LETTER

Performance Evaluation of Grid Computing with Parallel Routes Transmission

Hiroyuki MIYAGI[†a)], **Yusuke OKAZAKI**[†], **Ryota USUI**[†], **Yutaka ARAKAWA**[†], **Satoru OKAMOTO**[†], ***and*** **Naoaki YAMANAKA**[†],

SUMMARY In a grid computing environment, the network characteristics such as bandwidth and latency affect the task performance. The demands for bandwidth of wide-area networks become large and it reaches more than 100 Gbps. In this article, we focus on parallel routes transmission, such as link aggregation, to realize large bandwidth network. The performance of grid computing with parallel routes transmission is evaluated on the emulated wide-area network.

key words: *Grid Computing, Parallel Routes Transmission, Link Aggregation, Optical Network, Parallel and Distributed Computing*

1. Introduction

With the advance of network technologies and high performance computing, research on grid computing is very popular [1]. Grid computing is the technique by which a high performance virtual machine can be created by combining computers via networks. Conventional grid computing assumes LAN(Local Area Network) as network environment, which is a high speed network. In recent years, 10 Gbps LAN is popular and the 100 Gbps LAN technique [2] is emerging. On the other hand, the bandwidth of inter-LAN connection becomes large rapidly. Inter-LAN connection is assumed as dedicated Ethernet among LANs in this article. With the assumption that the network bandwidth of inter-LAN connection expands more and more at low cost and users can use it more securely, research on applied computational science and technology, and wide-area distributed storage system become popular. Large bandwidth makes it possible to execute large scale tasks by cluster computers connected by inter-LAN connections. Large scale tasks require inter-cluster bandwidth more than 100 Gbps. The wide-area distributed storage system also requires more larger bandwidth, because a large amount of data are transferred frequently. If the 100 Gbps bandwidth is available in inter-LAN connections, it is possible to make the wide-area distributed storage system on line. In grid environment, the effect that data transmission on grid computing cannot be negligible. Therefore, 100 Gbps bandwidth networks are required to improve grid computing performance. There are two techniques to realize 100 Gbps bandwidth network. One is serial route transmission and the other is parallel routes transmission. Serial route transmission transfers data with a single route, while parallel routes transmission transfers data with multiple routes. A serial route transmission of 100 Gbps bandwidth is one of the ideal solutions, since frame reordering is not be occurred in the data route. However, the economical cost becomes very high because of much expensive equipments. So, it is a distant idea to apply serial route transmission of 100 Gbps bandwidth into grid computing today. On the other hand, frame reordering may be occurred in parallel routes transmission, since it is realized by bonding multiple dedicated Ethernet among clusters. But, its economical cost is not high, because it can be realized using existing low cost 10 Gbps Ethernet techniques.

In this article, to realize 100 Gbps bandwidth networks, a parallel transmission technique is evaluated. Parallel routes transmission is realized by extending link aggregation to network wide. Link aggregation is the method to achieve larger bandwidth logical link by bonding of two or more links into a single link.

2. Flow Aggregation Methods for Making Parallel Routes Transmission

There are two typical aggregation methods as frame load distribution algorithms. One is flow-based aggregation and the other is round-robin aggregation.

2.1 Flow-Based Aggregation

Flow-based aggregation algorithm determines the route to transfer frames according to data flows. Figure 1 shows an example of flow-based aggregation. In this example, flow-based aggregation makes logical single route by bonding three routes. The data flow from node A to node C uses route 1, and the data flow from node B to node D uses route 3. If the number of data flows is lower than the number of aggregated routes, the utilization efficiency of aggregated routes becomes low. Although three routes are aggregated, the number of data flow is two in Fig.1.

2.2 Round-Robin Aggregation

When the switch receives frames, it forwards one frame out of each output port in the aggregated group using a round-robin scheme. Figure 2 shows an example of

[†]Department of Information and Computer Science, Keio University 3-14-1 Hiyoshi, Kohoku, Yokohama, 223-8522, JAPAN

a) E-mail: miyagi@yamanaka.ics.keio.ac.jp

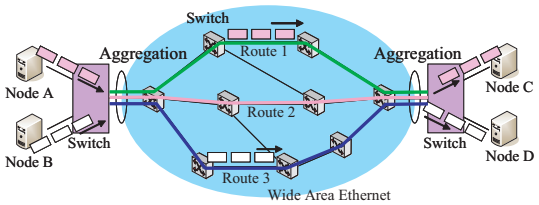

Fig. 1 Flow-Based Aggregation.

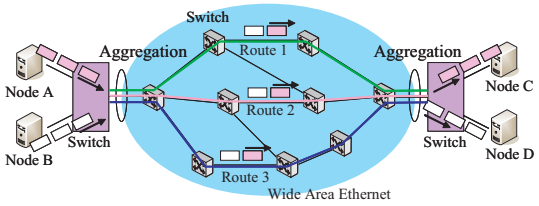

Fig. 2 Round-Robin Aggregation.

round-robin aggregation. When node A sends frames to node C, the frames from node A are transmitted using all routes in the aggregation group. When node B sends frames to node D, similarly, the frames from node B are transmitted using all routes in the aggregation group. Therefore, the utilization efficiency of network bandwidth is high. On the other hand, frame reordering may be occurred when the delay of each route is different.

3. Experimental System

To evaluate the performance of grid computing with parallel routes transmission, we have examined the task execution time and the file transfer time by experiments, since computer simulations can not reflect the burstness of grid applications such as data transfer. Figure 3 shows the experimental system. The environment that two clusters are connected by inter-LAN connection is assumed. So, the inter-LAN connection emulated network between two clusters is added some delay into the whole network. Two types of networks are prepared, one is serial route transmission and the other is parallel routes transmission. Since serial route transmission of 100 Gbps bandwidth is difficult to prepare, 10 GE is used in serial route transmission and GbE(Gigabit Ethernet) is used in parallel routes transmission. Parallel routes transmission is realized by using flow-based and round-robin aggregation in case of 2 parallel, 4 parallel and 8 parallel. Each cluster system is implemented with 4 computing nodes as shown in Table 1. Intel Xeon 2.33 GHz Quad Core CPU is used and the total number of CPUs is 32. 10 GE NIC(Network Interface Card) is attached to each node. GridMPI-1.1 [3] is used as MPI system. Inter-process communications of MPI use TCP/IP. BIC-TCP [4], which is a default TCP used in linux kernel 2.6, is used. The default, maximum and minimum TCP window size of sender

Table 1 The specifications of computing machines used in the experiments.

CPU	Intel Xeon 2.33 GHz(Quad Core)
Memory	8 GByte DDR400
NIC	Myricom Myri10GE
NIC Driver	Myricom1.3.1
OS	Fedora Core 6 (Linux-2.6.22)

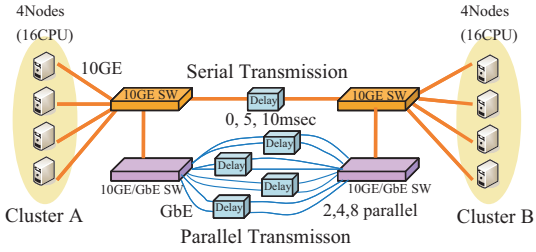

Fig. 3 Experimental System.

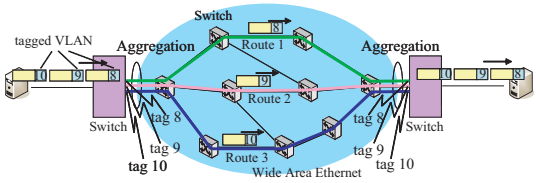

Fig. 4 Implementation of Round-Robin Aggregation.

side is 64 kByte, 16 MByte and 4 kByte, respectively. The default, maximum and minimum TCP window size of receiver side is 85 kByte, 16 MByte and 4 kByte, respectively. MTU(Maximum Transfer Unit) is set 9000 Byte in all devices.

3.1 Implementation of Round-Robin Aggregation

End-to-end round-robin aggregation is implemented for the experiment. Figure 4 shows the implemented round-robin aggregation. Tagged-VLAN of linux kernel is extended that the tags of the frames sent from each computing node are allocated by round-robin method as shown in Fig.4. The frame from each computing node is assigned VLAN tag value range from 8 to 15, since the number of aggregated routes is 2, 4, or 8 in the experiments. The switches which aggregate multiple routes are configured with VLAN. When the frames with a VLAN tag input to each switch, the frames are switched to the corresponding route. When the number of aggregated routes is 2, for example, the route 1 uses VLAN tags from 8 to 11 and the route 2 uses from 12 to 15.

3.2 Implementation of Network Emulator

Tasks are executed with computing nodes in two clusters. They are connected by inter-LAN connection. So, 5 msec and 10 msec delay are added into network routes by using PC based delay emulaters. 5msec delay approximately equals to RTT from Tokyo to Osaka.

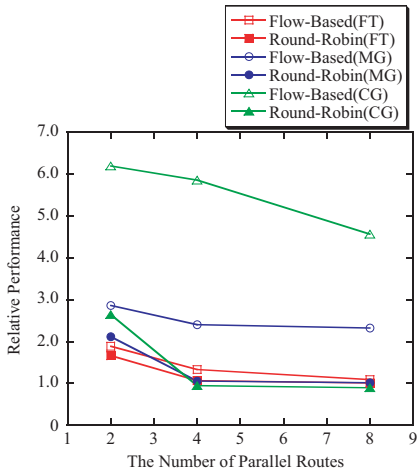

Fig. 5 The relative performance of execution time versus the number of parallel routes. The relative performance is defined as the ratio of execution time when using 10 GE serial route without delay. Network delay is zero.

Netem [5] which provides network emulation functionality of layer 2 and layer 3 is worked on linux machines.

4. Experimental Results

4.1 Performance Evaluation of Task Execution Time

NPB(NAS Parallel Benchmarks) [6] version 2.3 is used for the performance evaluation of task execution time. NPB benchmark suite consists of eight programs. FT(Fast Fourier Transform), MG(MultiGrid) and CG(conjugate Gradient) applications are focused on. Almost the same characteristics in other five applications are observed by the experiments. Figure 5 shows the relative performance of execution time versus the number of parallel routes transmission compared with that of 10 GE serial. Here, the relative performance of execution time is defined as the ratio of execution time when using 10 GE serial route transmission without any delay. Therefore, the relative performance of 10 GE serial route transmission without any dealy is 1.0. It is ideal to adopt serial route transmission, since frame reordering is not occurred. As the number of parallel routes transmission increases, the relative performance improves. This is because parallel routes transmission makes network bandwidth large. From Fig.5, round-robin aggregation also achieves ideal performance with 4 and 8 parallel routes transmission. Compared with round-robin and flow-based aggregation, round-robin aggregation improves performance more than that of flow-based aggregation. From Fig.5, the difference of relative performance spreads more than five times. This is because round-robin aggregation improves the utilization efficiency of aggregated routes, and abbreviates the data transfer time compared with flow-based aggregation. Figure 6 shows the time to transfer 8 frames using

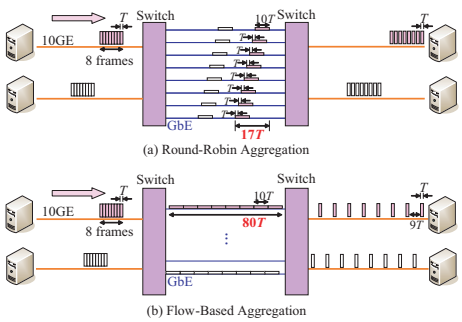

Fig. 6 The transfer time of round-robin and flow-based aggregation.

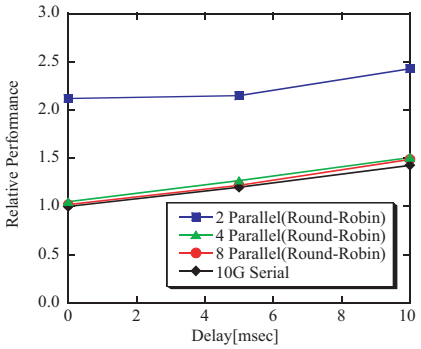

Fig. 7 The relative performance of execution time of MG versus network delay.

round-robin and flow-based aggregation. If the time to transfer one frame is defined as T on 10 GE, the time on GbE is $10T$. As shown in Fig6(a), the time to transfer 8 frames by using round-robin aggregation is $17T$, because round-robin aggregation can use all routes efficiently. On the other hand, the time by using flow-based aggregation is $80T$, because the output ports of the switches are determined by its data flow. In grid environment, processes change information by message passing frequently, therefore, a large difference of the relative performance appears between round-robin and flow-based aggregation.

Figure 7 shows the relative performance versus network delay of MG applications. As the network delay increases, the relative performance degrades in all methods. This is because the network delay affects message passing. As the number of parallel routes transmission is increases, the relative performance of round-robin aggregation is improved. Round-robin aggregation with 8 routes achieves almost the same performance as 10 GE serial. The same tendency is also confirmed in FT and CG applications. Therefore, parallel routes transmission using round-robin aggregation is effective when the number of routes becomes large.

4.2 Performance Evaluation of File Transfer Time

The file transfer time has an effect on the grid performance in the application such as the wide-area distributed storage system. In this experiment, we evalu-

ate the time to transfer a 100 GByte file between clusters. Because of the frame reordering, the throughput of parallel routes transmission using round-robin aggregation degrades if the delay of each route is different. To evaluate the effect of differential delay, network delay is added into a single route. In this experiment, GNET-1 [7] is used as the inter-LAN connection emulator, which is an FPGA-based configurable network testbed which can precisely control network delay in 10 nsec order while Netem can control network delay in msec order.

Figure 8 shows the file transfer time versus network delay. In Fig.8, the bold line and the dashed line show the theoretical time using 10 GE serial route transmission and flow-based aggregation, respectively. As serial route transmission and flow-based aggregation do not occur frame reordering, each theoretical time is constant. However, the theoretical time of flow-based aggregation is constant regardless of the number of aggregated routes. On the other hand, round-robin aggregation improves file transfer time, as the number of aggregated routes increases. However, its file transfer time is degraded as a single network delay increases. From Fig.8, it observed that the time of round-robin aggregation is not degraded within 300 μsec. However, the file transfer time of round-robin aggregation increases rapidly, when network delay becomes more than 300 μsec. TCP retransmission control makes throughput low because of frame reordering. The tendency is remarkable as the number of aggregated routes is large. Then, it takes 72 μsec to transfer a 9000 Byte frame on GbE. Therefore, frame reordering is often occured when network delay becomes more than 72 μsec. The default window size of receiver side is set to 85 kByte, a few frame reordering is tolerant. However, the number of frame reordering becomes more larger, TCP makes frame loss of frame reordering and the throughput becomes low.

Hence, there is an issue to improve throughput when round-robin aggregation is used. It is necessary to control the delay of each route within 300 μsec on GbE. It corresponds to 30 μsec on 10 GE. 30 μsec delay on 10 GE corresponds to 6 km in terms of a fiber length. 100 Gbps bandwidth route is realized by bonding several 10 Gbps routes. Such a delay is generated if aggregation is used for inter-LAN connections. Therefore, the mechanism to control a delay is an important issue and it is required few μsec order accuracy for 100 Gbps parallel transmission.

5. Conclusion

The performance of grid computing with parallel routes transmission has been evaluated by experiments. Parallel transmission is realised by bonding dedicated Ethernet among clusters. In the evaluation of task execution time using NPB, the results showed that round-robin aggregation can improve task execution time compared with that of flow-based aggregation. In the evaluation of file transfer time, the result showed that the file transfer time of round-robin aggregation can improve when the differential delay of each route is less than 300 μsec in the GbE based parallel transmission systems, which corresponds to 30 μsec in the 10 GE based system. The feasibility of 100 Gbps parallel transmission system was confirmed.

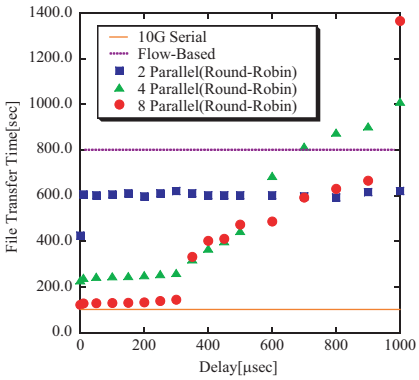

Fig. 8 File transfer time versus a single network delay. A 100 GByte file is transferred between two clusters.

Acknowledgment

This work was partially supported by “Lambda Access” Project funded by the National Institute of Information and Communications Technology (NICT).

References

[1] I.Foster , C.Kesselman, “The grid: blueprint for a new computing infrastructure,” Morgan Kufmann, Nov. 1998.
[2] Soichiro Araki, “Photonic Service Gateways in the Japan’s Lambda Utility Project,” The 2nd International OFC/NFOEC Workshop on the Future of Optical Networking (FON), Anaheim, CA, USA, Mar. 2007
[3] M. Matsuda, T. Kudoh, Y. Kodama, R. Takano and Y. Ishikawa, “TCP Adaptation for MPI on Long-and-Fat Networks,” Proc. of IEEE International Conference on Cluster Computing 2005, pp.1-10, Boston, Massachusetts, USA, Sep. 2005.
[4] S. Ha, L. Le, I. Rhee, and L. Xu, “Impact of background traffic on performance of high-speed TCP variant protocols,” Computer Networks: The International Journal of Computer and Telecommunications Networking, vol.51, issue 7, pp.1748-1762, May 2007.
[5] Net:Netem, http://www.linux-foundation.org/en/Net:Netem
[6] D. Bailey, T. Harris, W. Saphir, R. van der Wijngaart, A. Woo and M. Yarrow, “The NAS Parallel Benchmarks 2.0,” International Journal of Supercomputer Applications, 1995. http://www.nas.nasa.gov/Resources/Software/npb.html
[7] Y. Kodama, T. Kudoh, R. Takano, H. Sato, O. Tatebe and S. Sekiguchi, “GNET-1: Gigabit Ethernet Network Testbed,” Proc. of IEEE International Conference on Cluster Computing 2004, pp.185-192, San Diego, California, USA, Sep. 2004