A Reliable Multicast Framework for Light-weight Sessions and Application Level Framing

Sally Floyd, Van Jacobson, Steven McCanne
Lawrence Berkeley Laboratory, University of California, Berkeley, CA 94720
floyd, van, mccanne@ee.lbl.gov

Ching-Gung Liu
University of Southern California, Los Angeles, CA 90089
charley@carlsbad.usc.edu

Lixia Zhang
Xerox PARC, 3333 Coyote Hill Road, Palo Alto, CA 94304
lixia@parc.xerox.com

Abstract

This paper describes SRM (Scalable Reliable Multicast), a reliable multicast framework for application level framing and light-weight sessions. The algorithms of this framework are efficient, robust, and scale well to both very large networks and very large sessions. The framework has been prototyped in wb, a distributed whiteboard application, and has been extensively tested on a global scale with sessions ranging from a few to more than 1000 participants. The paper describes the principles that have guided our design, including the IP multicast group delivery model, an end-to-end, receiver-based model of reliability, and the application level framing protocol model. As with unicast communications, the performance of a reliable multicast delivery algorithm depends on the underlying topology and operational environment. We investigate that dependence via analysis and simulation, and demonstrate an adaptive algorithm that uses the results of previous loss recovery events to adapt the control parameters used for future loss recovery. With the adaptive algorithm, our reliable multicast delivery algorithm provides good performance over a wide range of underlying topologies.

1 Introduction

Several researchers have proposed generic reliable multicast protocols, much as TCP is a generic transport protocol for reliable unicast transmission. In this paper we take a different view: unlike the unicast case where requirements for reliable, sequenced data delivery are fairly general, different multicast applications have widely different requirements for reliability. For example, some applications require that delivery obey a total ordering while many others do not. Some applications have many or all the members sending data while others have only one data source. Some applications have replicated data, for example in an n-redundant file store, so several members are capable of transmitting a data item while for others all data originates at a single source. These differences all affect the design of a reliable multicast protocol. Although one could design a protocol for the worst-case requirements, e.g., guarantee totally ordered delivery of replicated data from a large number of sources, such an approach results in substantial overhead for applications with more modest requirements. One cannot make a single reliable multicast delivery scheme that simultaneously meets the functionality, scalability and efficiency requirements of all applications.

The weakness of "one size fits all" protocols has long been recognized. In 1990 Clark and Tennenhouse proposed a new protocol model called Application Level Framing (ALF) which explicitly includes an application's semantics in the design of that application's protocol [CT90]. ALF was later elaborated with a light-weight rendezvous mechanism based on the IP multicast distribution model, and with a notion of receiver-based adaptation for unreliable, real-time applications such as audio and video conferencing. The result, known as Light-Weight Sessions (LWS), has been very successful in the design of wide-area, large-scale, conferencing applications. This paper further evolves the principles of ALF and LWS to add a framework for scalable reliable multicast (SRM).

ALF says that the best way to meet diverse application requirements is to leave as much functionality and flexibility as possible to the application. Therefore our algorithms are designed to meet only the minimal definition of reliable multicast, i.e., eventual delivery of all the data to all the group members, without enforcing any particular delivery order. We believe that if the need arises, machinery to enforce a particular delivery order can be easily added on top of this reliable delivery service.

The design is also heavily based on the group delivery model that is the centerpiece of the IP multicast protocol [D91]. In IP multicast, data sources simply send to the group's multicast address (a normal IP address chosen from a reserved range of addresses) without needing any advance knowledge of the group membership. To receive any data sent to the group, receivers simply announce that they are interested (via a "join" message broadcast on the local sub-

[0] An earlier version of this paper will appear in ACM SIGCOMM 95. This version corrects errors in the graphs from the SIGCOMM paper.

net) — no knowledge of the group membership or active senders is required. Each receiver joins and leaves the group individually, without affecting the data transmission to any other member. Our multicast delivery framework further enhances the multicast group concept by maximizing information and data sharing among all the members, and strengthens the individuality of membership by making each member responsible for its own correct reception of all the data.

Finally, our design attempts to follow the core design principles of TCP/IP. First, we require only the basic IP delivery model — best-effort with possible duplication and reordering of packets — and build the reliability on an end-to-end basis. No change or special support is required from the underlying IP network. Second, in a fashion similar to TCP adaptively setting timers or congestion control windows, our algorithms dynamically adjust their control parameters based on the observed performance within a session. This allows applications using this model to adapt to a wide range of group sizes, topologies and link bandwidths while maintaining robust and high performance.

References

[AFM92] Armstrong, S., Freier, A., and Marzullo, K., "Multicast Transport Protocol", *Request for Comments (RFC) 1301*, Feb. 1992.

[B93] Birman, K., "The Process Group Approach to Reliable Distributed Computing", *Communications of the ACM*, Dec. 1993.

[BSS91] Birman, K., Schiper, A., and Stephenson, P., "Lightweight Casual and Atomic Group Multicast", *ACM Transactions on Computer Systems*, Vol.9, No. 3, pp. 272-314, Aug. 1991.

[BCS94] B. Braden, D. Clark, and S. Shenker, "Integrated Services in the Internet Architecture: an Overview", *Request for Comments (RFC) 1633*, IETF, June 1994.

[CM84] Chang, J., and Maxemchuk, N., "Reliable Broadcast Protocols", *ACM Transactions on Computer Systems*, Vol.2, No. 3, pp. 251-275, Aug. 1984.

[CT90] Clark, D., and Tennenhouse, D., "Architectural Considerations for a New Generation of Protocols", *Proceedings of ACM SIGCOMM '90*, Sept. 1990, pp. 201-208.

[CLZ87] Clark, D., Lambert, M., and Zhang, L., "NETBLT: A High Throughput Transport Protocol", *Proceedings of ACM SIGCOMM '87*, pp. 353-359, Aug. 1987.

[D91] Deering, S., "Multicast Routing in a Datagram Internetwork", PhD thesis, Stanford University, Palo Alto, California, Dec. 1991.

[FJLMZ95] Floyd, S., Jacobson, V., Liu, C., McCanne, S., and Zhang, L., "A Reliable Multicast Framework for Light-weight Sessions and Application Level Framing, Extended Report", LBL Technical Report, URL ftp://ftp.ee.lbl.gov/papers/wb.tech.ps.Z, Sept. 1995.

[J92] Jacobson, V., "A Portable, Public Domain Network 'Whiteboard' ", Xerox PARC, viewgraphs, April 28, 1992.

[J94] Jacobson, V., "A Privacy and Security Architecture for Lightweight Sessions", Sante Fe, NM, Sept. 94.

[J94a] Jacobson, V., "Multimedia Conferencing on the Internet", Tutorial 4, SIGCOMM 1994, Aug. 1994.

[KTHB89] Kaashoek, M., Tannenbaum, A., Hummel, and Bal, "An Efficient Reliable Broadcast Protocol", *Operating Systems Review*, Oct., 1989.

[M92] McCanne, S., "A Distributed Whiteboard for Network Conferencing", May 1992, UC Berkeley CS 268 Computer Networks term project.

[MJ95] McCanne, S., and Jacobson, V., "vic: A Flexible Framework for Packet Video", submitted to *ACM Multimedia 1995*.

[MMA90] Melliar-Smith, P., Moser, L., and Agrawala, V., "Broadcast Protocols for Distributed Systems", *IEEE Transactions on Parallel and Distributed Systems*, Vol. 1 No. 1, Jan. 1990, pp. 17-25.

[M84] Mills, D.L., "Network Time Protocol (Version 3)", RFC (Request For Comments) 1305, March 1992.

[Pa85] Palmer, E., *Graphical Evolution: An Introduction to the Theory of Random Graphs*, John Wiley & Sons, 1985.

[SCFJ94] Schulzrinne, H., Casner, S., Frederick, R., and Jacobson, V., "RTP: A Transport Protocol for Real-Time Applications", *Internet Draft* draft-ietf-avt-rtp-06.txt, work in progress, Nov. 1994.

[PTK94] Pingali, S., Towsley, D., and Kurose, J., "A Comparison of Sender-Initiated and Receiver-Initiated Reliable Multicast Protocols", *SIGMETRICS '94*.

[TS94] Thyagarajan, A., and Deering, S., IP Multicast release 3.3, Aug. 1994, available from ftp://parcftp.xerox.com/pub/ net-research /ipmulti3.3-sunos413x.tar.Z.

[WKM95] Whetten, B., Kaplan, S., Montgomery, T., "A High Performance Totally Ordered Multicast Protocol", submitted to *INFOCOM '95*.