Internet Engineering Task Force
INTERNET-DRAFT
draft-ietf-avt-profile-04.ps

Audio-Video Transport Working Group
H. Schulzrinne
GMD Fokus

March 23, 1995
Expires: 9/1/95

RTP Profile for Audio and Video Conferences with Minimal Control

Status of this Memo

This document is an Internet-Draft. 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."

To learn the current status of any Internet-Draft, please check the "1id-abstracts.txt" listing contained in the Internet-Drafts Shadow Directories on ftp.is.co.za (Africa), nic.nordu.net (Europe), munnari.oz.au (Pacific Rim), ds.internic.net (US East Coast), or ftp.isi.edu (US West Coast).

Distribution of this document is unlimited.

Abstract

This note describes a profile for the use of the real-time transport protocol (RTP) and the associated control protocol, RTCP, within audio and video multiparticipant conferences with minimal control. It provides interpretations of generic fields within the RTP specification suitable for audio and video conferences. In particular, this document defines a set of default mappings from payload type numbers to encodings.

The document also describes how audio and video data may be carried within RTP. It defines a set of standard encodings and their names when used within RTP. However, the definitions are independent of the particular transport mechanism used. The descriptions provide pointers to reference implementations and the detailed standards. This document is meant as an aid for implementors of audio, video and other real-time multimedia applications.

1 Introduction

This profile defines aspects of RTP left unspecified in the RTP protocol definition (RFC TBD). This profile is intended for the use within audio and video conferences with minimal session control. In particular, no support for the negotiation of parameters or membership control is provided. Other profiles may make different choices for the items specified here. The profile specifies the use of RTP over unicast and multicast UDP as well as ST-II.

> How to indicate usage of the profile? Port numbers are not likely to be well-defined.

2 RTP and RTCP Packet Forms and Protocol Behavior

This profile follows the default and/or recommended aspects of the RTP specification for these items:

> Maybe the main spec should number these items, so that they can be easily aligned between spec and profile?

- The standard format of the fixed RTP data header is used (one marker bit).
- No additional fixed fields are appended to the RTP data header.
- The suggested constants are to be used for the RTCP report interval calculation.
- No extension section is defined for the RTCP SR or RR packet.
- No additional RTCP packet types are defined by this profile specification.
- The RTP default security services are also the default under this profile.
- The standard mapping of RTP and RTCP to transport-level addresses is used.
- No encapsulation of RTP packets is specified.
- No RTP header extensions are defined, but applications operating under this profile may use such extensions. Thus, applications should not assume that the RTP header X bit is always zero and should be prepared to ignore the header extension. Extensions should register the content of the first 16 bits with IANA.

  Yet another IANA space? Other ideas?
- Applications may use any of the SDES items described.

New encodings are to be registered with the Internet Assigned Numbers Authority. When registering a new encoding, the following information should be provided:

- name and description of encoding, in particular the RTP timestamp clock rate;
- indication of who has change control over the encoding (for example, CCITT/ITU, other international standardization bodies, a consortium or a particular company or group of companies);
- any operating parameters;
- a reference to a further description, if available, for example (in order of preference) an RFC, a published paper, a patent filing, a technical report or a computer manual;
- for proprietary encodings, contact information (postal and email address).
- the payload type value for this profile.

3 Audio

3.1 Encoding-independent recommendations

The following recommendations are default operating parameters. Applications should be prepared to handle other values. The ranges given are meant to give guidance to application writers, allowing a set of applications conforming to these guidelines to interoperate without additional negotiation. These guidelines are not intended to restrict operating parameters for applications that can negotiate a set of interoperable parameters, e.g., through a conference control protocol.

For packetized audio, the default packetization interval should have a duration of 20 ms, unless otherwise noted when describing the encoding. The packetization interval determines the minimum end-to-end delay; longer packets introduce less header overhead but higher delay and make packet loss more noticeable. For non-interactive applications such as lectures or links with severe bandwidth constraints, a higher packetization delay may be appropriate.

For N-channel encodings, each sampling period (say, 1/8000 of a second) generates N samples. (This terminology is standard, but somewhat confusing, as the total number of samples generated per second is then the sampling rate times the channel count.)

If multiple audio channels are used, channels are numbered left-to-right, starting at one. In RTP audio packets, information from lower-numbered channels precedes that from higher-numbered channels. For more than two channels, the convention followed by the AIFF-C audio interchange format should be followed [1]. For two-channel stereo, the numbering sequence is left, right; for three channels, left, right, center; for quadrophonic systems, front left, front right, rear left, rear right; for four-channel systems, left, center, right, and surround sound; for six-channel systems left, left center, center, right, right center and surround sound. All channels belonging to a single sampling instance must be within the same packet.

The sampling frequency should be drawn from the set: 8000, 11025, 16000, 22050, 44100 and 48000 Hz. (The Apple Macintosh computers have native sample rates of 22254.54 and 11127.27, which can be converted to 22050 and 11025 with acceptable quality by dropping 4 or 2 samples in a 20 ms frame.)

A receiver should accept packets representing between 0 and 200 ms of audio data.[1] Receivers should be prepared to accept multi-channel audio, but may choose to only play a single channel.

3.2 Guidelines for Sample-Based Audio Encodings

In *sample-based* encodings, each audio sample is represented by a fixed number of bits. Within the compressed audio data, codes for individual samples may span octet boundaries. An RTP audio packet may contain any number of audio samples, subject to the constraint that the number of bits per sample times the number of samples per packet yields an integral octet count. *Fractional encodings* produce less than one octet per sample.

For sample-based encodings producing one or more octets per sample, samples from different channels, but the same sampling instant are consecutive. For example, for a two-channel encoding, the octet sequence is (left channel, first sample), (right channel, first sample), (left channel, second sample), (right channel, second sample), ... For multi-octet encodings, octets are transmitted in network byte order (i.e., most significant octet first).

The packing order for fractional encodings is that described for the IMA Wave types [2]. For audio encodings yielding four bits per sample, eight such compressed samples from channel 1 are packet into one 32-bit word, followed by eight compressed samples from channel 2, until all channels have been accomodated and the packing resumes at channel 1. For audio encodings yielding three bits per sample, 32 such compressed samples at three bits each from channel 1 are packed into 12 octets, followed by 32 samples from channel 2, etc.

[1] This restriction allows reasonable buffer sizing for the receiver.

3.3 Guidelines for Frame-Based Audio Encodings

Frame-based encodings encode a fixed-length block of audio into another block of compressed data, typically also of fixed length. For frame-based encodings, the sender may choose to combine several such frames into a single message. The receiver can tell the number of frames contained in a message since the frame duration is defined as part of the encoding.

For frame-based codecs, the channel order is defined for the whole block. That is, for two-channel audio, right and left samples are coded independently, with the encoded frame for the left channel preceding that for the right channel.

All frame-oriented audio codecs should be able to encode and decode several consecutive frames within a single packet. Since the frame size for the frame-oriented codecs is given, there is no need to use a separate designation for the same encoding, but with different number of frames per packet.

3.4 Audio Encodings

encoding	sample/frame	bits/sample	ms/frame
1016	frame		30
G721	sample	4	
G723	sample	3	
GSM	frame		20
IDVI	sample	4	
LPC	frame		20
L8	sample	8	
L16	sample	16	
MPA	frame		
PCMU	sample	8	
PCMA	sample	8	

Table 1: Properties of Audio Encodings

1016: Encoding 1016 is a frame based encoding using code-excited linear prediction (CELP) and is specified in Federal Standard FED-STD 1016 [3, 4, 5, 6].

The U. S. DoD's Federal-Standard-1016 based 4800 bps code excited linear prediction voice coder version 3.2 (CELP 3.2) Fortran and C simulation source codes are available for worldwide distribution at no charge (on DOS diskettes, but configured to compile on Sun SPARC stations) from: Bob Fenichel, National Communications System, Washington, D.C. 20305, phone +1-703-692-2124, fax +1-703-746-4960. and `ftp://ftp.super.org/pub/speech/celp_3.2a.tar.Z`.

G721: G721 is specified in ITU recommendation G.721. Reference implementations for G.721 and G.723 are available as part of the CCITT/ITU-T Software Tool Library (STL) from the ITU General Secretariat, Sales Service, Place du Nations, CH-1211 Geneve 20, Switzerland. The library is covered by a license and is available for anonymous ftp on `gaia.cs.umass.edu`, file `pub/hgschulz/ccitt/ccitt_tools.tar.Z`.

G723: G721 is specified in ITU recommendation G.723. See G721 for information about a reference implementation. bits per sample.

GSM: GSM (group speciale mobile) denotes the European GSM 06.10 provisional standard for full-rate speech transcoding, prI-ETS 300 036, which is based on RPE/LTP (residual pulse excitation/long term prediction) coding at a rate of 13 kb/s. A reference implementation was written by Carsten Borman and Jutta Degener (TU Berlin, Germany) and is available for anonymous ftp from `ftp.cs.tu-berlin.de`, directory `pub/local/kbs/tubmik/gsm`.

IDVI: IDVI is specified, with reference implemention, in [2]. Each packet contains a single DVI block.

The "header" word for each channel has the following structure:

```
int16  valpred;  /* previous predicted value, network byte order */
u_int8 index;    /* index into stepsize table */
```

Header words for all channels precede the compressed data. Note that the first 16 bits differ in definition from the IMA and Microsoft DVI ADPCM Wave type [7]. There, the first 16 bits contain the first (uncompressed) sample.

This discrepancy is unfortunate, creating all kinds of problems with hardware-based codecs common with PCs.

L8: L8 denotes linear audio data, using 8-bits of precision with an offset of 128, that is, the most negative signal is encoded as 0.

L16: L16 denotes uncompressed audio data, using 16-bit signed representation with 65535 equally divided steps between minimum and maximum signal level, ranging from -32768 to 32767. The value is represented in two's complement notation and network byte order.

MPA: MPA denotes MPEG-I or MPEG-II audio. The encoding is defined in ISO standards ISO/IEC 11172-3 and 13818-3. The encapsulation is specified in RFC TBD. Sampling rate and channel count are contained in the payload.

PCMU: PCMU is specified in CCITT/ITU-T recommendation G.711. Audio data is encoded as eight bits per sample, after companding. Code to convert between linear and mu-law companded data is available in [2].

PCMA: PCMA is specified in CCITT/ITU-T recommendation G.711. Audio data is encoded as eight bits per sample, after companding. Code to convert between linear and A-law companded data is available in [2].

LPC: LPC designates an experimental linear predictive encoding written by Ron Frederick, Xerox PARC, available from `ftp://parcftp.xerox.com/pub/net-research/lpc.tar.Z`.

VDVI: VDVI is a variable-rate version of IDVI, yielding speech bit rates of between 10 and 25 kbps. It is specified for single-channel operation only. It uses the following encoding:

IDVI codeword	VDVI bit pattern
0	00
1	010
2	1100
3	11100
4	111100
5	1111100
6	11111100
7	11111110
8	10
9	011
10	1101
11	11101
12	111101
13	1111101
14	11111101
15	11111111

TSP0: TSP0 designates the proprietary variable-rate, frame-based encoding called True Speech. The encoding is defined for a sampling rate of 7200 Hz and has an average data rate of 7200 bits per second. Further information is available by contacting VocalTec (see VSC encoding) or the address:

DSP Group, Inc.
email: tsplayer@dsgp.com

VSC: VSC designates the proprietary variable-rate encoding called Vocaltec Software Compression. The encoding is defined for a sampling rate of 5500 Hz and has an average data rate of 963 bytes per second. Further information is available by contacting

Alon Cohen
VocalTec Ltd.
Maskit 1, Herzliya
Israel
phone: +972-9-5612121
email: alon@vocaltec.com

The standard audio encodings and their payload types are listed in Table 2.

4 Video

The following video encodings are currently defined, with their abbreviated names used for identification:

CelB: The CELL-B encoding is a proprietary encoding proposed by Sun Microsystems. The byte stream format is described in RFC TBD.

CPV: This proprietary encoding, "Compressed Packet Video" is implemented by Concept, Bolter, and ViewPoint Systems video codecs. For further information, contact:

Glenn Norem, President
ViewPoint Systems, Inc.
2247 Wisconsin Street, Suite 110
Dallas, TX 75229-2037
United States
Phone: +1-214-243-0634

JPEG: The encoding is specified in ISO Standards 10918-1 and 10918-2. The RTP payload format is as specified in RFC TBD.

H261: The encoding is specified in CCITT/ITU-T standard H.261. The packetization and RTP-specific properties are described in RFC TBD.

HDCC: The HDCC encoding is a proprietary encoding used by Silicon Graphics. [TBD: Need contact information.]

MPV: The MPEG-I and MPEG-II video encoding are specified in ISO Standards ISO/IEC 11172 and 13818-2, respectively. The RTP payload format is as specified in RFC TBD.

nv: The encoding is implemented in the program 'nv' developed at Xerox PARC by Ron Frederick.

CUSM: The encoding is implemented in the program CU-SeeMe developed at Cornell University by Dick Cogger, Scott Brim, Tim Dorcey and John Lynn.

PicW: The encoding is implemented in the program PictureWindow developed at Bolt, Beranek and Newman (BBN).

RGB8: 8-bit encoding of RGB values, sequenced TBD. Each pixel can assume values from 0 to 255. Each frame is prefixed by a header containing TBD.

If there is no strong technical reason to the contrary, all video encodings use a timestamp frequency of 65536 Hz. The standard video encodings and their payload types are listed in Table 2.

5 Port Assignment

As specified in the RTP protocol definition, RTP data is to be carried on an even UDP port number and the corresponding RTCP packets are to be carried on the next higher (odd) port number.

Applications operating under this profile may use any such UDP port pair or ST-II SAP pair. For example, the port pair may be allocated randomly by a session management program. A single fixed port number pair cannot be required because multiple applications using this profile are likely to run on the same host, and there are some operating systems that do not allow multiple processes to use the same UDP port with different multicast addresses.

However, port numbers 5004 and 5005 have been registered for use with this profile for those applications that choose to use them as the default pair. Applications that operate under multiple

PT	encoding name	audio/video (A/V)	clock rate (Hz)	channels (audio)
0	PCMU	A	8000	1
1	1016	A	8000	1
2	G721	A	8000	1
3	GSM	A	8000	1
4	G723	A	8000	1
5	IDVI	A	8000	1
6	IDVI	A	16000	1
7	LPC	A	8000	1
8	unassigned	A		
9	unassigned	A		
10	L16	A	44100	2
11	L16	A	44100	1
12	TSP0	A	7200	1
13	VSC	A	5500	1
14	MPA	A	90000	(see text)
15–22	unassigned	A		
23	RGB8	V	65536	N/A
24	HDCC	V	65536	N/A
25	CelB	V	65536	N/A
26	JPEG	V	65536	N/A
27	CUSM	V	65536	N/A
28	nv	V	65536	N/A
29	PicW	V	65536	N/A
30	CPV	V	65536	N/A
31	H261	V	65536	N/A
32	MPV	V	90000	N/A
33–71	unassigned	V	65536	N/A
72–76	reserved	N/A	N/A	N/A
77–127	unassigned	?		N/A

Table 2: Payload types (PT) for standard audio and video encodings

profiles may use this port pair as an indication to select this profile if they are not subject to the constraint of the previous paragraph. Applications need not have a default and may require that the port pair be explicitly specified. The particular port numbers were chosen to lie in the range above 5000 to accomodate port number allocation practice within the Unix operating system, where port numbers below 1024 can only be used by privileged processes and port numbers between 1024 and 5000 are automatically assigned by the operating system.

References

[1] Apple Computer, "Audio interchange file format AIFF-C," Aug. 1991. (also ftp://ftp.sgi.com/sgi/aiff-c.9.26.91.ps.Z).

[2] IMA Digital Audio Focus and Technical Working Groups, "Recommended practices for enhancing digital audio compatibility in multimedia systems (version 3.00)," tech. rep., Interactive Multimedia Association, Annapolis, Maryland, Oct. 1992.

[3] Office of Technology and Standards, "Telecommunications: Analog to digital conversion of radio voice by 4,800 bit/second code excited linear prediction (celp)," Federal Standard FS-1016, GSA, Room 6654; 7th & D Street SW; Washington, D.C. 20407 (+1-202-708-9205), 1990.

[4] J. P. Campbell, Jr., T. E. Tremain, and V. C. Welch, "The proposed Federal Standard 1016 4800 bps voice coder: CELP," *Speech Technology*, vol. 5, pp. 58–64, April/May 1990.

[5] J. P. Campbell, Jr., T. E. Tremain, and V. C. Welch, "The federal standard 1016 4800 bps CELP voice coder," *Digital Signal Processing*, vol. 1, no. 3, pp. 145–155, 1991.

[6] J. P. Campbell, Jr., T. E. Tremain, and V. C. Welch, "The dod 4.8 kbps standard (proposed federal standard 1016)," in *Advances in Speech Coding* (B. Atal, V. Cuperman, and A. Gersho, eds.), ch. 12, pp. 121–133, Kluwer Academic Publishers, 1991.

[7] Microsoft, *DVI ADPCM wave type*. Microsoft, 1993. Microsoft Development Library – SDKs: Multimedia Standards Update.

6 Address of Author

Henning Schulzrinne
GMD Fokus
Hardenbergplatz 2
D-10623 Berlin
Germany
electronic mail: `hgs@fokus.gmd.de`