的 64kb 四路组相联片上高速缓冲存储器(cache). 当采用串行访问方式时, 该四路组相联 cache 的功耗比采用传统并行访问方式在 cache 命中时降低 26%, 在 cache 失效时降低 35%. 该 cache 的设计中还采用了高速电路模块如高速电流灵敏放大器和分裂式动态 tag 比较器等来提高电路工作速度. 电路仿真结果显示 cache 命中时从时钟输入到数据输出的延时为 2.7ns.

关键词: 高速缓冲存储器; 组相联; 顺序访问方式; 并行访问方式; 电流灵敏放大器; 比较器

EEACC: 2570D; 1265D

中图分类号: TN432　　**文献标识码**: A　　**文章编号**: 0253-4177(2004)04-0366-06

* 国家高技术研究发展计划资助项目(合同号: 2002AA1Z1060)

孙　慧　女, 1978 年出生, 硕士研究生, 研究方向为超大规模集成电路设计和低功耗设计等.

李文宏　男, 1967 年出生, 博士, 研究方向为专用集成电路和片上系统设计等.

章倩苓　女, 1936 年出生, 教授, 博士生导师, 研究方向为微电子、片上系统设计等.

2003-07-25 收到, 2003-11-26 定稿