社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2004－DPS－117 (8)
2004－CSEC－ 24 (8)
2004／3／4

Multi-Striping: Multicast Protocol with Network Striping Approach on Grid Computing Environment

Yasutaka Nishimura, Tomoya Enokido, and Makoto Takizawa
Dept. of Computers and Systems Engineering
Tokyo Denki University
{*yasu, eno, taki*}*@takilab.k.dendai.ac.jp*

Abstract

This paper discusses a novel type of high-performance data transmission protocol for multicasting multimedia messages in a Grid network. A Grid network is composed of various types of computers interconnected in types of networks. There can be multiple routes from a sender process to each of destination processes in a Grid network. Multimedia data can be in parallel transmitted to multiple processes by using multiple routes. In addition, messages can be replicated by not only in a same route but also in different routes. Some packets lost can be recovered from the redundant parity packets. We evaluate the protocol in terms of jitter and effective packet loss ratio, compared with traditional tree routing.

マルチストライピング: グリッドコンピューティング環境下におけるストライピング転送を用いたマルチキャストプロトコル

西村 康孝　榎戸 智也　滝沢 誠
東京電機大学理工学部情報システム工学科
E-mail {yasu, eno, taki}@takilab.k.dendai.ac.jp

現在の情報システムは複数のコンピュータがネットワークを通して相互接続され、協調動作を行う分散型のシステムとなっている。分散型のシステムにおいて、複数の組織間で資源（プロセッサ、メモリ、ストレージ）を共有し、使用者が資源を自由に使用できるグリッドコンピューティングが注目されている。また、分散マルチメディアシステムではルータの対応を必要としないアプリケーションレベル (応用層) マルチキャストが用いられる場合がある。しかし、アプリケーションレベルマルチキャストでは経路が最適化されておらず、IP マルチキャストと比べて遅延時間が大きくなることが考えられる。本論文ではグリッドコンピューティング環境下における分散マルチメディアシステムを構築し、グリッドの処理能力を利用したマルチキャストプロトコルを提案する。

1. Introduction

Large number and various types of peer processes like personal computers (PCs) are interconnected with various types of networks like the Internet and gigabit networks in a peer-to-peer (P2P) system [8]. Peer processes are cooperating by exchanging messages. Especially, large number of multimedia data like image and video are required to be transmitted. Traditional communication protocols like TCP [6] and RTP [9] support processes with reliable one-to-one or one-to-many transmission of data. Through a connection among a pair of processes, a messages are efficiently and reliably transmitted from a process to one or more than one destination process. Recently, multiple connections are used to in parallel transmit data from a process to another process in the *network striping* like GridFTP [1], SplitStream [3], and PSockets [10]. In PSockets [10], data is divided into partitions and the data is striped over the sockets. In SplitStream [3], data is splited and each of striped data is transmitted in a different tree. In GridFTP [1], a high-performance file transfer protocol (FTP) is discussed by using multiple connections.

One approach to realizing the high-speed real-time communications of multimedia data is to use a high-speed network which is composed of high-performance routers and computers with high-speed channels like WDM [7] and ATM [2]. However, such a high-performance network is too expensive for every application to use. Another approach is to take advantage of a Grid architecture [4]. A Grid network is composed of huge number of computers, especially personal computers which are interconnected with various types of networks [4]. A Grid network can support higher transmission speed and more reliable communication than the high-speed networks by taking usage of parallel transmission of data. Even if the Grid network is composed of low-performance computers like personal computers (PCs) interconnected in low-performance networks, the Grid networks can support applications with reliable high-performance data transmission.

In this paper, we discuss how to reliably and efficiently exchange multimedia messages among multiple processes in a Grid network. A multimedia message is decomposed into a sequence of packets. A packet is a unit of data transmission in the Grid network. Packets may be lost and be delayed due to congestions and faults in the network. In order to increase the bandwidth from the source process to each destination process, packets of the message can be in parallel transmitted. That is, the packets are transmitted by using multiple connections in the network striping.

A process sends each packet of a message to multiple destination processes to multicast the message. It takes time to send each packet to multiple destinations. In order to decrease the overhead to multicast packets, a multicast

tree is constructed in the Grid network. Here, each node forwards a packet to a limited number of nodes where the number of the nodes is decided by the processing rate of the sender node. The more number of nodes messages are sent to, the longer processing rate is required. The multicast tree is constructed so that the delay time is minimized.

Packets may be lost and delayed. A parity packet is transmitted for some number of packets. Even if one packet is lost or delayed, the packet lost and delayed can be obtained from the other packets.

In section 2, we present a system model. In section 3, we discuss protocol to multicast messages with the network striping and duplication in a Grid network. In section 4, we evaluated the protocol in term of jitter and effective packet loss ratio compared with types of multicast protocols.

2. System Model

A system is composed of multiple application processes interconnected with a Grid network. A Grid network G is composed of nodes $G_1, \ldots, G_m$ ($m \geq 1$) which are interconnected with communication channels. Nodes are interconnected with types of networks like 100base-TX, 1000base-T, and ATM. A node is realized by a type of computer, mainly personal computer (PC). Processes exchange messages through the Grid network. A process can communicate with one or more than one node in the Grid network. The nodes supporting a process are referred to as *service nodes* of the process. A message is a unit of data transmission among application processes. A process first decomposes a message to a sequence of packets. A packet is a unit of data transmission in the Grid network. Each node receives packets and then forwards the packets to one or more than one node in the Grid network.

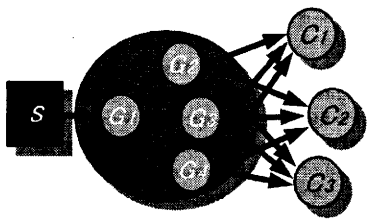

Figure 1. System model.

A process first requests some service node in the Grid network to deliver packets to multiple processes, i.e. multicast packets. Here, the service node for the process is referred to as *root* node. For example, a process S sends a sequence of packets of a message m to a node G_1 in a Grid network G as shown in Figure 1. Each node forwards packets received to one or more than one node. The root node G_1 forwards packets to three nodes G_2, G_3, and G_4. Then, the nodes deliver the packets to destination processes. A node which delivers packets to a destination process is referred to as *leaf* node. For example, the leaf nodes G_2, G_3, and G_4 deliver packets to processes C_1, C_2, and C_3. Thus, a process is connected with one or more than one node. For example, a process G_i receives packets from three leaf nodes G_2, G_3, and G_4.

In the Grid network, a multicast tree from the root node to the leaf nodes is constructed as shown in Figure 1. In traditional multicast trees [5], a same packet is forwarded to leaf nodes from a root node through root-to-leaf routes. That is, on receipt of a packet, a node forwards the packet to all the child nodes in the multicast tree. In other approaches, each packet may be carried by different root-to-leaf routes. That is, on receipt of a pair of packets p_1 and p_2, a node may forward p_1 and p_2 to different child nodes. In addition, the node forwards p_1 and p_2 in parallel. In this paper, we try to get high-performance real-time multicast communication of multimedia data by parallel transmission.

3. Multicast with Network Striping

Suppose that a process S multicasts a message to multiple processes $C_1, \ldots, C_n$ through a Grid network G. A message m is decomposed into a sequence of packets p_1, $\ldots$, p_l ($l \geq 1$). In traditional protocols, packets are sequentially transmitted to a destination process through a connection in a sending order. In our protocol, packets are duplicated and in parallel transmitted into each destination processes through multiple routes.

3.1. Network striping

In order to increase the throughput of data transmission, a message is transmitted from a process to each of destination processes by using multiple routes. This is referred to as *network striping* [10] which is used to transmit packets to realize the high bandwidth. Figure 1 shows a multicast tree in a Grid network G to multicast packets p_1, p_2, and p_3 from a source process S to there destination processes C_1, C_2, and C_3. Three are multiple routes from the sender process to each of the destination processes, i.e. G_1 to G_2, G_1 to G_3, and G_1 to G_4. For example, the root node G_1 can forward the first and second packets p_1 and p_2 to the node G_2, the third packet p_3 to G_3, and the fourth packet p_4 to G_4. Each of the node G_2, G_3, and G_4 forwards the packets to each destination process C_i. The destination process C_i ($i = 1, 2, 3$) receives the packets p_1, p_2, p_3, and p_4 from the leaf nodes G_2, G_3, and G_4. Thus, packets are in parallel transmitted in the multicast tree in the Grid network.

A Grid network is composed of various types of *node* computers like personal computers and workstations as nodes, which are interconnected with various types of communication channels from high-speed to low-speed channels. Each route from a root to a leaf node supports different quality of service (QoS), bandwidth, delay time,

and packet loss ratio. For example, a route from the root node to each leaf node G_i is realized in a connection from the root node G_1 to the nodes G_i ($i = 2, 3, 4$). Each route supports different bandwidth, i.e. the routes G_1 to G_2 with 15Mbps, G_1 to G_3 with 10Mbps, and G_1 to G_4 with 5Mbps, respectively. The node G_1 forwards a sequence of the packets $p_1, p_2, p_3, \ldots, p_l$ to the nodes G_2, G_3, and G_4 as shown in Figure 4. Here, a parity packet is transmitted for every three packets. For example, a parity packet R_{456} is transmitted for three packets p_4, p_5, and p_6. The node G_1 forwards six, four, and three packets to the nodes G_2, G_3, and G_4, respectively. The node G_2 delivers a sequence of six packets p_1, p_2, p_5, R_{123}, p_7, and p_9 to the destination processes C_1, C_2, and C_3. The node G_4 delivers three packets p_4, R_{456}, and p_{11} to the destination processes. The node G_1 distributes packets to each route channel on the bandwidth of the channel. Thus, the more number of packets are transmitted to the higher bandwidth channel. For example, G_2 forwards about 40%, 30%, and 20% of packets to G_2, G_3, and G_4, respectively, proportional depending to the bandwidth of each route.

A subsequence transmitted through a route is referred to as *stream* carried on the route. For example, $\langle p_1, p_2, p_5, R_{123}, p_7, p_9 \rangle$ is a stream on the route G_1 to G_2. Each destination process receives streams from multiple leaf nodes. The destination process reassembles a sequence of packets from the streams.

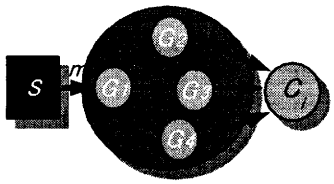

Figure 2. Striping transmission.

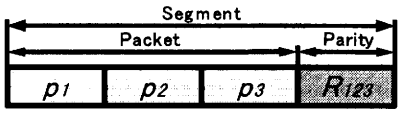

Figure 3. Attachment of parity data.

3.2. Redundant transmission

In order to increase the reliability of data transmission, packets are replicated by transmitting parity packets of the packets. A *parity* packet is transmitted for a subsequence of packets [11]. For example, one parity packet R_{123} is transmitted for a subsequence of three packets p_1, p_2, and p_3 as shown in Figure 3. A parity packet R_{456} is transmitted for three packets p_4, p_5, and p_6. A subsequence of packets with a parity packet is referred to as *segment*. A pair of subsequences $\langle p_1, p_2, p_3, R_{123} \rangle$ and $\langle p_4, p_5, p_6, R_{456} \rangle$ are segments. Here, as long as at most one packet is lost in a segment, the packet lost can be recovered from the other packets. For example, a packet p_2 can be obtained by computing the exclusive or (XOR) of the other packets p_1, p_3, and R_{123}. Let l be the number of non-parity packets in a segment. Here, the robustness ratio is given $1/(l+1)$. The redundant ratio of the segment $\langle p_1, p_2, p_3, R_{123} \rangle$ is 0.25. Even if at most 25% of packets in a segment are lost, all the data in the segment can be delivered.

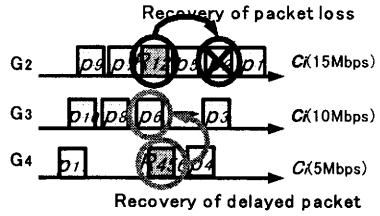

Figure 4. Recovery.

Suppose the packet p_2 is lost in the communication channel from the root node G_1 to the node G_2. A destination process C_i delivers packets p_1, p_3, p_4, p_5, and R_{123} but can not deliver the packet p_2 because any leaf node does not receive p_2. On receipt of the parity packet R_{123}, the packet p_2 can be obtained from the packets p_1, p_2, and R_{123} received as explained here. Thus, even if a packet is lost in a segment of packets, the packet lost can be recovered from the other packets without retransmission of the packet.

After receiving the packet p_5, the destination process C_i does not receive the packet p_6 due to congestions even if all the other packets are received, i.e. the packet p_6 is delayed. The destination process C_i has received the parity packet R_{456} and a pair of the packets p_4 and p_5. Here, the process C_i can obtain data to be carried by the packet p_6 from the packets p_4, p_5, and R_{456} without waiting for the packet p_6 delayed. Thus, even if a packet is delayed, the packet delayed can be delivered before the packet is received only if all the other packets in a segment are received. Since delayed packets can be recovered, the real-time constraint can be satisfied.

Packets in each segment are distributed in different streams. Since a segment is a unit of recovery, only destination process as can recover packets lost after collecting all packets in a segment from different leaf nodes. In addition to segments, each stream can be replicated so that packets lost and delayed can be recovered by each node. Let us take a stream $\langle p_1, p_2, p_3, R_{123}, p_4, p_5, \ldots \rangle$ as an example. Even if the packet p_2 is lost in between G_1 and G_2, the node G_2 cannot receive the packet p_2. One idea is that another parity packet is transmitted for some number of packets in each stream. For example, a parity packet S_{125} is transmitted for three packets p_1, p_2, and p_5 in the

stream. The node G_2 recovers the packet p_2 lost by using the packets p_1, p_5, and the parity packet S_{125}.

4. Evaluation

We evaluate the multicast protocol in a Grid network in terms of the number of packets lost and the jitter delay. There are two ways for multicasting messages, non-striping (NS) and striping (S) ones. In the non-striping (NS) way for multicasting packets, a sequence of packets are transmitted in a route from the root to every leaf node. That is, packets are serially transmitted in each route. In the striping (S) way, packets in a message are in parallel transmitted in multiple routes. Different packets can be transmitted in different routes.

There are other types of transmission, non-parity (NP) and parity (P) ones. In the non-parity (NP) one, messages are not replicated, i.e. no parity packet is transmitted. If some packet is lost, the packet is required to be retransmitted. On the other hand, messages are replicated by transmitting parity packets. Here, even if some packets are lost, the packets can be obtained by a destination process.

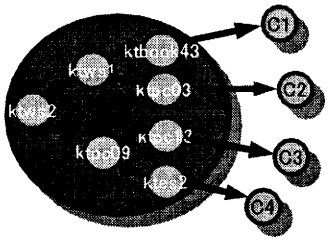

Figure 5. Evaluation of non-striping protocol.

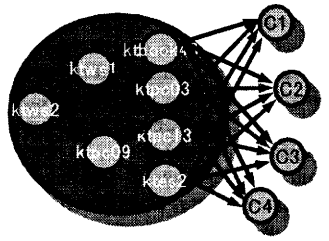

Figure 6. Evaluation of striping protocol.

In the evaluation, seven personal computers (PCs) are interconnected in a multicast tree as shown in Figures 5 and 6. Figure 5 shows a multicast tree for non-striping (NS) multicast and Figure 6 indicates a striping (S) multicast tree. By using the multicast tree, packets are multicast to four destination processes C_1, C_2, C_3, and C_4. A pair of computers are interconnected with a 100base-TX network. In the NP tree as shown in Figure 5, a same sequence of packets are transmitted in each of four root-to-leaf routes. On the other hand, each route carries a subsequence of packets, i.e. stream different from every other route.

Figure 7 shows the jitter and the packet loss ratio for four multicast trees, NS/NP, NS/P, S/NP, and S/P. A video data of 38M bytes is multicast to seven destination processes by using the multicast tree shown in Figures 5 and 6. Here, each packet is 10K bytes long. A parity packet is transmitted each time six packets are transmitted, i.e. the redundantly ratio is 16.7% in the type P multicast. The bandwidth of one stream is transmitted at 30Mbps.

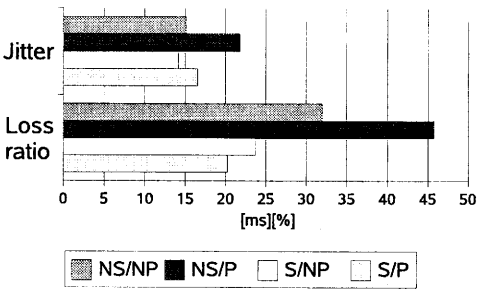

Figure 7. Evaluation of four transmission protocols.

Table 1. Evaluation result.

	NS/NP	NS/P	S/NP	S/P
Jitter[ms]	15.1	21.77	14.18	16.51
Loss ratio[%]	31.98	45.72	23.71	20.36

We measure jitter, i.e. how long it takes to receive six packets, i.e. 60K bytes. In the P type multicast, it takes a longer time to calculate the parity than the NP type. As shown in Figure 7 and Table 1, the jitter of the type S/P is about one [msec] longer than NS/NP while S/NP supports one [msec] shorter jitter than NS/NP. The jitter must be 15 [msec] to transmit continuous video data, i.e. 60K bytes in the half NTSC rate as Digital Video (DV). The our protocol. S/P, does not satisfy the constant , about 1 [msec] longer than the constant 15 [msec]. It take about two [msec] to calculate the parity to receive packets lost and delayed.

The packet loss ratio shows how many packets a destination process does not receive. In the type NP, the packet loss ratio is the same as one in the network. In the type P, some packets lost are recovered by each destination process. Figure 7 shows that 20.36% of packets are lost in the S/P type which is smaller than the S/NP.

In the NP type, only one connection, i.e. route is used to transmit packets to each destination process. If parity packets are transmitted, more number of packets are transmitted. Hence, the more number of parity packets are transmitted, the larger the packet loss ratio. On the other hand, multiple routes are used in the S types. Parity packets are transmitted in multiple routes in the S/P type. Even if parity packets are transmitted, the traffic in each connection is not so much increased as the NP type since the packet loss ratio is smaller.

In order to decrease the jitter, we need some high-performance computing mechanism to calculate the parity of multiple packets. We are now discussing how to reduce the jitter delay.

5. Concluding Remarks

We discussed the multicast protocol with network striping and redundant transmission for distributed multimedia systems on the Grid computing environment. We presented the high-performance application-level multicast based on the Grid network, multicast with network striping. We evaluated the protocol in terms of jitter and effective packet loss ratio through the experiment. The protocol implies smaller packet loss ratio than the other protocols. We are now disclosing how to reduce the jitter implied by the computation of parity packets.

References

[1] B. Allcock, J. Bester, J. Bresnahan, A. L. Chervenak, I. Foster, C. Kesselman, S. Meder, V. Nefedova, D. Quesnel, and S. Tuecke, Data management and transfer in high-performance computational grid environments. *Parallel Computing Journal*, 28(5):749-771, 2002.

[2] ATM Forum, *Traffic Management Specification Version4.0*, 1996.

[3] M. Castro, P. Druschel, A-M. Kermarrec, A. Nandi, A. Rowstron, and A. Singh, SplitStream: high-bandwidth multicast in a cooperative environment. *Proceedings of the nineteenth ACM symposium on Operating systems principles (SOSP2003)*, pages 298-313, 2003.

[4] I. Foster, C. Kasselman, and S. Tuecke, The anatomy of the grid: enabling scalable virtual organizations. *International Journal Supercomputer Applications*, 15(3), 2001.

[5] D. Pendarakis and S. Shi and D. Verma and M. Waldvogel. ALMI: An application level multicast infrastructure. *Proceedings of the 3rd USENIX Symposium on Internet Technologies and Systems (USITS)*, pages 49-60, 2001.

[6] J. Postel, Transmission control protocol, *Request for Comments*, 0793, 1992.

[7] B. Ramamurthy, *Design of Optical WDM Networks*, Kluwer Academic Publishers, 2001.

[8] N. S. Ross, L. Grahman, and G. Carroni, *Proc. of the Third International Conference on Peer-to-Peer Computing*, 2003.

[9] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, RTP: a transport protocol for real-time applications, *Request for Comments*, 1889, 1996.

[10] H. Sivakumar, S. Bailey, and R. L. Grossman, PSockets: the case for application-level network striping for data intensive applications using high speed wide area networks. *Proc. of the 2000 ACM/IEEE Conference on Supercomputing*, http://www.sc2000.org/proceedings/techpapr/papers/pap.pap240.pdf, 2000.

[11] T. Tojo, T. Enokido, and M. Takizawa. Notification-Based QoS Control Protocol for Multimedia Group Communication in High-Speed Networks. *accepted for publication in Proc. of the 24th International Conf. on Distributed Computing Systems (ICDCS 2004)*, 2004.