O. Hodson / ICSI
6 May 2002
Expires: November 2002

RTP Payload for Interleaved Audio

Status of this Document

This document is a product of the IETF AVT WG. Comments should be addressed to the author, or the WG's mailing list at avt@ietf.org.

Abstract

This document describes a payload format for use with the Real-time Transport Protocol (RTP) version 2 for interleaving encoded audio data. It is intended for use in audio streaming delay tolerant applications operating over best-effort packet networks. The goal of interleaving is to disperse burst losses into a series of shorter losses. The total amount of audio lost is not changed by interleaving, but the individual loss events are shorter and easier to conceal at the receiver.

Table of Contents

1. Introduction

The Real-time Transport Protocol (RTP) [1] is the standardized method for transporting between end-systems attached to the Internet. The standard RTP audio profiles [2] allow a number of consecutive audio frames to be encapsulated within a single packet. Encapsulating multiple audio frames within a single packet increases the latency of communication, but results in fewer packets being transmitted and a smaller amount of network bandwidth dedicated to IP/UDP/RTP headers.

When a packet containing multiple audio frames is lost, or a burst of packet losses occurs, the receiving system experiences a burst of audio frame losses. The receiver can apply loss concealment algorithms to mitigate the frame losses. However, the performance of receiver based audio loss concealment schemes varies inversely with the length of loss [4]. The greater the number of consecutive audio frames lost the lower the probability of successful concealment.

Interleaving is a technique for re-arranging the frames from an audio source. The technique introduces temporal separation between adjacent frames for the purposes of transmission. When burst frame losses occur in an interleaved stream, they are dispersed into a series of shorter and easier to conceal losses for the receiver to handle.

Interleaving is employed in several proprietary audio protocols used on the Internet and several payloads undergoing standardization support interleaving in their RTP framing. The format presented here is intended to provide interleaving support for audio codecs with fixed frames and those whose frame size is determinable by inspection of the payload. It's anticipated use is in broadcast style applications where quality is more important than latency.

2. Requirements

- To provide support for interleavers that re-arrange the ordering of audio frames within an RTP audio stream.
- To work with audio codecs that have fixed frame sizes or have self-describing frames that allow the frame size to be inferred.
- To support audio streams employing silence suppression as well as those that do not.
- To support codec changes mid-stream.

3. Interleaver Implementation

For the purpose of clarifying the Payload Format Description we describe the implementation of a model interleaver. The description is intended to be as straightforward as possible. There are alternative styles of interleaver implementation, some of which are provably optimal [5] with regard to latency, however these place constraints on the configuration parameters.

Suppose the interleaver module at the sender has two equally sized buffers: an input buffer and output buffer. The input buffer holds audio frames passed from the media encoder. The output buffer passes audio frames to the RTP encapsulator. When a frame is passed to the input buffer, a frame is removed from the output buffer. When the input buffer is full the output buffer is empty and they swap roles.

We assume throughout this document that frames enter the input buffer in order and are read from the output buffer out of order. The *interleaver cycle length* is the number of frames that can be stored in the input buffer. The *interleaver stride length* is the separation between frames originally adjacent in the output buffer. Consider a full output buffer with an interleaver cycle length of 12 and a stride length of 4. For an input buffer containing audio frames:

A B C D E F G H I J K L

the frames leave the output buffer in the order:

A E I B F J C G K D H L

If we denote the interleaver stride length as *SL* and the interleaver cycle length as *CL*, and assume the frames in the output buffer are labelled 0...*CL*-1, the buffer index of the *n*-th frame out of the interleaver will be:

$$II[n] = n * SL \bmod CL + (n * SL) / CL$$

The payload described in the next section describes how an RTP interleaver places re-ordered frames within an RTP packet. The RTP interleaver may encapsulate any number of frames within a single packet.

4. Payload Format Description

Since only a limited set of interleaver stride lengths and cycle lengths are likely to be of interest for a session, we rely on an external mechanism, such as the Session Description Protocol [6] , to communicate payload mappings describing these values. An SDP format is proposed in section 5.

The proposed payload format for interleaved audio is:

```
 0                   1
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|IC |     II      |     PT      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

IC: Interleaver Cycle (2 bits)
This is a counter that is incremented each time a complete cycle is completed at the sender. A receiver may have multiple decode buffers active and this facilitates placing the incoming frames into the correct buffer. The interleaver cycle has a range from 0 to 3 and is incremented by 1 with the complete transmission of a cycle.

II: Interleaver Index (7 bits)
This is the index of the first audio frame from output buffer, which is encapsulated in the current packet. The interleaver index has a range from 0 to the interleaver cycle length - 1.

PT: Audio Payload (7 bits)
This identifies the type of audio encoding of all the interleaved audio frames encapsulated.

This format allows a sender to interleave the audio frames of stream and encapsulate one or multiple frames in each packet. When multiple frames follow the interleaving header, the offset between each successive frame is the cycle length *CL*. When multiple frames follow the interleaving header, they should be packed according to the their default packing rules. If frames are normally octet aligned, then they MUST be octet aligned when interleaved.

The interleaver payload is only intended for codecs with fixed compressed frame sizes and codecs where the frame boundaries can be determined by examining the codec data. For sample based codecs the number of samples per frame should be the default for the codec concerned. In most cases, the number of samples is 160 per frame. This differs from the RTP A/V profile [2] which suggests sample based codecs should have 160 sample per frame, but frames of any length should be accepted. This restriction removes the need to specify the length of each audio frame in an interleaved packet.

The interleaved audio payload format only supports a single payload type field. All of the audio frames following the interleaved MUST be of the same type. For ease of implementation packets containing multiple interleaved frames MUST only contain frames from one interleaving cycle. Received packets that do not comply SHOULD be discarded.

An RTP packet carry interleaved audio frames SHALL have a standard RTP header with a payload indicating interleaved audio. All fields, with the exception of the timestamp, should be implemented according the methods layed out in RTP. The timestamp field merits special consideration because RTP uses the timestamp field to derive jitter estimates for reporting and applications may use this value in their playout calculation. In the example given in section 3 , frames leave the interleaver in the order:

A E I B F J C G K D H L

If the encapsulation function only places one or two frames in each packet there is a potential issue with the timestamp associated with each packet. If the timestamp is derived from the sampling time of each frame then the timestamps will not increase monotonically, e.g. for one frame per packet the timestamp of the fourth packet is less than the timestamp of the third packet, ie (t(I) <= t(B)).

For applications to be able to use interleaving without modification to their playout calculation we propose the timestamp of each outgoing packet is the time stamp of the frame that would have been in the packet if interleaving had not been applied, i.e. for an interleave with cycle length 12, stride length 4, and a packetizer encapsulating 2 frames per packet the packets are:

AE, IB, FJ, CG, KD, HL

and the timestamps of the outgoing packets are:

t(A), t(C), T(E), t(G), t(I), t(K)

which correspond to the timestamps of the packet had interleaving not been applied:

AB, CD, EF, GH, IJ, KL

This preserves the integrity of existing RTP playout and jitter calculations and allows interleaving to be implemented without modifying the RTP processing in existing applications.

A final point is the interaction with audio codecs using silence suppression. At the start of a new talkspurt, the Interleaver should reset it's cycle counter (IC) and interleaving index (II) to zero. If the codec normally sets the marker bit in the RTP header for new talkspurts, then it should do so when used in conjunction with interleaving.

5. Relation to SDP

The interleaved payload is used an external mapping mechanism may be required for end-systems to identify a particular RTP payload as interleaved audio. A common mechanism for performing this is through the Session Description Protocol (SDP) [6]. The proposed SDP mapping for an interleaved audio payload identifier is:

```
m=audio 10000 RTP/AVP 96 14
a=rtpmap:96 intl/64/8
```

This specifies an interleaved audio stream encapsulated in RTP. The specified port is 10000 and the payload identifier is 96 (selected from the dynamic payloads). The interleaved audio is MPEG-I/II audio (static payload 14). The term 'intl' indicates interleaving. The slash separated parameters are the interleaving cycle length and the stride length respectively. In the example, the interleaver has an interleaving cycle length of 64 and an interleaving stride length of 8.

6. Security Considerations

The security considerations and issues presented in the RTP protocol definition [1] and the RTP sampling document [3] apply to RTP streams carrying the interleaved audio payload.

An additional risk with interleaved stream comes from hostile senders transmitting an interleaved audio stream with randomly changing interleaver cycle number and interleaver index fields. This may cause a receiver to allocate buffer resources and store a large number of audio frames. As a result, implementations SHOULD constrain the number of de-interleaving buffers at the receiver.

7. Example Packet

For an interleaver with a cycle length of 8, stride length 4, and 2 audio frames per packet, the packetized frame sequence is:

AE, BF, CG, DH

As an example consider a stream encoded with G.723.1 audio (RTP A/V payload 4, frame duration 30ms, sample rate 8kHz, channels 1) that uses this interleaver. If the timestamp of first frame in an interleaver sequence is 100 and this is the interleavers first cycle, the second packet will be:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|V=2|P|X| CC=0  |M|     PT      |       sequence number         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        timestamp = 130                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           synchronization source (SSRC) identifier            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| 0 |    II = 1     |   PT = 4    |                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+                             +
|                                                               |
|                         G.723.1 Frame B                       |
|                                                               |
|                                                               |
+                               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                               |                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+                               |
|                                                               |
|                         G.723.1 Frame F                       |
|                                                               |
|                                                               |
+                               +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

8. Acknowledgements

This document derives from an unsubmitted draft that was markedly improved by feedback from Colin Perkins and Ross Finlayson.

9. Author's Address

Orion Hodson
International Computer Science Institute
1947 Center Street (Suite 600)
Berkeley CA94703 USA
hodson@icir.org

10. References

[1] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", RFC 1889.

[2] H. Schulzrinne, and S. Casner, "RTP Profile for Audio and Video Conferences with Minimal Control", Work In Progress, <draft-ietf-avt-profile-new-12.txt>, 2001.

[3] J. Rosenberg, and H. Schulzrinne, "Sampling of the Group Membership in RTP", RFC 2762.

[4] D.J. Goodman, G.B. Lockhard, O.J. Wasem, and W.-C. Wong, "Waveform Substitution Techniques for Recovering Missing Speech Segments in Packet Voice Communications", IEEE Transactions on Acoustics, Speech, and Signal Processing, pp. 1440-1448, vol. ASSP-34, no. 6, December 1986.

[5] J.L. Ramsey, "Realization of Optimium Interleavers", IEEE Transactions on Information Theory, pp. 338-345, vol. IT-16, May 1970.

[6] M. Handley, and V. Jacobson, "SDP: Session Description Protocol", RFC 2327.