A Simplified Communication Protocol on Dawning 2000

Ma Jie
National Research Center for Intelligent Computing Systems
Institute of Computing Technolog
Chinese Academy of Sciences
Beijing 100080, P.R.China
majie@ncic.ac.cn

Abstract

Point to point latency and bandwidth are two important factors to evulate a cluster communication system. SCP is a Simplified Communication Protocol on Dawning 2000. It is a low level communication system based on PowerPC/AIX. The latency of SCP is limited to 9.6μs. And the peak bandwidth is reached when message length is 4K. The paper mainly discusses the key points about design and implementation of SCP. The paper gives the idea to improve the performace of low level communication system by simplifying communication protocol, using zero memory copy, providing an efficient communication data path and balance the work on host and NIC. Multiple services are provided to optimize upper layers.

1. Introduction

In design and implementation of Dawning 2000, the cluster communication system, BCL(Basic Communication Library), is a key issue. Since we need a high reliability system, BCL's implementation is very complex. This affects the performance of the whole system. A simplified communication protocol, SCP, is provided in this paper to improve system performance. Our work is directly in the ideas that motivated the GM and BIP in order to get rid of the costly protocol stack. The design of SCP is mainly based on our BCL protocol using on Dawning 2000.

2. SCP protocol

SCP is a low level communication layer based on PowerPC/AIX platform. We achieve 9.6μs of latency and 65MB/s of bandwidth. The goal of SCP is to provide a low level communication protocol with high performance, and give the upper layers an efficient interface.

The key points of the design and implementation are:

- Zero memory copy
- Efficient communication data path
- Balance the work on host and NIC
- Multiple services to optimize upper layers

Since SCP is a low level communication layer, there are some limitations to users. Some such limitations are the following:

- SCP does not support multi-process
- SCP can only detect error while transfer
- SCP has no flow control on system buffer

2.1. Zero memory copy

Since the network bandwidth is much higher than memory copy bandwidth, avoid memory copy can improve the system's bandwidth. SCP is designed for zero memory copy transfers. Using zero memory copy means user buffer should be DMAable memory. User buffer should be pinned in the physical memory and the virtual address should be translated to physical address before start communication. This will decrease the communication bandwidth. In order to avoid this overhead, SCP separates pin-down memory operation from message-passing. Since SCP is a low level communication layer, the upper layers may use one buffer serval time with only one pin-down operation and address translation operation.

2.2. Efficient communication data path

With the improvement of PCI bus bandwidth and network bandwidth, cost of transfering one page (4K) message

Proceedings of the Fourth International Conference/Exhibition on
High Performance Computing in Asia-Pacific Region
0-7695-0589-2/00 $10.00 © 2000 IEEE

is decreased. This means every additional overhead added to the data path will have more effect on the final performance.

To decrease the additional system overhead, we can transfer more pages in one message-passing operation. In this case, the average overhead will decrease. But it needs to transfer very large message to reach the peak bandwidth. And the small message bandwidth is not very high. In SCP, it is concentrated to eliminate the system overhead on the communication data path to improve the small message bandwidth.

2.3. Balance the work on host and NIC

Today, most of the NICs used in cluster communication systems have their own processors. With the NIC's processor, host processor can do computing while NIC transfers messages. But it is not said that we can use NIC's processor to do anything because that the NIC's processor is not as fast as host processor. In a word, balance the work on host processor and NIC's processor will improve the system's performance.

In SCP, we simplified the NIC's program and use event driven technique to make the program suitable for the NIC's processor.

2.4. Multiple services to optimize upper layers

SCP is a low level communication layer. And it is designed for the upper layers. In SCP, there are some extra services which have the same function but have some different options. These services can be used to optimize the upper layer software's performance. There are two main options in SCP's services.

- Choice between I/O and DMA

  Using DMA to transfer data has a higher bandwidth than using I/O. But Using DMA need to pin-down user buffer and convert virtual address to physical address. Usually, I/O communication is used to transfer small messages and DMA communication is used to transfer large messages. The boundary between small message and large message is determined by the performance of these two kind of communication.

  In SCP, we separate pin-down memory operation from message-passing. This means we can use DMA to transfer small messages if the buffer has been pinned down and the address has been converted. So we did not define the boundary in SCP. It is the upper layer's duty to decide whether to use I/O or to use DMA. This can be used to optimize the upper layer software.

- Buffer alignment

  DMA operation needs the physical address of the DMAable memory to be contiguous. Since the buffer with contiguous virtual address may not have contiguous physical address, DMA operation can be used only within a page boundary. The maximum length of one DMA operation is 4K (one page). If the buffer was not page aligned, we need to start DMA twice to transfer one page data. This will affect DMA bandwidth.

  In SCP, we defined a group of special send/receive services to use page aligned buffer. They have the same function as the normal services but need a page aligned buffer. Since the upper layer software is not the end-user's program, it can be optimize to use page aligned buffers. The page-aligned services give the upper layer method to optimize their performance.

3. Performace and conclusion

Here we give the latency and bandwidth obtained by a user-level program written directly on top of the SCP protocol. We use two PowerPC 300 workstations, M2M-PCI32C NICs and M2M-OCT-SW8 switch to test the performance. The performance shows as below:

Latency	9.6μs
Bandwidth	65MB/s

Myrinet RAW DMA performance over this kind of machine is 71MB/s. SCP reached 92% of the total bandwidth. The peak bandwidth is reached with 4KB messages. And the 1/2 peak bandwidth is reached with less than 1K message. One way latency is 9.6μs.

References

[1] Kai Hwang, Zhiwei Xu, Sealable Parallel Computing: Technology, Architecture, Programming, WCB/McGraw-Hill, 1998

[2] Loïc Prylli, Bernard Tourancheau, BIP: a new protocol designed for high performance networking on Myrinet, In Workshop PC-NOW, IPPS/SPDP98, Orlando, USA, 1998

[3] Patrick Geoffray, Loïc Prylli, Bernard Tourancheau, BIP-SMP: High Performace Message Passing over a Cluster of Commodity SMPs, SC99

[4] Loïc Prylli, BIP user reference manual, Technical Report TR97-02, LIP/ENSLYON, 1997

[5] Gregory Henley, Nathan Doss, Thomas McMahon, Anthony Skjellum, BDM: A Multiprotocol Myrinet Control Program and Host Application Programmer Interface, Technical Report MSSU-EIRS-ERC-97-3

[6] Myricom. The GM Message Passing System, http://www.myri.com/GM/doc/, 1999