The MINC chip: Multistage Interconnection Network with Cache control mechanism chip

Takashi Midorikawa, Takayuki Kamei[1] , Toshihiro Hanawa[2] , Hideharu Amano
Dept. of Computer Science, Keio University
3-14-1, Hiyoshi Kohoku-ku, Yokohama, Kanagawa, 223-8522 Japan
Email: {midori, kamei, hanawa, hunga}@aa.cs.keio.ac.jp

Abstract

The Multistage Interconnection Network with Cache control mechanism (MINC) is a hardware mechanism to control the cache coherent in a switch-connected multiprocessors using a crossbar or Multistage Interconnection Network(MIN). In the MINC, the directory is located on the shared memory module, and the Reduced Hierarchical Bit-map Directory schemes(RHBDs) are used to reduce the directory. In order to reduce unnecessary packets caused by compacting the bit map in the RHBD, a small cache called the pruning cache is introduced in the switching element.

Using the 0.6 μm LPGA(Laser Programmable Gate Array) technology, the MINC chip(Network scale: 16-input/16-output) is implemented in a single chip(26477 cells including 60Kbit memory). The chip works with at least 50MHz clock, and the total throughput of the chip is 800Mbyte/sec(50Mbyte/sec for each port).

1 Introduction

Although bus-connected multiprocessors have been widely used as high-end workstations or servers, the number of connected processors is strictly limited by the maximum bandwidth of the shared bus. Instead of them, a switch-connected multiprocessor which uses a crossbar or Multistage Interconnection Networks(MINs) for connecting processors and memory modules is a hopeful candidate. However, in such a system, a snoop cache technique in bus-connected multiprocessors cannot be used, and consistency problems must be solved for providing the cache memory between a processor and the switch.

Then, hardware approaches by making the best use of advanced VLSI technology have been proposed[1][2]. However, traditional methods require a large memory outside the switching element and it causes not only a large additional hardware but also the extra latency by accessing the outside memory. Moreover, the complicated MIN with cache or directory must also treat data packet which should be transferred quickly.

To solve these problems, we proposed the MINC (MIN with Cache control mechanism)[3]. In the MINC, the MIN which only transfers a part of the address and cache coherent messages is separated from the data transfer network, and pushed into an LSI chip called the MINC chip. The coherent control is done based on the directory using the Reduced Hierarchical Bit-map Directory scheme(RHBD). In order to reduce unnecessary packets, the pruning cache which is a small cache enough to implement inside the chip is introduced in the MINC chip.

2 Overview of the MINC

2.1 The directory management method

The key idea of the MINC is a cache directory scheme called the RHBD[4] which was proposed for a massively parallel processor JUMP-1[4] which provides a hierarchical direct network. It can be easily applied to the MIN because of its embedded tree (hierarchical) structure.

Using the RHBD, since multicast does not require to access the directory in each hierarchy, quick message transfers can be performed. However, processors which don't share the cache line receive unnecessary messages and they may cause the congestion of the network. To reduce these packets, the pruning cache is introduced in switching elements of a stage.

RHBD scheme

The bit-map of the hierarchical directory is equipped only in the main memory module and reduced using two techniques.

[1] Now joint to Toshiba
[2] Now joint to Tokyo Enginnering University

- using the common bit-map for all nodes of the same level of hierarchy (tree), and
- a message is sent to all children of the node (thus, broadcasting) when the corresponding bit in the map is set.

By the combination of these techniques, several schemes are delivered[4]. We adopted the simplest scheme (SM:Single Map), since it is advantageous when the number of processors is not so large.

The reduced directory is not stored in each hierarchy but stored only in the root. Message multicast is done according to the reduced bit-map attached to the message header. Using this method, since multicast does not require to access the directory in each hierarchy, quick message transfers can be performed. However, unnecessary packets are transfered. Although they are just discarded in the cache controller of the processor, it may cause the congestion of the network.

Pruning cache

The major disadvantage of the RHBD is that it requires unnecessary message multicast since it only uses a single bit map or broadcasting in each level of the hierarchy. From the simple evaluation, the RHBD generates more than 100 times messages as really required messages in the worst case[4].

To cope with this problem, "the bit-map pruning cache" is proposed for the MINC chip[3]. The bit-map pruning cache is a small cache of the bit-map referred by the line address of multicasting packets.

When a processor issues the first read access of a cache line, the read request is transferred to the memory module. At the same time, the corresponding pruning cache entry and the bit-map is set. The bit-map is updated when other processors transferred read request of the same cache line to the memory module.

When the invalidation message or updating data is multicast, the bit-map pruning cache is checked. If there is an entry which matches the address, the bit-map of the pruning cache is utilized for multicasting. Otherwise, multicast is done according to usual RHBD schemes(Figure 1.) After the invalidation,the entry is immediately deleted.

2.2 Multiprocessors based on the MINC

Figure 2 illustrates a switch-connected cache coherent multiprocessor based on the MINC. This system consists of the following components.

Processing Unit (PU) with a private cache : The private cache is a simple write-through cache

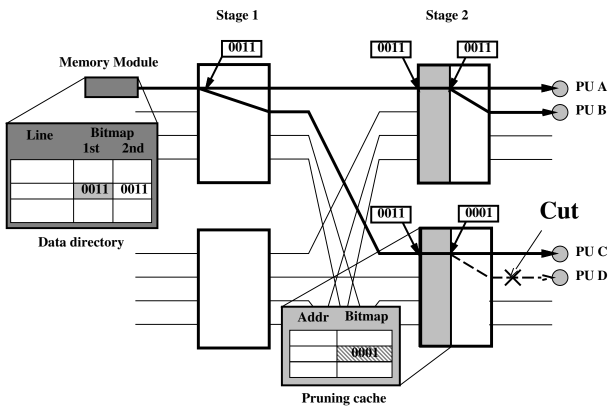

Figure 1: Pruning cache based on the RHBD

which stores copy of the shared memory module. The structure of the cache controller is almost same as the snoop cache controller for bus-connected multiprocessors.

Data Transfer Network : Cache lines, writing scalar data and vectors are transferred with this network. Any type of high bandwidth network like crossbars, MINs, or data exchanger chips can be used.

The MINC chip: Cache coherent messages must be multicast according to the bit-map of the RHBD. The MINC chip is a dedicated network chip which transfers only a part of address and messages to maintain cache consistency of the private cache.

Memory module : The shared memory and the bit-map of the RHBD are stored here. The memory controller manages the cache directory and generates packets for the MINC chip and the data transfer network.

2.3 Multicasting the invalidation messages

When a processor writes into the shared data on its cache, other copies must be invalidated. In this case, the invalidation message must be transferred to all caches which hold the copy of data. First, the write address is transferred to the memory module using the data transfer network, and the RHBD bit-map which indicates the location of copies is pushed into its packet header. Then, this invalidation packet is inserted into the MINC chip.

Each switching element of the MINC checks the bit-map within the packet header, and multicast the invalidation message. Since multiple multicasting messages may conflict each other, acknowledge bit-map is returned immediately to the input buffer(Figure 3.) In

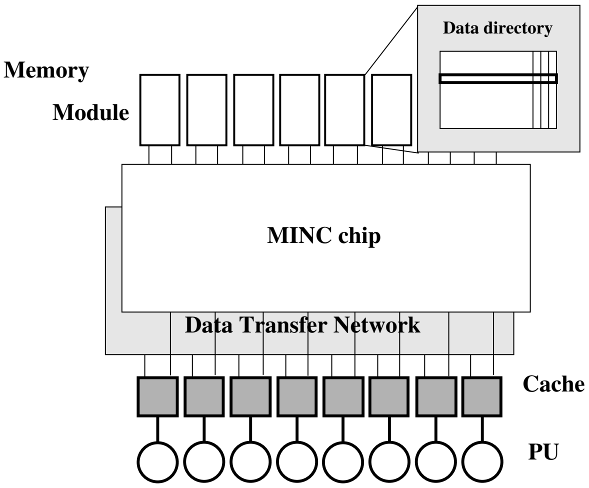

Figure 2: The multiprocessor with the MINC

the next frame, the packet is transferred again only to the destination corresponding acknowledge bit is 0.

3 Chip Design

As shown in Figure 4, the MINC chip consists of input buffers and 2-stage bi-directional omega networks. The operation of the MINC chip is based on the Simple, Serial, and Synchronized (SSS) style control mechanism[5].

Inserted packets are synchronized with a unique frame clock in the input buffer. When packets conflict each other inside a switching element, one packet is selected and transferred to the desired destination. Others are discarded and inserted again from in the next frame. Although this control mechanism causes the loss of synchronization, structure/control of switching elements can be simplified, and the high frequency clock can be used.

The bit-map pruning cache is omitted except the switching elements nearest to the processor. This is based on the result of the simulation study which shows the pruning cache works most effectively when it is provided in the nearest stage of the processor[3].

4 Implementation

The MINC chip was developed by the pilot program of the VDEC design curriculum. In this pilot program, the type of chip is limited to be the $0.6\mu m$ ChipExpress's LPGA(Laser Programmable Gate Array) which has 100k gates (recommended 50k gates and 64Kbit memory cells) at maximum in the PGA package with 391 pins (264 pins for signals). From these limitations, basic design of the MINC chip is decided as follows:

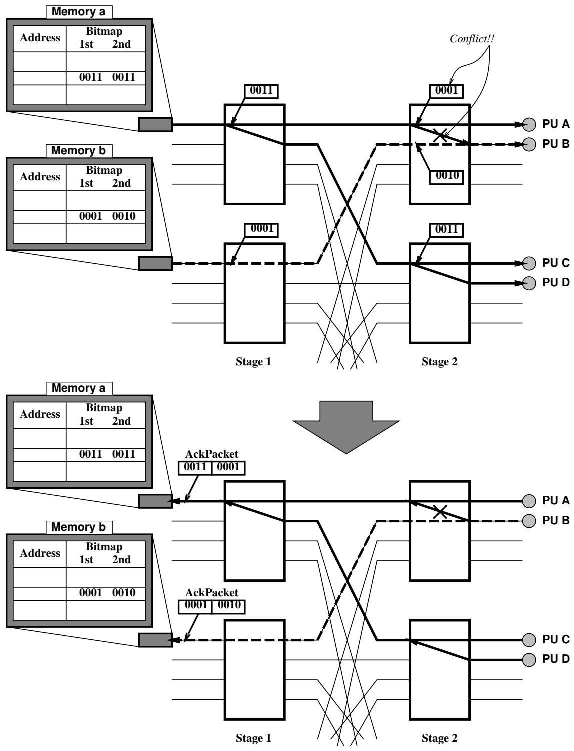

Figure 3: Multicast and Acknowledge

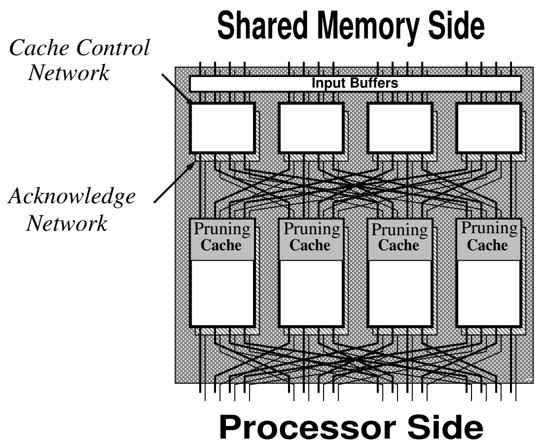

Figure 4: The structure of the MINC chip

Network Scale : Sixteen inputs/outputs are provided. Thus, at most sixteen processors and memory modules can be connected with this chip. Eight pins/wires are used for each link: four for forward packets and others for acknowledge.

Network structure : The MINC chip consists of small switching elements connected in the multiple stages. Considering the hardware requirement and performance the size of the switching element is set to be 4×4. Thus, the 2-stage omega network with 4 ×4 switching elements is used.

Pruning Cache : From the limitation of the RAM inside the chip, 256 entries two set associative cache is provided in every switching element of the stage connecting to processors. The capacity miss rarely occurs with this size of cache. Although the structure is almost the same as the common cache controller, the matching and accessing the bit-map is done without disturbing the packet transfer.

Table 1 shows the specification of MINC chip. Since a number of inverters were inserted to adjust the hold time error, used gates becomes larger than recommended. These introduce difficulty of wiring and degrade the maximum speed. However, 50MHz clock can be used and 400Mbyte/sec total throughput can be obtained.

These designs are described in Verilog-HDL, and synthesized with Synopsys Design Compiler. Place-and-route is done by the original CAD from ChipExpress Corp.

Clock	50MHz
Bit width	4bit(packet transfer)
	4bit(acknowledge)
Size	16-in/out
Max bandwidth	$400 Mbits/sec \times 16$
the number of cell(basic cell)	26477
the number of cell(memory)	60Kbit
signal pin	262pin
Technology	$0.6\mu m$ LPGA
1 Frame	30 clock

Table 1: the specification of the MINC chip

5 Conclusions

The MINC chip designed here multicasts cache coherent messages based on the RHBD quickly by using the SSS control structure and the pruning cache. Using the 0.6 μm LPGA technology, it works at least 50MHz clock and the total throughput becomes 400Mbyte/sec. A multiprocessor called SNAIL-2 is now under development by using this MINC chip and the PBSF chip[6] as the high speed data transfer network. The real performance and efficiency of the MINC chip will be evaluated on this multiprocessor.

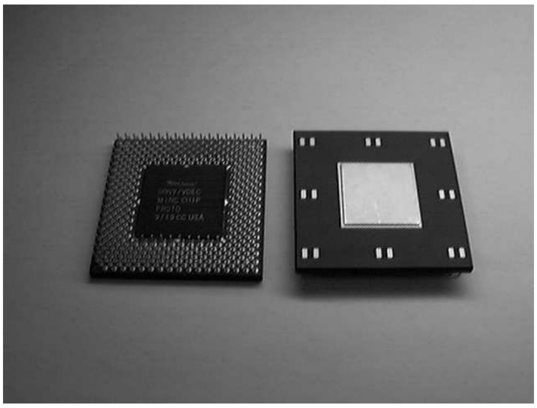

Figure 5: The MINC chip

Acknowlegements

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 with the collaboration by ChipExpress Corporation and Sony Corporation. The authors appreciate with the staff of the VDEC and thank Dr. Makoto Ikeda who is a staff of the VDEC for his help.

References

[1] L.N. Bhuyan, A.K. Nanda, T. Askar, "Performance and Reliability of the Multistage Bus Network", *Proc. of International Conference on Parallel Processing*, pp.I-26–I-33, 1994

[2] A.K. Nanda, L.N. Bhuyan, "Design and Analysis of Cache Coherent Multistage Interconnection Networks", *IEEE Trans. on Computers*,42(4) pp.458–470, 1993

[3] T. Hanawa, T. Kamei, H. Yasukawa, K. Nishimura, H. Amano:, "MINC: Multistage Interconnection Network with Cache control mechanism" *IEICE Transactions on Information and Systems*, Vol. E80-D, No. 9, pp. 863–870, September, 1997.

[4] T. Kudoh, H. Amano, T. Matsumoto, K. Hiraki, Y. Yang, K. Nishimura, K. Yoshimura, and Y. Fukushima. "Hierarchical bit-map directory schemes on the RDT interconnection network for a massively parallel processor JUMP-1," *Proc. of International Conference on Parallel Processing*, 1995.

[5] H. Amano, L. Zhou, K. Gaye, "SSS(Simple Serial Synchronized)-MIN: a novel multi stage interconnection architecture for multiprocessors," *Proc. of the IFIP 12th World Computer Congress*, Vol.1, pp.571-577, Sep. 1992

[6] T. Kamei, M. Sasahara, H. Amano, "An LSI implementation of the Simple Serial Synchronized Multistage Interconnection Network", *Proc. ASP-DAC'97*, pp.673-674, 1997