Multicast communication in large scale networks[1]

Markus Hofmann, Torsten Braun[2], Georg Carle
Institute of Telematics, University of Karlsruhe, Zirkel 2, 76128 Karlsruhe, Germany
E-mail: {hofmann, carle}@telematik.informatik.uni-karlsruhe.de
[2]INRIA, Sophia Antipolis, France (currently on leave from the University of Karlsruhe)
E-mail: Torsten.Braun@sophia.inria.fr

Abstract

Upcoming applications, such as computer supported cooperative work, distributed parallel processing and virtual shared memory depend on efficient data exchange among multiple computer systems. They often require a reliable multipoint communication service with low end-to-end delay and high throughput. Additional problems occur for the provision of a reliable multipoint service, where errors are more likely and where the sender has to deal with many receivers. In order to meet the required reliability, and in order to allow an efficient use of network resources, powerful error control mechanisms are essential. This paper presents a novel concept to support reliable multicast communication in large scale heterogeneous networks.

1 Introduction

The construction of large scale 'Information Superhighways' aims at communication without any spatial constraints, where participants may be spread worldwide. In such a scenario, data exchange within large groups causes additional problems. Mechanisms reducing acknowledgment implosion and efficient error correction schemes have to be developed. Various transport protocols providing a reliable multicast communication service have been developed [XTP91] [JSW91] [AFM92]. Most of them focus on information exchange in local environments, which is why their protocol mechanisms are not suitable to support group communication with a very large number of receivers and in large scale global networks. The Xpress Transfer Protocol, for example, defines damping and slotting algorithm to reduce control information exchange between origin and targets [XTP91] [SDW92]. Instead of addressing control packets exclusively to the multicast sender, receivers in XTP transmit them after a random delay to the whole group. Consequently, every group member receives this message and skips its status report if the incoming control packet corresponds to the own state. In large scale heterogeneous networks, the usefulness of multicasting control packets is very questionable [HJC93]. A large number of participants together with the mechanism described above cause a flood of control packets all over the global network. A second issue of high importance is to identify error correction schemes suitable to provide reliable multicast service on top of unreliable network service. Common protocols use Go-Back-N or selective repeat to retransmit lost and corrupted data. Receivers request missing data directly from the multicast sender without consideration of network topology and actual network load. In case of multicast, it may be preferable to contact another communication participant placed closer to the requesting endsystem.

This paper presents a novel error control mechanism suitable for global heterogeneous networks. It aims to reduce transfer delay, network load and the acknowledgement processing overhead at the sending side. Section 2 presents the proposed framework based on the concept of local groups. Section 3 gives results of a performance analysis with a simple example scenario. Finally, section 4 presents concluding remarks.

2 The Local Group Concept

For better scalability of point-to-multipoint services in respect of group sizes, distances and data rates, we propose to divide global multicast groups into separate subgroups. These subgroups will combine

[1] In *Proceedings of Third IEEE Workshop on High Performance Communication Subsystems (HPCS),* Mystic, Connecticut, USA, August 1995.

communication participants within a local region, defining so called *Local Groups*. Each of these subgroups is represented by a specific communication node, named *Local Group Controller*. These nodes support the provision of the following enhanced services:

- *Local Acknowledgement Processing*: The novel concept foresees the integration of acknowledgment processing capabilities into Local Group Controllers, as it reduces the acknowledgment implosion problem of reliable multicast for large groups. Local Group Controllers evaluate received control packets and inform the multicast sender about the status of the local group.
- *Local Retransmissions*: Local Group Controllers are able to perform retransmissions. This reduces the delay caused by retransmissions, and reduces the load for sender and network.

In each local region, a particular communication participant is determined to function as Local Group Controller. The dedicated station has to collect status messages from all members of its subgroup and has to forward them to the multicast sender in a single composite control packet. To state more precisely, the term 'local region' has to be clarified. A kind of metric is necessary to determine the distance between two communication nodes. The suitability of different metrics, such as delay, bandwidth, throughput, error probability, reliability or number of hops between two nodes, depends mainly on the application using a communication service. An additional requirement of using a specific metric is that the underlying network protocol supports the chosen unit of measurement. IP, for example, considers metrics like delay, throughput and reliability, while the ISO protocol IS-IS supports bandwidth, error probability, delay and financial costs [Per92]. It can also be thought about combinations of the mentioned metrics.

Controllers of subgroups are also responsible to perform local retransmissions. After evaluating received status messages, a Local Group Controller tries to transfer lost data packets directly to receivers that observed errors or losses. If a controller itself is missing some data packets, he tries to get them from another member of its local group. Therefore, a multicast sender has only to retransmit messages missed by all members of a subgroup. Local retransmissions lead to short delays and decrease the number of packets flowing through the global network.

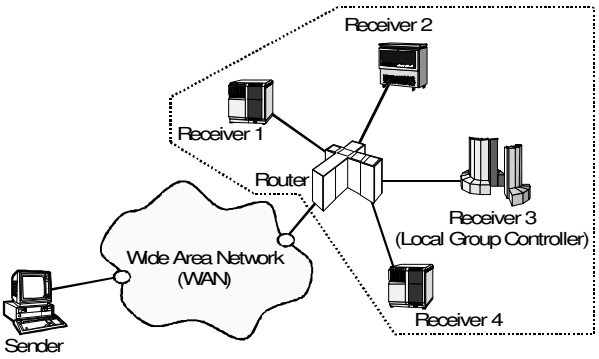

Figure 1 - Combining endsystems to a subgroup

In a scenario as the one illustrated in Figure 1, the four receivers may be combined into a single subgroup. One of the receivers has to function as controller of the subgroup. In this case, local retransmissions do not traverse the wide area network. This reduces transfer delay and network load within the wide area link.

The following sections will give a more detailed description of the protocol. A complete specification including some pseudo code for integrating the Local Group Concept into existing multicast protocols can be found in [Hof 94].

2.1 Establishing Local Groups

New communication participants have to subscribe to a certain local group first. The search for an appropriate Local Group Controller is based on routing information obtained from routing tables. If it is not possible to get this information directly from routers, the decision may be based on measured round trip times. Depending on the result of the search, different actions have to be performed:

- If no Local Group Controller is found within a defined distance, the joining endsystem establishes a new subgroup, appoints itself as group controller and informs the multicast sender about this event. Hence, the multicast sender has knowledge about all local groups and their representative controllers.
- If an appropriate local group exists, the joining endsystem registers itself at the corresponding controller. Therefore, a Local Group Controller always knows the identity of all its members. This is prerequisite for providing a reliable multicast service.

After subscription to a local group, the endsystem is allowed to participate at the group communication. The division of dynamic communication groups into local subgroups should be actualized after a certain time. Due to dynamic changes in group membership, it may be necessary to modify the group structure. It may be preferred, for example, to delegate the role of Local Group Controller to a newly joined participant that is located in the center of the subgroup. A reconfiguration mechanism is required to identify the optimal Local Group Controller.

Due to the knowledge of group membership, a multicast sender is able to guarantee correct data transfer to all Local Group Controllers. The controllers are responsible for correct transmission of user data to all members of their subgroup. This allows to provide a reliable multicast service, that scales well for both, large distances and a large number of receivers.

2.2 Data transfer Phase

The sender multicasts data packets to all destinations. After a certain period of time, he requests the status of all receivers. Instead of addressing control packets directly to the sender, receivers transmit them to the controller of their local group. The controller collects incoming status messages, processes them and calculates the status of the whole subgroup. Using the gained knowledge, a Local Group Controller requests corrupted data packets from other members of the subgroup and performs necessary local retransmissions. In addition, the controller combines the received status messages into a single acknowledgment packet and transmits it to the multicast sender. The acknowledgement includes the request for all data packets missed by the members of the local group.

After reception of user data, receivers deliver the information to the service user. Nevertheless, they are not allowed to release the corresponding buffer space. This is because the Local Group Controller may miss this data and request it at a later moment. Therefore, receivers have to wait either for a sufficient amount of time, or for an explicit permission by the Local Group Controller before releasing the receive buffer.

2.3 Leaving Local Groups

If normal receivers want to leave the group communication, they unsubscribe themselves at the corresponding Local Group Controller. After performing necessary retransmissions, the group controller confirms unsubscription. If an endsystem functioning as a Local Group Controller leaves communication, first of all an appropriate successor has to be determined, and all outstanding retransmissions have to be performed. Afterwards, group management information, like status and identification information migrate to the successor. If the leaving endsystem is the only one within its local group, it unregisters itself directly at the multicast sender.

2.4 Integration of Local Groups in XTP

As an example for integrating the novel concept into existing transport protocols, we have adopted the mechanisms described above to XTP. An additional packet type for migration of management information must be added. To implement a k-reliable multicast service, sender and group controllers must be able to distinguish between the single receivers. Therefore, an additional field identifying individual source addresses have to be integrated into CNTL packets.

2.5 Comparison with related work

The main goals of the Local Group Concept are reducing the acknowledgment implosion problem and decreasing delays by using local retransmissions instead of retransmissions by the sender.

Slotting and damping as proposed by XTP [SDW92] tries to avoid transmission of redundant acknowledgments. However, this approach is not useful for a reliable multicast service in large scale networks.

A similar approach as in this paper is presented in [PSK94]. The protocol called CP uses local exchanges, for example ATM switches, to combine status messages of receivers within a local region. CP requires modifications of the local exchanges, while in the Local Group Concept, a dedicated end system performs this task. The Local Group Concept does not require any modifications in switches or routers and can be used in all kinds of networks.

The DSP protocol [PSK94] combines status messages in end systems, which are selected at connection setup time. In contrast, the Local Group concept allows to move this functionality between end systems of a local group during the lifetime of a connection in order to support also dynamic multicast groups.

The generation of local multicast regions for CP and DSP is mainly based on addresses and the physical location of the end systems (for example end systems connected to a common switch). The Local Group concept allows also the (dynamic) creation of multicast groups dependent on other metrics like delay, costs, etc. Due to this fact, the Local Group concept also supports big MANs (e.g. DQDB, FDDI) and mobile environments.

3 Performance Evaluation

Analytical methods were applied and simulations were performed in order to evaluate the achievable performance and the benefits in reducing network traffic. The simulation scenario consists of a sender and four receivers, which are connected to a local area network. The sender is linked to the LAN over a wide area network, as illustrated in Figure 1.

A transfer delay of 20 ms for the wide area link was assumed, which is approximately the delay for transferring data between east and west coast of USA. A transfer delay within the local area network was assumed to 2 ms. The error probability varied from 10^{-4} to 10^{-1}, which is not uncommon for highly loaded internetworks. Other simulation parameters were data rate, processing delay, status request rate and burst length. A comparison of transfer delay for the Local Group Concept and two common techniques using multicast and unicast retransmissions is illustrated in Figure 2.

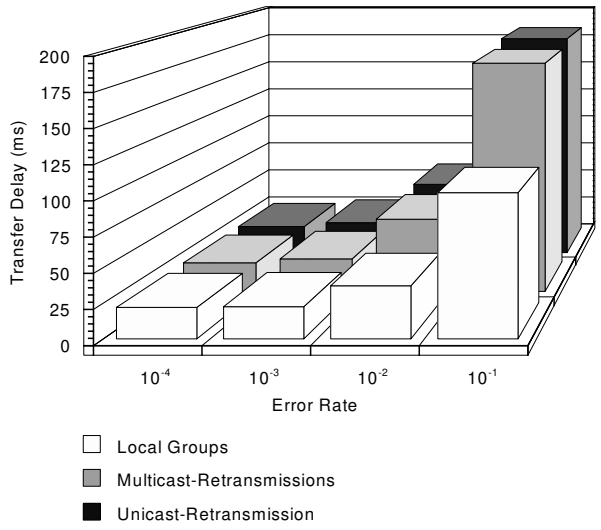

Figure 2 - Transfer Delay

One can see significant benefits while using links with an error rate of 10^{-1} and 10^{-2}. For lower error rates, there is just a little gain in comparison to both common techniques.

Analytical methods were applied to determine the dependency between network load and the number of communication participants. In a scenario similar to that one described above, the number of receivers varied from 1 up to 1000. In contrast to the scenario in Figure 1, we assumed 50 additional receivers at the sending side of the wide area network. The result of the calculation is illustrated in Figure 3. For both common techniques, the ratio of retransmissions and total data traffic over the wide area link depends directly on the number of recipients. The Local Group Concept, in contrast, is not influenced by the number of participants and shows very low values compared with the results of the other techniques. The independence between network load and the number of receivers simplifies resource reservation for highly dynamic communication groups. The bandwidth required on the wide are link may be reserved without care about the actual or future group size. Additional resource reservation during data transfer phase due to joined participants is not required.

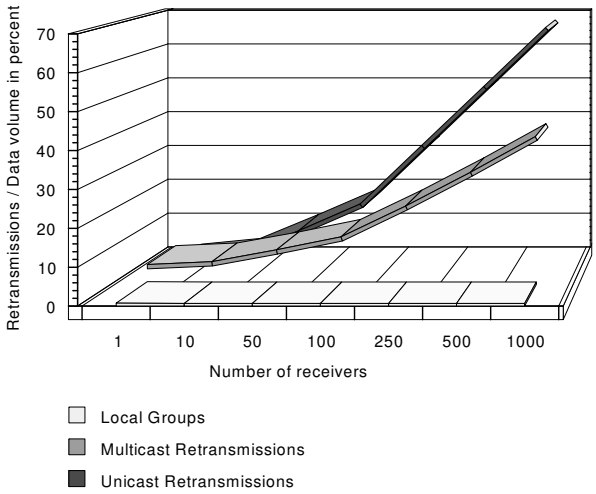

Figure 3 - Network Load

The Local Group Concept requires additional memory at receiving sides due to the delayed release of receive buffers. Simulations demonstrated that the amount of additional buffer space is independent from the error probability of the network and from the number of communication participants. More precisely, the difference between buffer occupancy for the common and for the novel error correction schemes remains almost constant and is about 1/3 of

the total buffer space filled with data packets. So, the need for more memory seems to be acceptable.

Additional results and detailed information about the analytical methods, the simulation tools and the simulation models can be found in [Hof 94].

4 Conclusion

The general goal of the developed concept has been the provision of protocol mechanisms appropriate for multicast communication in global heterogeneous networks. Performance evaluations and analysis of network load show significant improvements compared to common multipoint protocols. The benefits arise from replacing global retransmission by local data exchange and from parallel preprocessing of acknowledgments within local groups. The performance benefits are achieved without the necessity to modify network equipment such as ATM switches, and without a dramatically increased buffer overhead. We expect that efficient protocol implementation techniques such as parallelism or the use of hardware can avoid any protocol processing overhead.

Careful placement of Local Group Controllers assists in optimizing transfer delay and network load. Depending on particular group dynamics and network topology, it may be advantageous to place controllers near or at the center of a local group. There is no simple algorithm to find the centre of a dynamic group at all. So, some kind of heuristic has to be developed defining the right balance between simplicity and optimal placement of the controller.

References

[AFM92] Susan M. Armstrong, Alan O. Freier, Keith A. Marzullo: M*ulticast Transport Protocol.* RFC 1301, February 1992

[HJC93] Bernd Heinrichs, Kai Jakobs, Alessandro Carone: *High Performance transfer services to support multimedia group communications.* Computer Communications, Vol. 16, No. 9, September 1993

[Hof 94] Markus Hofmann: *Reliable Group Communication in Global Heterogeneous Networks.* Diploma Thesis at Institute of Telematics, University of Karlsruhe, August 1994 (in german)

[JSW91] Mark G. W. Jones, Søren-Aksel Sørensen, Steve R. Wilbur: *Protocol design for large group multicasting: the message distribution protocol.* Computer Communications, Vol. 14, No.5, June 1991

[Per92] Radia Perlmann: *Interconnections: Bridges and Routers*. Addison-Wesley Publishing Company, 1992

[PSK94] S. Paul, K. Sabnani, D. Kristol: *Multicast Transport Protocols for High Speed Networks*. Proceedings of International Conference on Network Protocols, Boston, October 1994.

[SDW92] W. Timothy Strayer, Bert. J. Dempsey, Alfred C. Weaver: *The Xpress Transfer Protocol.* Addison-Wesley Publishing Company, 1992

[XTP91] Protocol Emgines Inc.: *XTP Protocol Definition, Revision 3.6.* Protocol Engines Inc., Santa Barbara, 1991