5 N − 8

Wide-area Group Communication *

Takayuki Tachikawa and Makoto Takizawa †
Tokyo Denki University ‡
e-mail{tachi, taki}@takilab.k.dendai.ac.jp

1 Introduction

Distributed systems are composed of multiple computers interconnected by communication networks. In distributed applications like teleconferences, a group of multiple processes is established and the processes in the group are cooperated. Group communication protocols support a group of processes with the reliable and ordered delivery of messages to multiple destinations. Group communication protocols discussed so far assume that every pair of processes have almost the same delay time and reliability. In the Internet, it takes about 60 msec to propagate a message in Japan while taking about 240 msec between Tokyo and Europe. In addition, the longer the distance is, the more messages are lost. For example, about 20% of the messages are lost between Japan and Europe while less than 1% are lost in Japan. Thus, it is essential to consider a group communication where the delay times between the processes are significantly different [1], i.e. not neglectable compared with the processing speed. Such a group of processes is named a *wide-area* group. In the wide-area group, the time for delivering messages to the destinations is dominated by the longest delay between the processes.

2 System Model

A communication system is composed of three hierarchical layers, i.e. *application*, *transport*, and *network* layers. A group of n (≥ 2) application processes AP_1, ..., AP_n are cooperated. Each AP_i communicates with other processes in the group by using the underlying group communication service provided by transport processes TP_1, ..., TP_n. Here, let G be a group of the transport processes ($G = \{ TP_1, ..., TP_n \}$). The network layer provides the IP service for the transport layer. A transport process TP_i requests the network layer to transmit ICMP "Timestamp" packets to know the delay time with each TP_j in G. We make the following assumptions about packets sent by TP_i: Packets may be lost and duplicated. Packets sent by the same process may be received by the destination processes not in the sending order.

3 Reliable Receipt

3.1 Transmission and confirmation

In the group communication, a message m sent by one process TP_i is sent to multiple destination processes in a group G. m has to be *reliably* delivered to all the destinations in G. There are two points to be discussed to realize the reliable receipt of m:

(1) how to deliver m to the destinations of m, and
(2) how to deliver the receipt confirmation of m to the sender TP_i and the destinations.

In (1), there are two ways: *direct* and *hierarchical* ones. In the direct multicast, TP_i sends m directly to all the destinations. In the hierarchical multicast, TP_i sends m to a subset of the destinations. On receipt of m from TP_i, TP_j forwards m to other destinations.

In (2), there are two schemes to deliver the confirmation: *decentralized* and *distributed* ones. In the decentralized scheme, TP_i sends m to the destinations and the destinations send back the receipt confirmation of m to TP_i. If TP_i receives all the confirmations, TP_i informs all the destinations of the reliable receipt of m. In the distributed scheme [2], every destination TP_j sends the receipt confirmation of m to all the destinations and TP_i on receipt of m. If each TP_j receives the confirmations from all the destinations, TP_j reliably receives m.

3.2 Recovery from message loss

In the underlying network, messages are lost due to buffer overruns, unexpected delay, and congestion. Hence, the processes have to recover from the message loss. There are two ways to retransmit m if TP_j loses m sent by TP_i. (1) Sender retransmission: TP_i retransmits m to TP_j. (2) Destination retransmission: some destination process TP_k forwards m to TP_j.

3.3 Protocols

Suppose that a process TP_i sends m to a subset V_m of the destination processes in the group G. There are the following protocols.

(1) Basic (B) protocol: direct multicast and distributed confirmation with sender retransmission.
(2) Modified (M) protocol: direct multicast and distributed confirmation with destination retransmission.
(3) Nested group (N) protocol: hierarchical multicast and decentralized confirmation with destination retransmission.
(4) Decentralized (D) protocol: direct multicast and decentralized confirmation with sender retransmission.

[Basic (B) protocol]

(T1) TP_i sends m to every destination process in V_m.
(T2) On receipt of m, each process TP_j in V_m sends the receipt confirmation to TP_i.
(T3) On receipt of the confirmation messages from all the processes in V_m, TP_i reliably receives m.
(R) If some TP_j fails to receive m, TP_i sends m to TP_j again. □

The modified (M) protocol is the same as B except that the destination retransmission is adopted.

[Modified (M) protocol]

(R) If TP_j fails to receive m, some destination TP_k nearest to TP_j sends m to TP_j. If all the destinations lose m, T1 is executed again. □

In the N protocol, G is decomposed into disjoint subgroups G_1, ..., G_{sg} ($sg \geq 2$). Each G_i is composed of the processes TP_{i1}, ..., TP_{ih_i} ($h_i \geq 1$) where TP_{i1}

*広域グループ通信
†立川 敬行 滝沢 誠
‡東京電機大学

is a coordinator.

[Nested group (N) protocol]

(T1) TP_{ij} sends m to the coordinator TP_{i1}. Let DC_i be a set of the coordinators whose subgroups include the destinations of m. TP_{i1} forwards m to the coordinators in DC_i.

(T2) On receipt of m, the coordinator TP_{k1} sends m to the destinations in G_k. On receipt of m, the destination TP_{kh} sends the confirmation back to TP_{k1}. On receipt of the confirmations from all the destinations in G_k, TP_{k1} sends the confirmation to the coordinators in DC_i.

(T3) On receipt of the confirmations from all coordinators in DC_i, TP_{k1} sends the confirmation to the destinations in G_k. On receipt of the confirmation from TP_{k1}, TP_{kh} reliably receives m.

(R) If TP_{kh} fails to receive m, TP_{k1} resends m to TP_{kh}. □

In the D protocol, only the sender TP_i can know whether each destination receives m or not. T1 and R are the same as the B protocol.

[Decentralized (D) protocol]

(T2) On receipt of m, TP_j sends the confirmation back to TP_i.

(T3) On receipt of all the confirmations, TP_i sends the acceptance to all the processes in B.

(T4) On receipt of the acceptance, TP_j accepts m. □

4 Evaluation of Protocols

We evaluate the four protocols in terms of the delay time for delivering and reliably receiving messages. The prototypes of the protocols have been implemented to be a group G of seven UNIX processes in SPARC workstations, i.e. three in Hatoyama, one in Tokyo, Japan, two in the U.S. and one in Keele, UK. We consider two cases: (1) there is no message loss and (2) some processes lose m. We measure the delay time where a process in UK sends a message m of 128 bytes to three processes in Hatoyama. In the N protocol, G is composed of Keele and Hatoyama subgroups. The following kinds of delays are obtained from the times measured: receipt(R) delay = time(receive) − time(send), delivery(DL) delay = time(deliver) − time(send), reliable receipt(RR) delay = time(reliable receive) − time(send), and detect(DT) delay = time(detect) − time(send).

(1) of Table 1 indicates the R, DL, and RR delays for four protocols in the first case. The difference between R and DL shows time for the protocol processing. The difference between R and RR shows time for exchanging the confirmation messages of m. Every protocol supports almost the same delay.

(2) of Table 1 shows the R, DT, DL, and RR delays in the presence of lost messages. The difference between DL and DT shows time for recovering from the message loss by retransmission. In the N protocol, we consider two cases: messages are lost among coordinators and lost in subgroups. The delay times in the first case are marked * in Table 1.

Following Table 1, the processes can recover from message loss with shorter delay in the M protocol than the others. In the wide-area group, each channel is different in the delay time and message loss ratio. Hence, the messages can be delivered with shorter delay if the messages are sent through channels with the shorter delay and less loss ratio.

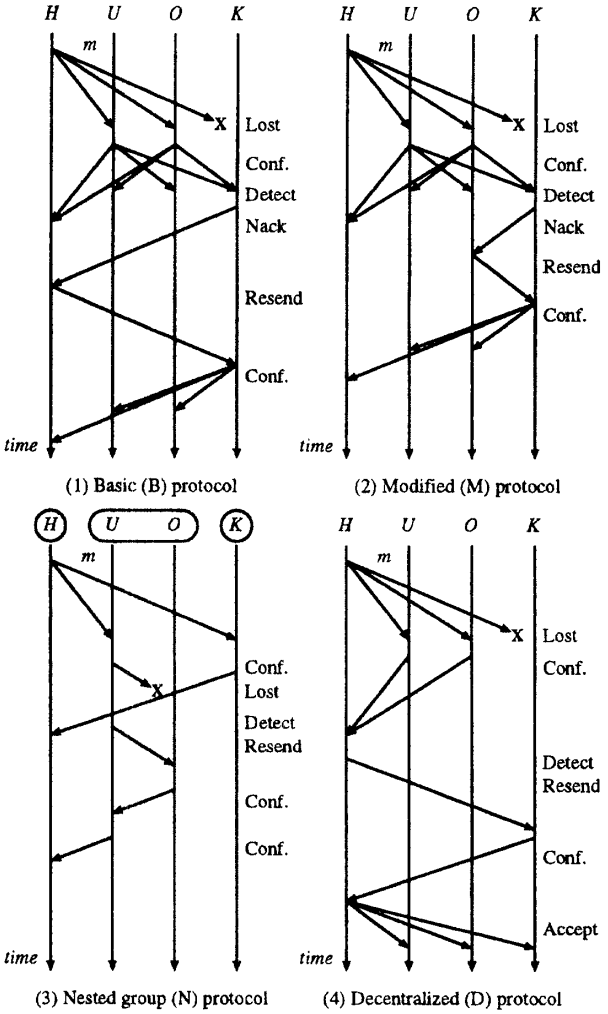

Figure 1: Protocols

5 Concluding Remarks

We have discussed the wide-area group communication which includes multiple processes interconnected by the Internet. Here, each logical channel between the processes in the group has a different delay time and message loss ratio. In this paper, we have presented ways to reduce the delay time of messages in the wide-area group.

Table 1: Delay [msec]

	Protocols	B	M	N		D
(1)	receipt(R)	376	376	377		376
	delivery(DL)	383	383	384		383
	rel. rec. (RR)	724	724	726		1128
(2)	detect (DT)	386	386	387	726*	762
	receipt (R)	1140	393	394	1103*	1135
	delivery (DL)	1141	394	395	1105*	1139
	rel. rec. (RR)	1527	735	736	1482*	1891

References

[1] Hofmann, M., Braun, T., and Carle, G., "Multicast Communication in Large Scale Netwoks," *Third IEEE Workshop on High Performance Communication Subsystems(HPCS)*, 1995.

[2] Nakamura, A. and Takizawa, M., "Causally Ordering Broadcast Protocol," *Proc. of IEEE ICDCS-14*, 1994, pp.48–55.