Extended Abstracts of the 2003 International Conference on Solid State Devices and Materials, Tokyo, 2003, pp. 152-153

G-1-5

Combined Data/Instruction Cache with Bank-Based Multi-Port Architecture

Koh Johguchi, Zhaomin Zhu, Tai Hirakawa[1], Tetsushi Koide,
Tetsuo Hironaka[1] and Hans Jürgen Mattausch
Research Center for Nanodevices and Systems, Hiroshima Univ., 1-4-2 Kagamiyama, Higashi-Hiroshima, 739-8527, Japan
[1]Grad. School of Infomation Sci., Hiroshima City Univ., 3-4-1 OzukaHigashi, AsaMinamiKu, Hiroshima, 731-3194, Japan
Phone: +81-824-24-6265, FAX : +81-824-22-7185, E-mail: {jouguchi, zzm, koide, hjm}@sxsys.hiroshima-u.ac.jp

1. Introduction

Modern processors simultaneously fetch, decode and execute many instructions. This results in the demand for a large access bandwidth of the processor's memory components and has already led to register files with many ports. However, for the cache memory the conventional solution of 1-port-data and instruction caches is still in use. Since the demand for parallelism tends to increase at a high rate, the cache system will become the bottleneck of processor performance and has to be innovated.

We propose to improve the low access bandwidth of the conventional 1-port-caches by multi-port caches utilizing a 1-port-bank based Hierarchical Multi-port memory Architecture (HMA). This results in the simultaneous realization of small area, high access bandwidth and low power dissipation [1,2]. Moreover, by combining instruction and data caches into a single multi-port cache, we are able to dynamically schedule the memory amount used for data and instructions, resulting in a more efficient usage of the caches storage capacity.

2. Multi-port Cache with HMA Structure

HMA is a 1-port-bank based multi-port memory architecture which further improves area consumption and performance of the conventional crossbar architecture. The crossbar's switch network is distributed into the bank structure, which decreases global wiring and transistor number. A two dimensional bank decoder reduces the overhead for bank selection and allows easy matrix arrangement of the banks [1].

Fig. 1 shows the structure-example of a direct mapped cache which uses HMA [2]. The cache index, consisting of line number (LN) and line offset (LO), is divided into two portions, a bank-internal address (BA) and a bank index (BI). BI is used for selecting a cache word or tag within memory banks, and BI is used for selecting the respective banks within data/instruction or tag memory. BI uses the lower rank bits in order to make sure that consecutive lines and words within lines are located in different banks, so that they can be accessed in parallel without access conflict.

3. Combination of Instruction and Data Cache

(a) Advantages of a Combined Cache

Using the proposed HMA cache, instruction and data cache can be combined without loss in access bandwidth, but with the advantage of a lower miss rate at the same storage capacity. On the other hand, using a bank-based multi-port cache, access to one bank is restricted to 1 port, and access-conflict rate may increase.

The miss-rate advantage of the combined cache and the required number of banks for sufficiently small access-conflict rate are examined for the example of a four way superscalar processor. The simulation is carried out with a modified version of SimpleScalar [3]. Dhrystone and SPEC95 (gcc, ijpeg, etc.) are used as benchmarks.

The results for splitted and combined direct-mapped cache are shown in Figs. 2 and 3. The storage capacity values in the figures show the total capacity of instruction and data cache, being the same for splitted and combined cache. According to Fig. 2, in the case that sufficient capacity is not prepared, the combined cache has higher miss rate than the splitted cache because data rewriting takes place frequently. However, if the total miss rate becomes lower than 10%, as required in real processors, the miss rate of the combined cache is clearly lower than that of the splitted cache. Moreover, it turns out that the miss rate of the combined cache is approximately equal to that of a splitted cache at 25% reduced storage capacity. We conclude from the result of Fig. 3, that the access-conflict rate becomes sufficiently low when more than 16 banks are provided.

(b) Optimum Combination of Instruction and Data Cache with HMA Structure

In general, the accesses to the instruction cache are consecutive. For a 4-way superscalar processor, it is therefore expected, that one instruction port with 4 times larger word length will deliver sufficient instruction-fetch performance. The optimum number of data-access ports is estimated to be 2 or 3.

Above considerations suggest that an optimized combined data/instruction cache should have different word length for instruction and data ports. Fig. 4 shows our HMA proposal of a combined write-through cache with 2 data ports and 1 instruction port, with 4 times larger word length, for 4-way superscalar processors. Although, it uses internally only a 1-to-3 port convertor with a relatively small area-overhead, the externally available access bandwidth corresponds to 6 ports, due to the 4 times increased word length of the instruction port. In the tag memory, we can achieve even a further simplification, because only one tag-data of the instruction port is necessary per 4-bank cluster. Therefore, 3 banks of each tag-memory cluster need only a 1-to-2 port converter.

4. Test-Chip Design for an HMA Cache

For the test chip of an HMA cache memory, a configuration with 4 ports was chosen and the design was carried out in a 0.18μm CMOS technology with 5 metal layers. Due to chip-space limitations, we could include only 4 of the 32 banks of the complete cache in the design. However, the chip-layout shown in Fig. 5 contains all needed new functional units. The design data are summarized in Table I. Small area and short delay are achieved with a dynamic CMOS circuit technology. The area-overhead of the 1:4-port convertor for the 1Kbyte bank of Fig. 6 is less than 10%. We also applied a new memory access method which overlaps bank-conflict management and bank decoding with the precharging phase of the banks. As a result, simulated bank-access time and complete cache-access time could be reduced to 0.91ns and 1.4ns, respectively.

5. Conclusions

In this paper, a bank-based combined data/instruction cache with multiple ports has been proposed and the advantages have been verified by simulation. Especially important is our method of providing a different word length for data and instruction ports, which takes advantage of the internal bank structure. To minimize bank conflicts, we use an addressing method, which insures that the words in one cache-line and also consecutive cache-lines are located in different banks. A test-chip design of a 4-port bank-based cache in a 0.18μm CMOS technology showed, that the area-overhead for the 4 ports is of the order of 10%. An internal bank-access time of 0.91ns and a cache access time of 1.4ns could be achieved with a dynamic CMOS circuit technology

and by overlapping the external bank access with the bank-internal precharge.

The proposed bank-based multi-port cache is also very attractive for low power dissipation, because the number of activated banks, determining power dissipation, corresponds to the port number and is independent of the total number of banks in the cache.

Acknowledgements

This research is supported by Semiconductor Technology Academic Research Center (STARC).

The VLSI chip in this study has been fabricated in the chip fabrication program of VLSI Design and Education Center (VDEC), the University of Tokyo in collaboration with Hitachi Ltd., Dai Nippon Printing Corporation and Cadence Design Systems, Inc.

References

[1] S. Fukae, N. Omori, H. J. Mattausch, T. Koide, T. Inoue and T. Hironaka, *Optimized Bank-Based Multi-Port Memories through a Hierarchical Multi-Bank Structure, Proc. of SASIMI2003*, pp. 323-330 (2003).

[2] H. J. Mattausch, K. Kishi and T. Gyoten, *Area-efficient multi-port SRAMs for on-chip data-storage with high random access bandwidth and large storage capacity, IEICE Trans. Electron., Vol. E84-C, No. 37*, pp. 410-417 (2001).

[3] D. Burger and T. Austin, *The SimpleScalar tool set Version 2.0, Univ. of Wisconsin-Madison Compt. Sci. Dept. Tech. Rep. #1342*, (1997).

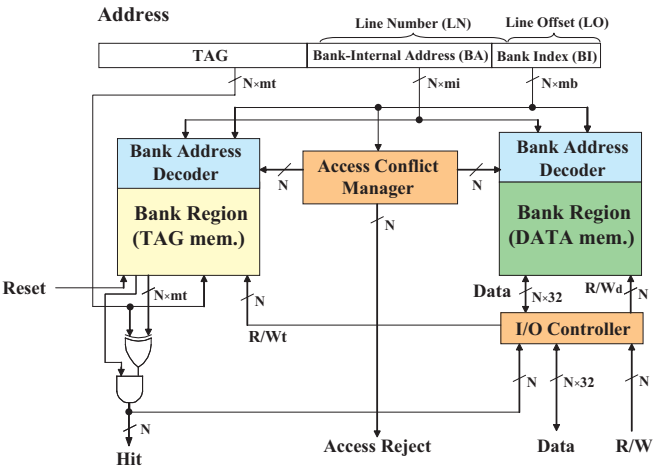

Fig. 1 Block diagram of direct mapped data cache with N ports in bank-based HMA structure. "R/Wt", "R/Wd" are Read/Write enable signals for tag or data memory, and "mt" , "mi", "mb" are bit length of tag, Bank-Internal Address (BA), Bank Index (BI), respectively.

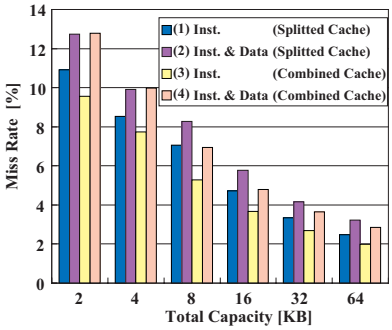

Fig. 2 Miss rate of splitted and combined direct-mapped cache. (1), (3) are instruction miss rate for splitted or combined caches, and (2), (4) are total miss rate for splitted or combined caches, respectively.

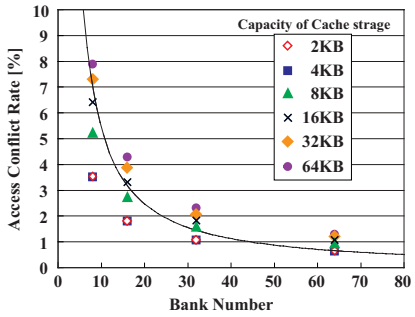

Fig. 3 Access conflict rate of combined cache.

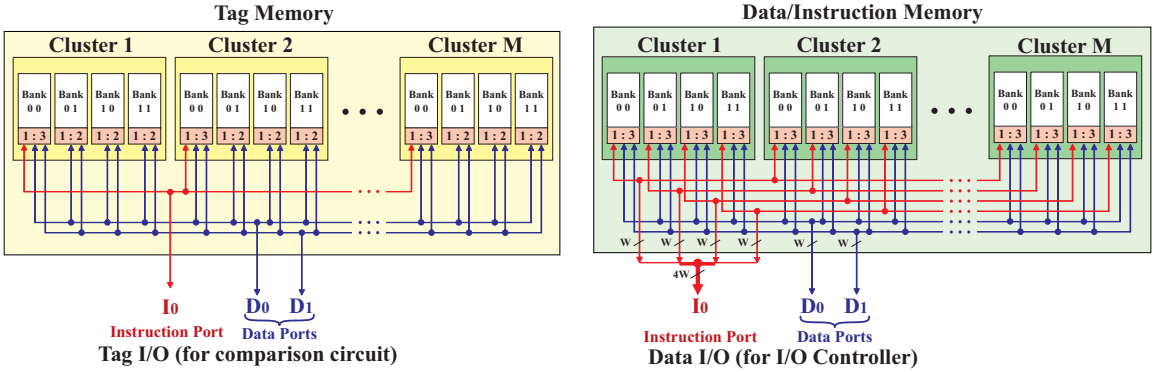

Fig. 4 Proposal of a combined data/instruction cache for a 4-way superscalar processor. The instruction port has 4 times the word length of the 2 data ports.

Table I Datasheet of the test-chip for an HMA cache design.

Technology, Supply Voltage	0.18 µm CMOS with 5 Al layers , 1.8 V
Chip size	2.8 mm × 2.8 mm
Port number, Data word length	4 ports , 32 bits
Bank number	2 Tag banks and 2 Data banks
Bank capacity, Bank size	1KByte , 390 µm × 435 µm
Bank-access time (Simulated)	< 1.0ns
Power dissipation (Simulated)	65 mW @500MHz

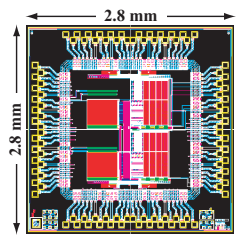

Fig. 5 Layout of the test chip for 4-port cache.

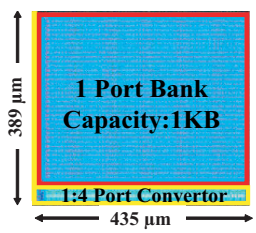

Fig. 6 Layout of a bank with 1:4-port convertor.