AVT   A. Begen
Internet-Draft   Cisco
Intended status: Standards Track   C. Perkins
Expires: April 26, 2012   University of Glasgow
                                                    October 24, 2011

                        Duplicating RTP Streams
                  draft-begen-avtcore-rtp-duplication-00

Abstract

   Packet loss is undesirable for real-time multimedia sessions, but it
   is not avoidable due to congestion or other unplanned network
   outages. This is especially the case for IP multicast networks. One
   technique to recover from packet loss without incurring unbounded
   delay for all the receivers is to duplicate the packets and send them
   in separate redundant streams. This document explains how RTP
   streams can be duplicated without breaking RTP and RTP Control
   Protocol (RTCP) rules.

Status of this Memo

   This Internet-Draft is submitted in full conformance with the
   provisions of BCP 78 and BCP 79.

   Internet-Drafts are working documents of the Internet Engineering
   Task Force (IETF). Note that other groups may also distribute
   working documents as Internet-Drafts. The list of current Internet-
   Drafts is at http://datatracker.ietf.org/drafts/current/.

   Internet-Drafts are draft documents valid for a maximum of six months
   and may be updated, replaced, or obsoleted by other documents at any
   time. It is inappropriate to use Internet-Drafts as reference
   material or to cite them other than as "work in progress."

   This Internet-Draft will expire on April 26, 2012.

Table of Contents

1. Introduction

   RTP [RFC3550] transport is widely used today for delivering real-time
   multimedia streams. Most of the applications also rely on IP
   multicast to reach many receivers efficiently.

   While the combination proves successful, there does exist a weakness.
   As [RFC2354] noted, packet loss is not avoidable. This might be due
   to congestion, it might also be a result of an unplanned outage
   caused by a flapping link, link or interface failure, a software bug,
   or a maintenance person accidentally cutting the wrong fiber. Since
   UDP does not provide any means for detecting loss and retransmitting
   packets, it leaves up to the RTP or the applications to detect and
   recover from the loss. For retransmission-based recovery, one
   example is described in [RFC4588].

   One technique to recover from packet loss without incurring unbounded
   delay for all the receivers is to duplicate the packets and send them
   in separate redundant streams. Variations of this technique have
   already been implemented and deployed today [IC2011]. However,
   duplication of RTP streams without breaking the RTP and RTCP
   functionality has not been documented properly. This document
   explains how duplication can be achieved for RTP streams.

2. Terminology and Requirements Notation

   The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT",
   "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and
   "OPTIONAL" in this document are to be interpreted as described in
   [RFC2119].

3. Dual Streaming Use Cases

   Dual streaming refers to a technique that involves transmitting two
   redundant (often RTP) streams of the same content, with each stream
   itself capable of supporting the playback when there is no packet
   loss. Therefore, adding an additional stream provides a protection
   against packet loss. The level of protection depends on how the
   packets are sent and transmitted inside the network.

   It is important to note that redundant streaming can easily be
   extended to support cases when more than two streams are desired.
   But triple, quadruple, or more, streaming is rarely used in practice.

3.1. Temporal Redundancy

   From a routing perspective, two streams are considered identical if
   their following two fields are the same since they will be both
   routed over the same path:

   o IP Source Address

   o IP Destination Address

   Two routing-plane identical RTP streams might carry the same payload
   but they could use different Synchronization Sources (SSRC) to
   differentiate the RTP packets belonging to each stream. In the
   context of dual streaming, we assume that the source duplicates the
   RTP packets and put them into separate RTP streams each with a unique
   SSRC identifier. All the redundant streams are transmitted in the
   same RTP session.

   For example, two redundant RTP streams can be sent to the same IP
   destination address and UDP destination port with a certain delay
   between them [I-D.begen-mmusic-temporal-interleaving]. The streams
   carry the same payload in their respective RTP packets with identical
   sequence numbers. This allows the receiver (or any other node
   responsible for duplicate suppression) to identify and suppress the
   duplicate packets, and subsequently produce a hopefully loss-free and
   duplication-free output stream (called stream merging).

3.2. Spatial Redundancy

3.2.1. Using Separate Source Interfaces

   An RTP source might have multiple network interfaces associated with
   it and it can send two redundant streams from two separate
   interfaces. Such streams can be routed over diverse or identical
   paths depending on the routing algorithm inside the network. At the
   receiving end, the node responsible for duplicate suppression can
   look into various RTP related fields to identify and suppress the
   duplicate packets.

   If source-specific multicast (SSM) transport is used to carry such
   redundant streams, there will be a separate SSM session for each
   redundant stream since the streams are sourced from different
   interfaces (i.e., IP addresses). The receiving host has to join each
   SSM session separately.

3.2.2. Using Separate Destination Addresses and/or Ports

   An RTP source might send the redundant streams to separate IP
   destination addresses and/or UDP ports.

3.3. Dual Streaming over a Single Path or Multiple Paths

   Having described the characteristics of the streams, one can reach
   the following conclusions:

   1. When two routing-plane identical streams are used, the two
       streams will have identical IP headers. This makes it
       impractical to forward the packets onto different paths. In
       order to minimize packet loss, the packets belonging to one
       stream are often interleaved with packets belonging to the other,
       and with a delay, so that if there is a packet loss, such a delay
       would allow the same packet from the other stream to reach the
       receiver because the chances that the same packet is lost in
       transit again is often small. This is what is also known as
       Time-shifted Redundancy, Temporal Redundancy or simply Delayed
       Duplication [I-D.begen-mmusic-temporal-interleaving] [IC2011].
       This approach can be used with all three types of dual streaming
       described in Section 3.1, Section 3.2.1 and Section 3.2.2.

   2. If the two streams have different IP headers, an additional
       opportunity arises in that one is able to build a network, with
       physically diverse paths, to deliver the two streams concurrently
       to the intended receivers. This reduces the delay when packet
       loss occurs and needs to be recovered. Additionally, it also
       further reduces chances for packet loss. An unrecoverable loss
       happens only when two network failures happen in such a way that
       the same packet is affected on both paths. This is referred to
       as Spatial Diversity or Spatial Redundancy [IC2011]. The
       techniques used to build diverse paths are beyond the scope of
       this document.

       Note that spatial redundancy often offers less delay in
       recovering from packet loss provided that the forwarding delay of
       the network paths are more or less the same. For both temporal
       and spatial redundancy approaches, packet misordering might still
       happen and needs to be handled using the RTP sequence numbers.

   To summarize, dual streaming allows an application and a network to
   work together to provide a near zero-loss transport with a bounded or
   minimum delay. The additional advantage includes a predictable
   bandwidth overhead that is proportional to the minimum bandwidth
   needed for the multimedia session, but independent of the number of
   receivers experiencing a packet loss and requesting a retransmission.

For a survey and comparison of similar approaches, refer to [IC2011].

4. Use of RTP and RTCP with Temporal Redundancy

   To achieve temporal redundancy, the main and redundant RTP streams
   are sent using the same source and destination IP addresses and ports
   (that is the 5-tuple of transport protocol, source and destination IP
   addresses, and source and destination transport ports is the same for
   both main and redundant RTP streams). This is perhaps overly
   restrictive, but with the possible presence of network address and
   port translation (NAPT) devices, using anything other than an
   identical 5-tuple can also cause spatial redundancy.

   Since main and redundant RTP streams follow an identical path, they
   are part of the same RTP session. Accordingly, the sender MUST
   choose a different SSRC for the redundant RTP stream than it chose
   for the main RTP stream, following the rules in [RFC3550] section 8.

4.1. RTCP Considerations

   If RTCP is being sent for the main RTP stream, then the sender MUST
   also generate RTCP for the redundant RTP stream. The RTCP for the
   redundant RTP stream is generated exactly as-if the redundant RTP
   stream were a regular media stream; the sender MUST NOT duplicate the
   RTCP packets sent for the main RTP stream. The sender MUST use the
   same RTCP CNAME in the RTCP reports it sends for the main and
   redundant streams, so that the receiver can synchronize them.

   Both main and redundant streams, and their corresponding RTCP, will
   be received. If RTCP is used, receivers MUST generate RTCP reports
   for both main and redundant streams in the usual way, treating them
   as entirely separate media streams.

   Editor's note: The receiving node can also produce a new XR report
   to report on the (loss/delay/jitter/etc.) performance of the output
   stream after the stream merging process. This is TBD.

4.2. Signaling Considerations

   Signaling is needed to allow the receiver to determine that an RTP
   stream is a redundant copy of another, rather than a separate stream
   that needs to be rendered in parallel. We need an SDP attribute to
   ensure that the receiver supports temporal redundancy, plus a new
   RTCP SDES item to indicate that this is a redundant stream that
   should not be directly rendered.

   Editor's notes:

   o How should we correlate the duplicate streams? Grouping is
      straightforward when streams are SSRC-muxed but what if there are
      non-duplicated RTP streams in the same session? Maybe also use
      Magnus' srcname proposal?

   The required SDP grouping semantics and SDP attribute have been
   defined in [I-D.begen-mmusic-redundancy-grouping] and
   [I-D.begen-mmusic-temporal-interleaving], respectively.

5. Use of RTP and RTCP with Spatial Redundancy

   When using spatial redundancy, the redundant RTP stream is sent on
   using a different source and/or destination address/port pair. This
   will be a separate RTP session to the session conveying the main RTP
   stream.

   SSRC for the redundant stream chosen randomly, following the rules in
   Section 8 of [RFC3550] and will almost certainly not match that of
   the main RTP stream. Sender MUST use the same RTCP CNAME for both
   main and redundant streams, in their separate sessions. Also the
   sender uses the new SDES item to indicate that this is a redundant
   stream. This is how the receiver can correlate the flows (can use
   srcname if appropriate).

5.1. RTCP Considerations

   If RTCP is being sent for the main RTP stream, then the sender MUST
   also generate RTCP for the redundant RTP stream. The RTCP for the
   redundant RTP stream is generated exactly as-if the redundant RTP
   stream were a regular media stream; the sender MUST NOT duplicate the
   RTCP packets sent for the main RTP stream. The sender MUST use the
   same RTCP CNAME in the RTCP reports it sends for the main and
   redundant streams, so that the receiver can synchronize them.

   Both main and redundant streams, and their corresponding RTCP, will
   be received. If RTCP is used, receivers MUST generate RTCP reports
   for both main and redundant streams in the usual way, treating them
   as entirely separate media streams.

   Editor's note: The receiving node can also produce a new XR report
   to report on the (loss/delay/jitter/etc.) performance of the output
   stream after the stream merging process. This is TBD.

5.2. Signaling Considerations

   The required SDP grouping semantics and SDP attribute have been
   defined in [I-D.begen-mmusic-redundancy-grouping] and

    [I-D.begen-mmusic-temporal-interleaving], respectively.

6. Use of RTP and RTCP with Temporal and Spatial Redundancy

    Editor's note: Nothing new here. This should use the same RTP/RTCP
    mechanisms, plus a combination of both sets of signaling.

7. Security Considerations

    The security considerations of [RFC3550] apply to this memo.

    Additional security considerations are TBC.

    Editor's note: Email from csp. For the stream de-duplication
    device: it seems that this would work with SRTP encryption
    [RFC3711], since the headers are in the clear, but would break the
    authentication when the SSRC is rewritten. You could just re-
    authenticate the packets, and avoid re-encryption, with appropriate
    signaling of who authenticates the packets.

8. IANA Considerations

    TBC.

9. Acknowledgments

    Thanks to Magnus Westerlund for his suggestions.

10. References

10.1. Normative References

    [RFC3550] Schulzrinne, H., Casner, S., Frederick, R., and V.
               Jacobson, "RTP: A Transport Protocol for Real-Time
               Applications", STD 64, RFC 3550, July 2003.

    [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate
               Requirement Levels", BCP 14, RFC 2119, March 1997.

    [I-D.begen-mmusic-temporal-interleaving]
               Begen, A., Cai, Y., and H. Ou, "Delayed Duplication
               Attribute in the Session Description Protocol",
               draft-begen-mmusic-temporal-interleaving-03 (work in

               progress), October 2011.

   [I-D.begen-mmusic-redundancy-grouping]
               Begen, A., Cai, Y., and H. Ou, "Duplication Grouping
               Semantics in the Session Description Protocol",
               draft-begen-mmusic-redundancy-grouping-02 (work in
               progress), October 2011.

   [RFC3711]   Baugher, M., McGrew, D., Naslund, M., Carrara, E., and K.
               Norrman, "The Secure Real-time Transport Protocol (SRTP)",
               RFC 3711, March 2004.

10.2. Informative References

   [RFC2354]   Perkins, C. and O. Hodson, "Options for Repair of
               Streaming Media", RFC 2354, June 1998.

   [RFC4588]   Rey, J., Leon, D., Miyazaki, A., Varsa, V., and R.
               Hakenberg, "RTP Retransmission Payload Format", RFC 4588,
               July 2006.

   [IC2011]   Evans, J., Begen, A., Greengrass, J., and C. Filsfils,
               "Toward Lossless Video Transport (to appear in IEEE
               Internet Computing)", November 2011.

Authors' Addresses

   Ali Begen
   Cisco
   181 Bay Street
   Toronto, ON M5J 2T3
   CANADA

   Email: abegen@cisco.com

   Colin Perkins
   University of Glasgow
   School of Computing Science
   Glasgow,   G12 8QQ
   UK

   Email: csp@csperkins.org