

A Multicast Transport Protocol for Cooperative Media Distribution

Dimitris Vyzovitis and Andrew Lippman
MIT Media Laboratory
{vyzo,lip}@media.mit.edu

Abstract

This paper describes the main ideas of DRMTP (Distributed Real-time Multicast Transport Protocol), an adaptive application-level protocol which allows cooperative multicast of real-time streams. Our protocol is designed for peer-to-peer operation, where multiple sources residing on end-user machines may be present in the network, and is based on a dynamic stream aggregation and congestion control scheme. The protocol solves the problem of network asymmetry that plagues residential broadband networks and allows for fast fail-over and adaptation to departure of source nodes, mitigating the reliability and upstream capacity problems of end-user nodes. Coupled with traffic localization, stream patching, and TCP-friendly congestion control, we design a protocol that enables scalable soft real-time media distribution in a completely decentralised, serverless fashion.

1. Introduction

Real-time media distribution is a resource and bandwidth intensive operation, exacerbated by network effects. The majority of media distribution systems and protocols today are server centric, with a single entity controlling the distribution, from either a single concentrated location or by a distributed networks of servers. Unfortunately, server centric approaches face severe scalability problems as the number of media files and clients increase.

A lot of research effort has been put into scalable and congestion adaptive [1] multicast transport protocols [2, 6], oftentimes tuned for real-time distribution with a particular class of coding schemes [3]. While these protocols alleviate many scalability problems at large, they still rely on centralized distribution and significant infrastructure investment from the distributor, even when a distributed network of servers is used.

The explosive popularity of peer-to-peer systems in the recent years has indicated that a decentralised distribution system may be applicable to the problem. Despite the ques-

tionable business models of early peer-to-peer systems like Napster, the ability to use an ad-hoc distributed network of peer nodes may be able to provide a scalable and low-cost solution for the problem. Unfortunately, there are currently no transport protocols designed for scalable real-time distribution in a peer-to-peer environment.

We discuss the main ideas behind the design of such a protocol, the Distributed Real-time Multicast Transport Protocol (DRMTP), in the remainder of the paper.

2. Cooperative Media Distribution with Lightweight Sessions and Stream Aggregation

The protocol is based on the idea of multiple nodes, normal Internet hosts, cooperating for the distribution of media files. Each node maintains a local store of media files, which can be locally accessible by the owner or transmitted in part or whole to other nodes. Each file has a unique immutable identifier, a well defined length, and for the purposes of distribution a well defined *framing*. The framing is a subdivision of the file in smaller parts, each with a well-defined length. Individual frames can be locally stored at a node, with local store including storage devices of the node and temporary memory. The real-time *schedule* for the file is defined as the relative times of delivery of individual frame.

Files are transferred in the system in *sessions*. A session is specific to a file, and includes all participating nodes. Data flows in the session within *streams* of the DRMTP protocol. Each stream represents a *localized* flow of frames within a single multicast address. The frames that comprise the stream are a subset of the file frame set. Each stream has a single source and is dynamically established.

A node can participate in a stream either as a *source* or a *sink*. A node however can participate in more than one streams at the same time, acting as a source in some streams and as a sink in others. Thus a node can *aggregate* data from multiple streams.

The stream aggregation mechanism is the core of the protocol. It allows us to consistently handle late joins, lo-

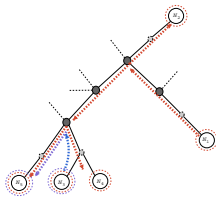


Figure 1. Local stream patching for late joins. Initially N_1 provides S_1 (red) to N_2 , N_3 , and N_4 . Later N_5 joins S_1 and N_3 locally patches with S_2 (blue).

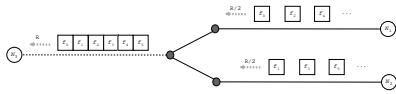


Figure 2. Real-time operation with aggregation of two half-rate streams.

calize traffic and provide a congestion control mechanism which can achieve the real-time rate of the stream, while reacting in a TCP-friendly manner. Figure 1 illustrates how late joins in a stream, are handled with local patching. Figure 2 illustrates how we can aggregate a number of slow interleaved streams to provide real-time operation for a node, even though none of the sources of each stream can support it.

3. Overview of the DRMTP Protocol

The protocol provides algorithms for dealing with implosion control, stream establishment, congestion avoidance and control, and failure recovery.

The basic mechanism for controlling implosion is coordination among nodes. Feedback coordination in all cases that may cause implosion is handled using a randomized feedback suppression algorithm. The protocol uses timers generated from a truncated exponential distribution, an optimal timer for large multicast group as shown in [4].

For each DRMTP stream, there is a single source and a primary controller elected during on stream establishment. The source is passive and transmits frames in bursts according to a schedule provided to it by the primary controller. The primary controller is solely responsible for scheduling bursts of frames, by multicasting schedule messages to the stream group. Hence all sinks compute distance samples from the source and the primary controller, used in selecting the schedule transmission times, setting time-outs for detecting packet losses and failures.

Streams are established using multiple rounds and expanding ring search. The sink will engage up to a maximum number of rounds or until enough streams have been found to provide the entire frame set. Stream establishment is receiver initiated, and implosion feedback suppression is used at all phases. At any time, actively flowing streams are reused with any node in the stream cooperatively announcing relevant information in response to a request.

During the lifetime of a stream there are four classes of error conditions that can arise: frame loss for some sinks, persistent congestion for some sinks, and controller or source failures. Frame loss occurs when packets are lost in some network links, but with a rate that does not signify congestion. In general, sinks will perceive different frame loss rates, as the paths from the source may differ. If the loss rate exceeds that of an equivalent source-sink TCP connection, the remaining stream frame set is partitioned into two interleaved sets by the controller and is handled with a new stream by a different source. Finally, failure recovery and error correction is also handled by each sink individually establishing a new stream.

4. Conclusion

In this paper we described the main ideas of DRMTP, an adaptive application-level protocol designed for scalable real-time media distribution. The protocol provides a novel multi-source TCP-friendly congestion control algorithm, supports complete traffic localization and topologically sensitive stream patching, and operates in a completely decentralised fashion. The protocol algorithms and performance are analyzed in detail in [5].

References

- [1] S. Floyd and K. Fall. Promoting end-to-end congestion control in the internet. *IEEE/ACM Transactions on Networking*, 1999.
- [2] S. Floyd, V. Jacobson, and S. McCanne. A reliable multicast framework for light-weight sessions and application level framing. *IEEE/ACM Trans. Networking*, 5(6):784–803, December 1997.
- [3] S. McCanne, V. Jacobson, and M. Vetterli. Receiver driven layered multicast. In *Proc. ACM SIGCOMM'96*, 1996.
- [4] J. Nonnenmacher and E. Biersack. Scalable feedback for large groups. *IEEE/ACM Transactions on Networking*, (August), 1999.
- [5] D. Vyzovitis. An active protocol architecture for collaborative media distribution. Master's thesis, Massachusetts Institute of Technology, 2002.
- [6] J. Widmer and M. Handley. Extending equation-based congestion control to multicast applications. In *ACM SIGCOMM'01*, 2001.