FireCluster: PC Cluster System Employing IEEE 1394

Kazuki Hyoudou and Yasuichi Nakayama
Department of Computer Science,
The University of Electro-Communications
Chofu, Tokyo 182-8585 Japan
E-mail: hyoudo-k@igo.cs.uec.ac.jp

1. Introduction

PC cluster systems are attracting attention for their flexibility and cost-effectiveness. In parallel cluster computing, low latency and high bandwidth networks are required. It is also important that these networks are cost-effective.

In conventional PC cluster system research, high-performance networks such as Myrinet have been used. However, these networks are expensive and not so commonly used. On the other hand, Fast Ethernet is inexpensive commercially available hardware. However, as a network for cluster systems, its performance is not so good. Therefore, it is difficult to find a network satisfying the above-mentioned requirements. We are interested in IEEE 1394 as network device to possibly satisfy these requirements.

We propose a PC cluster system called the FireCluster which uses IEEE 1394 to connect nodes. We have constructed a PC cluster system using IEEE 1394 that supports a 400 Mbps data transfer rate. We have designed and implemented the communication facility with IEEE 1394 in the system. This facility is implemented without any modification of the operating system. It provides application with low latency communication. In addition, it provides a reliable group communication. Even though the FireCluster is constructed at a very low cost, it provides a good latency time performance.

2. Over view of IEEE 1394

The IEEE 1394 standard is a low cost high-performance serial bus. IEEE 1394 is well known as the FireWire or the iLink, and it is regarded as the next-generation standard for connecting PCs to multimedia devices.

It has two data transfer types, adopts memory-based communication, provides 400 Mbps data transfer rate and adopts hot-plug connectors.

Two data transfer types are as follows:

- *Isynchronous transfer* that requires data delivery at constant intervals. These transfers provide guaranteed quality of service at the user-level, but do not perform flow control at the physical layer.

- *Asynchronous transfer* that does not require delivery at a constant data rate. Because these transfers perform flow control at the physical layer, packet disappearance does not happen.

It is reported that IEEE 1394 switches will appear soon. IEEE 1394 is inexpensive commercially available hardware and provides a good performance. Therefore, it is possible to construct a PC cluster system which is more cost-effective by using IEEE 1394.

3. Design of communication facility

We design a communication facility considered the features of IEEE 1394 for our system. This facility supports low-latency peer-to-peer communication and reliable group communication. See [1] for technical details.

We adopt the following models for the peer-to-peer communication:

- *Data transfer using remote memory write*
  We adopt the connection-oriented communication model and use the remote memory write that is supported by hardware. It reduces software overhead because a target can be identified by only its address in memory space. We use the *Physical Write* which is supported by the Open Host Controller Interface for the IEEE 1394 specification.
for remote memory write. By using this feature, reception node can receive data without interruptions to host processors. Therefore, communication latency time is reduced.

- **User-level communication**
  Our communication facility provides applications with direct access to the network interface. By removing kernel trapping and data copy to buffer in kernel, it can reduce software overhead.

- **Flow control using double-buffering**
  The sliding-window flow control, which has some buffers of fixed size, is not so practical on IEEE 1394. The reason is that host memory is not used efficiently, because maximum packet size for IEEE 1394 is large.

  Therefore, we suggest a flow control using double-buffering. In this method, the sending and reception node allocates buffers of same size. It sends a control message only when switching buffers.

4. Experiment

The system used in our experiment consisted of eight computation nodes and a front-end computer. We used the IOI-1394TTO 1394-to-PCI host adapters. This adapter is Open HCI compliant and supports a 400 Mbps data transfer rate. We used Fast Ethernet adapters and a switching hub to connect each node. Fast Ethernet is used for working of the system, such as a sharing filesystem. We measured round-trip time (RTT), peer-to-peer bandwidth and multicast bandwidth used the system.

Our facility achieved 21.8 microseconds round-trip time in 4 bytes data transfer (Fig. 1) and 251 Mbps bandwidth in 4096 bytes data transfer (Fig. 2). In addition, the multicast supported by our facility reached about 190 Mbps bandwidth regardless of the number of reception nodes. Compared with Fast Ethernet, it provided good performance.

5. Conclusion

We have constructed a PC cluster system using IEEE 1394 which supports a 400 Mbps data transfer rate to connect nodes. We have designed and implemented the communication facility with IEEE 1394 for use in the system.

This facility provides low-latency communication and supports a reliable group communication. Even though our system is constructed at a very low cost, compared with Fast Ethernet, it provided a good performance. Especially in minimum latency time, our communication facility is not inferior to Myrinet. Therefore, it is useful for parallel applications which communicate small data frequently.

Böszörmenyi et al. also reported on research of the cluster system employing IEEE 1394 [2]. We have, for the first time, implemented and evaluated the system as long as we know.

References
