RTP Payload Format for AC-3 Streams

Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC2026.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/1id.shadow.html.

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [1].

Abstract

This document describes an RTP payload format for transporting AC-3 encoded audio data. AC-3 is a high quality multichannel audio coding system fully described in [2] by the Advanced Television Standards Committee (ATSC). The RTP payload format presented in this document provides mechanisms for interleaving redundant data, which can increase packet loss resilience. An intelligent method for fragmenting AC-3 frames that exceed the maximum transfer unit (MTU) is also described.

1. Introduction

AC-3 is a high quality audio codec designed to encode multiple channels of audio into a low bit-rate format. AC-3 achieves its large compression ratio via encoding a multiplicity of channels as a single entity. Dolby digital, which is a branded version of AC-3 encodes up to 5.1 channels of audio.

AC-3 has been adopted as an audio compression scheme for many consumer and professional applications. AC-3 is the mandatory codec for DVD-video, ATSC digital terrestrial television, laser disc, and DVD-audio (as an optional multichannel audio format). AC-3 is also a common audio format for film.

Presently there exists a tremendous amount of content encoded in AC-3. The majority of AC-3 content is comprised of more than two channels. It is highly likely that people may wish to stream AC-3 data over computer networks. Applications for streaming AC-3 range from video on demand to multichannel Internet radio. RTP provides a mechanism for stream synchronization and hence serves as the best transport solution for AC-3, which is a codec primarily used in audio for video applications. The RTP payload described in this document also provides a method of ensuring a continuous high quality AC-3 stream.

1.1 Overview of AC-3

AC-3 can deliver upwards of 5.1 channels of audio at data rates approximately equal to half of one PCM channel [2], [4]. The ".1" refers to a band limited optional low-frequency enhancement channel AC-3 was designed for signals sampled at rates of 32, 44.1, or 48 kHz. Data rates can vary between 64 kbps and 640 kbps depending the number of channels and desired quality.

AC-3 exploits psychoacoustic phenomenon that reveal large amounts of inaudible information contained in a typical audio signal. Substantial data reduction occurs via the removal of all inaudible information contained in an audio stream. Source coding techniques are further used to reduce the data used to code an audio signal. Like most perceptual coders, AC-3 operates in the frequency domain. A 512-point TDAC transform is take with 50% overlap, providing 256 new frequency samples. Frequency samples are then converted to exponents and mantissas. Exponents are differentially encoded. Mantissas are allocated a varying number of bits depending on the audibility of the spectral component associated with it. Audibility is determined via a masking curve. Bits for mantissas are allocated from a global bit pool.

1.2 AC-3 Bitstream

AC-3 bitstreams are organized into synchronization frames. Each AC-3 sync frame contains a Sync Information (SI) field, a Bit Stream Information (BSI) field, and 6 audio blocks (AB) representing 1536 PCM samples for each channel. The entire frame represents a time duration of 1536 PCM samples across all coded channels (32 msec @ 48kHz) [2]. Figure 1 shows the AC-3 frame format.

```
+-----------------+-----------------+-----------------+
| S1 | B21 | AB0 | AB1 | AB2 | AB3 | AB4 | AB5 | AUX | CRC |
+-----------------+-----------------+-----------------+
```

The Synchronization Information field contains information needed to acquire and maintain synchronization. The Bit Stream Information field contains parameters that describe the coded audio service [2]. Each audio block also contains fields that determine the usage of block switching, dither, dynamic range control, coupling, and exponent strategy. Figure 2 shows the format of an AC-3 audio block.
2. RTP AC-3 Payload Format

According to [5] RTP payload formats should contain an integral number of application data units (ADUs). With compression algorithms an ADU typically coincides with codec frame boundaries. In this case an ADU is equivalent to an AC-3 sync frame. Hence each RTP packet will contain an integral number of AC-3 frames unless the AC-3 frame exceeds the maximum transfer unit (MTU) of the underlying network.

\[
\text{RTP Payload} = x \cdot \text{AC-3 Frame},
\]

Where \( x \) belongs to \( \mathbb{Z} \) (set of all positive integers), and \( \text{RTP Payload} < \text{MTU} \)

2.1 RTP Header Extension

The following header extension should be at the front of every AC-3 RTP payload. The fields should aid in maintaining order when multiple AC-3 frames are sent in a single payload, or when an AC-3 is fragmented over several frames. A field is also defined to indicate the addition of redundant data.

\[
0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 0
\]

Number of frames (\( NF \)): Number of AC-3 frames present in the RTP payload. This should be set to 0 if the frame is fragmented.

Fragment sequence number (\( FS \)): This number indicates the sequence number of the fragment contained in this RTP payload.

Redundant Data Bit (\( R \)): This bit is set to 1 if the packet contains redundant data for correcting possible lost or corrupted data.

Reserved (\( RSV \)): This field is reserved for a later date.

Figure 4 shows how a full AC-3 RTP payload format should appear.

\[
0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 0
\]

2.2 Fragmentation of AC-3 Frames

The size of AC-3 frames are consistent throughout an encode procedure of a particular piece of audio, but the initial frame size selected can be chosen from large number of possibilities. According to table 5.13 in [2] frames sizes range from 128 bytes to 3840 bytes dependent upon the initial desired bit rate and the sample rate of the uncompressed audio.

AC-3 frame sizes can be quite large, which may require fragmentation. For example an audio file sampled at 32 kHz and compressed with a desired bit rate of 640 kbps would have a frame size of 3840 bytes. This exceeds the standard 1500 byte MTU of an Ethernet network, and would require fragmentation. In [6] it is specified that fragmentation should not be left to IP layer, but instead should be handled by the application itself.

AC-3 frames were designed with possibility of buffers being smaller than an entire AC-3 frame. For this reason each AC-3 frame contains two 16-bit CRC words. CRC1 is contained in the synchronization information (SI) header located at the beginning of each AC-3 frame. CRC1 is the second 16-bit word of the frame. Figure 2 shows the structure of the SI header.

\[
0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 0
\]

CRC2 is the last 16-bit word of an AC-3 frame. CRC1 applies to the first 5/8ths of the frame excluding the sync word. CRC2 covers the remaining 3/8ths of the frame as well as the entire frame (excluding the sync word). All AC-3 encoders enforce specific block size restrictions that guarantee blocks 0 and 1 are completely covered by CRC1 [2]. This allows decoders to immediately begin processing block 0 when the 5/8ths point is reached.

This 5/8ths split in all AC-3 frames, which was intended for the possibility of smaller input buffers, provides a very logical fragmentation unit. Using the 5/8ths point provides two gains:

1) A CRC check can be done on the beginning of the frame providing early detection of a corrupted data
2) Presuming the remaining data in the frame arrives in a timely fashion, immediate processing can be performed on block 0 of the AC-3 frame decreasing any delays in having to concatenate the frame before sending it to the decoder.

In [2] the 5/8ths point is defined to be:

\[
5/8\text{-framesize} = \text{truncate}(\text{framesize}/2) = \text{truncate}(\text{framesize}/8)
\]

According to table 7.34 in [2], 5/8ths frame sizes can range from 80 bytes to 2400 bytes. Hence there are still instances where the 5/8ths boundary may exceed the MTU of the underlying network. In an Ethernet network this would be rare because the majority of AC-3 data publicly
available is sampled at 48kHz and is encoded at a data rate of 384kbps or 448kbps. This provides a 5/8ths point of 960 bytes and 1120 bytes respectively, which would be less than the MTU of a typical Ethernet network. In the rare instances where even the 5/8ths point exceeds the MTU, AC-3 frames should be arbitrarily fragmented to a length that is less than the MTU.

2.3 Data Resiliency

In a previously defined AC-3 RTP payload format a method for data resiliency is presented. The paper suggests that AC-3 frames encoded at 32 kbps should be interleaved with the higher quality AC-3 frames, allowing the AC-3 decoder to decode the lower quality frame if the high quality packet is dropped, lost, or arrives with errors. The method described above is a suitable method for trying to send redundant data. However it may be bandwidth intensive and the redundant data can be extremely low quality, especially in cases where a large number of channels are used.

AC-3 data is often used for film audio. The audio track is stored between the sprocket holes of the film. Over time wear can render sections of the AC-3 track unreadable. When no other error corrections techniques can recover the lost data the two-channel audio track will be used in its place. We present a similar method here for multichannel audio.

When encoding multichannel audio a secondary two-channel version of the audio can also be encoded at a lower bit rate. Since the audio is reduced to two channels, it is still possible to maintain high quality even at a lower bit rate. The lower bit-rate two-channel version can be interleaved with the multichannel audio, and when a packet is lost or corrupted the two-channel version can be used in its place. The redundant data shall be interleaved such that for some RTP Packet(N) with Multichannel AC-3 Frame(M), then RTP Packet(N-1) will contain the two-channel AC-3 Frame(M). Figure 5 shows how redundant data should be interleaved

```
+-----------------+-+------------------+-+
| RTP(N-1)  |       | Two-channel AC-3 Frame(M)   |
+-----------------+-+------------------+-+
| RTP(N)         |       | Two-channel AC-3 Frame(M+1) |
+-----------------+-+------------------+-+
```

Continuously sending redundant data can unnecessarily increase the bandwidth. Therefore in certain instances one may wish to send this redundant data when it is absolutely necessary. One example may be to only send the redundant data when a transient is involved. This would require a transient detector before the encode process.

### RTP Header Fields

Payload Type (PT): It is expected that the RTP profile for a particular class of applications will assign a payload type for this encoding, or alternatively a payload type in the dynamic range [96,127] shall be chosen.

Marker (M) bit: The M bit is set for last fragment of an AC-3 frame. In instances where one or more full AC-3 frames is encapsulated in an RTP packet the M bit will be set, and the full frame itself will be considered the last fragment.

Extension (X) bit: Defined by the RTP profile used.

Timestamp: A 32-bit word that corresponds to the sampling instant for the first AC-3 frame in an RTP packet. AC-3 encodes data sampled at 32kHz, 44.1kHz, and 48kHz. Fragmented frames shall maintain the same timestamp until the last fragment is sent. The starting timestamp is selected at random.

### Types and Names

4.1 MIME type registration

MIME media type name: audio
MIME subtype name: ac3
Required parameters: none
Optional parameters: channels, ptime, maxptime

Encoding considerations:
The AC-3 bitstream shall be generated according to the AC-3 specification [2]. This bitstream is binary data and MUST be encoded for non-binary transport (for Email or any transport that cannot accommodate binary directly, the Base64 encoding is sufficient). This type is also defined for transfer via RTP. ALL RTP packets MUST be packetized using the RTP payload format described in this document.

Security considerations: see section 5 of this document

Interoperability considerations: none

Published specification: see [2]

Applications:
Multichannel audio compression for audio and audio for video

Additional Information: none
Magic number(s): none
File extension(s): .ac3
Macintosh File Type Code(s): none
Object Identifier(s) or OID(s): none

Personal information: Jason Flaks
Email: jsf@dolby.com

Intended Usage: COMMON

Author/Change controller: jsf@dolby.com
Change Controller: IETF AVT WG
4.2 SDP usage

The encoding name when using SDP [6] SHALL be "ac3" (MIME subtype). An example of the media representation in SDP is given below.

m = audio 49000 RTP/AVP 100
a = rtpmap:100 ac3/48000

5. Security considerations

In order to protect copyrighted material, certain security precautions may be necessary. The payload format described in this document is subject to the security considerations defined in the RTP specification [7]. The security considerations discussed in [7] imply the usage of encryption to protect the confidentiality of content. Such an encryption scheme is harmless to the encoded audio data presuming the data is decrypted before being sent to the decoder.

6. References


7. Authors’ Addresses

Jason Flaks
Dolby Laboratories
100 Potrero Ave
San Francisco, CA 94103
Email: jsf@dolby.com
www.dolby.com