

Internet Draft  
Document: draft-ietf-avt-rtp-h264-03.txt  
Expires: December 2003

S. Wenger  
M.M. Hannuksela  
T. Stockhammer  
M. Westerlund  
D. Singer

October 2003  
Expires April 2004

## RTP payload Format for H.264 Video

### Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC 2026. 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/lid-abstracts.txt>

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

### Copyright Notice

Copyright (C) The Internet Society (2003). All Rights Reserved.

### Abstract

This memo describes an RTP Payload format for the ITU-T Recommendation H.264 video codec. This codec was designed as a Wenger et. al. Expires August 2003 [Page 1]

Internet Draft

26 June, 2003

joint project of the Video Coding Experts Group (VCEG) of ITU-T and the Moving Picture Experts Group (MPEG) of ISO/IEC. Recommendation H.264 was approved by ITU-T on May 2003, and the approved draft specification is available for public review. ISO/IEC International Standard 14496-10 will be technically identical to ITU-T Recommendation H.264.

Wenger et. al. Expires December 2003

[Page 2]

## Table of Contents

1. Introduction.....	5
1.1. The H.264 codec.....	5
1.2. Parameter Set Concept.....	6
1.3. Network Abstraction Layer Unit Types.....	7
2. Conventions.....	8
3. Scope.....	9
4. Definitions and Abbreviations.....	9
4.1. Definitions.....	9
4.2. Abbreviations.....	11
5. RTP Payload Format.....	11
5.1. RTP Header Usage.....	11
5.2. Common structure of the RTP payload format.....	14
5.3. NAL Unit Octet Usage.....	15
5.4. Packetization Modes.....	17
5.5. Decoding Order Number (DON).....	18
5.6. Single NAL Unit Packet.....	20
5.7. Aggregation Packets.....	21
5.8. Fragmentation Units (FUs).....	29
6. Packetization Rules.....	33
6.1. Common Packetization Rules.....	33
6.2. Single NAL Unit Mode.....	35
6.3. Non-Interleaved Mode.....	35
6.4. Interleaved Mode.....	35
7. De-Packetization Process (Informative).....	35
7.1. Single NAL Unit and Non-Interleaved Mode.....	36
7.2. Interleaved Mode.....	36
7.3. Additional De-Packetization Guidelines.....	39
8. Payload Format Parameters.....	40
8.1. MIME Registration.....	40
8.2. SDP Parameters.....	47
9. Security Considerations.....	48
10. Informative Appendix: Application Examples.....	49
10.1. Video Telephony according to ITU-T Recommendation H.241 Annex A.....	49
10.2. Video Telephony, No Slice Data Partitioning, No NAL Unit Aggregation.....	50
10.3. Video Telephony, Interleaved Packetization Using NAL Unit Aggregation.....	50
10.4. Video Telephony, with Data Partitioning.....	51
10.5. Video Telephony or Streaming, with FUs and Forward Error Correction.....	52
10.6. Low-Bit-Rate Streaming.....	54

10.7. Robust Packet Scheduling in Video Streaming.....55  
11. Informative Appendix: Rationale for Decoding Order Number..56  
11.1. Introduction.....56  
11.2. Example of Multi-Picture Slice Interleaving.....56  
11.3. Example of Robust Packet Scheduling.....58  
11.4. Robust Transmission Scheduling of Redundant Coded Slic..62  
11.5. Remarks on Other Design Possibilities.....63  
12. Open Issues.....64  
13. Full Copyright Statement.....64  
14. Intellectual Property Notice.....65  
15. References.....66  
15.1. Normative References.....66  
15.2. Informative References.....66  
Annex A: Changes relative to draft-ietf-avt-rtp-h264-02.txt....68

## 1. Introduction

## 1.1. The H.264 codec

This memo specifies an RTP payload specification for the video coding standard known as ITU-T Recommendation H.264 [1] and ISO/IEC International Standard 14496 Part 10 (also known as MPEG-4 Advanced Video Coding) [2]. Recommendation H.264 was approved by ITU-T on May 2003, and the approved draft specification is available for public review [8]. In this memo the H.264 acronym is used for the codec and the standard, but the memo is equally applicable to the ISO/IEC counterpart of the coding standard.

The H.264 video codec has a very broad application range that covers all forms of digital compressed video from low bit rate Internet Streaming applications to HDTV broadcast and Digital Cinema applications with near loss-less coding. Most, if not all, relevant companies in all of these fields (including Video-Conferencing, Streaming, TV broadcast, and Digital Cinema) have participated in the standardization, which gives hope that this wide application range is more than an illusion and may materialize, probably in a relatively short time frame. The overall performance of H.264 is as such that bit rate savings of 50% or more, compared to the current state of technology, are reported. Digital Satellite TV quality, for example, was reported to be achievable at 1.5 Mbit/s, compared to the current operation point of MPEG 2 video at around 3.5 Mbit/s [9].

The codec specification [1] itself distinguishes conceptually between a video coding layer (VCL), and a network abstraction layer (NAL). The VCL contains the signal processing functionality of the codec, things such as transform, quantization, motion search/compensation, and the loop filter. It follows the general concept of most of today's video codecs, a macroblock-based coder that utilizes inter picture prediction with motion compensation, and transform coding of the residual signal. The VCL encoder outputs slices: a bit string that contains the macroblock data of an integer number of macroblocks, and the information of the slice header (containing the spatial address of the first macroblock in the slice, the initial quantization parameter, and similar).

Macroblocks in slices are ordered in scan order unless a different  
Wenger et. al. Expires December 2003 [Page 5]

macroblock allocation is specified, using the so-called Flexible Macroblock Ordering syntax. In-picture prediction is used only within a slice. More information is provided in [8].

The NAL encoder encapsulates the slice output of the VCL encoder into Network Abstraction Layer Units (NAL units), which are suitable for the transmission over packet networks or the use in packet oriented multiplex environments. Annex B of H.264 defines an encapsulation process to transmit such NAL units over byte-stream oriented networks. In the scope of this memo Annex B is not relevant.

Internally, the NAL uses NAL units. A NAL unit consists of a one-byte header and the payload byte string. The header co-serves as the RTP payload header and indicates the type of the NAL unit, the (potential) presence of bit errors or syntax violations in the NAL unit payload, and information regarding the relative importance of the NAL unit for the decoding process. This RTP payload specification is designed to be unaware of the bit string in the NAL unit payload.

One of the main properties of H.264 is the complete decoupling of the transmission time, the decoding time, and the sampling or presentation time of slices and pictures. The decoding process specified in H.264 is unaware of time, and the H.264 syntax does not carry information such as the number of skipped frames (as common in the form of the Temporal Reference in earlier video compression standards). Also, there are NAL units that are affecting many pictures and are, hence, inherently time-less. For this reason, the handling of the RTP timestamp requires some special considerations for those NAL units for which the sampling or presentation time is not defined, or, at transmission time, unknown.

## 1.2. Parameter Set Concept

One very fundamental design concept of H.264 is to generate self-contained packets, to make mechanisms such as the header duplication of RFC 2429 [11] or MPEG-4's HEC [12] unnecessary. The Wenger et. al. Expires December 2003 [Page 6]

way how this was achieved is to decouple information that is relevant to more than one slice from the media stream. This higher layer meta information should be sent reliably, asynchronously and in advance from the RTP packet stream that contains the slice packets. (Provisions for sending this information in-band are also available for such applications that do not have an out-of-band transport channel appropriate for the purpose). The combination of the higher-level parameters is called a parameter set. The H.264 specification includes two types of parameter sets: sequence parameter set and picture parameter set. An active sequence parameter set remains unchanged throughout a coded video sequence, and an active picture parameter set remains unchanged within a coded picture. The sequence and picture parameter set structures contain information such as picture size, optional coding modes employed, and macroblock to slice group map.

In order to be able to change picture parameters (such as the picture size), without having the need to transmit parameter set updates synchronously to the slice packet stream, the encoder and decoder can maintain a list of more than one sequence and picture parameter set. Each slice header contains a codeword that indicates the sequence and picture parameter set to be used.

This mechanism allows to decouple the transmission of parameter sets from the packet stream, and transmit them by external means, e.g. as a side effect of the capability exchange, or through a (reliable or unreliable) control protocol. It may even be possible that they get never transmitted but are fixed by an application design specification.

### 1.3. Network Abstraction Layer Unit Types

Tutorial information on the NAL design can be found in [13], [14] and [15].

All NAL units consist of a single NAL unit type octet, which also co-serves as the payload header. The payload of a NAL unit follows immediately.

The syntax and semantics of the NAL unit type octet are specified in [1], but the essential properties of the NAL unit type octet are summarized below. The NAL unit type octet has the following format:

```
+-----+
|0|1|2|3|4|5|6|7|
+---+---+---+---+---+---+
|F|NRI|  Type  |
+-----+
```

The semantics of the components of the NAL unit type octet, as specified in the H.264 specification, are described briefly below.

F: 1 bit

forbidden\_zero\_bit. The H.264 specification declares a value of 1 as a syntax violation.

NRI: 2 bits

nal\_ref\_idc. A value of 00 indicates that the content of the NAL unit is not used to reconstruct reference pictures for inter picture prediction. Such NAL units can be discarded without risking the integrity of the reference pictures. Values greater than 00 indicate that the decoding of the NAL unit is required to maintain the integrity of the reference pictures.

Type: 5 bits

nal\_unit\_type. The NAL unit payload type as defined in table 7-1 of [1], and later within this memo. For a reference of all currently defined NAL unit types and their semantics please refer to section 7.4.1 in [1].

This memo introduces new NAL unit types, which are introduced in Section 5.2. Note that the NAL unit types defined in this memo are marked as unspecified in [1]. Moreover, this specification extends the semantics of F and NRI as described in section 5.3.

## 2. Conventions

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 [3].

This specification uses the notion of setting and clearing a bit when handling bit fields. Setting a bit is the same as assigning that bit the value of 1 (On). Clearing a bit is the same as assigning that bit the value of 0 (Off).

### 3. Scope

This payload specification can only be used to carry the "naked" H.264 NAL unit stream over RTP. Likely, the first applications of this specification will be in the conversational multimedia field, video telephone or video conference. The draft is not intended for the use in conjunction with the byte stream format of Annex B of H.264.

### 4. Definitions and Abbreviations

#### 4.1. Definitions

This document uses the definitions of [1]. The following terms defined in [1] are summed up below for convenience:

access unit: A set of NAL units always containing a primary coded picture. In addition to the primary coded picture, an access unit may also contain one or more redundant coded pictures or other NAL units not containing slices or slice data partitions of a coded picture. The decoding of an access unit always results in a decoded picture.

coded video sequence: A sequence of access units that consists, in decoding order, of an IDR access unit followed zero or more non-IDR access units including all subsequent access units up to but not including any subsequent IDR access unit.

instantaneous decoding refresh (IDR) access unit: An access unit in which the primary coded picture is an IDR picture.

instantaneous decoding refresh (IDR) picture: A coded picture containing only slices with I or SI slice types that causes a "reset" in the decoding process. After the decoding of an IDR picture all following coded pictures in decoding order can be decoded without inter prediction from any picture decoded prior to the IDR picture.

primary coded picture: The coded representation of a picture to be used by the decoding process for a bitstream conforming to H.264. The primary coded picture contains all macroblocks of the picture.

redundant coded picture: A coded representation of a picture or a part of a picture. The content of a redundant coded picture shall not be used by the decoding process for a bitstream conforming to H.264. The content of a redundant coded picture may be used by the decoding process for a bitstream that contains errors or losses.

VCL NAL unit: A collective term used to refer to coded slice and coded data partition NAL units.

In addition, the following definitions apply:

decoding order number (DON): A field in the payload structure or a derived variable indicating NAL unit decoding order. Values of DON are in the range of 0 to 65535, inclusive. After reaching the maximum value, the value of DON wraps around to 0.

NAL unit decoding order: A NAL unit order that conforms to the constraints on NAL unit order given in section 7.4.1.2 in [1].

transmission order: The order of packets in ascending RTP sequence number order (in modulo arithmetic). Within an aggregation packet, the NAL unit transmission order is the same as the order of appearance of NAL units in the packet.

## 4.2. Abbreviations

DON: Decoding Order Number  
DONB: Decoding Order Number Base  
DOND: Decoding Order Number Difference  
FU: Fragmentation Unit  
IDR: Instantaneous Decoding Refresh  
IEC: International Engineering Consortium  
ISO: International Organization for Standardization  
ITU-T: International Telecommunication Union,  
Telecommunication Standardization Sector  
MTAP: Multi-Time Aggregation Packet  
MTAP16: MTAP with 16-bit timestamp offset  
MTAP24: MTAP with 24-bit timestamp offset  
NAL: Network Adaptation Layer  
NALU: NAL Unit  
SEI: Supplemental Enhancement Information  
STAP: Single-Time Aggregation Packet  
STAP-A: STAP type A  
STAP-B: STAP type B  
TS: Timestamp  
VCL: Video Coding Layer

## 5. RTP Payload Format

## 5.1. RTP Header Usage

The format of the RTP header is specified in RFC 3550 [4] and reprinted in Figure 1 for convenience. This payload format uses the fields of the header in a manner consistent with that specification.

When encapsulating one NAL unit per RTP packet, the RECOMMENDED RTP payload format is specified in section 5.6. The RTP payload (and the settings for some RTP header bits) for aggregation packets and fragmentation units are specified in sections 5.7 and 5.8, respectively.

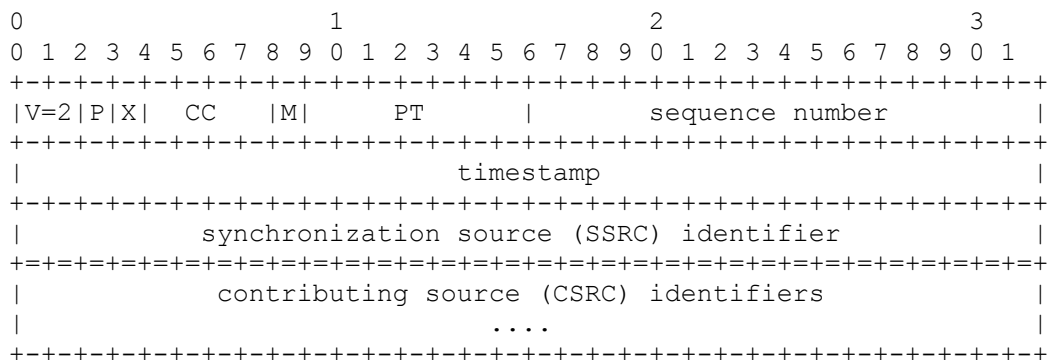


Figure 1: RTP header according RFC 3550.

The RTP header information is set as follows:

Version (V): 2 bits  
 Set to 2 according to RFC 3550.

Padding (P): 1 bit  
 Used according to RFC 3550.

Extension (X): 1 bit  
 Used according to RFC 3550 and profile definitions.

CSRC count (CC): 4 bits  
 Used according to RFC 3550.

Marker bit (M): 1 bit  
 Set for the very last packet of the access unit indicated by the RTP timestamp, in line with the normal use of the M bit in video formats and to allow an efficient playout buffer handling. Decoders MAY use this bit as an early indication of the last packet of an access unit, but MUST NOT rely on this property. Informative note: Only one M bit is associated with an aggregation packet carrying multiple NAL units, and thus if a gateway has re-packetized an aggregation packet into several packets, it cannot reliably set the M bit of those packets.

## Payload type (PT): 7 bits

The assignment of an RTP payload type for this new packet format is outside the scope of this document, and will not be specified here. The assignment of a payload type needs to be performed either through the profile used or in a dynamic way.

## Sequence number (SN): 16 bits

Increased by one for each sent packet. Set to a random value during startup as per RFC 3550

## Timestamp: 32 bits

The RTP timestamp is set to the sampling timestamp of the content. A 90 kHz clock rate MUST be used.

If the NAL unit has no own timing properties (e.g. parameter set and SEI NAL units), the RTP timestamp is set to the RTP timestamp of the primary coded picture of the access unit to which the NAL unit is included according to section 7.4.1.2 of [1].

The setting of the RTP Timestamp for MTAPs is defined in section 5.7.2.

If the content is a part of a coded frame that was sampled as two fields having distinct sampling times and that is supposed to be displayed as fields having distinct display times, the RTP timestamp MUST be set to the sampling timestamp of the latest sampled field. In addition, the picture timing supplemental enhancement information (SEI) message (subclauses D.1.2 and D.2.2 of [1]) SHOULD be used to convey the timestamps for display, and the last clock timestamp in decoding order conveyed in a picture timing SEI message MUST correspond to the RTP timestamp of the primary coded picture of the same access unit.

Informative note: Displaying coded frames as fields is needed commonly in an operation known as 3:2 pulldown where film content that consists of coded frames is displayed on a display using interlaced scanning. The picture timing SEI message enables carriage of multiple timestamps for the same coded

picture, and therefore the 3:2 pulldown process is perfectly controlled. The picture timing SEI message mechanism is necessary, because only one timestamp per coded frame can be conveyed in the RTP timestamp.

Receivers SHOULD ignore any picture timing SEI messages included in access units that have only one display timestamp. Instead, receivers SHOULD use the RTP timestamp for synchronizing the display process.

RTP senders SHOULD NOT transmit picture timing SEI messages for pictures that are not supposed to be displayed as multiple fields.

Synchronization source (SSRC) identifier: 32 bits  
Used according to RFC 3550.

Contributing source (CSRC) identifiers: 0 to 15 items, 32 bits each  
Used according to RFC 3550.

## 5.2. Common structure of the RTP payload format

The payload format is defined as a number of different payload structures depending on need. However, which structure a received RTP packet contains is evident from the first byte of the payload. This byte will always be structured as a NAL unit header. The NAL unit type field indicates which structure is present. The possible structures are:

Single NAL Unit Packet: Contains only a single NAL unit in the payload. The NAL header type field will be equal to the original NAL unit type, i.e., in the range of 1 to 23, inclusive. Specified in section 5.6.

Aggregation packet: Packet type used to aggregate multiple NAL units into a single RTP payload. This packet exists in four versions, the Single-Time Aggregation Packet type A (STAP-A), the Single-Time Aggregation Packet type B (STAP-B), Multi-Time Aggregation Packet (MTAP) with 16 bit offset (MTAP16), and Multi-

Wenger et. al. Expires December 2003 [Page 14]

Time Aggregation Packet (MTAP) with 24 bit offset (MTAP24). The NAL unit type numbers assigned for STAP-A, STAP-B, MTAP16, and MTAP24 are 24, 25, 26, and 27 respectively. Specified in section 5.7.

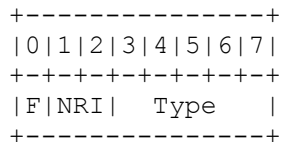
Fragmentation unit: Used to fragment a single NAL unit over multiple RTP packets. Exists with two versions identified with the NAL unit type numbers 28 and 29. Specified in section 5.8.

Table 1. Summary of NAL unit types and their payload structures.

Type	Packet	Type name	Section
1-23	NAL unit A	single NAL unit packet	5.6
24	STAP-A	Single-time aggregation packet	5.7.1
25	STAP-B	Single-time aggregation packet	5.7.1
26	MTAP16	Multi-time aggregation packet	5.7.2
27	MTAP24	Multi-time aggregation packet	5.7.2
28	FU-A	Fragmentation unit	5.8
29	FU-B	Fragmentation unit	5.8

### 5.3. NAL Unit Octet Usage

The structure and semantics of the NAL unit octet were introduced in section 1.3. For convenience, the format of the NAL unit type octet is reprinted below:



This section specifies the semantics of F and NRI according to this specification.

F: 1 bit

forbidden\_zero\_bit. A value of 0 indicates that the NAL unit type octet and payload SHOULD not contain bit errors or other

syntax violations. A value of 1 indicates that the NAL unit type octet and payload MAY contain bit errors or other syntax violations.

Network elements, such as gateways, MAY set the F bit to indicate detected bit errors in the NAL unit. The H.264 specification requires that the F bit is equal to 0. Thus, receivers MUST NOT pass NAL units in which the F bit is equal to 1 to the decoder, when the decoder is incapable of handling erroneous bitstreams. Otherwise, when the decoder is capable of handling erroneous bitstreams, receivers SHOULD pass NAL unit in which the F bit is equal to 1 to the decoder. When the F bit is set, the decoder is advised that bit errors or any other syntax violation may be present in the payload or in the NAL unit type octet. The simplest decoder reaction to respond to a NAL unit in which the F bit is equal to 1 is to discard such a NAL unit and to conceal the lost data in the discarded NAL unit.

#### NRI: 2 bits

nal\_ref\_idc. The semantics of value 00 and a non-zero value remain unchanged compared to the H.264 specification. In other words, a value of 00 indicates that the content of the NAL unit is not used to reconstruct reference pictures for inter picture prediction. Such NAL units can be discarded without risking the integrity of the reference pictures. Values above 00 indicate that the decoding of the NAL unit is required to maintain the integrity of the reference pictures.

In addition to the specification above, according to this RTP payload specification, values of NRI greater than 00 indicate the relative transport priority, as determined by the encoder. Intelligent network elements can use this information to protect more important NAL units better than less important NAL units. 11 is the highest transport priority, followed by 10, then by 01 and, finally, 00 is the lowest.

Informative note: Any non-zero value of NRI is handled identically in H.264 decoders. Therefore, receivers need not manipulate the value of NRI when passing NAL units to the decoder.

## 5.4. Packetization Modes

This memo specifies three cases of packetization modes:

- o Single NAL unit mode
- o Non-interleaved mode
- o Interleaved mode

The single NAL unit mode is targeted for conversational systems that comply with ITU-T Recommendation H.241 [16] (see section 10.1). The non-interleaved mode is targeted for conversational systems that may not comply with ITU-T Recommendation H.241. In the non-interleaved mode NAL units are transmitted in NAL unit decoding order. The interleaved mode is targeted for systems that do not require very low end-to-end latency. The interleaved mode allows transmission of NAL units out of NAL unit decoding order.

The packetization mode in use MAY be signaled by the value of the optional packetization-mode MIME parameter or by external means. The used packetization mode governs which NAL unit types are allowed in RTP payloads. Table 2 summarizes the allowed NAL unit types for each packetization mode. "No" in the "type 1-23" row indicates that an RTP payload cannot contain a single NAL unit whose type is in the range of 1 to 23, inclusive. Packetization modes are explained in detail in section 6.

Table 2. Summary of allowed NAL unit types for each packetization mode (yes = allowed, no = disallowed).

Type	Packet	Single NAL Unit Mode	Non-Interleaved Mode	Interleaved Mode
1-23	NAL unit	yes	yes	no
24	STAP-A	no	yes	no
25	STAP-B	no	no	yes
26	MTAP16	no	no	yes
27	MTAP24	no	no	yes
28	FU-A	no	yes	yes
29	FU-B	no	no	yes

## 5.5. Decoding Order Number (DON)

In the interleaved packetization mode, the transmission order of NAL units is allowed to differ from the decoding order of the NAL units. Decoding order number (DON) is a field in the payload structure or a derived variable that indicates the NAL unit decoding order. Rationale and example use cases for transmission out of decoding order and for the use of DON are given in section 11.

The coupling of transmission and decoding order is controlled by the optional interleaving-depth MIME parameter as follows. When the value of the optional interleaving-depth MIME parameter is equal to 0 and transmission of NAL units out of their decoding order is disallowed by external means, the transmission order of NAL units MUST conform to the NAL unit decoding order. When the value of the optional interleaving-depth MIME parameter is greater than 0 or transmission of NAL units out of their decoding order is allowed by external means,

- o the order of NAL units in an MTAP16 and an MTAP24 is NOT REQUIRED to be the NAL unit decoding order, and
- o the order of NAL units composed by decapsulating STAP-Bs, MTAPs, and FUs in two consecutive packets is NOT REQUIRED to be the NAL unit decoding order.

The RTP payload structures for a single NAL unit packet, an STAP-A, and an FU-A do not include DON. STAP-B and FU-B structures include DON, and the structure of MTAPs enables derivation of DON as specified in section 5.7.2.

Informative note: If a transmitter wants to encapsulate one NAL unit per packet and transmit packets out of their decoding order, STAP-B packet type can be used.

In the single NAL unit packetization mode, the transmission order of NAL units MUST be the same as their NAL unit decoding order. In the non-interleaved packetization mode, the transmission order of NAL units in single NAL unit packets and STAP-As, and FU-As MUST be

Wenger et. al. Expires December 2003 [Page 18]

the same as their NAL unit decoding order. The NAL units within an STAP MUST appear in the NAL unit decoding order.

Informative note: Due to the fact that H.264 allows the decoding order to be different from the display order, values of RTP timestamps may not be monotonically non-decreasing as a function of RTP sequence numbers.

Signaling of the value of DON for NAL units carried in STAP-B, MTAP, and a series of fragmentation units starting with an FU-B is specified in sections 5.7.1, 5.7.2, and 5.8 respectively. The DON value of the first NAL unit in transmission order MAY be set to any value. Values of DON are in the range of 0 to 65535, inclusive. After reaching the maximum value, the value of DON wraps around to 0.

The decoding order of two NAL units contained in any STAP-B, MTAP, or a series of fragmentation units starting with an FU-B is determined as follows. Let the value of DON of one NAL unit be  $D1$  and the value of DON of another NAL unit be  $D2$ . If  $D1$  equals to  $D2$ , then the NAL unit decoding order of the two NAL units can be whichever. If  $D1 < D2$  and  $D2 - D1 < 32768$ , or if  $D1 > D2$  and  $D1 - D2 \geq 32768$ , then the NAL unit having a value of DON equal to  $D1$  precedes the NAL unit having a value of DON equal to  $D2$  in NAL unit decoding order. If  $D1 < D2$  and  $D2 - D1 \geq 32768$ , or if  $D1 > D2$  and  $D1 - D2 < 32768$ , then the NAL unit having a value of DON equal to  $D2$  precedes the NAL unit having a value of DON equal to  $D1$  in NAL unit decoding order.

Values of DON related fields (DON, DONB, and DOND, see section 5.7) MUST be such that the decoding order determined by the values of DON as specified above conforms to the NAL unit decoding order. If the order of two consecutive NAL units in the NAL unit stream is switched and the new order still conforms to the NAL unit decoding order, the NAL units MAY have the same value of DON. For example, when arbitrary slice order is allowed by the video coding profile in use, all the coded slice NAL units of a coded picture are allowed to have the same value of DON. Consequently, NAL units having the same value of DON can be decoded in any order, and two NAL units having a different value of DON should be passed to the Wenger et. al. Expires December 2003 [Page 19]

decoder in the order specified above. When two consecutive NAL units in the NAL unit decoding order have a different value of DON, the value of DON for the second NAL unit in decoding order SHOULD be the value of DON for the first NAL unit in decoding order incremented by one.

An example decapsulation process to recover the NAL unit decoding order is given in section 7.

Informative note: Receivers SHOULD not expect that the absolute difference of values of DON for two consecutive NAL units in the NAL unit decoding order is equal to one even in case of error-free transmission. An increment by one is not required, because at the time of associating values of DON to NAL units, it may not be known, whether all NAL units are delivered to the receiver. For example, a gateway may not forward coded slice NAL units of non-reference pictures or SEI NAL units, when there is a shortage of bitrate in the network to which the packets are forwarded. In another example a live broadcast is interrupted by pre-encoded content such as commercials from time to time. The first intra picture of a pre-encoded clip is transmitted in advance to ensure that it is readily available in the receiver. At the time of transmitting the first intra picture, the originator does not exactly know how many NAL units are going to be encoded before the first intra picture of the pre-encoded clip follows in decoding order. Thus, the values of DON for the NAL units of the first intra picture of the pre-encoded clip have to be estimated at the time of transmitting them and gaps in values of DON may occur.

#### 5.6. Single NAL Unit Packet

The single NAL unit packet defined here MUST contain one and only one NAL unit of the types defined in [1]. This means that neither an aggregation packet nor a fragmentation unit can be used within a single NAL unit packet. A NAL unit stream composed by decapsulating single NAL unit packets in RTP sequence number order MUST conform to the NAL unit decoding order. The structure of the single NAL unit packet is shown in Figure 2.

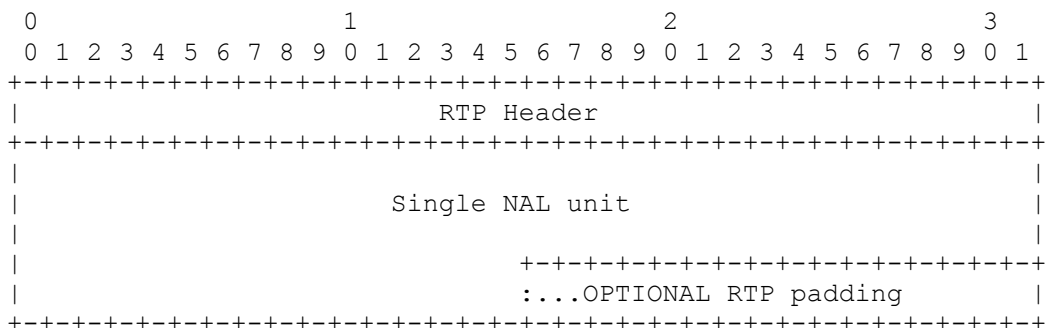


Figure 2. RTP payload format for single NAL unit packet.

5.7. Aggregation Packets

Aggregation packets are the NAL unit aggregation scheme of this payload specification. The scheme is introduced to reflect the dramatically different MTU sizes of two key target networks -- wireline IP networks (with an MTU size that is often limited by the Ethernet MTU size -- roughly 1500 bytes), and IP or non-IP (e.g. ITU-T H.324/M) based wireless communication systems with preferred transmission unit sizes of 254 bytes or less. In order to prevent media transcoding between the two worlds, and to avoid undesirable packetization overhead, a NAL unit aggregation scheme is introduced.

Two types of aggregation packets are defined by this specification:

- o Single-time aggregation packet (STAP) aggregates NAL units with identical NALU-time. Two types of STAPs are defined, one without DON (STAP-A) and another one including DON (STAP-B).
- o Multi-time aggregation packet (MTAP) aggregates NAL units with potentially differing NALU-time. Two different MTAPs are defined that differ in the length of the NAL unit timestamp offset.

The term NALU-time is defined as the value that the RTP timestamp would have if that NAL unit would be transported in its own RTP packet.

Each NAL unit to be carried in an aggregation packet is encapsulated in an aggregation unit. Please see below for the three different aggregation units and their characteristics.

The structure of the RTP payload format for aggregation packets is presented in Figure 3.

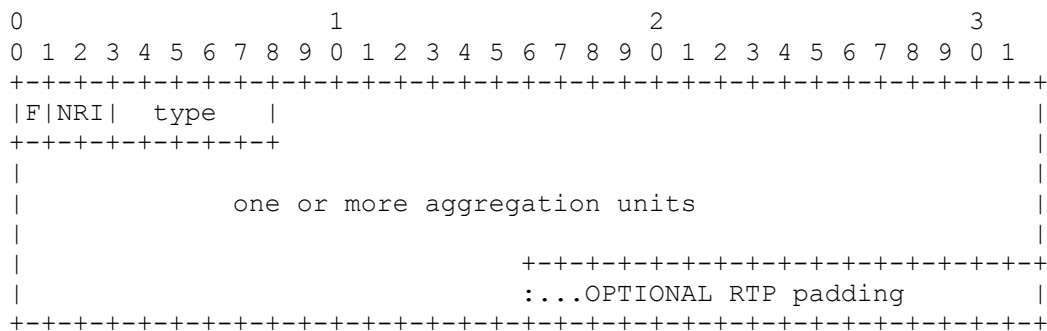


Figure 3. RTP payload format for aggregation packets.

MTAPs and STAPs share the following packetization rules: The RTP timestamp MUST be set to the earliest of the NALU times of all the NAL units to be aggregated. The type field of the NAL unit type octet MUST be set to the appropriate value as indicated in Table 3. The F bit MUST be cleared if all F bits of the aggregated NAL units are zero, otherwise it MUST be set. The value of NRI MUST be the maximum of all the NAL units carried in the aggregation packet.

Table 3. Type field for STAPs and MTAPs

Type	Packet	Timestamp offset field length (in bits)	DON related fields (DON, DONB, DOND) present
24	STAP-A	0	no
25	STAP-B	0	yes
26	MTAP16	16	yes
27	MTAP24	24	yes

The marker bit in the RTP header MUST be set to the value the marker bit of the last NAL unit of the aggregated packet would have if it were transported in its own RTP packet.

The payload of an aggregation packet consists of one or more aggregation units. See section 5.7.1 and 5.7.2 for the three different types of aggregation units. An aggregation packet can carry as many aggregation units as necessary, however the total amount of data in an aggregation packet obviously MUST fit into an IP packet, and the size SHOULD be chosen such that the resulting IP packet is smaller than the MTU size. An aggregation packet MUST NOT contain fragmentation units specified in section 5.8.

#### 5.7.1. Single-Time Aggregation Packet

Single-time aggregation packet (STAP) SHOULD be used whenever aggregating NAL units that share the same NALU-time. The payload of an STAP-A does not include DON and consists of at least one single-time aggregation unit as presented in Figure 4. The payload of an STAP-B consists of a 16-bit unsigned decoding order number (DON) followed by at least one single-time aggregation unit as presented in Figure 5.

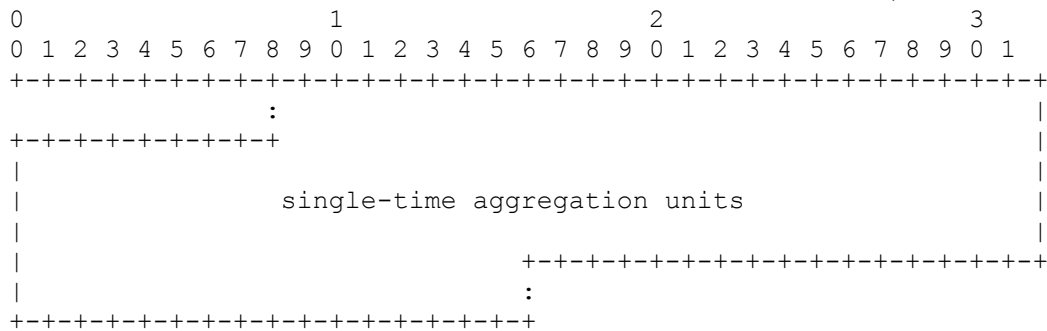


Figure 4. Payload format for STAP-A.

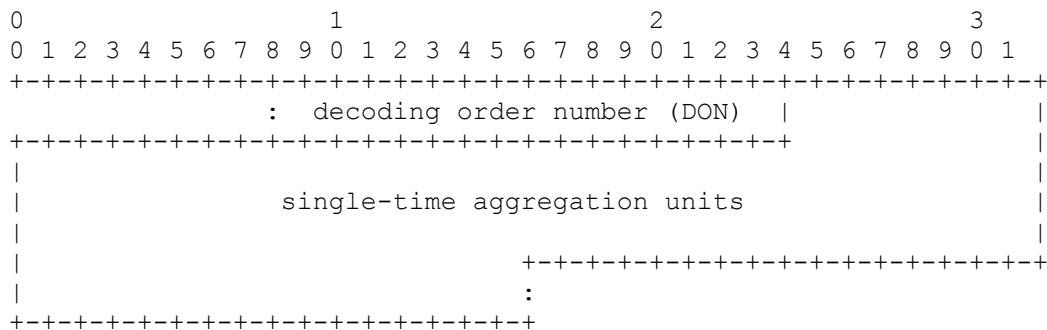


Figure 5. Payload format for STAP-B.

A single-time aggregation unit consists of 16-bit unsigned size information that indicates the size of the following NAL unit in bytes (excluding these two octets, but including the NAL unit type octet of the NAL unit), followed by the NAL unit itself including its NAL unit type byte. A single-time aggregation unit is byte-aligned within the RTP payload but it may not be aligned on a 32-bit word boundary. Figure 6 presents the structure of the single-time aggregation unit.

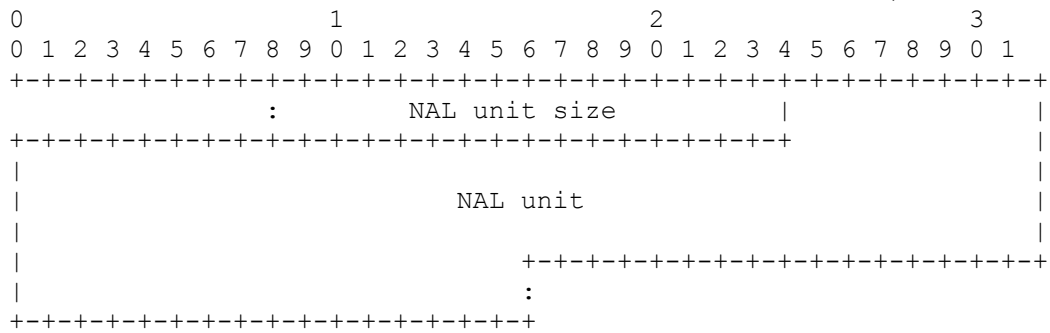


Figure 6. Structure for single-time aggregation unit.

The DON field specifies the value of DON for the first NAL unit in an STAP-B in transmission order. The value of DON for each successive NAL unit in appearance order in an STAP-B is equal to (the value of DON of the previous NAL unit in the STAP-B + 1) % 65536, in which '%' stands for the modulo operation.

Figure 7 presents an example of an RTP packet that contains an STAP-B. The STAP contains two single-time aggregation units, labeled as 1 and 2 in the figure.

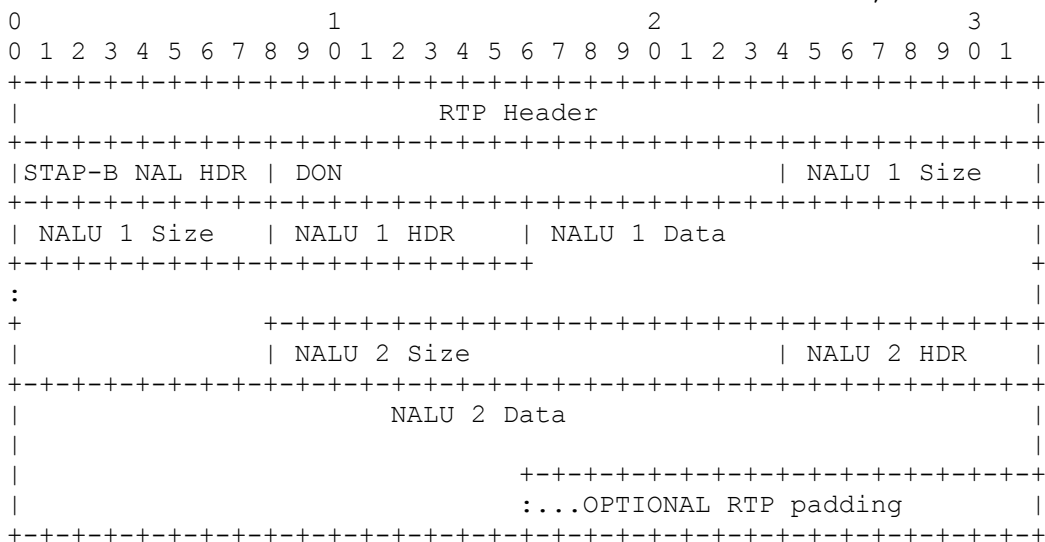


Figure 7. An example of an RTP packet including an STAP-B and two single-time aggregation units.

5.7.2. Multi-Time Aggregation Packets (MTAPs)

The NAL unit payload of MTAPs consists of a 16-bit unsigned decoding order number base (DONB) and one or more multi-time aggregation units as presented in Figure 8. DONB MUST contain the value of DON for the first NAL unit in the NAL unit decoding order among the NAL units of the MTAP.

Informative note: The first NAL unit in the NAL unit decoding order is not necessarily the first NAL unit in the order the NAL units are encapsulated in an MTAP.

The choice between the different MTAP types (MTAP16 and MTAP24) is application dependent -- the larger the timestamp offset is, the higher is the flexibility of the MTAP, but the higher is also the overhead.

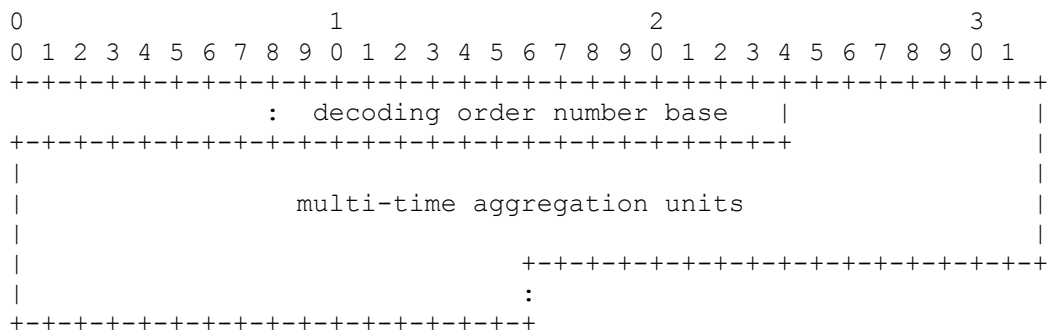


Figure 8. NAL unit payload format for MTAPs.

Two different multi-time aggregation units are defined in this specification. Both of them consist of 16 bits unsigned size information of the following NAL unit, an 8-bit unsigned decoding order number delta (DOND), and n bits of timestamp offset (TS offset) for this NAL unit, whereby n can be 16 or 24. The structure of the multi-time aggregation units for MTAP16 and MTAP24 are presented in Figure 9 and Figure 10 respectively. Note that the starting or ending position of an aggregation unit within a packet is NOT REQUIRED to be on a 32-bit word boundary. DON of the following NAL unit is equal to  $(DONB + DOND) \% 65536$ , in which % denotes the modulo operation. This memo does not specify how the NAL units within an MTAP are ordered, but, in most cases, NAL unit decoding order SHOULD be used. The timestamp offset field MUST be set to a value equal to the value of the following formula: (the NALU-time of the NAL unit - the RTP timestamp of the packet).

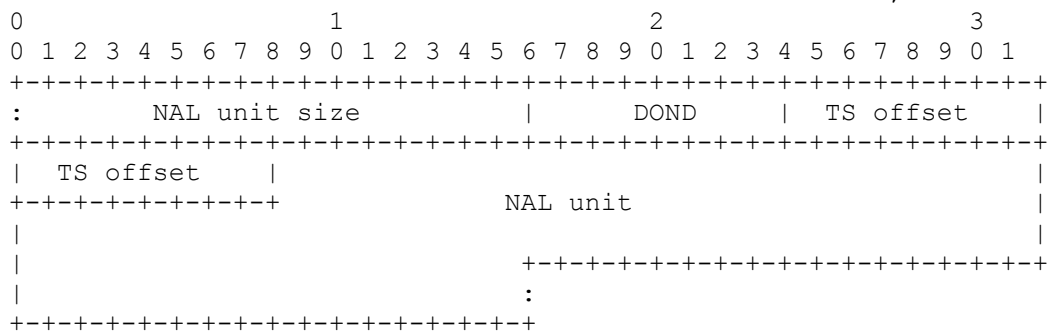


Figure 9. Multi-time aggregation unit for MTAP16

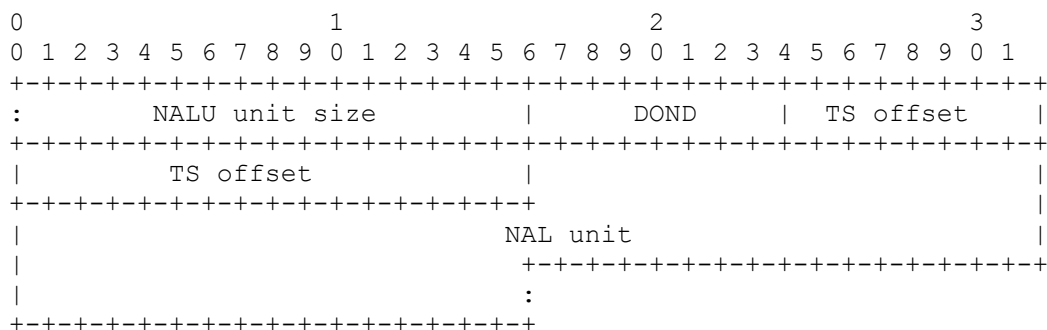


Figure 10. Multi-time aggregation unit for MTAP24

For the "earliest" multi-time aggregation unit in an MTAP the timing offset MUST be zero. Hence, the RTP timestamp of the MTAP itself is identical to the earliest NALU-time.

Informative note: The "earliest" multi-time aggregation unit is such that has the smallest RTP timestamp among all the aggregation units of an MTAP if the aggregation units were encapsulated in single NAL unit packets. Such an "earliest" aggregation unit may not be the first one in the order the aggregation units are encapsulated in an MTAP. The "earliest" NAL unit need not be the same as the first NAL unit in the NAL unit decoding order either.



- o The payload format is capable of transporting NAL units bigger than 64 kbytes over an IPv4 network that may be present in pre-recorded video, particularly in High Definition formats (there is a limit of the number of slices per picture, which results in a limit of NAL units per picture, which may result in big NAL units)
- o The fragmentation mechanism allows fragmenting a single picture and applying generic forward error correction as described in section 10.5.

Fragmentation is defined only for a single NAL unit, and not for any aggregation packets. A fragment of a NAL unit consists of an integer number of consecutive octets of that NAL unit. Each octet of the NAL unit MUST be part of exactly one fragment of that NAL unit. Fragments of the same NAL unit MUST be sent in consecutive order with ascending RTP sequence numbers (with no other RTP packets within the same RTP packet stream being sent between the first and last fragment). Similarly, a NAL unit MUST be reassembled in RTP sequence number order.

When a NAL unit is fragmented and conveyed within fragmentation units (FUs), it is referred to as fragmented NAL unit. STAPs and MTAPs MUST NOT be fragmented. FUs MUST NOT be nested, i.e., an FU MUST NOT contain another FU.

The RTP timestamp of an RTP packet carrying an FU is set to the NALU time of the fragmented NAL unit.

Figure 12 presents the RTP payload format for FU-As. An FU-A consists of a fragmentation unit indicator of one octet, a fragmentation unit header of one octet, and a fragmentation unit payload.

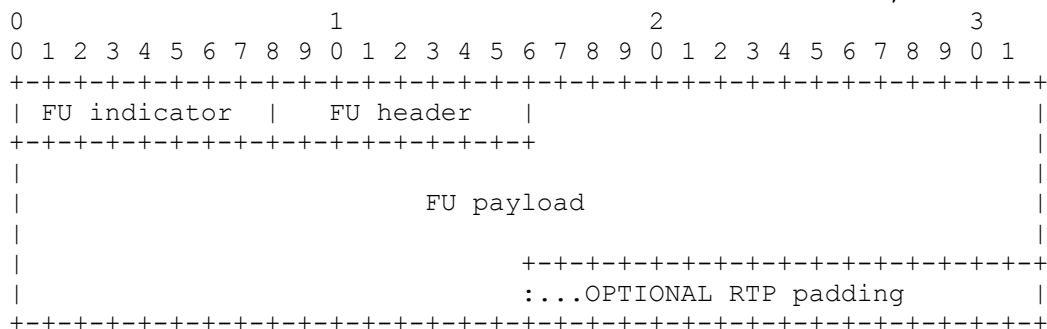


Figure 12. RTP payload format for FU-A.

Figure 13 presents the RTP payload format for FU-Bs. An FU-B consists of a fragmentation unit indicator of one octet, a fragmentation unit header of one octet, a decoding order number (DON), and a fragmentation unit payload. In other words, the structure of FU-B is the same as the structure of FU-A except for the additional DON field.

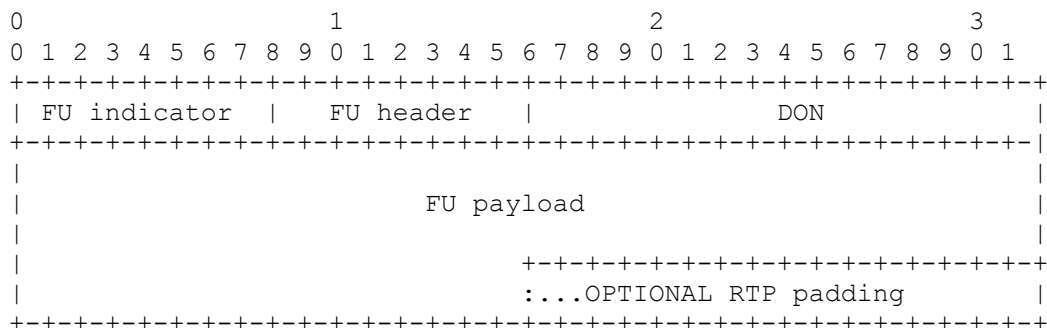


Figure 13. RTP payload format for FU-B.

NAL unit type FU-B MUST be used in the interleaved packetization mode for the first fragmentation unit of a fragmented NAL unit. NAL unit type FU-B MUST NOT be used in any other case.

The FU indicator octet has the following format:

```
+-----+
|0|1|2|3|4|5|6|7|
+---+---+---+---+---+---+
|F|NRI|  Type  |
+-----+
```

Values equal to 28 and 29 in the Type field of the FU indicator octet identify an FU-A and an FU-B respectively. The use of the F bit is described in section 1.3. The value of the NRI field MUST be set according to the value of the NRI field in the fragmented NAL unit.

The FU header has the following format:

```
+-----+
|0|1|2|3|4|5|6|7|
+---+---+---+---+---+---+
|S|E|R|  Type  |
+-----+
```

S: 1 bit

The Start bit, when one, indicates the start of a fragmented NAL unit. Otherwise, when the following FU payload is not the start of a fragmented NAL unit payload, the Start bit is set to zero.

E: 1 bit

The End bit, when one, indicates the end of a fragmented NAL unit, i.e., the last byte of the payload is also the last byte of the fragmented NAL unit. Otherwise, when the following FU payload is not the last fragment of a fragmented NAL unit, the End bit is set to zero.

R: 1 bit

The Reserved bit MUST be equal to 0 and MUST be ignored by the receiver.

Type: 5 bits

The NAL unit payload type as defined in table 7-1 of [1].

The value of DON in FU-Bs is selected as described in section 5.5.

Informative note: The DON field in FU-Bs allows gateways to fragment NAL units to FU-Bs without organizing the incoming NAL units to the NAL unit decoding order.

A fragmented NAL unit MUST NOT be transmitted in one FU, i.e., Start bit and End bit MUST NOT both be set to one in the same FU header.

The FU payload consists of fragments of the payload of the fragmented NAL unit such that if the fragmentation unit payloads of consecutive FUs are sequentially concatenated, the payload of the fragmented NAL unit is reconstructed. Note that the NAL unit type octet of the fragmented NAL unit is not included as such in the fragmentation unit payload, but rather the information of the NAL unit type octet of the fragmented NAL unit is conveyed in F and NRI fields of the FU indicator octet of the fragmentation unit and in the type field of the FU header. A FU payload MAY have any number of octets and MAY be empty.

If a fragmentation unit is lost, the receiver SHOULD discard all following fragmentation units in transmission order corresponding to the same fragmented NAL unit.

## 6. Packetization Rules

The packetization modes are introduced in section 5.2. The packetization rules that are common to more than one of the packetization modes are specified in section 6.1. The packetization rules for the single NAL unit mode, the non-interleaved mode, and the interleaved mode are specified in sections 6.2, 6.3, and 6.4 respectively.

### 6.1. Common Packetization Rules

All senders MUST enforce the following packetization rules regardless of the packetization mode in use:

- o Coded slice NAL units or coded slice data partition NAL units belonging to the same coded picture (and hence sharing the same RTP timestamp value) MAY be sent in any order permitted by the applicable profile defined in [1], although, for delay-critical systems, they SHOULD be sent in their original coding order to minimize the delay. Note that the coding order is not necessarily the scan order, but the order the NAL packets become available to the RTP stack.
- o Sequence and picture parameter set NAL units MUST NOT be sent in an RTP session whose parameter sets were already changed by control protocol messages during the lifetime of the RTP session.
- o Network elements such as gateways MUST NOT duplicate any NAL unit except for sequence or picture parameter set NAL units, because neither this memo nor the H.264 specification provides means to identify duplicated NAL units. Sequence and picture parameter set NAL units MAY be duplicated to make their correct reception more probable, but any such duplication MUST NOT affect the contents of any active sequence or picture parameter set.

Senders according to the non-interleaved mode and the interleaved mode MUST enforce the following packetization rule:

- o Network elements such as gateways MAY convert single NAL unit packets into one aggregation packet, convert an aggregation packet into several single NAL unit packets, or mix both concepts. However, when doing so they SHOULD take into account at least the following parameters: path MTU size, unequal protection mechanisms (e.g. through packet-based FEC according to RFC 2733 [21], carried by RFC 2198 [20], especially for sequence and picture parameter set NAL units and coded slice data partition A NAL units), bearable latency of the system, and buffering capabilities of the receiver.

## 6.2. Single NAL Unit Mode

This mode is in use when the value of the optional packetization-mode MIME parameter is equal to 0 or packetization-mode is not present or no other packetization mode is signaled by external means. All receivers MUST support this mode. It is primarily intended for low-delay applications that are compatible with systems using ITU-T Recommendation H.241 [16] (see section 10.1). Only single NAL unit packets MAY be used in this mode. STAPs, MTAPs, and FUs MUST NOT be used. The transmission order of single NAL unit packets MUST comply with the NAL unit decoding order.

## 6.3. Non-Interleaved Mode

This mode is in use when the value of the optional packetization-mode MIME parameter is equal to 1 or the mode is turned on by external means. This mode SHOULD be supported. It is primarily intended for low-delay applications. Only single NAL unit packets, STAP-As and FU-As MAY be used in this mode. STAP-Bs, MTAPs, and FU-Bs MUST NOT be used. The transmission order of NAL units MUST comply with the NAL unit decoding order.

## 6.4. Interleaved Mode

This mode is in use when the value of the optional packetization-mode MIME parameter is equal to 2 or the mode is turned on by external means. Some receivers MAY support this mode. STAP-Bs, MTAPs, FU-As, and FU-Bs MAY be used. STAP-As and single NAL unit packets MUST NOT be used. The transmission order of packets and NAL units is constrained as specified in section 5.5.

## 7. De-Packetization Process (Informative)

The de-packetization process is implementation dependent. Hence, the following description should be seen as an example of a suitable implementation. Other schemes may be used as well. Optimizations relative to the described algorithms are likely possible. Section 7.1 presents the de-packetization process for Wenger et. al. Expires December 2003 [Page 35]

the single NAL unit and non-interleaved packetization modes, whereas section 7.2 describes the process for the interleaved mode. Section 7.3 includes additional decapsulation guidelines for intelligent receivers.

#### 7.1. Single NAL Unit and Non-Interleaved Mode

The receiver includes a receiver buffer to compensate transmission delay jitter. The receiver stores incoming packets in reception order into the receiver buffer. Packets are decapsulated in RTP sequence number order. If a decapsulated packet is a single NAL unit packet, the NAL unit contained in the packet is passed to the decoder immediately after decapsulation. If a decapsulated packet is an STAP-A, the NAL units contained in the packet are passed to the decoder in the order they are encapsulated in the packet immediately after decapsulation. If a decapsulated packet is an FU-A, all the fragments of the fragmented NAL unit are concatenated and passed to the decoder. Note: If the decoder supports Arbitrary Slice Order, coded slices of a picture can be passed to the decoder in any order regardless of their reception and transmission order.

#### 7.2. Interleaved Mode

The general concept behind these de-packetization rules is to reorder NAL units from transmission order to the NAL unit decoding order.

The receiver includes a receiver buffer, which is used to reorder packets from transmission order to the NAL unit decoding order. The receiver may use the following guidelines when determining the size of the receiver buffer. The optional interleaving-depth MIME parameter indicates the size of the receiver buffer as the number of VCL NAL units. The size of the receiver buffer in bytes may be estimated by multiplying the optional init-buf-time MIME parameter with the bandwidth-value of the corresponding media level SDP description parameter, if available (and taking into account the necessary conversions between the units of init-buf-time and bandwidth-value). The receiver should also take buffering for

transmission delay jitter into account and either reserve a separate buffer for transmission delay jitter buffering or combine the buffer for transmission delay jitter with the receiver buffer.

The receiver stores incoming NAL units in reception order into the receiver buffer as follows. NAL units of aggregation packets are stored into the receiver buffer individually. The value of DON is calculated and stored for all NAL units.

Hereinafter, let  $N$  be the value of the optional interleaving-depth MIME type parameter (see section 8.1) incremented by 1.

Furthermore, let function  $\text{AbsDON}$  be the same as specified in section 8.1 and function  $\text{don\_diff}(m,n)$  be specified as follows:

If  $\text{DON}(m) == \text{DON}(n)$ ,  $\text{don\_diff}(m,n) = 0$

If  $(\text{DON}(m) < \text{DON}(n) \text{ and } \text{DON}(n) - \text{DON}(m) < 32768)$ ,  
 $\text{don\_diff}(m,n) = \text{DON}(n) - \text{DON}(m)$

If  $(\text{DON}(m) > \text{DON}(n) \text{ and } \text{DON}(m) - \text{DON}(n) \geq 32768)$ ,  
 $\text{don\_diff}(m,n) = 65536 - \text{DON}(m) + \text{DON}(n)$

If  $(\text{DON}(m) < \text{DON}(n) \text{ and } \text{DON}(n) - \text{DON}(m) \geq 32768)$ ,  
 $\text{don\_diff}(m,n) = -(\text{DON}(m) + 65536 - \text{DON}(n))$

If  $(\text{DON}(m) > \text{DON}(n) \text{ and } \text{DON}(m) - \text{DON}(n) < 32768)$ ,  
 $\text{don\_diff}(m,n) = -(\text{DON}(m) - \text{DON}(n))$

where  $\text{DON}(i)$  is the decoding order number of the NAL unit having index  $i$  in the transmission order. The decoding order number is specified in section 5.5 of this RTP payload specification.

A positive value of  $\text{don\_diff}(m,n)$  indicates that the NAL unit having transmission order index  $n$  follows, in decoding order, the NAL unit having transmission order index  $m$ .

There are two buffering states in the receiver: initial buffering and buffering while playing. Initial buffering occurs when the RTP session is initialized. After initial buffering, decoding and playback is started and the buffering-while-playing mode is used.

Initial buffering lasts until one of the following conditions is fulfilled:

- o There are N VCL NAL units in the receiver buffer.
- o If max-don-diff is present, don\_diff(m,n) is greater than the value of max-don-diff, in which n corresponds to the NAL unit having the greatest value of AbsDON among the received NAL units and m corresponds to the NAL unit having the smallest value of AbsDON among the received NAL units.
- o Initial buffering has lasted for the duration equal to or greater than the value of the optional init-buf-time MIME parameter.

The NAL units to be removed from the receiver buffer are determined as follows:

- o If the receiver buffer contains at least N VCL NAL units, NAL units are removed from the receiver buffer and passed to the decoder in the order specified below until the buffer contains N-1 VCL NAL units.
- o If max-don-diff is present, all NAL units m for which don\_diff(m,n) is greater than max-don-diff are removed from the receiver buffer and passed to the decoder in the order specified below. Herein, n corresponds to the NAL unit having the greatest value of AbsDON among the received NAL units.
- o Variable ts is set to the value of system timer that was initialized to 0 when the first packet of the NAL unit stream was received. If the receiver buffer contains a NAL unit whose reception time tr fulfills the condition that  $ts - tr > \text{init-buf-time}$ , NAL units are passed to the decoder (and removed from the receiver buffer) in the order specified below until the receiver buffer contains no NAL unit whose reception time tr fulfills the specified condition. Note that transmission delay jitter should be taken into account in the calculations with timestamps.

The order that NAL units are passed to the decoder is specified as follows:

- o Let PDON be a variable that is initialized to 0 at the beginning of the an RTP session.

- o For each NAL unit associated with a value of DON, a DON distance is calculated as follows. If the value of DON of the NAL unit is larger than the value of PDON, the DON distance is equal to  $DON - PDON$ . Otherwise, the DON distance is equal to  $65535 - PDON + DON + 1$ .
- o NAL units are delivered to the decoder in ascending order of DON distance. If several NAL units share the same value of DON distance, they can be passed to the decoder in any order.
- o When a desired number of NAL units have been passed to the decoder, the value of PDON is set to the value of DON for the last NAL unit passed to the decoder.

### 7.3. Additional De-Packetization Guidelines

The following additional de-packetization rules may be used to implement an operational H.264 de-packetizer:

- o Intelligent RTP receivers (e.g. in gateways) may identify lost coded slice data partitions A (DPAs). If a lost DPA is found, a gateway may decide not to send the corresponding coded slice data partitions B and C, as their information is meaningless for H.264 decoders. In this way a network element can reduce network load by discarding useless packets, without parsing a complex bitstream.
- o Intelligent RTP receivers (e.g. in gateways) may identify lost FUs. If a lost FU is found, a gateway may decide not to send the following FUs of the same NAL unit, as their information is meaningless for H.264 decoders. In this way a network element can reduce network load by discarding useless packets, without parsing a complex bitstream.
- o Intelligent receivers may discard all packets in which the value of the NRI field of the NAL unit type octet is equal to 0. However, they process those packets if possible, because the user experience may suffer if the packets are discarded.

## 8. Payload Format Parameters

This section specifies the parameters that MAY be used to select optional features of the payload format. The parameters are specified here as part of the MIME subtype registration for the ITU-T H.264 | ISO/IEC 14496-10 codec. A mapping of the parameters into the Session Description Protocol (SDP) [5] is also provided for those applications that use SDP. Equivalent parameters could be defined elsewhere for use with control protocols that do not use MIME or SDP.

### 8.1. MIME Registration

The MIME subtype for the ITU-T H.264 | ISO/IEC 14496-10 codec is allocated from the IETF tree.

The receiver MUST ignore any unspecified parameter.

Media Type name: video

Media subtype name: H264

Required parameters: none

Optional parameters:

profile-level-id: A string of profile-level elements without any delimiters, in which each profile-level element is a base16 [6] (hexadecimal) representation of the following three bytes in the sequence parameter set NAL unit specified in [1]: 1) profile\_idc, 2) a byte herein referred to as profile-iop, composed of the values of constrained\_set0\_flag, constrained\_set1\_flag, constrained\_set2\_flag, and reserved\_zero\_5bits in bit-significance order starting from the most significant bit, and 3) level\_idc. Note that reserved\_zero\_5bits is required to be equal to

0 in [1], but other values for it may be specified in the future by ITU-T or ISO/IEC. If the profile-level-id parameter is used for indicating properties of a NAL unit stream, it indicates the profiles that are in use in the stream and the highest level that is in use for each signaled profile. The profile-iop byte for each signaled profile indicates whether the NAL unit stream also obeys all constraints of the indicated profiles as follows. If bit 7 (the most significant bit), bit 6, or bit 5 of profile-iop is equal to 1, all constraints of the Baseline profile, the Main profile, or the Extended profile, respectively, are obeyed in the NAL unit stream. If the profile-level-id parameter is used for capability exchange or session setup procedure, it indicates the profiles that the codec supports and the highest level that is supported for each signaled profile. The profile-iop byte for each signaled profile indicates whether the codec has such additional limitations that only the common subset of the algorithmic features and limitations of the profiles signaled with the profile-iop byte and the profile indicated by profile\_idc is supported by the codec. For example, if a codec supports the Baseline Profile at level 3 and below and the Main Profile at level 2.1 and below without any additional limitations, the profile-level-id becomes 42A01E4D4015. If a codec supports only the common subset of the coding tools of the Baseline profile and the Main profile at level 2.1 and below, the profile-level-id becomes 42E015. If no profile-level-id is present, the Baseline Profile without additional constraints at Level 1 MUST be implied.

parameter-sets: This parameter MAY be used to convey such sequence and picture parameter set NAL units, herein referred to as the initial parameter set NAL units, that MUST precede any other NAL units in decoding order. The parameter MUST NOT be used to indicate codec capability in any capability exchange procedure. The value of the parameter is the base64 [6] representation of the initial parameter set NAL units as specified in sections 7.3.2.1 and 7.3.2.2 of [1]. The parameter sets are conveyed in decoding order and no framing of the parameter set NAL units takes place. A comma is used to separate any pair of parameter sets in the list. Note that the number of bytes in a parameter set NAL unit is typically less than 10 bytes, but a picture parameter set NAL unit can contain several hundreds of bytes.

packetization-mode: When the value of packetization-mode is equal to 0 or packetization-mode is not present, single NAL unit packets MUST be present in the stream, but STAPs, MTAPs, and FUs MUST NOT be present in the stream. This mode is in use in standards using ITU-T Recommendation H.241 [16] (see section 10.1). When the value of packetization-mode is equal to 1, single NAL unit packets, STAP-As and FU-As MAY be present in the stream, but STAP-Bs, MTAPs, and FU-Bs MUST NOT be present in the stream. When the value of packetization-mode is equal to 2, STAP-Bs, MTAPs, FU-As, and FU-Bs MAY be present in the stream, but single NAL unit packets and STAP-As MUST NOT be present in the stream. The value of packetization mode MUST be an integer in the range of 0 to 2, inclusive.

interleaving-depth: This parameter MUST NOT be present when packetization-mode is not present or the value of packetization-mode is equal to 0 or 1. This parameter MUST be present when the value of packetization-mode is equal to 2.

This parameter signals the properties of a NAL unit stream or the capabilities of a receiver implementation. When the parameter is used to signal the properties of a NAL unit stream, it specifies the maximum number of VCL NAL units that precede any VCL NAL unit in the NAL unit stream in transmission order and follow the VCL NAL unit in decoding order. Consequently, it is guaranteed that receivers can reconstruct NAL unit decoding order, when the buffer size for NAL unit decoding order recovery is at least the value of interleaving-depth + 1 in terms of VCL NAL units. When the parameter is used to signal the capabilities of a receiver implementation, the receiver is able to correctly reconstruct the NAL unit decoding order of NAL unit streams that are characterized by the same value of interleaving-depth. When the receiver buffers such number of VCL NAL units that equals to or is greater than the value of interleaving-depth, it is able to reconstruct NAL unit decoding order from the transmission order.

If the parameter is not present, then a value of 0 MUST be used for interleaving-depth. The value of interleaving-depth MUST be an integer in the range of 0 to 32767, inclusive.

init-buf-time: This parameter MAY be used to signal the properties of a NAL unit stream or the capabilities of a receiver implementation.

When the parameter is used to signal the properties of a NAL unit stream, it signals the initial buffering time that a receiver MUST buffer before starting decoding to recover the NAL unit decoding order from the transmission order. The parameter is the maximum value of (transmission time of a NAL unit - decoding time of the NAL unit) assuming reliable and instantaneous transmission, the same timeline for transmission and decoding, and starting of decoding when the first packet arrives.

An example of specifying the value of init-buf-time follows: A NAL unit stream is sent in the following interleaved order, in which the value corresponds to the decoding time and the transmission order is from left to right:

0 2 1 3 5 4 6 8 7 ...

Assuming a steady transmission rate of NAL units, the transmission times are:

0 1 2 3 4 5 6 7 8 ...

Subtracting the decoding time from the transmission time column-wise results into the following series:

0 -1 1 0 -1 1 0 -1 1 ...

Thus, the value of init-buf-time in this example is 1 in terms of intervals of NAL unit transmission times.

When the parameter is used to signal the capabilities of a receiver implementation, it signals the duration of initial buffering that the receiver is capable of handling in any circumstances.

The parameter is coded as a decimal representation in clock ticks of a 90-kHz clock. If the parameter is not present, then a value of 0 MUST be used for init-buf-time. The value of initial-init-buf-time MUST be an integer in the range of 0 to 4 294 967 295, inclusive.

Receivers SHOULD take transmission delay jitter buffering, including buffering for the delay jitter caused by mixers, translators, gateways, proxies, traffic-shapers and other network elements, into account in addition to the signaled init-buf-time.

max-don-diff: This parameter MAY be used to signal the properties of a NAL unit stream. It MUST NOT be used to signal transmitter or receiver or codec capabilities. The parameter MUST NOT be present, if the value of packetization-mode is equal to 0 or 1. max-don-diff is an integer in the range of 0 to 32767, inclusive. If max-don-diff is not present, the value of the parameter is unspecified. max-don-diff is calculated as follows:

$$\text{max-don-diff} = \max\{\text{AbsDON}(i) - \text{AbsDON}(j)\},$$

for any  $i$  and any  $j > i$ ,

where  $i$  and  $j$  indicate the index of the NAL unit in the transmission order and AbsDON denotes such decoding order number of the NAL unit that does not wrap around to 0 after 65535. In other words, AbsDON is calculated as follows: Let  $m$  and  $n$  are consecutive NAL units in transmission order. For the very first NAL unit in transmission order (whose index is 0),  $\text{AbsDON}(0) = \text{DON}(0)$ . For other NAL units, AbsDON is calculated as follows:

If  $DON(m) == DON(n)$ ,  $AbsDON(n) = AbsDON(m)$

If  $(DON(m) < DON(n) \text{ and } DON(n) - DON(m) < 32768)$ ,  
 $AbsDON(n) = AbsDON(m) + DON(n) - DON(m)$

If  $(DON(m) > DON(n) \text{ and } DON(m) - DON(n) >= 32768)$ ,  
 $AbsDON(n) = AbsDON(m) + 65536 - DON(m) + DON(n)$

If  $(DON(m) < DON(n) \text{ and } DON(n) - DON(m) >= 32768)$ ,  
 $AbsDON(n) = AbsDON(m) - (DON(m) + 65536 - DON(n))$

If  $(DON(m) > DON(n) \text{ and } DON(m) - DON(n) < 32768)$ ,  
 $AbsDON(n) = AbsDON(m) - (DON(m) - DON(n))$

where  $DON(i)$  is the decoding order number of the NAL unit having index  $i$  in the transmission order. The decoding order number is specified in section 5.5 of this RTP payload specification.

Informative note: Receivers MAY use max-don-diff to trigger which NAL units in the receiver buffer can be passed to the decoder.

Encoding considerations:

This type is only defined for transfer via RTP (RFC 3550).

Security considerations:

See section 9 of RFC XXXX. [Ed.Note: to be replaced with the RFC number of this specification]

Public specification:

Please refer to RFC XXXX [Ed.Note: to be replaced with the RFC number of this specification] and its section 15.

Additional information:

None

File extensions: none

Macintosh file type code: none

Object identifier or OID: none

Person & email address to contact for further information:

stewe@cs.tu-berlin.de

Intended usage: COMMON.

Author/Change controller:

stewe@cs.tu-berlin.de

IETF Audio/Video transport working group

## 8.2. SDP Parameters

The MIME media type video/H264 string is mapped to fields in the Session Description Protocol (SDP) [5] as follows:

- o The media name in the "m=" line of SDP MUST be video.
- o The encoding name in the "a=rtpmap" line of SDP MUST be H264 (the MIME subtype).
- o The clock rate in the "a=rtpmap" line MUST be 90000.
- o The optional parameters "profile-level-id", "parameter-sets", "packetization-mode", "interleaving-depth", "init-buf-time", and "max-don-diff", if any, SHALL be included in the "a=fmtp" line of SDP. These parameters are expressed as a MIME media type string, in the form of as a semicolon separated list of parameter=value pairs.

An example of media representation in SDP is as follows (Baseline Profile, Level 3.0, more than one slice group, arbitrary slice ordering, and redundant slices are in use):

```
m=video 49170/2 RTP/AVP 98
a=rtpmap:98 H264/90000
a=fmtp:98 profile-level-id=42A01E
```

## 9. Security Considerations

RTP packets using the payload format defined in this specification are subject to the security considerations discussed in the RTP specification [4], and any appropriate RTP profile (for example [17]). This implies that confidentiality of the media streams is achieved by encryption. Because the data compression used with this payload format is applied end-to-end, encryption may be performed after compression so there is no conflict between the two operations.

A potential denial-of-service threat exists for data encodings using compression techniques that have non-uniform receiver-end computational load. The attacker can inject such pathological datagrams into the stream that are complex to decode and cause the receiver to be overloaded. H.264 is particularly vulnerable to such attacks because it is extremely simple to generate datagrams containing NAL units that affect the decoding process of many future NAL units.

As with any IP-based protocol, in some circumstances a receiver may be overloaded simply by the receipt of too many packets, either desired or undesired. Network-layer authentication may be used to discard packets from undesired sources, but the processing cost of the authentication itself may be too high. In a multicast environment, pruning of specific sources may be implemented in future versions of IGMP [18] and in multicast routing protocols to allow a receiver to select which sources are allowed to reach it.

Decoders MUST exercise caution with respect to the handling of user data SEI messages, particularly if they contain active elements, and MUST restrict their domain of applicability to the presentation containing the stream.

## 10. Informative Appendix: Application Examples

This payload specification is very flexible in its use, to cover the extremely wide application space that is anticipated for the H.264. However, such a great flexibility also makes it difficult for an implementer to decide on a reasonable packetization scheme. Some information how to apply this specification to real-world scenarios is likely to appear in the form of academic publications and a test model software and description in the near future. However, some preliminary usage scenarios are described here as well.

### 10.1. Video Telephony according to ITU-T Recommendation H.241 Annex A

H.323-based video telephony systems that use H.264 as an optional video compression scheme are required to support H.241 Annex A [16] as a packetization scheme. The packetization mechanism defined in this Annex is technically identical with a small subset of this specification.

When operating according to H.241 Annex A, parameter sets NAL units are sent in-band. Only Single NAL unit packets are used. A typical packet stream generated by such a system consists of all sequence and picture parameter sets used for the future video sequence, possibly sent in more than one copy to raise the likeliness of their arrival at the receiver, followed by the packets carrying the NAL units of the IDR picture, and followed by packets carrying the subsequent pictures. Many such systems are not sending IDR pictures regularly, but only when required by user interaction or by control protocol means, e.g. when switching between video channels in an Multipoint Control Unit.

## 10.2. Video Telephony, No Slice Data Partitioning, No NAL Unit Aggregation

The RTP part of this scheme is implemented and tested (though not the control-protocol part, see below).

In most real-world video telephony applications, the picture parameters such as picture size or optional modes never change during the lifetime of a connection. Hence, all necessary parameter sets (usually only one) are sent as a side effect of the capability exchange/announcement process e.g. according to the SDP syntax specified in section 8.2 of this document. Since all necessary parameter set information is established before the RTP session starts, there is no need for sending any parameter set NAL units. Slice data partitioning is not used either. Hence, the RTP packet stream consists basically of NAL units that carry single coded slices.

The encoder chooses the size of coded slice NAL units such that they offer the best performance. Often, this is done by adapting the coded slice size to the MTU size of the IP network. For small picture sizes this may result in a one-picture-per-one-packet strategy. Intra refresh algorithms clean up the loss of packets and the resulting drift-related artifacts.

## 10.3. Video Telephony, Interleaved Packetization Using NAL Unit Aggregation

This scheme allows better error concealment and is widely used in H.263 based designed using RFC 2429 packetization [11]. It is also implemented and good results were reported [13].

The VCL encoder codes the source picture such that all macroblocks (MBs) of one MB line are assigned to one slice. All slices with even MB row addresses are combined into one STAP, and all slices with odd MB row addresses into another STAP. Those STAPs are transmitted as RTP packets. The establishment of the parameter sets is performed as discussed above.

Note that the use of STAPs is essential here, because the high number of individual slices (18 for a CIF picture) would lead to unacceptably high IP/UDP/RTP header overhead (unless the source coding tool FMO is used, which is not assumed in this scenario). Furthermore, some wireless video transmission systems, such as H.324M and the IP-based video telephony specified in 3GPP, are likely to use relatively small transport packet size. For example, a typical MTU size of H.223 AL3 SDU is around 100 bytes [19]. Coding individual slices according to this packetization scheme provides a further advantage in communication between wired and wireless networks, as individual slices are likely to be smaller than the preferred maximum packet size of wireless systems. Consequently, a gateway can convert the STAPs used in a wired network to several RTP packets with only one NAL unit that are preferred in a wireless network and vice versa.

#### 10.4. Video Telephony, with Data Partitioning

This scheme is implemented and was shown to offer good performance especially at higher packet loss rates [13].

Data Partitioning is known to be useful only when some form of unequal error protection is available. Normally, in single-session RTP environments, even error characteristics are assumed, i.e., the packet loss probability of all packets of the session is the same statistically. However, there are means to reduce the packet loss probability of individual packets in an RTP session. RFC 2198 [20], for example, allows carrying a redundant copy of an essential packet in the next RTP packet. Packet-based Forward Error Correction [21] carried in RFC 2198 is also an appropriate means to protect high priority information.

In all cases, the incurred overhead is substantial, but in the same order of magnitude as the number of bits that have otherwise be spent for intra information. However, this mechanism is not adding any delay to the system.

Again, the complete parameter set establishment is performed through control protocol means.

### 10.5. Video Telephony or Streaming, with FUs and Forward Error Correction

This scheme is implemented and was shown to provide good performance especially at higher packet loss rates [22].

The most efficient means to combat packet-losses for scenarios where retransmissions are not applicable is forward error correction (FEC). Although end-to-end solutions are usually not preferable, they are unavoidable in some scenarios. For example, RFC2733 [21] provides means to use generic FEC in packet-loss environments. A binary forward error correcting code is generated by applying the XOR operation to the bits at the same bit position in different packets. The binary code can be specified by the parameters  $(n,k)$  in which  $k$  is the number of information packets used in the connection and  $n$  is the total number of packets generated for  $k$  information packets, i.e.,  $n-k$  parity packets are generated for  $k$  information packets.

When using a code with parameters  $(n,k)$  within the RFC2733 framework, the following properties are well-known:

- a) RFC2733 can only be applied over a sequence of RTP packets, not over one RTP packet.
- b) RFC2733 is most bit-rate efficient if XOR-connected packets have equal length.
- c) At the same packet loss probability  $p$  and for a fixed  $k$ , the greater the value of  $n$  is, the smaller the residual error probability becomes. For example, for packet loss probability 10%,  $k=1$ , and  $n=2$ , the residual error probability is about 1%, whereas for  $n=3$ , the residual error probability is about 0.1%.
- d) At the same packet loss probability  $p$  and for a fixed code rate  $k/n$ , the greater the value of  $n$  is, the smaller the residual error probability becomes. For example, at a packet loss probability of  $p=10%$ ,  $k=1$  and  $n=2$ , the residual error rate is about 1%, whereas for an extended Golay code with  $k=12$  and  $n=24$ , the residual error rate is about 0.01%.

For applying RFC2733 in combination with H.264 baseline coded video without using FUs several options might be considered:

- 1) The video encoder produces NAL units where each video frame is coded in a single slice. Applying FEC, one could use a simple code, e.g.  $(n=2, k=1)$ , i.e., each NAL unit would basically just be repeated. The disadvantage is obviously the bad code performance according to (d) and the low flexibility as only  $(n, k=1)$  codes can be used.
- 2) The video encoder produces NAL units where each video frame is encoded in a single slice. Applying FEC, one could use a better code, e.g.  $(n=24, k=12)$ , over a sequence of NAL units. The disadvantage is obviously that in case of losses a significant delay is introduced and packets of completely different length might be connected, which decreases bit-rate efficiency according to (b)
- 3) The video encoder produces NAL units, where a certain frame contains  $k$  slices of possibly almost equal length. Then, applying FEC, a better code, e.g.  $(n=24, k=12)$ , over the sequence of NAL units for each frame can be used. The delay compared to (2) is reduced, but several disadvantages are obvious. Firstly, the coding efficiency of the encoded video is lowered significantly as slice-structured coding reduces intra-frame prediction and additional slice overhead is necessary. Secondly, pre-encoded content or, when operating over a gateway, the video is usually not appropriately coded with  $k$  slices such that FEC can be applied. Finally, the encoding of video producing  $k$  slices of equal length is not straightforward and might require more than one encoding pass.

Many of the mentioned disadvantages can be avoided by applying FUs in combination with FEC. Each NAL unit can be split into any number of FUs of basically equal length, and therefore FEC with a reasonable  $k$  and  $n$  can be applied even if the encoder made no effort of producing slices of equal length. For example, a coded slice NAL unit containing an entire frame can be split to  $k$  FUs and a parity check code  $(n=k+1, k)$  can be applied.

The presented technique makes it possible to achieve good transmission error tolerance even if no additional source coding layer redundancy, such as periodic intra frames, is present.

Consequently, the same coded video sequence can be used for achieving the maximum compression efficiency and quality over error-free transmission and for transmission over error-prone networks. Furthermore, the technique allows the application of FEC to pre-encoded sequences without adding delay. In addition, in this case pre-encoded sequences that are not encoded for error-prone networks can still be transmitted almost reliably without adding extensive delays. In addition, FUs of equal length result in a bit-rate efficient use of RFC2733.

In case that the error probability depends on the length of the transmitted packet, e.g. in case of mobile transmission [15], the benefits of applying FUs with FEC are even more obvious. Basically, the flexibility of the size of FUs allows applying appropriate FEC for each NAL unit and even unequal error protection of NAL units.

The incurred overhead when using FUs and FEC is substantial, but in the same order of magnitude as the number of bits that have to be spent for intra coded macroblocks if no FEC is applied. In [22] it was shown that the overall performance at the same error rate and the same overall bit-rate including the overhead, the FEC-based approach can enhance the quality.

#### 10.6. Low-Bit-Rate Streaming

This scheme has been implemented with H.263 and non-standard RTP packetization and gave good results [23]. There is no technical reason why similarly good results could not be achievable with H.264.

In today's Internet streaming, some of the offered bit-rates are relatively low in order to allow terminals with dial-up modems to access the content. In wired IP networks, relatively large packets, say 500 - 1500 bytes, are preferred to smaller and more frequently occurring packets in order to reduce network congestion. Moreover, use of large packets decreases the amount of RTP/UDP/IP header overhead. For low-bit-rate video, the use of large packets

Wenger et. al. Expires December 2003 [Page 54]

means that sometimes up to few pictures should be encapsulated in one packet.

However, loss of a packet including many coded pictures would have drastic consequences in visual quality, as there is practically no other way to conceal a loss of an entire picture than to repeat the previous one. One way to construct relatively large packets and maintain possibilities for successful loss concealment is to construct MTAPs that contain slices from several pictures in an interleaved manner. An MTAP should not contain spatially adjacent slices from the same picture or spatially overlapping slices from any picture. If a packet is lost, it is likely that a lost slice is surrounded by spatially adjacent slices of the same picture and spatially corresponding slices of the temporally previous and succeeding pictures. Consequently, concealment of the lost slice is likely to succeed relatively well.

#### 10.7. Robust Packet Scheduling in Video Streaming

This scheme has been implemented with MPEG-4 Part 2 and simulated in a wireless streaming environment [24]. There is no technical reason why similar or better results could not be achievable with H.264.

Streaming clients typically have a receiver buffer that is capable of storing a relatively large amount of data. Initially, when a streaming session is established, a client does not start playing the stream back immediately, but rather it typically buffers the incoming data for a few seconds. This buffering helps to maintain continuous playback, because, in case of occasional increased transmission delays or network throughput drops, the client can decode and play buffered data. Otherwise, without initial buffering, the client has to freeze the display, stop decoding, and wait for incoming data. The buffering is also necessary for either automatic or selective retransmission in any protocol level. If any part of a picture is lost, a retransmission mechanism may be used to resend the lost data. If the retransmitted data is received before its scheduled decoding or playback time, the loss is perfectly recovered. Coded pictures can be ranked according to

their importance in the subjective quality of the decoded sequence. For example, non-reference pictures, such as conventional B pictures, are subjectively least important, because their absence does not affect decoding of any other pictures. In addition to non-reference pictures, the ITU-T H.264 | ISO/IEC 14496-10 standard includes a temporal scalability method called sub-sequences [25]. Subjective ranking can also be made on coded slice data partition or slice group basis. Coded slices and coded slice data partitions that are subjectively the most important can be sent earlier than their decoding order indicates, whereas coded slices and coded slice data partitions that are subjectively the least important can be sent later than their natural coding order indicates. Consequently, any retransmitted parts of the most important slices and coded slice data partitions are more likely to be received before their scheduled decoding or playback time compared to the least important slices and slice data partitions.

## 11. Informative Appendix: Rationale for Decoding Order Number

### 11.1. Introduction

The Decoding Order Number (DON) concept was introduced mainly to enable efficient multi-picture slice interleaving (see section 10.6) and robust packet scheduling (see section 10.7). In both of these applications NAL units are transmitted out of decoding order. DON indicates the decoding order of NAL units and should be used in the receiver to recover the decoding order. Example use cases for efficient multi-picture slice interleaving and for robust packet scheduling are given in sections 11.2 and 11.3 respectively. Section 11.4 describes the benefits of the DON concept in error resiliency achieved by redundant coded pictures. Section 11.5 summarizes considered alternatives to DON and justifies why DON was chosen to this RTP payload specification.

### 11.2. Example of Multi-Picture Slice Interleaving

An example of multi-picture slice interleaving follows. A subset of a coded video sequence is depicted below in output order. R  
Wenger et. al. Expires December 2003 [Page 56]

denotes a reference picture, N denotes a non-reference picture, and the number indicates a relative output time.

... R1 N2 R3 N4 R5 ...

The decoding order of these pictures is from left to right as follows:

... R1 R3 N2 R5 N4 ...

The NAL units of pictures R1, R3, N2, R5, and N4 are marked with a DON equal to 1, 2, 3, 4, and 5, respectively.

Each reference picture consists of three slice groups that are scattered as follows (a number denotes the slice group number for each macroblock in a QCIF frame):

```

0 1 2 0 1 2 0 1 2 0 1
2 0 1 2 0 1 2 0 1 2 0
1 2 0 1 2 0 1 2 0 1 2
0 1 2 0 1 2 0 1 2 0 1
2 0 1 2 0 1 2 0 1 2 0
1 2 0 1 2 0 1 2 0 1 2
0 1 2 0 1 2 0 1 2 0 1
2 0 1 2 0 1 2 0 1 2 0
1 2 0 1 2 0 1 2 0 1 2

```

For the sake of simplicity, we assume that all the macroblocks of a slice group are included in one slice. Three MTAPs are constructed from three consecutive reference pictures so that each MTAP contains three aggregation units, each of which contains all the macroblocks from one slice group. The first MTAP contains slice group 0 of picture R1, slice group 1 of picture R2, and slice group 2 of picture R3. The second MTAP contains slice group 1 of picture R1, slice group 2 of picture R2, and slice group 0 of picture R3. The third MTAP contains slice group 2 of picture R1, slice group 0 of picture R2, and slice group 1 of picture R3. Each non-reference picture is encapsulated into an STAP-B.

Consequently, the transmission order of NAL units is the following:

R1, slice group 0, DON 1, carried in MTAP, RTP SN: N  
Wenger et. al. Expires December 2003 [Page 57]

R3, slice group 1, DON 2, carried in MTAP,	RTP SN: N
R5, slice group 2, DON 4, carried in MTAP,	RTP SN: N
R1, slice group 1, DON 1, carried in MTAP,	RTP SN: N+1
R3, slice group 2, DON 2, carried in MTAP,	RTP SN: N+1
R5, slice group 0, DON 4, carried in MTAP,	RTP SN: N+1
R1, slice group 2, DON 1, carried in MTAP,	RTP SN: N+2
R3, slice group 1, DON 2, carried in MTAP,	RTP SN: N+2
R5, slice group 0, DON 4, carried in MTAP,	RTP SN: N+2
N2, DON 3, carried in STAP-B,	RTP SN: N+3
N4, DON 5, carried in STAP-B,	RTP SN: N+4

The receiver is able to organize the NAL units back in decoding order based on the value of DON associated with each NAL unit.

If one the MTAPs is lost, the spatially adjacent and temporally co-located macroblocks are received and can be used to conceal the loss efficiently. If one of the STAPs is lost, the effect of the loss does not propagate temporally.

### 11.3. Example of Robust Packet Scheduling

An example of robust packet scheduling follows. The communication system used in the example consists of the following components in the order that the video is processed from source to sink:

- o camera and capturing
- o pre-encoding buffer
- o encoder
- o encoded picture buffer
- o transmitter
- o transmission channel
- o receiver
- o receiver buffer
- o decoder
- o decoded picture buffer
- o display

The video communication system used in the example operates as follows. Note that processing of the video stream happens gradually and at the same time in all components of the system.

The source video sequence is shot and captured to a pre-encoding buffer. The pre-encoding buffer can be used to order pictures from sampling order to encoding order or to analyze multiple uncompressed frames for bitrate rate control purposes, for example. In some cases the pre-encoding buffer may not exist, but rather the sampled pictures are encoded right away. The encoder encodes pictures from the pre-encoding buffer and stores the output, i.e., coded pictures, to the encoded picture buffer. The transmitter encapsulates the coded pictures from the encoded picture buffer to transmission packets and sends them to a receiver through a transmission channel. The receiver stores the received packets to the receiver buffer. The receiver buffering process typically includes buffering for transmission delay jitter. The receiver buffer can also be used to recover correct decoding order of coded data. The decoder reads coded data from the receiver buffer and produces decoded pictures as output into the decoded picture buffer. The decoded picture buffer is used to recover the output (or display) order of pictures. Finally, pictures are displayed.

In the following example figures, I denotes an IDR picture, R denotes a reference picture, N denotes a non-reference picture, and the number after I, R, or N indicates a relative sampling time proportional to the previous IDR picture in decoding order. Values below the sequence of pictures indicate scaled system clock timestamps. The system clock is initialized arbitrarily in this example, and time runs from left to right. Each I, R, and N picture is mapped into the same timeline compared to the previous processing step, if any, assuming that encoding, transmission, and decoding take no time. Thus, events happening at the same time are located in the same column throughout all example figures.

A subset of a sequence of coded pictures is depicted below in sampling order.

```

... N58 N59 I00 N01 N02 R03 N04 N05 R06 ... N58 N59 I00 N01 ...
... --|---|---|---|---|---|---|---|---|  ... -|---|---|---|  ...
... 58 59 60 61 62 63 64 65 66 ... 128 129 130 131 ...

```

The sampled pictures are buffered in the pre-encoding buffer to arrange them in encoding order. In this example, we assume that the non-reference pictures are predicted from both the previous and the next reference picture in output order. Thus, the pre-encoding buffer has to contain at least two pictures and the buffering causes a delay of two picture intervals. The output of the pre-encoding buffering process and the encoding (and decoding) order of the pictures are as follows:

```

... N58 N59 I00 R03 N01 N02 R06 N04 N05 ...
... -|---|---|---|---|---|---|---|---|---|- ...
... 60 61 62 63 64 65 66 67 68 ...

```

The encoder or the transmitter can set the value of DON for each picture to a value of DON for the previous picture in decoding order plus one.

For the sake of simplicity, let us assume that:

- o the frame rate of the sequence is constant,
- o each picture consists of only one slice,
- o each slice is encapsulated in a single NAL unit packet,
- o pictures are transmitted in decoding order, and
- o pictures are transmitted at constant intervals (that is equal to 1 / frame rate).

Thus, pictures are received in decoding order:

```

... N58 N59 I00 R03 N01 N02 R06 N04 N05 ...
... -|---|---|---|---|---|---|---|---|---|- ...
... 60 61 62 63 64 65 66 67 68 ...

```

The optional interleaving-depth MIME type parameter is set to 0, because the transmission (or reception) order is identical to the decoding order.

The decoder has to buffer for one picture interval initially in its decoded picture buffer to organize pictures from decoding order to output order as depicted below:

```

... N58 N59 I00 N01 N02 R03 N04 N05 R06 ...

```

```

... -|---|---|---|---|---|---|---|---|---|
... 61 62 63 64 65 66 67 68 69 ...

```

The amount of required initial buffering in the decoded picture buffer can be signaled in the buffering period SEI message or with the `num_reorder_frames` syntax element of H.264 video usability information. `num_reorder_frames` indicates the maximum number of frames, complementary field pairs, or non-paired fields that precede any frame, complementary field pair, or non-paired field in the sequence in decoding order and follow it in output order. For the sake of simplicity, we assume that `num_reorder_frames` is used to indicate the initial buffer in the decoded picture buffer. In this example, `num_reorder_frames` is equal to 1.

It can be observed that if the IDR picture I00 is lost during transmission and a retransmission request is issued when the value of the system clock is 62, there is one picture interval of time (until the system clock reaches timestamp 63) to receive the retransmitted IDR picture I00.

Let us then assume that IDR pictures are transmitted two frame intervals earlier than their decoding position, i.e., the pictures are transmitted as follows:

```

... I00 N58 N59 R03 N01 N02 R06 N04 N05 ...
... --|---|---|---|---|---|---|---|---|---|
... 62 63 64 65 66 67 68 69 70 ...

```

The optional interleaving-depth MIME type parameter is set equal to 1 according to its definition. (The value of interleaving-depth in this example can be derived as follows: Picture I00 is the only picture preceding picture N58 or N59 in transmission order and following it in decoding order. Except for pictures I00, N58, and N59, the transmission order is the same as the decoding order of pictures. Since a coded picture is encapsulated into exactly one NAL unit, the value of interleaving-depth is equal to the maximum number of pictures preceding any picture in transmission order and following the picture in decoding order.)

The receiver buffering process contains two pictures at a time according to the value of the interleaving-depth parameter and orders pictures from the reception order to the correct decoding order based on the value of DON associated with each picture. The output of the receiver buffering process is the following:

```

... N58 N59 I00 R03 N01 N02 R06 N04 N05 ...
... -|---|---|---|---|---|---|---|---|---|
... 63 64 65 66 67 68 69 70 71 ...

```

Again, an initial buffering delay of one picture interval is needed to organize pictures from decoding order to output order as depicted below:

```

... N58 N59 I00 N01 N02 R03 N04 N05 ...
... -|---|---|---|---|---|---|---|---|---|
... 64 65 66 67 68 69 70 71 ...

```

It can be observed that the maximum delay that IDR pictures can undergo during transmission, including possible application, transport, or link layer retransmission, is equal to three picture intervals. Thus, the loss resiliency of IDR pictures is improved in systems supporting retransmission compared to the case in which pictures were transmitted in their decoding order.

#### 11.4. Robust Transmission Scheduling of Redundant Coded Slices

A redundant coded picture is a coded representation of a picture or a part of a picture that is not used in the decoding process if the corresponding primary coded picture is correctly decoded. There should be no noticeable difference between any area of the decoded primary picture and a corresponding area that would result from application of the H.264 decoding process for any redundant picture in the same access unit. A redundant coded slice is a coded slice that is a part of a redundant coded picture.

Redundant coded pictures can be used to provide unequal error protection in error-prone video transmission. If a primary coded representation of a picture is decoded incorrectly, a corresponding

Wenger et. al. Expires December 2003 [Page 62]

redundant coded picture can be decoded. Examples of applications and coding techniques utilizing the redundant codec picture feature include the video redundancy coding [26] and protection of "key pictures" in multicast streaming [27].

One property of many error-prone video communications systems is that transmission errors are often bursty and therefore they may affect more than one consecutive transmission packets in transmission order. In low bitrate video communication it is relatively common that an entire coded picture can be encapsulated into one transmission packet. Consequently, a primary coded picture and the corresponding redundant coded pictures may be transmitted in consecutive packets in transmission order. In order to make the transmission scheme more tolerant of bursty transmission errors, it is beneficial to transmit a primary coded picture apart from the corresponding redundant coded pictures. The DON concept enables this.

#### 11.5. Remarks on Other Design Possibilities

The slice header syntax structure of the H.264 coding standard contains the `frame_num` syntax element that can indicate the decoding order of coded frames. However, the usage of the `frame_num` syntax element is not feasible or desirable to recover the decoding order due to the following reasons:

- o The receiver is required to parse at least one slice header per coded picture (before passing the coded data to the decoder).
- o Coded slices from multiple coded video sequences cannot be interleaved, because the frame number syntax element is reset to 0 in each IDR picture.
- o The coded fields of a complementary field pair share the same value of the `frame_num` syntax element. Thus, the decoding order of the coded fields of a complementary field pair cannot be recovered based on the `frame_num` syntax element or any other syntax element of the H.264 coding syntax.

The RTP payload format for transport of MPEG-4 elementary streams [28] enables interleaving of access units and transmission of multiple access units in the same RTP packet. An access unit is  
Wenger et. al. Expires December 2003 [Page 63]

specified in the H.264 coding standard to consist of all NAL units that are associated with a primary coded picture according to subclause 7.4.1.2 of [1]. Consequently, slices of different pictures cannot be interleaved and the multi-picture slice interleaving technique (see section 10.6) for improved error resilience cannot be used.

## 12. Open Issues

- o Security section needs review.
- o Is max-don-diff necessary?
- o ITU-T H.241 provides a way for decoders to signal capability for greater processing speed or memory amount than required in the profile and level that is used. H.241 specifies CustomMaxMBPS, CustomMaxFS, CustomMaxDPB, and CustomMaxBrandCPB. Should similar parameters be specified as optional MIME/SDP parameters to enhance the capability exchange procedure of SIP-based video conferencing?

## 13. Full Copyright Statement

Copyright (C) The Internet Society (2003). All Rights Reserved.

This document and translations of it may be copied and furnished to others, and derivative works that comment on or otherwise explain it or assist in its implementation may be prepared, copied, published and distributed, in whole or in part, without restriction of any kind, provided that the above copyright notice and this paragraph are included on all such copies and derivative works.

However, this document itself may not be modified in any way, such as by removing the copyright notice or references to the Internet Society or other Internet organizations, except as needed for the purpose of developing Internet standards in which case the procedures for copyrights defined in the Internet Standards process must be followed, or as required to translate it into languages other than English.

The limited permissions granted above are perpetual and will not be revoked by the Internet Society or its successors or assigns.

This document and the information contained herein is provided on an "AS IS" basis and THE INTERNET SOCIETY AND THE INTERNET ENGINEERING TASK FORCE DISCLAIMS ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

#### 14. Intellectual Property Notice

The IETF takes no position regarding the validity or scope of any intellectual property or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; neither does it represent that it has made any effort to identify any such rights. Information on the IETF's procedures with respect to rights in standards-track and standards-related documentation can be found in BCP-11. Copies of claims of rights made available for publication and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementors or users of this specification can be obtained from the IETF Secretariat.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights which may cover technology that may be required to practice this standard. Please address the information to the IETF Executive Director.

The IETF has been notified of intellectual property rights claimed in regard to some or all of the specification contained in this document. For more information consult the online list of claimed rights at <http://www.ietf.org/ipr>.

15.1. Normative References

- [1] ITU-T Recommendation H.264, "Advanced video coding for generic audiovisual services", May 2003.
- [2] ISO/IEC International Standard 14496-10:2003.
- [3] S. Bradner, "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [4] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", RFC 3550, July 2003.
- [5] M. Handley and V. Jacobson, "SDP: Session Description Protocol", RFC 2327, April 1998.
- [6] S. Josefsson, "The Base16, Base32, and Base64 Data Encodings", RFC 3548, July 2003.
- [7] ITU-T Recommendation T.35, "Procedure for the allocation of ITU-T defined codes for non-standard facilities", February 2000.

15.2. Informative References

- [8] "Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264 | ISO/IEC 14496-10 AVC)", available from [ftp://ftp.imtc-files.org/jvt-experts/2003\\_03\\_Pattaya/JVT-G50r1.zip](ftp://ftp.imtc-files.org/jvt-experts/2003_03_Pattaya/JVT-G50r1.zip), May 2003.
- [9] A. Luthra, G.J. Sullivan, and T. Wiegand (eds.), Special Issue on H.264/AVC. IEEE Transactions on Circuits and Systems on Video Technology, July 2003.
- [10] P. Borgwardt, "Handling Interlaced Video in H.26L", VCEG-N57r2, available from [ftp://standard.pictel.com/video-site/0109\\_San/VCEG-N57r2.doc](ftp://standard.pictel.com/video-site/0109_San/VCEG-N57r2.doc), September 2001.
- [11] C. Borman et. Al., "RTP Payload Format for the 1998 Version of ITU-T Rec. H.263 Video (H.263+)", RFC 2429, October 1998.
- [12] ISO/IEC IS 14496-2.
- [13] S. Wenger, "H.26L over IP", IEEE Transaction on Circuits and Systems for Video technology, July 2003.
- [14] S. Wenger, "H.26L over IP: The IP Network Adaptation Layer", Proceedings Packet Video Workshop 02, April 2002

- [15] T. Stockhammer, M.M. Hannuksela, and S. Wenger, "H.26L/JVT Coding Network Abstraction Layer and IP-based Transport" in Proc. ICIP 2002, Rochester, NY, September 2002.
- [16] ITU-T Recommendation H.241, "Extended video procedures and control signals for H.300 series terminals", July 2003.
- [17] H. Schulzrinne and S. Casner, "RTP Profile for Audio and Video Conferences with Minimal Control", RFC 3551, July 2003.
- [18] B. Cain, S. Deering, I. Kouvelas, B. Fenner, and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, October 2002.
- [19] ITU-T Recommendation H.223, "Multiplexing protocol for low bit rate multimedia communication", July 2001.
- [20] C. Perkins et. al., "RTP Payload for Redundant Audio Data", RFC 2198, September 1997.
- [21] J. Rosenberg, H. Schulzrinne, "An RTP Payload Format for Generic Forward Error Correction", RFC 2733, December 1999.
- [22] T. Stockhammer, T. Wiegand, T. Oelbaum, and F. Obermeier, "Video Coding and Transport Layer Techniques for H.264/AVC-Based Transmission over Packet-Lossy Networks", IEEE International Conference on Image Processing (ICIP 2003), Barcelona, Spain, September 2003.
- [23] V. Varsa, M. Karczewicz, "Slice interleaving in compressed video packetization", Packet Video Workshop 2000.
- [24] S.H. Kang and A. Zakhor, "Packet scheduling algorithm for wireless video streaming," International Packet Video Workshop 2002, available <http://www.pv2002.org>.
- [25] M.M. Hannuksela, "Enhanced concept of GOP", JVT-B042, available [ftp://standard.pictel.com/video-site/0201\\_Gen/JVT-B042.doc](ftp://standard.pictel.com/video-site/0201_Gen/JVT-B042.doc), January 2002.
- [26] S. Wenger, "Video Redundancy Coding in H.263+", 1997 International Workshop on Audio-Visual Services over Packet Networks, September 1997.
- [27] Y.-K. Wang, M.M. Hannuksela, and M. Gabbouj, "Error Resilient Video Coding Using Unequally Protected Key Pictures", in Proc. International Workshop VLBV03, September 2003.
- [28] J. van der Meer, D. Mackie, V. Swaminathan, D. Singer, and P. Gentric, "RTP Payload Format for Transport of MPEG-4 Elementary Streams", draft-ietf-avt-mpeg4-simple-08.txt, August 2003.

Author's Addresses

Stephan Wenger  
TU Berlin / Teles AG  
Franklinstr. 28-29  
D-10587 Berlin  
Germany

Phone: +49-172-300-0813  
Email: stewe@cs.tu-berlin.de

Miska M. Hannuksela  
Nokia Corporation  
P.O. Box 100  
33721 Tampere  
Finland

Phone: +358-7180-73151  
Email: miska.hannuksela@nokia.com

Thomas Stockhammer  
Institute for Communications Eng.  
Munich University of Technology  
D-80290 Munich  
Germany

Phone: +49-89-28923474  
Email: stockhammer@ei.tum.de

Magnus Westerlund  
Multimedia Technologies  
Ericsson Research EAB/TVA/A  
Ericsson AB  
Torshamsgatan 23  
S-164 80 Stockholm  
Sweden

Phone: +46-8-4048287  
Email:  
magnus.westerlund@ericsson.com

David Singer  
QuickTime Engineering  
Apple  
1 Infinite Loop MS 302-3MT  
Cupertino  
CA 95014  
USA

Phone +1 408 974-3162  
Email: singer@apple.com

Annex A: Changes relative to draft-ietf-avt-rtp-h264-02.txt

[This section will be removed in a future version of this draft.]

This memo contains the following technical changes relative to the previous I-D:

- o Assignment of DON values for NAL units in an STAP-B and decoding order of NAL units in an STAP-B corrected and clarified. De-packetization process changed accordingly.
- o Derivation of DON for MTAPs changed to allow wraparound of DON values within one MTAP.
- o The use of RTP timestamp and picture timing SEI message is clarified.
- o Single NAL unit packetization mode introduced for compatibility with ITU-T Recommendation H.241.
- o Packetization modes simplified. Single-picture and multi-picture mode changed to non-interleaved and interleaved modes. Packets including DON cannot be mixed with packets not including DON anymore, and therefore the derivation of the decoding order becomes easier and more tolerant to transmission delay jitter.
- o The optional packetization-mode MIME parameter introduced to reflect the new packetization modes. The previous parameter for selecting the packetization mode, i.e., mtap-allowed, was deleted.
- o Created two types of fragmentation units, FU-A (not including DON) and FU-B (including DON).
- o Base64 encoding used in the optional parameter-sets MIME parameter instead of hexadecimal encoding to improve compression efficiency.
- o Added an informative note clarifying why values of DON in consecutive NAL units in decoding order are not required to be incremented by one.

o Section 1.3 ("Network Abstraction Layer Unit Types") appeared in the introduction section but actually specified such semantics of Wenger et. al. Expires December 2003 [Page 69]

F and NRI that are specific only to the RTP payload format. These semantics are now specified in section 5.3.

- o A third option, max-don-diff, was added as an option to control the receiver buffering in the interleaved packetization mode. max-don-diff is specified similarly to the maximum displacement parameter in the draft-ietf-avt-mpeg4-simple Internet Draft, but instead of using a maximum difference in terms of RTP timestamps a maximum difference in terms of decoding order numbers is used. This design decision was made due to the following facts:
  - 1) RTP timestamp indicates the capture/display timestamp.
  - 2) H.264/AVC allows decoding order different from output order.
  - 3) The receiver buffer is used to reorder packets from transmission/reception order to decoding order.
  - 4) Thus, displacement specified between differences in RTP timestamps cannot be used to reception-to-decoding-order reorganization.
  
- o Editorial changes and new informative notes.

