Network Working Group
Request for Comments: 2190
Category: Standards Track

C. Zhu
Intel Corp.

September 1997

RTP Payload Format for H.263 Video Streams

Status of This Memo

This document specifies an Internet standards track protocol for the Internet community, and requests discussion and suggestions for improvements. Please refer to the current edition of the "Internet Official Protocol Standards" (STD 1) for the standardization state and status of this protocol. Distribution of this memo is unlimited.

Abstract

This document specifies the payload format for encapsulating an H.263 bitstream in the Real-Time Transport Protocol (RTP). Three modes are defined for the H.263 payload header. An RTP packet can use one of the three modes for H.263 video streams depending on the desired network packet size and H.263 encoding options employed. The shortest H.263 payload header (mode A) supports fragmentation at Group of Block (GOB) boundaries. The long H.263 payload headers (mode B and C) support fragmentation at Macroblock (MB) boundaries.

1. Introduction

This document describes a scheme to packetize an H.263 video stream for transport using RTP [1]. H.263 video stream is defined by ITU-T Recommendation H.263 (referred to as H.263 in this document) [4] for video coding at very low data rates. RTP is defined by the Internet Engineering Task Force (IETF) to provide end-to-end network transport functions suitable for applications transmitting real-time data over multicast or unicast network services.

2. Definitions

The following definitions apply in this document:

CIF: Common Intermediate Format. For H.263, a CIF picture has 352 x 288 pixels for luminance, and 176 x 144 pixels for chrominance.

QCIF: Quarter CIF source format with 176 x 144 pixels for luminance and 88 x 72 pixels for chrominance.

Sub-QCIF: picture source format with 128 x 96 pixels for luminance and 64 x 48 pixels for chrominance.

4CIF: Picture source format with 704 x 576 pixels for luminance and 352 x 288 pixels for chrominance.

16CIF: Picture source format with 1408 x 1152 pixels for luminance and 704 x 576 pixels for chrominance.

GOB: For H.263, a Group of Blocks (GOB) consists of k*16 lines, where k depends on the picture format (k=1 for QCIF, CIF and sub-QCIF; k=2 for 4CIF and k=4 for 16CIF).

MB: A macroblock (MB) contains four blocks of luminance and the spatially corresponding two blocks of chrominance. Each block consists of 8x8 pixels. For example, there are eleven MBs in a GOB in QCIF format and twenty two MBs in a GOB in CIF format.

3. Design Issues for Packetizing H.263 Bitstreams

H.263 is based on the ITU-T Recommendation H.261 [2] (referred to as H.261 in this document). Compared to H.261, H.263 employs similar techniques to reduce both temporal and spatial redundancy, but there are several major differences between the two algorithms that affect the design of packetization schemes significantly. This section summarizes those differences.

3.1 Optional Features of H.263

In addition to the basic source coding algorithms, H.263 supports four negotiable coding options to improve performance: Advanced Prediction, PB-frames, Syntax-based Arithmetic Coding, and Unrestricted Motion Vectors. They can be used in any combination.

Advanced Prediction(AP): One or four motion vectors can be used for some macroblocks in a frame. This feature makes recovery from packet loss difficult, because more redundant information has to be preserved at the beginning of a packet when fragmenting at a macroblock boundary.

PB-frames: Two frames (a P frame and a B frame) are coded into one bitstream with macroblocks from the two frames interleaved. From a packetization point of view, a MB from the P frame and a MB from the B frame must be treated together because each MB for the B frame is coded based on the corresponding MB for the P frame. A means must be provided to ensure proper rendering of two frames in the right order. Also, if part of this combined bitstream is lost, it will affect both frames, and possibly more.

Syntax-based Arithmetic Coding (SAC): When the SAC option is used, the resultant run-value pair after quantization of Discrete Cosine Transform (DCT) coefficients will be coded differently from Huffman codes, but the macroblock hierarchy will be preserved. Since context variables are only synchronized after fixed length codes in the bitstream, any fragmentation starting at variable length codes will result in difficulty in decoding in the presence of packet loss without carrying the values of all the context variables in each H.263 payload header.

The Unrestricted motion vectors feature allows large range of motion vectors to improve performance of motion compensation for inter-coded pictures. This option also affects packetization because it uses larger range of motion vectors than normal.

To enable proper decoding of packets received, without dependency on previous packets, the use of these optional features is signaled in the H.263 payload header, as described in Section 5.

3.2 GOB Numbering

In H.263, each picture is divided into groups of blocks (GOB). GOBs are numbered according to a vertical scan of a picture, starting with the top GOB and ending with the bottom GOB. In contrast, a GOB in H.261 is composed of three rows of 16x16 MB for QCIF, and three half-rows of MBs for CIF. A GOB is divided into macroblocks in H.263 and the definition of the macroblocks are the same as in H.261.

Each GOB in H.263 can have a fixed GOB header, but the use of the header is optional. If the GOB header is present, it may or may not start on a byte boundary. Byte alignment can be achieved by proper bit stuffing by the encoder, but it is not required by the H.263 bitstream specification [4].

In summary, a GOB in H.263 is defined and coded with finer granularity but with the same source format, resulting in more flexibility for packetization than with H.261.

3.3 Motion Vector Encoding

Differential coding is used to code motion vectors as variable length codes. Unlike in H.261, where each motion vector is predicted from the previous MB in the GOB, H.263 employs a more flexible prediction scheme, where one or three candidate predictors could be used depending on the presence of GOB headers.

If the GOB header is present in a GOB, motion vectors are coded with reference to MBs in the current GOB only. If a GOB header is not present in the current GOB, three motion vectors must be available to decode one macroblock, where two of them might come from the previous GOB. To correctly decode a whole inter-coded GOB, all the motion vectors for MBs in the previous GOB must be available to compute the predictors or the predictors themselves must be present. The optional use of three motion vector predictors can be a major problem for a packetization scheme like the one defined for H.261 when packetizing at MB boundaries [5].

Consider the case that a packet starts with a MB but the GOB header is not present. If the previous packet is lost, then all the motion vectors needed to predict the motion vectors for the MBs in the current GOB are not available. In order to decode the received MBs correctly, all the motion vectors for the previous GOB or the motion vector predictors would have to be duplicated at the beginning of the packet. This kind of duplication would be very expensive and unacceptable in terms of bandwidth overhead.

The encoding strategy of each H.263 CODEC (CODer and DECoder) implementation is beyond the scope of this document, even though it has significant effect on visual quality in the presence of packet loss. However, we strongly recommend use of the GOB header for every GOB at the beginning of a packet to address this problem.

Similar problems exist because of cross-GOB data dependency related to motion vectors, but they can not be addressed by using the GOB header. For 16CIF and 4CIF pictures, a GOB contains more than one row of MBs. If a GOB can not fit in one RTP packet, and the first packet containing the GOB header is lost, then MBs in the second packet can not compute motion vectors correctly, because they are coded relative to data in the lost packet. Similarly, when OBMC (Overlapped Block Motion Compensation) [4] in Advanced Prediction mode is used, motion compensation for some MBs in one GOB could use motion vectors of MBs in previous GOB regardless of the presence of GOB header. When MBs that are used to decode received MBs are lost, those received MBs can not be decoded correctly. Each implementation of the method described in this document should take these limitations into account.

3.4 Macroblock Address

As specified by H.261, a macroblock address (MBA) is encoded with a variable length code to indicate the position of a macroblock within a group of MBs in H.261 bitstreams. H.263 does not code the MBA explicitly, but the macroblock address within a GOB is necessary to recover from packet loss when fragmenting at MB boundaries. Therefore, this information must be included in the H.263 payload header for modes (mode B and mode C as described in Section 5) that allow packetization at MB boundaries.

4. Usage of RTP

When transmitting H.263 video streams over the Internet, the output of the encoder can be packetized directly. For every video frame, the H.263 bitstream itself is carried in the RTP payload without alteration, including the picture start code, the entire picture header, in addition to any fixed length codes and variable length codes. In addition, the output of the encoder is packetized without adding the framing information specified by H.223 [6]. Therefore multiplexing audio and video signals in the same packet is not accommodated, as UDP and RTP provide a much more efficient way to achieve multiplexing.

RTP does not guarantee a reliable and orderly data delivery service, so a packet might get lost in the network. To achieve a best-effort recovery from packet loss, the decoder needs assistance to proceed with decoding of other packets that are received. Thus it is desirable to be able to process each packet independent of other packets. Some frame level information is included in each packet, such as source format and flags for optional features to assist the decoder in operating correctly and efficiently in presence of packet loss. The flags for H.263 optional features also provide information about coding options used in H.263 video bitstreams that can be used by session management tools.

H.263 video bitstreams will be carried as payload data within RTP packets. A new H.263 payload header is defined in section 5 on the H.263 payload header. This section defines the usage of RTP fixed header and H.263 video packet structure.

4.1 RTP Header Usage

Each RTP packet starts with a fixed RTP header [1]. The following fields of the RTP fixed header are used for H.263 video streams:

Marker bit (M bit): The Marker bit of the RTP fixed header is set to 1 when the current packet carries the end of current frame; set to 0 otherwise.

Payload Type (PT): The Payload Type shall specify H.263 video payload format using the value specified by the RTP profile in use, for example RFC 1890 [3].

Timestamp: The RTP timestamp encodes the sampling instant of the video frame contained in the RTP data packet. The RTP timestamp may be the same on successive packets if a video frame occupies more than one packet. For H.263 video streams, the RTP timestamp is based on a 90 kHz clock, the same as the RTP timestamp for H.261 video streams [5].

4.2 Video Packet Structure

For each RTP packet, the RTP fixed header is followed by the H.263 payload header, which is followed by the standard H.263 compressed bitstream [4].

The size of the H.263 payload header is variable depending on modes used as detailed in the next section. The layout of an RTP H.263 video packet is shown as:

```
    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
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                 RTP header                                    |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                 H.263 payload header                          |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |                 H.263 bitstream                               |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

5. H.263 Payload Header

For H.263 video streams, each RTP packet carries only one H.263 video packet. The H.263 payload header is always present for each H.263 video packet.

Three formats (mode A, mode B and mode C) are defined for H.263 payload header. In mode A, an H.263 payload header of four bytes is present before actual compressed H.263 video bitstream in a packet. It allows fragmentation at GOB boundaries. In mode B, an eight byte H.263 payload header is used and each packet starts at MB boundaries without the PB-frames option. Finally, a twelve byte H.263 payload header is defined in mode C to support fragmentation at MB boundaries for frames that are coded with the PB-frames option.

The mode of each H.263 payload header is indicated by the F and P fields in the header. Packets of different modes can be intermixed. All client application are required to be able to receive packets in any mode, but decoding of mode C packets is optional because the PB-frames feature is optional.

In this section, the H.263 payload format is shown as rows of 32-bit words. Each word is transmitted in network byte order. Whenever a field represents a numeric value, the most significant bit is at the left of the field.

5.1 Mode A

In this mode, an H.263 bitstream will be packetized on a GOB boundary or a picture boundary. Mode A packets always start with the H.263 picture start code [4] or a GOB, but do not necessarily contain complete GOBs. Four bytes are used for the mode A H.263 payload header. The H.263 payload header definition for mode A is shown as follows with F=0. Mode A packets are allowed to start at a GOB boundary even if no GOB header is present in the bitstream for the GOB. However, such use is discouraged due to the dependencies it creates across GOB boundaries, as described in Section 3.3.

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|F|P|SBIT |EBIT | SRC |I|U|S|A|R      |DBQ| TRB |    TR         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

F: 1 bit
The flag bit indicates the mode of the payload header. F=0, mode A; F=1, mode B or mode C depending on P bit defined below.

P: 1 bit
Optional PB-frames mode as defined by the H.263 [4]. "0" implies normal I or P frame, "1" PB-frames. When F=1, P also indicates modes: mode B if P=0, mode C if P=1.

SBIT: 3 bits
Start bit position specifies number of most significant bits that shall be ignored in the first data byte.

EBIT: 3 bits
End bit position specifies number of least significant bits that shall be ignored in the last data byte.

SRC : 3 bits
Source format, bit 6,7 and 8 in PTYPE defined by H.263 [4], specifies the resolution of the current picture.

I: 1 bit.
Picture coding type, bit 9 in PTYPE defined by H.263[4], "0" is intra-coded, "1" is inter-coded.

U: 1 bit
Set to 1 if the Unrestricted Motion Vector option, bit 10 in PTYPE defined by H.263 [4] was set to 1 in the current picture header, otherwise 0.

S: 1 bit
Set to 1 if the Syntax-based Arithmetic Coding option, bit 11 in PTYPE defined by the H.263 [4] was set to 1 for current picture header, otherwise 0.

A: 1 bit
Set to 1 if the Advanced Prediction option, bit 12 in PTYPE defined by H.263 [4] was set to 1 for current picutre header, otherwise 0.

R: 4 bits
Reserved, must be set to zero.

DBQ: 2 bits
Differential quantization parameter used to calculate quantizer for the B frame based on quantizer for the P frame, when PB-frames option is used. The value should be the same as DBQUANT defined by H.263 [4]. Set to zero if PB-frames option is not used.

TRB: 3 bits
Temporal Reference for the B frame as defined by H.263 [4]. Set to zero if PB-frames option is not used.

TR: 8 bits
Temporal Reference for the P frame as defined by H.263 [4]. Set to zero if the PB-frames option is not used.

5.2 Mode B

In this mode, an H.263 bitstream can be fragmented at MB boundaries. Whenever a packet starts at a MB boundary, this mode shall be used without PB-frames option. Mode B packets are intended for a GOB whose size is larger than the maximum packet size allowed in the underlying protocol, thus making it impossible to fit one or more complete GOBs in a packet. This mode can only be used without the PB-frames option. Mode C as defined in the next section can be used to fragment H.263

bitstreams at MB boundaries with the PB-frames option. The H.263 payload header definition for mode B is shown as follows with F=1 and P=0:

```
    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
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |F|P|SBIT |EBIT | SRC | QUANT   |  GOBN   |   MBA           |R  |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |I|U|S|A| HMV1        | VMV1        | HMV2        | VMV2        |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The following fields are defined the same as in mode A: F, P, SBIT, EBIT, SRC, I, U, S and A. Other fields are defined as follows:

QUANT: 5 bits
Quantization value for the first MB coded at the starting of the packet. Set to 0 if the packet begins with a GOB header. This is the equivalent of GQUANT defined by the H.263 [4].

GOBN: 5 bits
GOB number in effect at the start of the packet. GOB number is specified differently for different resolutions. See H.263 [4] for details.

MBA: 9 bits
The address within the GOB of the first MB in the packet, counting from zero in scan order. For example, the third MB in any GOB is given MBA = 2.

HMV1, VMV1: 7 bits each.
Horizontal and vertical motion vector predictors for the first MB in this packet [4]. When four motion vectors are used for current MB with advanced prediction option, these would be the motion vector predictors for block number 1 in the MB. Each 7 bits field encodes a motion vector predictor in half pixel resolution as a 2’s complement number.

HMV2, VMV2: 7 bits each.
Horizontal and vertical motion vector predictors for block number 3 in the first MB in this packet when four motion vectors are used with the advanced prediction option. This is needed because block number 3 in the MB needs different motion vector predictors from other blocks in the MB. These two fields are not used when the MB only has one motion vector. See the H.263 [4] for block organization in a macroblock. Each 7 bits field encodes a motion vector predictor in half pixel resolution as a 2’s complement number.

R : 2 bits
Reserved, must be set to zero.

5.3 Mode C

In this mode, an H.263 bitstream is fragmented at MB boundaries of P frames with the PB-frames option. It is intended for those GOBs whose sizes are larger than the maximum packet size allowed in the underlying protocol when PB-frames option is used. The H.263 payload header definition for mode C is shown as follows with F=1 and P=1:

```
    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
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |F|P|SBIT |EBIT | SRC | QUANT   |  GOBN   |   MBA           |R  |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |I|U|S|A| HMV1        | VMV1        | HMV2        | VMV2        |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   | RR                                  |DBQ| TRB |    TR         |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The following fields are defined the same as in mode B: F, P, SBIT, EBIT, SRC, QUANT, GOBN, MBA, R, I, U, S, A, HMV1, VMV1, HMV2, VMV2. The rest of the fields (TR, DBQ, TRB) are defined the same as in mode A, except field RR. The RR field takes 19 bits, and is currently reserved. It must be set to zero.

5.4 Selection of Modes for the H.263 Payload Header

Packets carrying H.263 video streams with different modes can be intermixed. The modes shall be selected carefully based on network packet size, H.263 coding options and underlying network protocols. More specifically, mode A shall be used for packets starting with a GOB or the H.263 picture start code [4], and mode B or C shall be used whenever a packet has to start at a MB boundary. Mode B or C are necessary for those GOBs with sizes larger than network packet size.

We strongly recommend use of mode A whenever possible. The major advantage of mode A over mode B and C is its simplicity. The H.263 payload header is smaller than mode B and C. Transmission overhead is reduced and the savings may be very significant when working with very low data rates or relatively small packet sizes.

Another advantage of mode A is that it simplifies error recovery in the presence of packet loss. The internal state of a decoder can be recovered at GOB boundaries instead of having to synchronize with MBs as in mode B and C. The GOB headers and the picture start code are easy to identify, and their presence will normally cause a H.263

decoder to re-synchronize its internal states.

Finally, we would like to stress that recovery from packet loss depends on a decoder’s ability to use the information provided in the H.263 payload header within RTP packets.

6. Limitations

The packetization method described in this document applies to the 1996 version of H.263. It may not be applicable to bitstreams with features added after that.

Security Considerations

Security issues are addressed by RTP [1]. This memo does not bring up any additional security issues.

7. Acknowledgments

The author would like to thank the following people for their valuable comments: Linda S. Cline, Christian Maciocco, Mojy Mirashrafi, Phillip Lantz, Steve Casner, Gary Sullivan, and Sassan Pejhan.

8. References

[1] Schulzrinne, H., Casner, S., Frederick, R., and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", RFC 1889, January 1996.

[2] International Telecommunication Union. Video Codec for Audiovisual Services at p x 64 kbits/s, ITU-T Recommendation H.261, 1993.

[3] Schulzrinne, H., "RTP Profile for Audio and Video Conference with Minimal Control", RFC 1890, January 1996.

[4] International Telecommunication Union. Video Coding for Low Bitrate Communication, ITU-T Recommendation H.263, 1996

[5] Turletti, T., and C. Huitema, "RTP Payload Format for H.261 Video Streams", RFC 2032, October 1996.

[6] International Telecommunication Union.
Multiplexing Protocol for Low Bitrate Multimedia Communication,
ITU-T Recommendation H.223, 1995.

7. Author’s Address

C. "Chad" Zhu
Mail Stop: JF3-202
Intel Corporation
2111 N.E. 25th Avenue
Hillsboro, OR 97124
USA

EMail: czhu@ibeam.intel.com
Phone: (503) 264-6008
Fax: (503) 264-1805