Unified Data/Instruction Cache with Distributed Crossbar, Hidden Precharge Pipeline and Dynamic CMOS Logic

K. Johguchi[1], Z. Zhu[1], K. Aoyama[1], Y. Mukuda[1], H. J. Mattausch[1], T. Koide[1] and T. Hironaka[2]

[1]Hiroshima Univ., Research Center for Nanodevices and Systems,1-4-2 Kagamiyama, Higashi-Hiroshima, 739-8527, Japan
[2]Hiroshima City Univ., Faculty of Computer Sciences, 3-4-1 OzukaHigashi, AsaMinami-Ku, Hiroshima, 731-3194, Japan
Phone: +81-82-424-6265, FAX : +81-82-424-3499, E-mail: {jouguchi, hjm, koide}@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 access bandwidth of the cache with a 1-port-bank-based Hierarchical Multi-port memory Architecture (HMA), which can simultaneously realize small area, high 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 dimentional 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 (BI) and a bank number (BN). BI is used for selecting a cache word or tag within memory banks, and BN is used for selecting the respective banks within data/instruction or tag memory. BN 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. Unification of Instruction and Data Cache

Conventional cache systems adopt split instruction and data caches to enable independent and parallel access to data and instructions, although only 1-port caches are used. A unified data/instruction cache can provide lower miss rate, because dynamical allocation of the effective storage capacity for data and instructions becomes possible. However, the accesses to the instruction cache are nomally consecutive. For a 4-way superscalar processor, it is therefore expected, that one instruction port with 4-time larger word length will deliver sufficient instruction-fetch performance. The optimum number of data-access ports is estimated to be 2 or 3. Then, we propose an HMA unified write-through cache which have different word length for data and instruction ports[3].

4. Circuit Methodology and Test-Chip Design

4.1. Dynamic CMOS Circuits and Hidden Precharge

Increased access time is a general problem for multi-port memories, because larger integration area leads to larger capacitances in the critical access path and port-switching as well as access-conflict management are required. The conventional access method of a synchronous memory divides the clock cycle into two phases: the memory access phase and the precharge phase, the latter to prepare the memory for the access in the following clock cycle. Therefore, the cycle time becomes the sum of the actual memory-access time and the precharge time. We propose to exploit the hierarchical structure of the multi-bank cache for a 2-stage synchronous access scheme, which hides the precharge by overlapping the access phase of one stage with the precharge phase of the other stage as illustrated in Fig. 2. Bank-conflict management, bank selection, port conversion and bank-internal wordline decoding constitute the 1st stage of the access path and are carried out simultaneously with the bank precharge when the clock is "0". After the clock becomes "1", the second stage of the access path, starting with the wordline-driver activation is executed, while the first stage of the access path is precharged. In this way equal clock-cycle time and actual multi-port cache access time are realized.

Dynamic CMOS technology is used for the design of the cache circuits in order to reduce in particular the capacitive loads of the global routing to the banks. Fig. 3 explains the construction of a bank and its 1-to-*N* port converter. In the shown example of the dynamic data-transfer subcircuit within the 1-to-*N* port converter, simple NMOS transfer gates and a PMOS transistor for precharging the internal node are used. Read and write path are divided in order to enable the proposed 2-stage synchronous access scheme. Reduced parasitic capacitances and smaller area of the 1-to-*N* port converter are thus achieved.

4.2. Test Chip Design

For the test chip of an HMA cache memory, a configuration with 4 ports was chosen and the design was carried out in a 5 metal, 0.18μm CMOS technology. The chip-layout shown in Fig. 4 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 and effective floor planning. The area-overhead of the 1:4-port convertor for the 1Kbyte bank of Fig. 5 is less than 25%. We also applied a new access method which overlaps bank-conflict management and bank decoding with the precharging phase of the banks. The timing diagram and spice simulation results of the netlist extracted from the layout are shown in Fig. 6 for the critical access cases. In the 1st stage the write access is critical, which takes longer than the read access because write data latching in the 1-to4 port converter completes after address latching. The 2nd access stage is limited by the read access because the read data must be transmitted from the bank to a cache-output port. Simulated critical delay times for 1st and 2nd stage are both slightly below 1.7 ns. This leads to a duty ratio of 50%, which is beneficial for the processor implementation. Simulated power dissipation is 247 mW at 250 MHz when all ports actively access the cache.

5. Conclusions

In this paper, a bank-based unified data/instruction cache with multiple ports has been proposed and its 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, so that access performance close to an ideal multi-port cache can be realized. Adopting an efficient floorplan without routing-only

areas restricts the area-overhead including 2nd level circuits and a conflict manager for the 4 ports to only 25% in comparison to a 1-port cache. A minimum clock cycle time of 3.4 ns 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, correponds to the port number and is independent of the total number of banks in the cache.

Acknowledgements

This research is sponsored 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.

This work was supported in part by the 21st Century COE program and a JSPS research fellowship, No.1605246, Japanese Government.

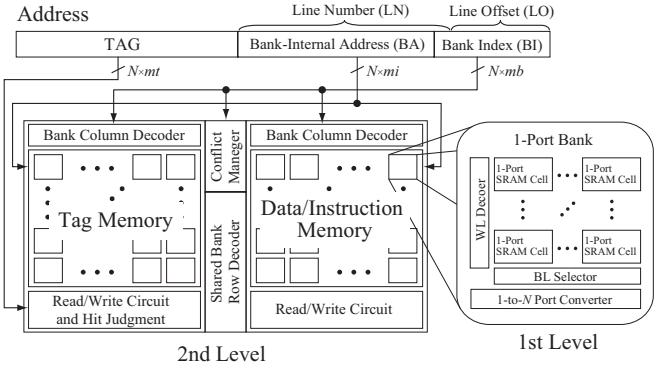


Fig. 1 Block diagram of direct mapped data cache with *N* ports in bank-based HMA structure.

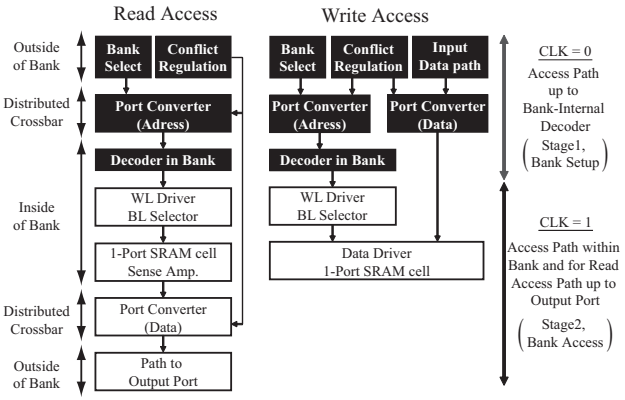


Fig. 2 Synchronous 2-stage access mode for hiding precharge phases.

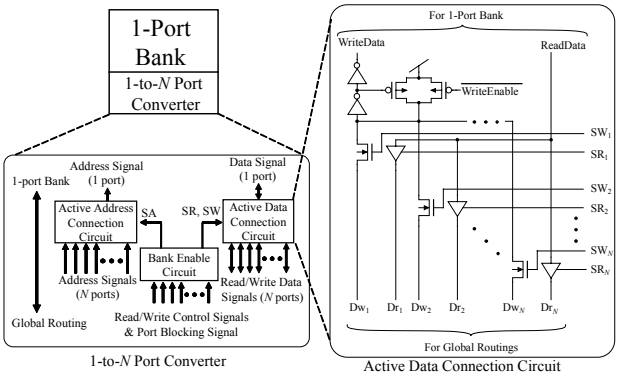


Fig. 3 Block diagram and partial schematic of 1-to-*N* port converter.

References

[1] S. Fukae, *et. al.*, "Optimized Bank-Based Multi-Port Memories through a Hierarchical Multi-Bank Structure," Proc. of SASIMI2003, pp.323-330, 2003.

[2] H. J. Mattausch, *et. al.*, "Area-effcient 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] K. Johguchi, *et. al.*, "Distributed-crossbar architecture for area-efficient combined data/instruction caches with multiple ports," Elec. Lett. 40, pp. 160-162, 2004.

Table I Datasheet of the design.

Technology	Logic CMOS w/o SRAM rules
Minimum Gate Length	200 nm
Routing Layer	1-Poly, 5-Al layers
Si-Area	6.2 mm^2
Total Storage Capacity	20.5 KByte
Port Number	4 ports
Minimum Access Cycle Time	3.4 nsec
Power Dissipation	247 mW at 250 MHz
Instruction & Service Port	
Port Number	2
Wordlength	64 bit
Data Ports	
Port Number	2
Wordlength	16 bit
Tag Memory	
Storage Capacity	4.5 KByte
Bank Number	16
Bank Capacity	2304 bit
Data/Instruction Memory	
Storage Capacity	16 KByte
Bank Number	64
Bank Capacity	2Kbit

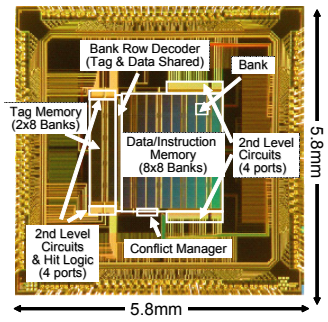


Fig. 4 Microphotograph of the test chip for 4-port cache.

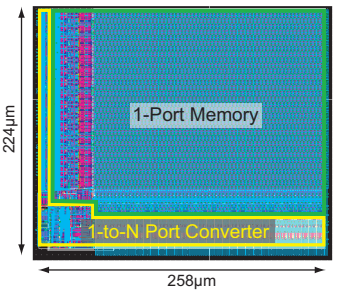


Fig. 5 Layout of a bank with 1-to-4-port convertor.

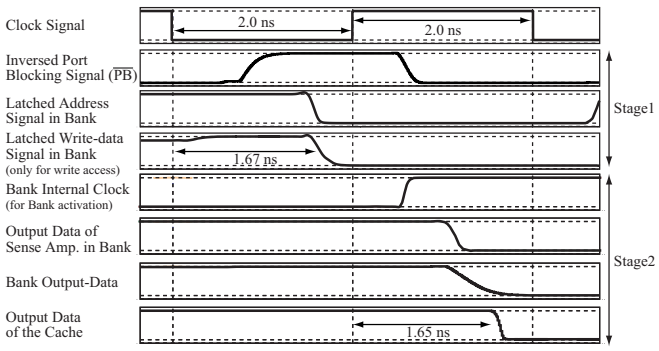


Fig. 6 Simulated access waveforms from extracted layout.

Unified Data/Instruction Cache with Distributed Crossbar, Hidden Precharge Pipeline and Dynamic CMOS Logic

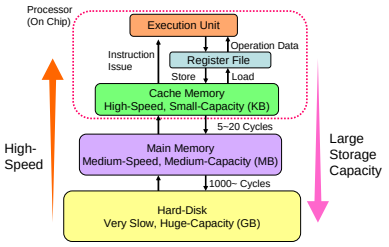
○Koh Johguchi[1], Zhaomin Zhu[1], Ken-ichi Aoyama[1], Yuya Mukuda[1], Hans Jürgen Mattausch[1], Tetsushi Koide[1], Tetsuo Hironaka[2]

[1] Research Center for Nanodevices and Systems (RCNS), Hiroshima University
[2] Department of Computer Engineering, Hiroshima City University

Hiroshima University
Nanoelectronics for Tera-Bit Information Processing

Background & Motivation

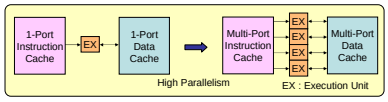
Memory Hierarchy of Processors

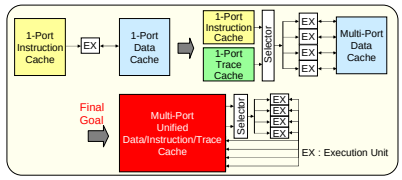


Multi-Port Cache Applications

Multiple-Issue Processor ⇨ Instruction Cache : Multiple instruction fetches per cycle
Data Cache : Multiple load/store operations per cycle

High performance multi-port cache is needed.





Motivation
Designing multi-port cache memories with small-area and high-performance

Combination of Instruction and Data Caches ⇨ Flexible cache-line usage, Reduced miss rate and storage capacity

Special Issues of Conventional Multi-Port Cache & Proposed Solution

Area Efficiency Problem

Multi-Port SRAM Cell
- Multi-port switches are connected directly to SRAM cells.
- Area increase is proportional to square of port number.
- Enough capacity for low miss rate cannot be achieved.

Multi-Bank Type Multi-Port Memory
- Multi-port switches are connected to banks with 1-port SRAM cells.
- High area efficiency, because of 1-port SRAM cells.
- High access conflict rate, because of arbitration.

High performance multi-port cache request :
Enough bank number is necessary for bank structure with high area efficiency

Solution for Multi-Port Caches

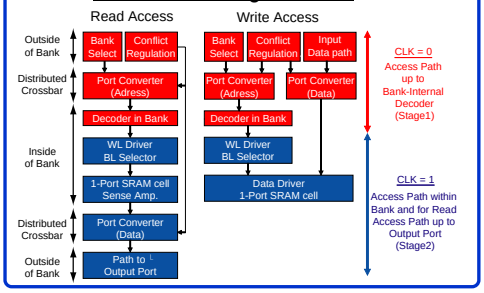
(1) Low area efficiency
- Application of 1-port SRAM bank based multi-port memory (HMA).
- Dynamic CMOS technology can be used for port-converters and conflict manager.

(2) Increase of delay time for multi-port memory
- Synchronous 2-stage access mode for hiding precharge phases with dynamic CMOS logic.

(3) Difference of the access pattern between data and instruction caches (in order to unify of two caches)
- Development of a new multi-port cache with the port-dependent word-length and interleaved cache-line words

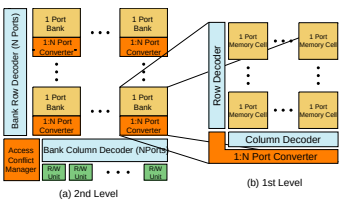
Solution for Delay Time

Hidden Precharge Method



HMA Cache Architecture

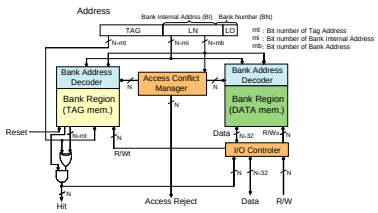
Hierarchical Multi-port memory Architecture (HMA)



- 1:N port converter is integrated into the banks to achieve modularity.
- Bank Decoders are two dimensional.

⇨ Applicable to for large bank numbers

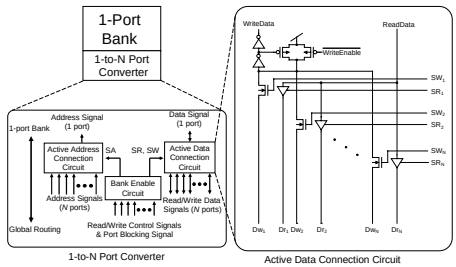
HMA Application of Multi-Port Caches



- Tag memory and data memory both use HMA.
- Bank number is determined by the lower bits of the index, so that conflicts will not take place when addresses are consecutive.

Dynamic CMOS Logic

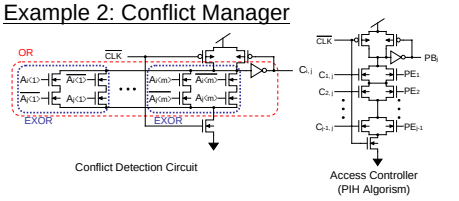
Example 1: 1-to-N port Converter



(1) Separated signal lines for read and write access within the bank
(2) Simple NMOS switch in write-access path (no tri-state buffer)
(3) Additional precharge circuit and keeper transistor (PMOS)

⇨ High speed and low power dissipation because of reduced parasitic capacitances

Example 2: Conflict Manager



- Designing conflict detection circuit by using multi-input EXNOR gates in dynamic CMOS logic
- Port-disable signal (PE) and conflict-detection signal ($C_{i,j}$) are evaluated by dynamic OR-AND gates.

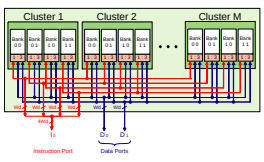
Unified Data/Instruction HMA Cache Design

Unified Data/Instruction Cache

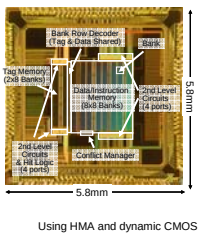
Deference of access pattern between instruction and data caches
- Instruction Cache : It accesses a consecutive address. → Large wordlength is needed.
- Data Cache : The pattern is comparatively random. → Large port number is needed.

Improvement of connection of port and bank

Port-dependent wordlength is achieved.



4-Port Unified Cache Test Chip



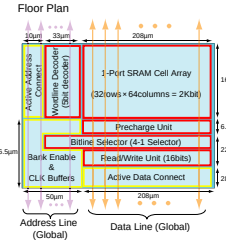
Specification

Technology	Logic CMOS w/o SRAM rules
Minimum Gate Leng	200 nm
Routing Layer	1-Poly, 5-Al layers
Si-Area	6.2 mm²
Total Storage Capacity	20.5 KByte
Port Number	4 ports
Minimum Access Cycle Time	3.4 nsec
Power Dissipation	247mW at 250MHz
Instruction & Service Port	
Port Number	2
Wordlength	64 bit
Data Ports	
Port Number	2
Wordlength	16 bit
Tag Memory	
Storage Capacity	4.5 KByte
Bank Number	16
Bank Capacity	2304 bit
Data/Instruction Memory	
Storage Capacity	16 KByte
Bank Number	64
Bank Capacity	2 Kbit

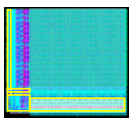
Using HMA and dynamic CMOS technology, we have achieved high speed and area-efficiency, simultaneously.

Layout Design of Bank (1st Level)

Floor Plan

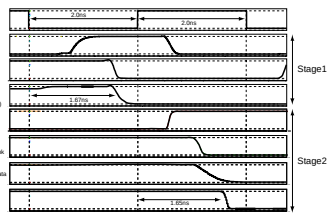


Layout Result



Height = 224 μm
Width = 258 μm
The area-overhead for multi-port capability is about 16%!!
(Considering only 1st level)

Simulated Waveform



Conclusions

- Test-chip design of a unified data/instruction HMA cache with 4 ports, 16KByte and 64 data-banks.
  - Cycle time is 3.4ns and area overhead for multi-port capability is less then 25%.
- Unified data/instruction cache advantages.
  - Reduced miss rate
  - High area efficiency
- Concept proposal for High speed multiport memory access
  - Hidden precharge pipeline method
  - Dynamic CMOS Logic