The MINC(Multistage Interconnection Network with Cache control mechanism) chip

Takashi Midorikawa Takayuki Kamei* Toshihiro Hanawa Hideharu Amano

Keio University
Dept. of Computer Science
3-14-1, Hiyoshi Yokohama, Japan
Tel: +81-45-560-1063
Fax: +81-45-560-1064
e-mail: {midori,kamei,hanawa,hunga}@aa.cs.keio.ac.jp

I. 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.

To address this problem, hardware approaches by making the best use of advanced VLSI technology have been proposed. 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.

In order to solve these problems, we proposed the MINC (MIN with Cache control mechanism)[1]. 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.

II. Overview of the MINC

A. The directory management method

The key idea of the MINC is a cache directory scheme called the RHBD[2]. The bit map of the hierarchical directory is reduced and equipped only in the main memory module. Although the RHBD was proposed for a massively parallel processor JUMP-1[2] with a hierarchical direct network, it can be easily applied to the MIN because of its embedded tree (hierarchical) structure. In this scheme, the bit map is reduced using two techniques.

- 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 techniques, several schemes are delivered[2]. We adopted the simplest scheme (SM:Single Map), since it is advantageous when the number of processors is not so large.

Using the RHDB, 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 message and it may cause the congestion of the network.

B. Multiprocessors based on the MINC

Figure 1 illustrates a switch connected cache coherent multiprocessor based on the MINC. This system consists of the following components.

Processing Unit (PU) providing a private cache : The private cache is a simple write-through cache which stores copy of the shared memory module.

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

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

Memory module : the bit map of the RHBD is stored here. Thus, the memory controller manages the cache directory and generates packets for the MINC chip and the data transfer network.

For a high speed data also transfer network, commercial crossbars or data exchanger chips can be used. Thus, the most important component of this system is the MINC chip which multicasts the coherent messages.

*Now joint to Toshiba

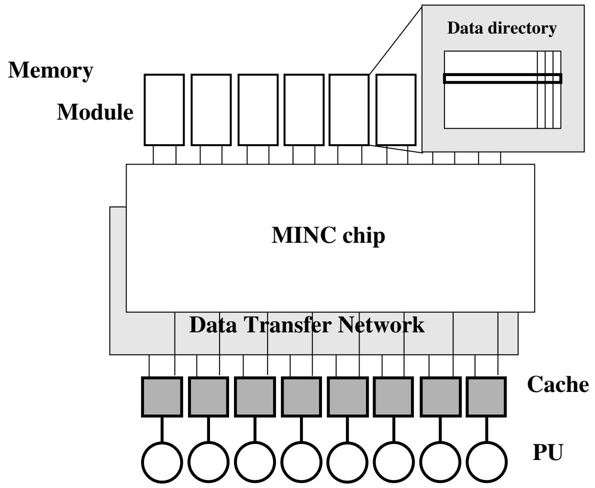

Fig. 1. The multiprocessor with the MINC

III. Chip Design and Implementation

A. Chip Design

As shown in Figure 2, 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[3].

In this style, inserted packets are synchronized with a unique frame clock in the input buffer. When packets conflict each other inside a switching element, a 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.

B. Implementation

The MINC chip was developed by the pilot program of the VDEC design curriculum. In the 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:

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 4x4. Thus, the 2-stage omega network with 4x4 switching elements is used.

Pruning Cache : From the limitation of the RAM inside the chip, 256 entries two set associative cache is provided in each switching element. The capacity miss rarely occurs with this size of cache.

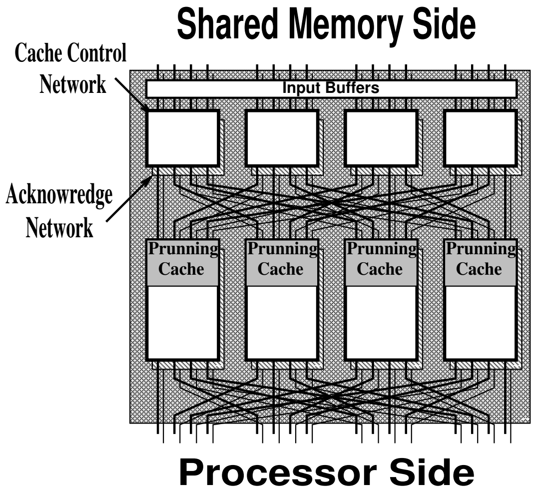

Fig. 2. The structure of the MINC chip

TABLE I
THE SPECIFICATION OF THE MINC CHIP

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 I 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. This introduce difficulty of wiring and degrades the maximum speed. However, 50MHz clock can be used and 400Mbyte/sec total throughput can be obtained.

The design is described in Verilog-HDL, and synthesized with Design compiler. Place-and-route is done by the original CAD from ChipExpress Corp.

Acknowlegement

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 Makoto Ikeda who is the staff of the VDEC for his help.

References

[1] T. Hanawa, H. Yasukawa, K. Nishimura, H. Amano, "MINC: Multistage Interconnection Network with Cache control mechanism." *Proc. of International Conference on Parallel and Distributed Computing Systems '96*, pp.310-317, 1996.

[2] 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.

[3] 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