Network Working Group
Internet Draft
Intended status: Standards track
Expires: December 2013

T. Schierl
Fraunhofer HHI
S. Wenger
Vidyo
Y.-K. Wang
Qualcomm
M. M. Hannuksela
Nokia
Y. Sanchez
Fraunhofer HHI
June 11, 2013

RTP Payload Format for High Efficiency Video Coding
draft-schierl-payload-rtp-h265-03.txt

Status of this Memo

This Internet-Draft will expire on December 11, 2013.

Copyright and License Notice

Abstract

This memo describes an RTP payload format for the video coding standard ITU-T Recommendation H.265 and ISO/IEC International Standard 23008-2, both also known as High Efficiency Video Coding (HEVC) [HEVC], developed by the Joint Collaborative Team on Video Coding (JCT-VC). The RTP payload format allows for packetization of one or more Network Abstraction Layer (NAL) units in each RTP packet payload, as well as fragmentation of a NAL unit into multiple RTP packets. Furthermore, it supports transmission of an HEVC stream over a single as well as multiple RTP flows. The payload format has wide applicability in videoconferencing, Internet video streaming, and high bit-rate entertainment-quality video, among others.

Table of Contents

1. Introduction

1.1. Overview of the HEVC Codec

High Efficiency Video Coding [HEVC], formally known as ITU-T Recommendation H.265 and ISO/IEC International Standard 23008-2 was ratified by ITU-T in April 2013 and reportedly provides significant coding efficiency gains over H.264 [H.264].

As both H.264 [H.264] and its RTP payload format [RFC6184] are widely deployed and generally known in the relevant implementer community, frequently only the differences between those two specifications are highlighted in non-normative, explanatory parts of this memo. Basic familiarity with both specifications is assumed for those parts. However, the normative parts of this memo do not require study of H.264 or its RTP payload format.

H.264 and HEVC share a similar hybrid video codec design. Conceptually, both technologies include a video coding layer (VCL), which is often used to refer to the coding-tool features, and a network abstraction layer (NAL), which is often used to refer to the systems and transport interface aspects of the codecs.

1.1.1 Coding-Tool Features

Similarly to earlier hybrid-video-coding-based standards, including H.264, the following basic video coding design is employed by HEVC. A prediction signal is first formed either by intra or motion compensated prediction, and the residual (the difference between the original and the prediction) is then coded. The gains in coding efficiency are achieved by redesigning and improving almost all parts of the codec over earlier designs. In addition, HEVC includes several tools to make the implementation on parallel architectures easier. Below is a summary of HEVC coding-tool features.

Quad-tree block and transform structure

One of the major tools that contribute significantly to the coding efficiency of HEVC is the usage of flexible coding blocks and transforms, which are defined in a hierarchical quad-tree manner. Unlike H.264, where the basic coding block is a macroblock of fixed

size 16x16, HEVC defines a Coding Tree Unit (CTU) of a maximum size of 64x64. Each CTU can be divided into smaller units in a hierarchical quad-tree manner and can represent smaller blocks down to size 4x4. Similarly, the transforms used in HEVC can have different sizes, starting from 4x4 and going up to 32x32. Utilizing large blocks and transforms contribute to the major gain of HEVC, especially at high resolutions.

Entropy coding

HEVC uses a single entropy coding engine, which is based on Context Adaptive Binary Arithmetic Coding (CABAC), whereas H.264 uses two distinct entropy coding engines. CABAC in HEVC shares many similarities with CABAC of H.264, but contains several improvements. Those include improvements in coding efficiency and lowered implementation complexity, especially for parallel architectures.

In-loop filtering

H.264 includes an in-loop adaptive deblocking filter, where the blocking artifacts around the transform edges in the reconstructed picture are smoothed to improve the picture quality and compression efficiency. In HEVC, a similar deblocking filter is employed but with somewhat lower complexity. In addition, pictures undergo a subsequent filtering operation called Sample Adaptive Offset (SAO), which is a new design element in HEVC. SAO basically adds a pixel-level offset in an adaptive manner and usually acts as a de-ringing filter. It is observed that SAO improves the picture quality, especially around sharp edges contributing substantially to visual quality improvements of HEVC.

Motion prediction and coding

There have been a number of improvements in this area that are summarized as follows. The first category is motion merge and advanced motion vector prediction (AMVP) modes. The motion information of a prediction block can be inferred from the spatially or temporally neighboring blocks. This is similar to the DIRECT mode in H.264 but includes new aspects to incorporate the flexible quad-tree structure and methods to improve the parallel implementations. In addition, the motion vector predictor can be

signaled for improved efficiency. The second category is high-precision interpolation. The interpolation filter length is increased to 8-tap from 6-tap, which improves the coding efficiency but also comes with increased complexity. In addition, interpolation filter is defined with higher precision without any intermediate rounding operations to further improve the coding efficiency.

Intra prediction and intra coding

Compared to 8 intra prediction modes in H.264, HEVC supports angular intra prediction with 33 directions. This increased flexibility improves both objective coding efficiency and visual quality as the edges can be better predicted and ringing artifacts around the edges can be reduced. In addition, the reference samples are adaptively smoothed based on the prediction direction. To avoid contouring artifacts a new interpolative prediction generation is included to improve the visual quality. Furthermore, discrete sine transform (DST) is utilized instead of traditional discrete cosine transform (DCT) for 4x4 intra transform blocks.

Other coding-tool features

HEVC includes some tools for lossless coding and efficient screen content coding, such as skipping the transform coding for certain blocks. These tools are particularly useful for example when streaming the user-interface of a mobile device to a large display.

1.1.2 Systems and Transport Interfaces

HEVC inherited the basic systems and transport interfaces designs, such as the NAL-unit-based syntax structure, the hierarchical syntax and data unit structure from sequence-level parameter sets, multi-picture-level or picture-level parameter sets, slice-level header parameters, lower-level parameters, the supplemental enhancement information (SEI) message mechanism, the hypothetical reference decoder (HRD) based video buffering model, and so on. In the following, a list of differences in these aspects compared to H.264 is summarized.

Video parameter set

A new type of parameter set, called video parameter set (VPS), was introduced. For the first (2013) version of [HEVC], the video parameter set NAL unit is required to be available prior to its activation, while the information contained in the video parameter set is not necessary for operation of the decoding process. For future HEVC extensions, such as the 3D or scalable extensions, the video parameter set is expected to include information necessary for operation of the decoding process, e.g. decoding dependency or information for reference picture set construction of enhancement layers. The VPS provides a "big picture" of a bitstream, including what types of operation points are provided, the profile, tier, and level of the operation points, and some other high-level properties of the bitstream that can be used as the basis for session negotiation and content selection, etc. (see section 7.1).

Profile, tier and level

The profile, tier and level syntax structure that can be included in both VPS and sequence parameter set (SPS) includes 12 bytes data to describe the entire bitstream (including all temporally scalable layers, which are referred to as sub-layers in the HEVC specification), and can optionally include more profile, tier and level information pertaining to individual temporally scalable layers. The profile indicator indicates the "best viewed as" profile when the bitstream conforms to multiple profiles, similar to the major brand concept in the ISO base media file format (ISOBMFF) [ISOBMFF] and file formats derived based on ISOBMFF, such as the 3GPP file format [3GP]. The profile, tier and level syntax structure also includes the indications of whether the bitstream is free of frame-packed content, whether the bitstream is free of interlaced source content and free of field pictures, i.e., contains only frame pictures of progressive source, such that clients/players with no support of post-processing functionalities for handling of frame-packed or interlaced source content or field pictures can reject those bitstreams.

Bitstream and elementary stream

HEVC includes a definition of an elementary stream, which is new compared to H.264. An elementary stream consists of a sequence of one or more bitstreams. An elementary stream that consists of two or more bitstreams has typically been formed by splicing together two or more bitstreams (or parts thereof). When an elementary stream contains more than one bitstream, the last NAL unit of the last access unit of a bitstream (except the last bitstream in the elementary stream) must contain an end of bitstream NAL unit and the first access unit of the subsequent bitstream must be an intra random access point (IRAP) access unit. This IRAP access unit may be a clean random access (CRA), broken link access (BLA), or instantaneous decoding refresh (IDR) access unit.

Random access support

HEVC includes signaling in NAL unit header, through NAL unit types, of IRAP pictures beyond IDR pictures. Three types of IRAP pictures, namely IDR, CRA and BLA pictures are supported, wherein IDR pictures are conventionally referred to as closed group-of-pictures (closed-GOP) random access points, and CRA and BLA pictures are those conventionally referred to as open-GOP random access points. BLA pictures usually originate from splicing of two bitstreams or part thereof at a CRA picture, e.g. during stream switching. To enable better systems usage of IRAP pictures, altogether six different NAL units are defined to signal the properties of the IRAP pictures, which can be used to better match the stream access point (SAP) types as defined in the ISOBMFF [ISOBMFF], which are utilized for random access support in both 3GP-DASH [3GPDASH] and MPEG DASH [MPEGDASH]. Pictures following an IRAP picture in decoding order and preceding the IRAP picture in output order are referred to as leading pictures associated with the IRAP picture. There are two types of leading pictures, namely random access decodable leading (RADL) pictures and random access skipped leading (RASL) pictures. RADL pictures are decodable when the decoding started at the associated IRAP picture, and RASL pictures are not decodable when the decoding started at the associated IRAP picture and are usually discarded. HEVC provides mechanisms to enable the specification of conformance of bitstreams with RASL pictures being discarded, thus

to provide a standard-compliant way to enable systems components to discard RASL pictures when needed.

Temporal scalability support

HEVC includes an improved support of temporal scalability, by inclusion of the signaling of TemporalId in the NAL unit header, the restriction that pictures of a particular temporal sub-layer cannot be used for inter prediction reference by pictures of a higher temporal sub-layer, the sub-bitstream extraction process, and the requirement that each sub-bitstream extraction output be a conforming bitstream. Media-aware network elements (MANEs) can utilize the TemporalId in the NAL unit header for stream adaptation purposes based on temporal scalability.

Temporal sub-layer switching support

HEVC specifies, through NAL unit types present in the NAL unit header, the signaling of temporal sub-layer access (TSA) and stepwise temporal sub-layer access (STSA). A TSA picture and pictures following the TSA picture in decoding order do not use pictures prior to the TSA picture in decoding order with TemporalId greater than or equal to that of the TSA picture for inter prediction reference. A TSA picture enables up-switching, at the TSA picture, to the sub-layer containing the TSA picture or any higher sub-layer, from the immediately lower sub-layer. An STSA picture does not use pictures with the same TemporalId as the STSA picture for inter prediction reference. Pictures following an STSA picture in decoding order with the same TemporalId as the STSA picture do not use pictures prior to the STSA picture in decoding order with the same TemporalId as the STSA picture for inter prediction reference. An STSA picture enables up-switching, at the STSA picture, to the sub-layer containing the STSA picture, from the immediately lower sub-layer.

Sub-layer reference or non-reference pictures

The concept and signaling of reference/non-reference pictures in HEVC are different from H.264. In H.264, if a picture may be used by any other picture for inter prediction reference, it is a reference picture; otherwise it is a non-reference picture, and this

is signaled by two bits in the NAL unit header. In HEVC, a picture is called a reference picture only when it is marked as "used for reference". In addition, the concept of sub-layer reference picture was introduced. If a picture may be used by another other picture with the same TemporalId for inter prediction reference, it is a sub-layer reference picture; otherwise it is a sub-layer non-reference picture. Whether a picture is a sub-layer reference picture or sub-layer non-reference picture is signaled through NAL unit type values.

Extensibility

Besides the TemporalId in the NAL unit header, HEVC also includes the signaling of a six-bit layer ID in the NAL unit header, which must be equal to 0 for a single-layer bitstream. Extension mechanisms have been included in VPS, SPS, PPS, SEI NAL unit, slice headers, and so on. All these extension mechanisms enable future extensions in a backward compatible manner, such that bitstreams encoded according to potential future HEVC extensions can be fed to then-legacy decoders (e.g. HEVC version 1 decoders) and the then-legacy decoders can decode and output the base layer bitstream.

Bitstream extraction

HEVC includes a bitstream extraction process as an integral part of the overall decoding process, as well as specification of the use of the bitstream extraction process in description of bitstream conformance tests as part of the hypothetical reference decoder (HRD) specification.

Reference picture management

The reference picture management of HEVC, including reference picture marking and removal from the decoded picture buffer (DPB) as well as reference picture list construction (RPLC), differs from that of H.264. Instead of the sliding window plus adaptive memory management control operation (MMCO) based reference picture marking mechanism in H.264, HEVC specifies a reference picture set (RPS) based reference picture management and marking mechanism, and the RPLC is consequently based on the RPS mechanism. A reference picture set consists of a set of reference pictures associated with

a picture, consisting of all reference pictures that are prior to the associated picture in decoding order, that may be used for inter prediction of the associated picture or any picture following the associated picture in decoding order. The reference picture set consists of five lists of reference pictures; RefPicSetStCurrBefore, RefPicSetStCurrAfter, RefPicSetStFoll, RefPicSetLtCurr and RefPicSetLtFoll. RefPicSetStCurrBefore, RefPicSetStCurrAfter and RefPicSetLtCurr contains all reference pictures that may be used in inter prediction of the current picture and that may be used in inter prediction of one or more of the pictures following the current picture in decoding order. RefPicSetStFoll and RefPicSetLtFoll consists of all reference pictures that are not used in inter prediction of the current picture but may be used in inter prediction of one or more of the pictures following the current picture in decoding order. RPS provides an "intra-coded" signaling of the DPB status, instead of an "inter-coded" signaling, mainly for improved error resilience. The RPLC process in HEVC is based on the RPS, by signaling an index to an RPS subset for each reference index. The RPLC process has been simplified compared to that in H.264, by removal of the reference picture list modification (also referred to as reference picture list reordering) process.

Ultra low delay support

HEVC specifies a sub-picture-level HRD operation, for support of the so-called ultra-low delay. The mechanism specifies a standard-compliant way to enable delay reduction below one picture interval. Sub-picture-level coded picture buffer (CPB) and DPB parameters may be signaled, and utilization of these information for the derivation of CPB timing (wherein the CPB removal time corresponds to decoding time) and DPB output timing (display time) is specified. Decoders are allowed to operate the HRD at the conventional access-unit-level, even when the sub-picture-level HRD parameters are present.

New SEI messages

HEVC inherits many H.264 SEI messages with changes in syntax and/or semantics making them applicable to HEVC. The active parameter sets SEI message includes the IDs of the active video parameter set and the active sequence parameter set and can be used to activate VPSs and SPSs. In addition, the SEI message includes the following

indications: 1) An indication of whether "full random accessibility" is supported (when supported, all parameter sets needed for decoding of the remaining of the bitstream when random accessing from the beginning of the current coded video sequence by completely discarding all access units earlier in decoding order are present in the remaining bitstream and all coded pictures in the remaining bitstream can be correctly decoded); 2) An indication of whether there is any parameter set within the current coded video sequence that updates another parameter set of the same type preceding in decoding order. An update of a parameter set refers to the use of the same parameter set ID but with some other parameters changed. If this property is true for all coded video sequences in the bitstream, then all parameter sets can be sent out-of-band before session start. The region refresh information SEI message can be used together with the recovery point SEI message (present in both H.264 and HEVC) for improved support of gradual decoding refresh (GDR). This supports random access from inter-coded pictures, wherein complete pictures can be correctly decoded or recovered after an indicated number of pictures in output/display order.

1.1.3 Parallel Processing Support

The reportedly significantly higher computational demand of HEVC over H.264 (especially with respect to encoders, where a complexity increase of a factor of ten has often been reported), in conjunction with the ever increasing video resolution (both spatially and temporally) required by the market, led to the adoption of VCL coding tools specifically targeted to allow for parallelization on the sub-picture level. That is, parallelization occurs, at the minimum, at the granularity of an integer number of CTUs. The targets for this type of high-level parallelization are multicore CPUs and DSPs as well as multiprocessor systems. In a system design, to be useful, these tools require signaling support, which is provided in Section 7 of this memo. This section provides a brief overview of the tools available in [HEVC].

Many of the tools incorporated in HEVC were designed keeping in mind the potential parallel implementations in multi-core/multi-processor architectures. Specifically, for parallelization, four picture partition strategies are available.

Slices are segments of the bitstream that can be reconstructed independently from other slices within the same picture (though there may still be interdependencies through loop filtering operations). Slices are the only tool that can be used for parallelization that is also available, in virtually identical form, in H.264. Slices based parallelization does not require much inter-processor or inter-core communication (except for inter-processor or inter-core data sharing for motion compensation when decoding a predictively coded picture, which is typically much heavier than inter-processor or inter-core data sharing due to in-picture prediction), as slices are designed to be independently decodable. However, for the same reason, slices can require some coding overhead. Further, slices (in contrast to some of the other tools mentioned below) also serve as the key mechanism for bitstream partitioning to match Maximum Transfer Unit (MTU) size requirements, due to the in-picture independence of slices and the fact that each regular slice is encapsulated in its own NAL unit. In many cases, the goal of parallelization and the goal of MTU size matching can place contradicting demands to the slice layout in a picture. The realization of this situation led to the development of the more advanced tools mentioned below. This payload format does not contain any specific mechanisms aiding parallelization through slices.

Dependent slice segments allow for fragmentation of a coded slice into fragments at CTU boundaries without breaking any in-picture prediction mechanism. They are complementary to the fragmentation mechanism described in this memo in that they need the cooperation of the encoder. As a dependent slice segment necessarily contains an integer number of CTUs, a decoder using multiple cores operating on CTUs can process a dependent slice segment without communicating parts of the slice segment's bitstream to other cores. Fragmentation, as specified in this memo, in contrast, does not guarantee that a fragment contains an integer number of CTUs.

In wavefront parallel processing (WPP), the picture is partitioned into rows of CTUs. Entropy decoding and prediction are allowed to use data from CTUs in other partitions. Parallel processing is possible through parallel decoding of CTU rows, where the start of the decoding of a row is delayed by two CTUs, so to ensure that data related to a CTU above and to the right of the subject CTU is

available before the subject CTU is being decoded. Using this staggered start (which appears like a wavefront when represented graphically), parallelization is possible with up to as many processors/cores as the picture contains CTU rows.

Because in-picture prediction between neighboring CTU rows within a picture is allowed, the required inter-processor/inter-core communication to enable in-picture prediction can be substantial. The WPP partitioning does not result in the creation of more NAL units compared to when it is not applied, thus WPP cannot be used for MTU size matching, though slices can be used in combination for that purpose.

Tiles define horizontal and vertical boundaries that partition a picture into tile columns and rows. The scan order of CTUs is changed to be local within a tile (in the order of a CTU raster scan of a tile), before decoding the top-left CTU of the next tile in the order of tile raster scan of a picture. Similar to slices, tiles break in-picture prediction dependencies (including entropy decoding dependencies). However, they do not need to be included into individual NAL units (same as WPP in this regard), hence tiles cannot be used for MTU size matching, though slices can be used in combination for that purpose. Each tile can be processed by one processor/core, and the inter-processor/inter-core communication required for in-picture prediction between processing units decoding neighboring tiles is limited to conveying the shared slice header in cases a slice is spanning more than one tile, and loop filtering related sharing of reconstructed samples and metadata. Insofar, tiles are less demanding in terms of inter-processor communication bandwidth compared to WPP due to the in-picture independence between two neighboring partitions.

1.1.4 NAL Unit Header

HEVC maintains the NAL unit concept of H.264 with modifications. HEVC uses a two-byte NAL unit header, as shown in Figure 1. The payload of a NAL unit refers to the NAL unit excluding the NAL unit header.

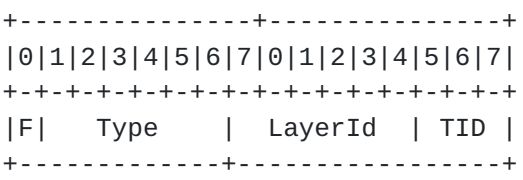

Figure 1 The structure of HEVC NAL unit header

The semantics of the fields in the NAL unit header are as specified in [HEVC] and described briefly below for convenience. In addition to the name and size of each field, the corresponding syntax element name in [HEVC] is also provided.

F: 1 bit
forbidden_zero_bit. MUST be zero. HEVC declares a value of 1 as a syntax violation. Note that the inclusion of this bit in the NAL unit header is to enable transport of HEVC video over MPEG-2 transport systems (avoidance of start code emulations) [MPEG2S].

Type: 6 bits
nal_unit_type. This field specifies the NAL unit type as defined in Table 7-1 of [HEVC]. For a reference of all currently defined NAL unit types and their semantics, please refer to Section 7.4.1 in [HEVC].

LayerId: 6 bits
nuh_layer_id. MUST be equal to zero. It is anticipated that in future scalable or 3D video coding extensions of this specification, this syntax element will be used to identify additional layers that may be present in the coded video sequence, wherein a layer may be, e.g. a spatial scalable layer, a quality scalable layer, a texture view, or a depth view.

TID: 3 bits
nuh_temporal_id_plus1. This field specifies the temporal identifier of the NAL unit plus 1. The value of TemporalId is equal to TID minus 1. A TID value of 0 is illegal to ensure that there is at least one bit in the NAL unit header equal to 1, so

to enable independent considerations of start code emulations in the NAL unit header and in the NAL unit payload data.

1.2. Overview of the Payload Format

This payload format defines the following processes required for transport of HEVC coded data over RTP [RFC3550]:

- Usage of RTP header with this payload format
- Packetization of HEVC coded NAL units into RTP packets using three types of payload structures, namely single NAL unit packet, aggregation packet, and fragment unit
- Transmission of HEVC NAL units of the same bitstream within a single RTP session or multiple RTP sessions
- Media type parameters to be used with the Session Description Protocol (SDP) [RFC4566]

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 BCP 14, RFC 2119 [RFC2119].

This specification uses the notion of setting and clearing a bit when bit fields are handled. 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. Definitions and Abbreviations

3.1 Definitions

This document uses the terms and definitions of [HEVC]. Section 3.1.1 lists relevant definitions copied from [HEVC] for convenience. Section 3.1.2 gives definitions specific to this memo.

3.1.1 Definitions from the HEVC Specification

access unit: A set of NAL units that are associated with each other according to a specified classification rule, are consecutive in decoding order, and contain exactly one coded picture.

BLA access unit: An access unit in which the coded picture is a BLA picture.

BLA picture: An IRAP picture for which each VCL NAL unit has nal_unit_type equal to BLA_W_LP, BLA_W_RADL, or BLA_N_LP.

coded video sequence: A sequence of access units that consists, in decoding order, of an IRAP access unit with NoRaslOutputFlag equal to 1, followed by zero or more access units that are not IRAP access units with NoRaslOutputFlag equal to 1, including all subsequent access units up to but not including any subsequent access unit that is an IRAP access unit with NoRaslOutputFlag equal to 1.

> Informative note: An IRAP access unit may be an IDR access unit, a BLA access unit, or a CRA access unit. The value of NoRaslOutputFlag is equal to 1 for each IDR access unit, each BLA access unit, and each CRA access unit that is the first access unit in the bitstream in decoding order, is the first access unit that follows an end of sequence NAL unit in decoding order, or has HandleCraAsBlaFlag equal to 1.

CRA access unit: An access unit in which the coded picture is a CRA picture.

CRA picture: A RAP picture for which each slice has nal_unit_type equal to CRA_NUT.

IDR access unit: An access unit in which the coded picture is an IDR picture.

IDR picture: A RAP picture for which each slice has nal_unit_type equal to IDR_W_RADL or IDR_N_LP.

IRAP access unit: An access unit in which the coded picture is an IRAP picture.

IRAP picture: A coded picture for which each VCL NAL unit has nal_unit_type in the range of BLA_W_LP to RSV_IRAP_VCL23, inclusive.

layer: A set of VCL NAL units that all have a particular value of nuh_layer_id and the associated non-VCL NAL units, or one of a set of syntactical structures having a hierarchical relationship.

operation point: bitstream created from another bitstream by operation of the sub-bitstream extraction process with the another bitstream, a target highest TemporalId, and a target layer identifier list as inputs.

random access: The act of starting the decoding process for a bitstream at a point other than the beginning of the stream.

sub-layer: A temporal scalable layer of a temporal scalable bitstream consisting of VCL NAL units with a particular value of the TemporalId variable, and the associated non-VCL NAL units.

tile: A rectangular region of coding tree blocks within a particular tile column and a particular tile row in a picture.

tile column: A rectangular region of coding tree blocks having a height equal to the height of the picture and a width specified by syntax elements in the picture parameter set.

tile row: A rectangular region of coding tree blocks having a height specified by syntax elements in the picture parameter set and a width equal to the width of the picture.

3.1.2 Definitions Specific to This Memo

media aware network element (MANE): A network element, such as a middlebox or application layer gateway that is capable of parsing certain aspects of the RTP payload headers or the RTP payload and reacting to their contents.

> Informative note: The concept of a MANE goes beyond normal routers or gateways in that a MANE has to be aware of the signaling (e.g., to learn about the payload type mappings of the media streams), and in that it has to be trusted when working

with SRTP. The advantage of using MANEs is that they allow packets to be dropped according to the needs of the media coding. For example, if a MANE has to drop packets due to congestion on a certain link, it can identify and remove those packets whose elimination produces the least adverse effect on the user experience. After dropping packets, MANEs must rewrite RTCP packets to match the changes to the RTP packet stream as specified in Section 7 of [RFC3550].

NAL unit decoding order: A NAL unit order that conforms to the constraints on NAL unit order given in Section 7.4.2.4 in [HEVC].

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

RTP packet stream: A sequence of RTP packets with increasing sequence numbers (except for wrap-around), identical PT and identical SSRC (Synchronization Source), carried in one RTP session. Within the scope of this memo, one RTP packet stream is utilized to transport one or more temporal sub-layers.

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.

base session: an RTP session in Multi-Session Transmission mode that transports a bitstream subset which the rest of RTP sessions in the Multi-Session Transmission depends on. [Ed. (YK): Check the need of this definition after the draft is more complete.]

3.2 Abbreviations

AP Aggregation Packet

BLA Broken Link Access

CRA Clean Random Access

CTB Coding Tree Block

CTU Coding Tree Unit

CVS	Coded Video Sequence
FU	Fragmentation Unit
GDR	Gradual Decoding Refresh
HRD	Hypothetical Reference Decoder
IDR	Instantaneous Decoding Refresh
IRAP	Intra Random Access Point
MANE	Media Aware Network Element
MST	Multi-Session Transmission
MTU	Maximum Transfer Unit
NAL	Network Abstraction Layer
NALU	Network Abstraction Layer Unit
PPS	Picture Parameter Set
RADL	Random Access Decodable Leading (Picture)
RASL	Random Access Skipped Leading (Picture)
RPS	Reference Picture Set
SEI	Supplemental Enhancement Information
SPS	Sequence Parameter Set
SST	Single-Session Transmission
STSA	Step-wise Temporal Sub-layer Access
TSA	Temporal Sub-layer Access
VCL	Video Coding Layer
VPS	Video Parameter Set

4. RTP Payload Format

4.1 RTP Header Usage

The format of the RTP header is specified in [RFC3550] and reprinted in Figure 2 for convenience. This payload format uses the fields of the header in a manner consistent with that specification.

The RTP payload (and the settings for some RTP header bits) for aggregation packets and fragmentation units are specified in Sections 4.6 and 4.7, respectively.

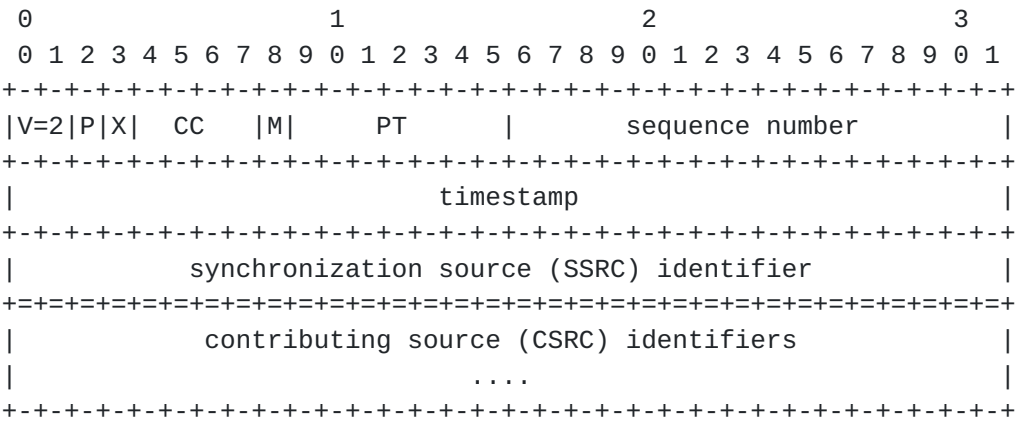

Figure 2 RTP header according to [RFC3550]

The RTP header information to be set according to this RTP payload format is set as follows:

Marker bit (M): 1 bit

Set for the last packet of the access unit indicated by the RTP timestamp, in line with the normal use of the M bit in video formats, to allow an efficient playout buffer handling. Decoders can use this bit as an early indication of the last packet of an access unit.

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 has to be performed either through the profile used or in a dynamic way.

Sequence number (SN): 16 bits

Set and used in accordance with 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 timing properties of its own (e.g., parameter set and SEI NAL units), the RTP timestamp is set to the RTP timestamp of the coded picture of the access unit in which the NAL unit is included, according to Section 7.4.2.4.4 of [HEVC].

Receivers SHOULD ignore the picture output timing information in any picture timing SEI messages or decoding unit information SEI messages as specified in [HEVC]. Instead, receivers SHOULD use the RTP timestamp for the display process. Receivers MUST pass picture timing SEI messages and decoding unit information SEI messages to the decoder and MAY use the field/frame related information for the display process e.g. when frame doubling or frame tripling is indicated by the field/frame related information.

4.2 Payload Structures

The first two bytes of the payload of an RTP packet are referred to as the payload header. The payload header consists of the same fields (F, Type, LayerId, and TID) as the NAL unit header as shown in section 1.1.4, irrespective of the type of the payload structure.

Three different types of RTP packet payload structures are specified. A receiver can identify the type of an RTP packet payload through the Type field in the payload header.

The three different payload structures are as follows:

- Single NAL unit packet: Contains a single NAL unit in the payload, and the NAL unit header of the NAL unit also serves as the payload header. This payload structure is specified in section 4.6.

- Aggregation packet (AP): Contains one or more NAL units within one access unit. This payload structure is specified in section 4.6.

- Fragmentation unit (FU): Contains a subset of a single NAL unit. This payload structure is specified in section 4.7.

4.3 Transmission Modes

This memo enables transmission of an HEVC bitstream over a single RTP session or multiple RTP sessions. The concept and working principle is inherited from [RFC6190] and follows a similar design. If only one RTP session is used for transmission of the HEVC bitstream, the transmission mode is referred to as single-session transmission (SST); otherwise (more than one RTP session is used for transmission of the HEVC bitstream), the transmission mode is referred to as multi-session transmission (MST).

[Ed. (YK): Unify the style of abbreviated words throughout the document.]

SST SHOULD be used for point-to-point unicast scenarios, while MST SHOULD be used for point-to-multipoint multicast scenarios where different receivers require different operation points of the same HEVC bitstream, to improve bandwidth utilizing efficiency.

> Informative note: A multicast may degrade to a unicast after all but one receivers have left (this is a justification of the first "SHOULD" instead of "MUST"), and there might be scenarios where MST is desirable but not possible e.g. when IP multicast is not

deployed in certain network (this is a justification of the second "SHOULD" instead of "MUST").

The transmission mode is indicated by the tx-mode media parameter (see section 7.1). If tx-mode is equal to "SST", SST MUST be used. Otherwise (tx-mode is equal to "MST"), MST MUST be used.

4.4 Decoding Order Number

For each NAL unit, the variable AbsDon is derived, representing the decoding order number that is indicative of the NAL unit decoding order.

Let NAL unit n be the n-th NAL unit in transmission order within an RTP session.

If tx-mode is equal to "SST" and sprop-depack-buf-nalus is equal to 0, AbsDon[n], the value of AbsDon for NAL unit n, is derived as equal to n.

Otherwise (tx-mode is equal to "MST" or sprop-depack-buf-nalus is greater than 0), AbsDon[n] is derived as follows, where DON[n] is the value of the variable DON for NAL unit n:

- If n is equal to 0 (i.e. NAL unit n is the very first NAL unit in transmission order), AbsDon[0] is set equal to DON[0].

- Otherwise (n is greater than 0), the following applies for derivation of AbsDon[n]:

```
If DON[n] == DON[n-1],
    AbsDon[n] = AbsDon[n-1]

If (DON[n] > DON[n-1] and DON[n] - DON[n-1] < 32768),
    AbsDon[n] = AbsDon[n-1] + DON[n] - DON[n-1]

If (DON[n] < DON[n-1] and DON[n-1] - DON[n] >= 32768),
    AbsDon[n] = AbsDon[n-1] + 65536 - DON[n-1] + DON[n]

If (DON[n] > DON[n-1] and DON[n] - DON[n-1] >= 32768),
    AbsDon[n] = AbsDon[n-1] - (DON[n-1] + 65536 - DON[n])
```

```
If (DON[n] < DON[n-1] and DON[n-1] - DON[n] < 32768),
    AbsDon[n] = AbsDon[n-1] - (DON[n-1] - DON[n])
```

For any two NAL units m and n, the following applies:

- AbsDon[n] greater than AbsDon[m] indicates that NAL unit n follows NAL unit m in NAL unit decoding order.
- When AbsDon[n] is equal to AbsDon[m], the NAL unit decoding order of the two NAL units can be in either order.
- AbsDon[n] less than AbsDon[m] indicates that NAL unit n precedes NAL unit m in decoding order.

When two consecutive NAL units in the NAL unit decoding order have different values of AbsDon, the value of AbsDon for the second NAL unit in decoding order MUST be greater than the value of AbsDon for the first NAL unit, and the absolute difference between the two AbsDon values MAY be greater than or equal to 1.

> Informative note: There are multiple reasons to allow for the absolute difference of the values of AbsDon for two consecutive NAL units in the NAL unit decoding order to be greater than one. An increment by one is not required, as at the time of associating values of AbsDon to NAL units, it may not be known whether all NAL units are to be delivered to the receiver. For example, a gateway may not forward coded slice NAL units of higher sub-layers or some SEI NAL units when there is congestion in the network. In another example, the first intra picture of a pre-encoded clip is transmitted in advance to ensure that it is readily available in the receiver, and when transmitting the first intra picture, the originator does not exactly know how many NAL units will be encoded before the first intra picture of the pre-encoded clip follows in decoding order. Thus, the values of AbsDon for the NAL units of the first intra picture of the pre-encoded clip have to be estimated when they are transmitted, and gaps in values of AbsDon may occur. Another example is MST where the AbsDon values must indicate cross-layer decoding order for NAL units conveyed in all the RTP sessions.

4.5 Single NAL Unit Packets

A single NAL unit packet contains exactly one NAL unit, and consists of a payload header (denoted as PayloadHdr), an optional 16-bit DONL field (in network byte order), and the NAL unit payload data (the NAL unit excluding its NAL unit header) of the contained NAL unit, as shown in Figure 3.

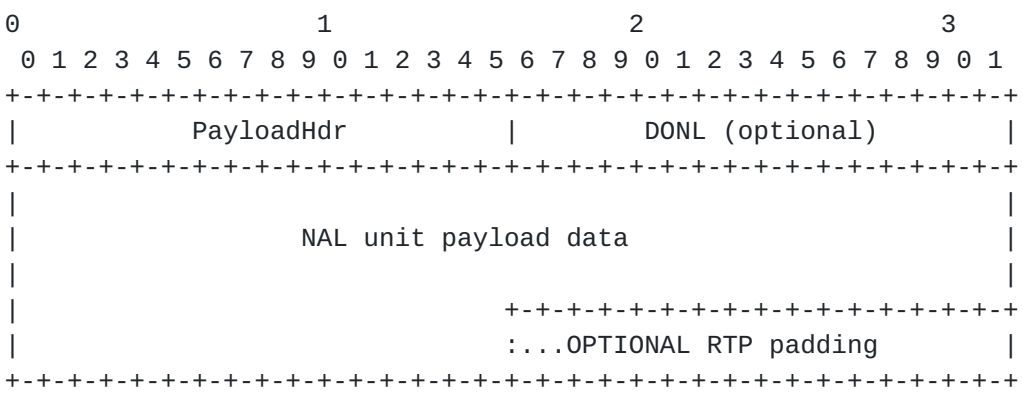

Figure 3 The structure of the first aggregation unit in an AP

The payload header MUST be an exact copy of the NAL unit header of the contained NAL unit.

The DONL field, when present, specifies the value of the 16 least significant bits of the decoding order number of the contained NAL unit.

If tx-mode is equal to "MST" or sprop-depack-buf-nalus is greater than 0, the DONL field MUST be present, and the variable DON for the contained NAL unit is derived as equal to the value of the DONL field. Otherwise (tx-mode is equal to "SST" and sprop-depack-buf-nalus is equal to 0), the DONL field MUST NOT be present.

4.6 Aggregation Packets (APs)

Aggregation packets (APs) are introduced to enable the reduction of packetization overhead for small NAL units, such as most of the non-VCL NAL units, which are often only a few octets in size.

An AP aggregates NAL units within one access unit. Each NAL unit to be carried in an AP is encapsulated in an aggregation unit. NAL units aggregated in one AP are in NAL unit decoding order.

An AP consists of a payload header (denoted as PayloadHdr) followed by one or more aggregation units, as shown in Figure 4.

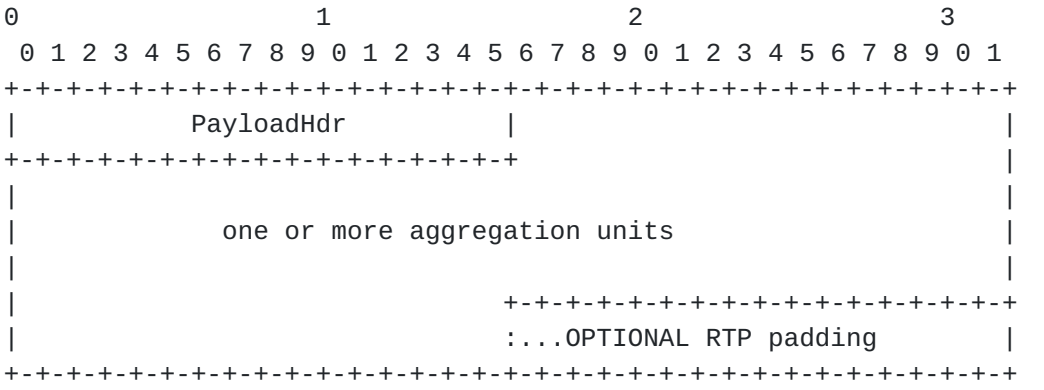

Figure 4 The structure of an aggregation packet

The fields in the payload header are set as follows. The F bit MUST be equal to 0 if the F bit of each aggregated NAL unit is equal to zero; otherwise, it MUST be equal to 1. The Type field MUST be equal to 48. The value of LayerId MUST be equal to the lowest value of LayerId of all the aggregated NAL units. The value of TID MUST be the lowest value of TID of all the aggregated NAL units.

> Informative Note: All VCL NAL units in an AP have the same TID value since they belong to the same access unit. However, an AP may contain non-VCL NAL units for which the TID value in the NAL unit header may be different than the TID value of the VCL NAL units in the same AP.

An AP can carry as many aggregation units as necessary; however, the total amount of data in an AP obviously MUST fit into an IP packet, and the size SHOULD be chosen so that the resulting IP packet is smaller than the MTU size so to avoid IP layer fragmentation. An AP MUST NOT contain Fragmentation Units (FUs) specified in section 4.7. APs MUST NOT be nested; i.e., an AP MUST NOT contain another AP.

The first aggregation unit in an AP consists of an optional 16-bit DONL field (in network byte order) followed by a 16-bit unsigned size information (in network byte order) that indicates the size of the NAL unit in bytes (excluding these two octets, but including the NAL unit header), followed by the NAL unit itself, including its NAL unit header, as shown in Figure 5.

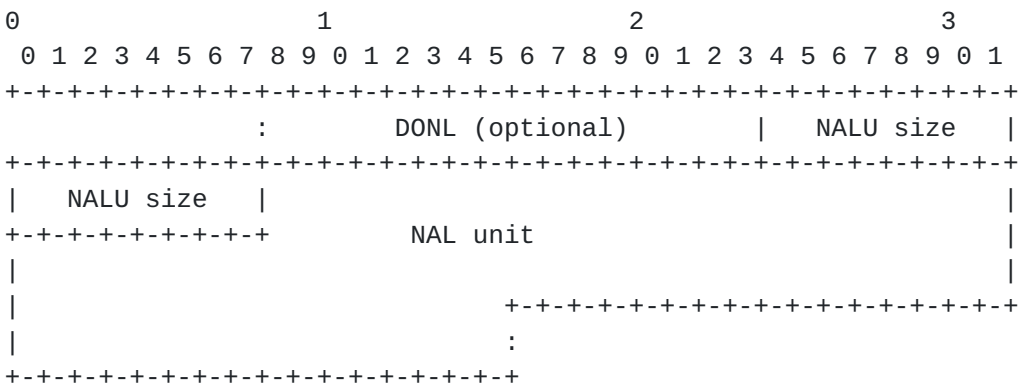

Figure 5 The structure of the first aggregation unit in an AP

The DONL field, when present, specifies the value of the 16 least significant bits of the decoding order number of the aggregated NAL unit.

If tx-mode is equal to "MST" or sprop-depack-buf-nalus is greater than 0, the DONL field MUST be present in an aggregation unit that is the first aggregation unit in an AP, and the variable DON for the aggregated NAL unit is derived as equal to the value of the DONL field. Otherwise (tx-mode is equal to "SST" and sprop-depack-buf-nalus is equal to 0), the DONL field MUST NOT be present in an aggregation unit that is the first aggregation unit in an AP.

An aggregation unit that is not the first aggregation unit in an AP consists of an optional 8-bit DOND field followed by a 16-bit unsigned size information (in network byte order) that indicates the size of the NAL unit in bytes (excluding these two octets, but including the NAL unit header), followed by the NAL unit itself, including its NAL unit header, as shown in Figure 6.

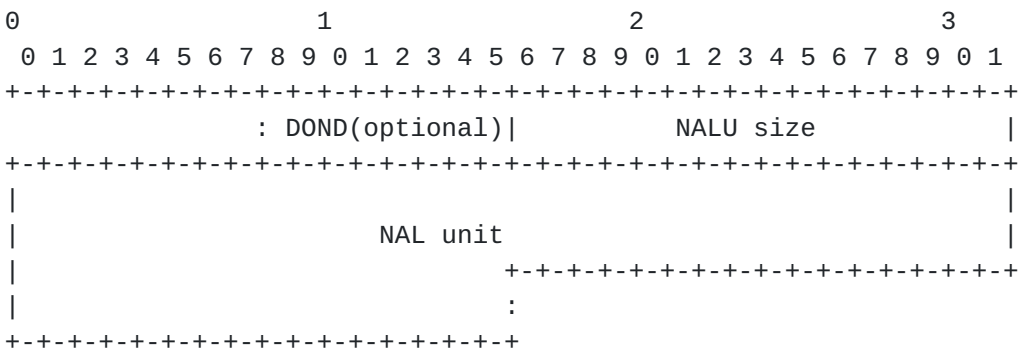

Figure 6 The structure of an aggregation unit that is not the first aggregation unit in an AP

When present, the DOND field plus 1 specifies the difference between the decoding order number values of the current aggregated NAL unit and the preceding aggregated NAL unit in the same AP.

If tx-mode is equal to "MST" or sprop-depack-buf-nalus is greater than 0, the DOND field MUST be present in an aggregation unit that is not the first aggregation unit in an AP, and the variable DON for the aggregated NAL unit is derived as equal to the DON of the preceding aggregated NAL unit in the same AP plus the value of the DOND field plus 1 modulo 65536. Otherwise (tx-mode is equal to "SST" and sprop-depack-buf-nalus is equal to 0), the DOND field MUST NOT be present in an aggregation unit that is not the first aggregation unit in an AP.

Figure 7 presents an example of an AP that contains two aggregation units, labeled as 1 and 2 in the figure, without the DONL and DOND fields being present.

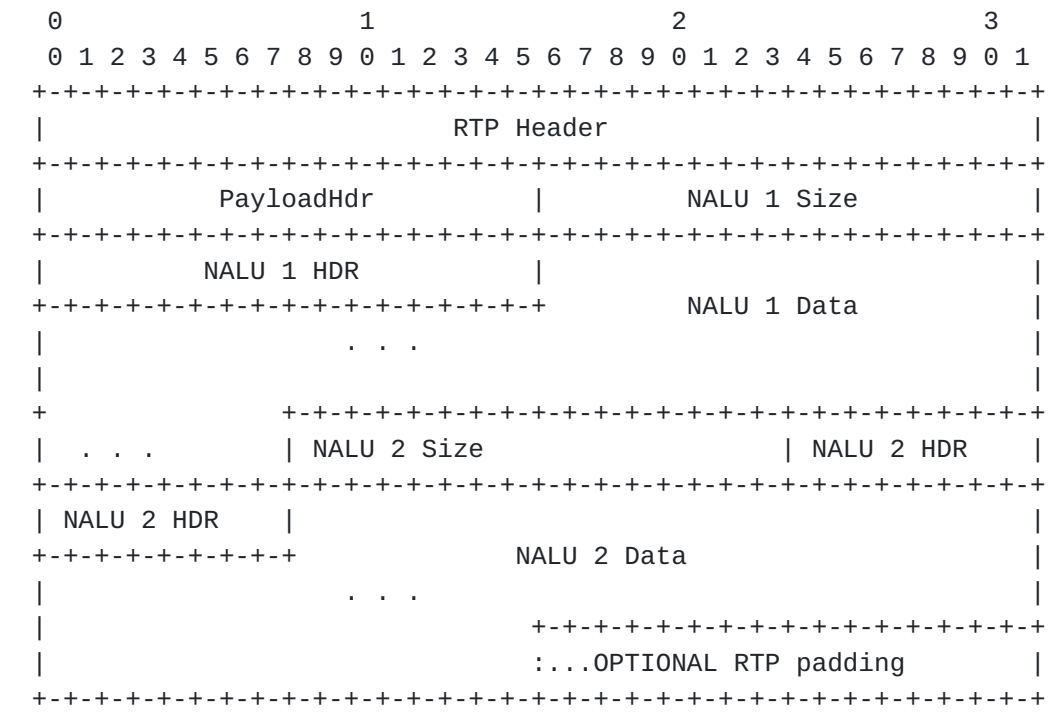

Figure 7 An example of an AP packet containing two aggregation units without the DONL and DOND fields

Figure 8 presents an example of an AP that contains two aggregation units, labeled as 1 and 2 in the figure, with the DONL and DOND fields being present.

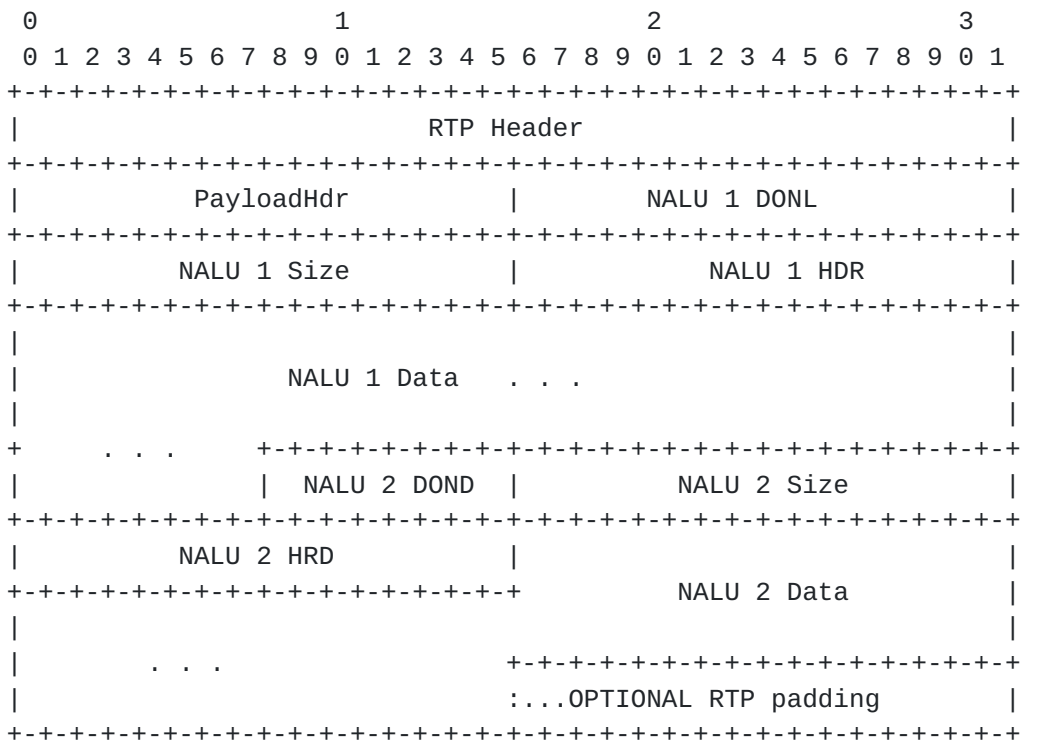

Figure 8 An example of an AP containing two aggregation units with the DONL and DOND fields

4.7 Fragmentation Units (FUs)

Fragmentation units (FUs) are introduced to enable fragmenting a single NAL unit into multiple RTP packets, possibly without cooperation or knowledge of the HEVC encoder. A fragment of a NAL unit consists of an integer number of consecutive octets 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).

When a NAL unit is fragmented and conveyed within FUs, it is referred to as a fragmented NAL unit. APs 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.

An FU consists of a payload header (denoted as PayloadHdr), an FU header of one octet, an optional 16-bit DONL field (in network byte order), and an FU payload, as shown in Figure 9.

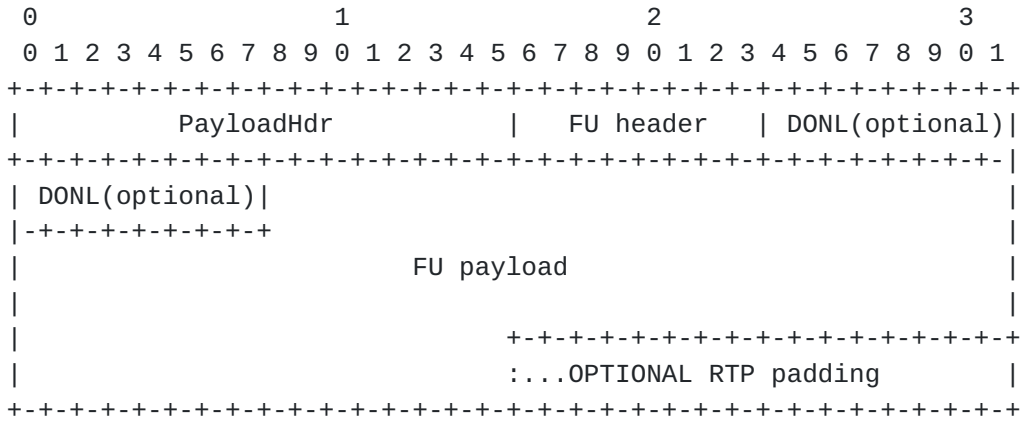

Figure 9 The structure of an FU

The fields in the payload header are set as follows. The Type field MUST be equal to 49. The fields F, LayerId, and TID MUST be equal to the fields F, LayerId, and TID, respectively, of the fragmented NAL unit.

The FU header consists of an S bit, an E bit, and a 6-bit Type field, as shown in Figure 10.

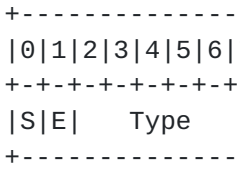

Figure 10 The structure of FU header

The semantics of the FU header fields are as follows:

S: 1 bit
When set to one, the S bit indicates the start of a fragmented NAL unit i.e., the first byte of the FU payload is also the first byte of the payload of the fragmented NAL unit. When the FU payload is not the start of the fragmented NAL unit payload, the S bit MUST be set to zero.

E: 1 bit
When set to one, the E bit 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. When the FU payload is not the last fragment of a fragmented NAL unit, the E bit MUST be set to zero.

Type: 6 bits
The field Type MUST be equal to the field Type of the fragmented NAL unit.

The DONL field, when present, specifies the value of the 16 least significant bits of the decoding order number of the fragmented NAL unit.

If tx-mode is equal to "MST" or sprop-depack-buf-nalus is greater than 0, and the S bit is equal to 1, the DONL field MUST be present in the FU, and the variable DON for the fragmented NAL unit is derived as equal to the value of the DONL field. Otherwise (tx-mode is equal to "SST" and sprop-depack-buf-nalus is equal to 0, or the S bit is equal to 0), the DONL field MUST NOT be present in the FU.

A non-fragmented NAL unit MUST NOT be transmitted in one FU; i.e., the 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 so that if the FU payloads of consecutive FUs, starting with an FU with the S bit equal to 1 and ending with an FU with the E bit equal to 1, are sequentially concatenated, the payload of the fragmented NAL unit can be reconstructed. The NAL unit header of the fragmented NAL unit is not included as such in

the FU payload, but rather the information of the NAL unit header of the fragmented NAL unit is conveyed in F, LayerId, and TID fields of the FU payload headers of the FUs and the Type field of the FU header of the FUs. An FU payload MAY have any number of octets and MAY be empty.

> Informative note: Empty FU payloads are allowed to reduce the latency of a certain class of senders in nearly lossless environments. These senders can be characterized in that they packetize fragments of a NAL unit before the NAL unit is completely generated and, hence, before the NAL unit size is known. If zero-length FU payloads were not allowed, the sender would have to generate at least one bit of data of the following fragment of the NAL unit before the current FU could be sent. Due to the characteristics of HEVC, where sometimes several CTUs occupy zero bits, this is undesirable and can add delay. However, the (potential) use of zero-length FU payloads should be carefully weighted against the increased risk of the loss of at least a part of the fragmented NAL unit because of the additional packets employed for its transmission.

If an FU is lost, the receiver SHOULD discard all following fragmentation units in transmission order corresponding to the same fragmented NAL unit, unless the decoder in the receiver is known to be prepared to gracefully handle incomplete NAL units.

A receiver in an endpoint or in a MANE MAY aggregate the first n-1 fragments of a NAL unit to an (incomplete) NAL unit, even if fragment n of that NAL unit is not received. In this case, the forbidden_zero_bit of the NAL unit MUST be set to one to indicate a syntax violation.

5. Packetization Rules

The following packetization rules apply:

- If tx-mode is equal to "MST" or sprop-depack-buf-nalus is greater than 0 for an RTP session, the transmission order of NAL units carried in the RTP session MAY be different than the NAL unit decoding order. Otherwise (tx-mode is equal to "SST" and sprop-depack-buf-nalus is equal to 0 for an RTP session), the transmission order of NAL units carried in the RTP session MUST be the same as the NAL unit decoding order.

- A NAL unit of a small size SHOULD be encapsulated in an aggregation packet together with one or more other NAL units in order to avoid the unnecessary packetization overhead for small NAL units. For example, non-VCL NAL units such as access unit delimiters, parameter sets, or SEI NAL units are typically small and can often be aggregated with slice NAL units without violating MTU size constraints.

- Each non-VCL NAL unit SHOULD be encapsulated in an aggregation packet together with its associated VCL NAL unit, as typically a non-VCL NAL unit would be meaningless without the associated VCL NAL unit being available.

- The TID value is designed to indicate (among other things) the relative importance of an RTP packet, for example because NAL units belonging to higher temporal sub-layers are not used for the decoding of lower temporal sub-layers. A lower value of TID indicates a higher importance. More important NAL units MAY be better protected against transmission losses than less important NAL units.

- FUs SHOULD NOT be applied in live-encoding scenarios such as video telephony, video conferencing, live streaming and live broadcast, in which cases dependent slice segments SHOULD be used when a slice should be transported in multiple RTP packets. For pre-encoded content where using of dependent slice segments is not possible without transcoding, FUs SHOULD be used for transporting of one NAL unit in multiple RTP packets for MTU size matching.

6. De-packetization Process

The general concept behind de-packetization is to get the NAL units out of the RTP packets in an RTP session and all the dependent RTP sessions, if any, and pass them to the decoder in the NAL unit decoding order.

The de-packetization process is implementation dependent. Therefore, the following description should be seen as an example of a suitable implementation. Other schemes may be used as well as long as the output for the same input is the same as the process described below. The output is the same when the set of NAL units and their order are both identical. Optimizations relative to the described algorithms are possible.

All normal RTP mechanisms related to buffer management apply. In particular, duplicated or outdated RTP packets (as indicated by the RTP sequences number and the RTP timestamp) are removed. To determine the exact time for decoding, factors such as a possible intentional delay to allow for proper inter-stream synchronization must be factored in.

NAL units with NAL unit type values in the range of 0 to 47, inclusive may be passed to the decoder. NAL-unit-like structures with NAL unit type values in the range of 48 to 63, inclusive, MUST NOT be passed to the decoder.

The receiver includes a receiver buffer, which is used to compensate for transmission delay jitter, to reorder NAL units from transmission order to the NAL unit decoding order, and to recovery the NAL unit decoding order in MST, when applicable. In this section, the receiver operation is described under the assumption that there is no transmission delay jitter. To make a difference from a practical receiver buffer that is also used for compensation of transmission delay jitter, the receiver buffer is here after called the de-packetization buffer in this section. Receivers SHOULD also prepare for transmission delay jitter; i.e., either reserve separate buffers for transmission delay jitter buffering and de-packetization buffering or use a receiver buffer for both transmission delay jitter and de-packetization. Moreover, receivers SHOULD take transmission delay jitter into account in the buffering

operation; e.g., by additional initial buffering before starting of decoding and playback.

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

Regardless of the buffering state, the receiver stores incoming NAL units, in reception order, into the de-packetization buffer. NAL units carried in single NAL unit packets, APs, and FUs are stored in the de-packetization buffer individually, and the value of AbsDon is calculated and stored for each NAL unit. When MST is in use, NAL units of all RTP packet streams are stored in the same de-packetization buffer.

Initial buffering lasts until condition A (the number of NAL units in the de-packetization buffer is greater than the value of sprop-depack-buf-nalus of the highest RTP session) is true.

After initial buffering, whenever condition A is true, the following operation is repeatedly applied until condition A becomes false:

- The NAL unit in the de-packetization buffer with the smallest value of AbsDon is removed from the de-packetization buffer and passed to the decoder.

When no more NAL units are flowing into the de-packetization buffer, all NAL units remained in the de-packetization buffer are removed from the buffer and passed to the decoder in the order of increasing AbsDon values.

7. Payload Format Parameters

This section specifies the parameters that MAY be used to select optional features of the payload format and certain features or properties of the bitstream. The parameters are specified here as part of the media type registration for the HEVC codec. A mapping of the parameters into the Session Description Protocol (SDP) [RFC4566] is also provided for applications that use SDP.

Equivalent parameters could be defined elsewhere for use with control protocols that do not use SDP.

7.1 Media Type Registration

The media subtype for the HEVC codec is allocated from the IETF tree.

The receiver MUST ignore any unspecified parameter.

Media Type name: video

Media subtype name: H265

Required parameters: none

OPTIONAL parameters:

In the following definitions of parameters, "the stream" or "the NAL unit stream" refers to all NAL units conveyed in the current RTP session in SST, and all NAL units conveyed in the current RTP session and all NAL units conveyed in other RTP sessions that the current RTP session depends on in MST.

profile-space, profile-id:

The profile-space parameter indicates the context for interpretation of the profile-id parameter value. The profile, which specifies the subset of coding tools that may have been used to generate the stream or that the receiver supports, as specified in [HEVC], is defined by the combination of profile-space and profile-id. Note that profile-space is required to be equal to 0 in [HEVC], but other values for it may be specified in the future by ITU-T or ISO/IEC.

If the profile-space and profile-id parameters are used to indicate properties of a NAL unit stream, it indicates that, to decode the stream, the minimum subset of coding tools a decoder has to support is the profile specified by both parameters.

If the profile-space and profile-id parameters are used for capability exchange or session setup, it indicates the subset of coding tools, which is equal to the profile, that the codec supports for both receiving and sending.

If no profile-space is present, a value of 0 MUST be inferred and if no profile-id is present the Main profile MUST be inferred.

The profile-space and profile-id parameters are derived from the sequence parameter set or video parameter set NAL units, as specified in [HEVC], as follows.

For SST or for the stream corresponding to the highest RTP session of MST when MST is applied, the following applies:

- profile_space = general_profile_space
- profile_id = general_profile_idc

For streams not corresponding to the highest RTP session of MST when MST is applied, the following applies, with j being the value of the sub-layer-id parameter:

- profile_space = sub_layer_profile_space[j]
- profile_id = sub_layer_profile_idc[j]

tier-flag, level-id:

The tier-flag parameter indicates the context for interpretation of the level-id value. The default level, which limits values of syntax elements or on arithmetic combinations of values of syntax elements, as specified in [HEVC], is defined by the combination of tier-flag and level-id.

If the tier-flag and level-id parameters are used to indicate properties of a NAL unit stream, it indicates that, to decode the stream the lowest level the decoder has to support is the default level.

If the tier-flag and level-id parameters are used for capability exchange or session setup, the following applies. If max-recv-level-id is not present, the default level defined by tier-flag and level-id indicates the highest level the codec wishes to support. Otherwise, tier-flag and max-recv-level-id indicate the highest level the codec supports for receiving. For either receiving or sending, all levels that are lower than the highest level supported MUST also be supported.

If no tier-flag is present, a value of 0 MUST be inferred and if no level-id is present, a value of 1 MUST be inferred.

The tier-flag and level-id parameters are derived from the sequence parameter set or video parameter set NAL units, as specified in [HEVC], as follows.

For SST or for the stream corresponding to the highest RTP session of MST when MST is applied, the following applies:

- tier-flag = general_tier_flag
- level-id = general_level_idc

For streams not corresponding to the highest RTP session of MST when MST is applied, the following applies, with j being the value of the sub-layer-id parameter:

- tier-flag = sub_layer_tier_flag[j]
- level-id = sub_layer_leve_idc[j]

interop-constraints:

A base16 [RFC4648] (hexadecimal) representation of the six bytes derived from the sequence parameter set or video parameter set NAL units as specified in [HEVC] consisting of progressive_source_flag, interlaced_source_flag, non_packed_constraint_flag, frame_only_constraint_flag, and reserved_zero_44bits. Note that reserved_zero_44bits is required to be equal to 0 in [HEVC], but other values for it may be specified in the future by ITU-T or ISO/IEC.

If no interop-constraints are present, the following MUST be inferred:

- progressive_source_flag = 1
- interlaced_source_flag = 0
- non_packed_constraint_flag = 1
- frame_only_constraint_flag = 1
- reserved_zero_44bits = 0

For SST or for the stream corresponding to the highest RTP session of MST when MST is applied, the following applies:

- progressive_source_flag = general_progressive_source_flag
- interlaced_source_flag = general_interlaced_source_flag
- non_packed_constraint_flag = general_non_packed_constraint_flag
- frame_only_constraint_flag = general_frame_only_constraint_flag
- reserved_zero_44bits = general_reserved_zero_44bits

For streams not corresponding to the highest RTP session of MST when MST is applied, the following applies, with j being the value of the sub-layer-id parameter:

- progressive_source_flag = sub_layer_progressive_source_flag[j]
- interlaced_source_flag = sub_layer_interlaced_source_flag[j]
- non_packed_constraint_flag = sub_layer_non_packed_constraint_flag[j]
- frame_only_constraint_flag = sub_layer_frame_only_constraint_flag[j]
- reserved_zero_44bits = sub_layer_reserved_zero_44bits[j]

profile-compatibility-indicator:

A base16 [RFC4648] representation of the four bytes representing the 32 profile compatibility flags in the sequence parameter set or video parameter set NAL units. A decoder conforming to a certain profile may be able to decode bitstreams conforming to other profiles. The profile-

compatibility-indicator provides exact information of the ability of a decoder conforming to a certain profile to decode bitstreams conforming to another profile. More concretely, if the profile compatibility flag corresponding to the profile, which a decoder conforms to, is set, then the decoder is able to decode that bitstream with the flag set, irrespective of the profile, which a bitstream conforms to (provided that the decoder supports the highest level of the bitstream).

For SST or for the stream corresponding to highest RTP session of MST when MST is used with temporal scalability the following applies with j = 0..31:

- The 32 flags = general_profile_compatibility_flag[j]

For streams not corresponding to the highest RTP session (the RTP session which no other RTP session depends on) of MST when MST is used with temporal scalability the following applies with i being the value of the sub-layer-id parameter and j = 0..31:

- The 32 flags = sub_layer_profile_compatibility_flag[i][j]

sub-layer-id:

This parameter MAY be used to indicate the TID of the highest sub-layer of the stream. When not present, the value of sub-layer-id is inferred to be equal to vps_max_sub_layers_minus1+1 and sps_max_sub_layers_minus1+1 in the video parameter set and sequence parameter set as defined in [HEVC].

recv-sub-layer-id:

This parameter MAY be used to signal a receiver's choice of the offers or declared sub-layers in the sprop-vps. The value of recv-sub-layer-id indicates the index of the highest sub-layer of the stream that a receiver supports. When not present, the value of recv-sub-layer-id is inferred to be equal to sub-layer-id.

max-recv-level-id:

This parameter MAY be used, together with tier-flag, to indicate the highest level a receiver supports. The highest level the receiver supports is equal to the value of max-recv-level-id divided by 30 for the Main or High tier (as determined by tier-flag equal to 0 or 1, respectively).

When max-recv-level-id is not present, the value is inferred to be equal to level-id.

max-recv-level-id MUST NOT be present when the highest level the receiver supports is not higher than the default level.

sprop-vps:

This parameter MAY be used to convey any video parameter set NAL unit of the stream. When present, the parameter MAY be used to indicate codec capability and sub-stream characteristics (i.e. properties of representations of sub-layers as defined in [HEVC]) as well as for out-of-band transmission of video parameter sets. The value of the parameter is a comma-separated (',') list of base64 [RFC4648] representations of the video parameter set NAL units as specified in Section 7.3.2.1 of [HEVC].

sprop-sps:

This parameter MAY be used to convey sequence parameter set NAL units of the stream for out-of-band transmission of sequence parameter sets. The value of the parameter is a comma-separated (',') list of base64 [RFC4648] representations of the sequence parameter set NAL units as specified in Section 7.3.2.2 of [HEVC].

sprop-pps:

This parameter MAY be used to convey picture parameter set NAL units of the stream for out-of-band transmission of picture parameter sets. The value of the parameter is a comma-separated (',') list of base64 [RFC4648] representations of

the picture parameter set NAL units as specified in Section 7.3.2.3 of [HEVC].

max-ls, max-lps, max-cpb, max-dpb, max-br:

These parameters MAY be used to signal the capabilities of a receiver implementation. These parameters MUST NOT be used for any other purpose. The highest level (specified by tier-flag and max-recv-level-id) MUST be such that the receiver is fully capable of supporting. max-ls, max-lps, max-cpb, max-dpb, and max-br MAY be used to indicate capabilities of the receiver that extend the required capabilities of the signaled highest level, as specified below.

When more than one parameter from the set (max-ls, max-lps, max-cpb, max-dpb, max-br) is present, the receiver MUST support all signaled capabilities simultaneously. For example, if both max-ls and max-br are present, the signaled highest level with the extension of both the frame rate and bitrate is supported. That is, the receiver is able to decode NAL unit streams in which the luma sample rate is up to max-ls (inclusive), the bitrate is up to max-br (inclusive), the coded picture buffer size is derived as specified in the semantics of the max-br parameter below, and the other properties comply with the highest level specified by tier-flag and max-recv-level-id.

Informative note: When the OPTIONAL media type parameters are used to signal the properties of a NAL unit stream, max-ls, max-lps, max-cpb, max-dpb, and max-br are not present, and the value of profile-space, profile-id, tier-flag and level-id must always be such that the NAL unit stream complies fully with the specified profile and level.

max-ls:
The value of max-ls is an integer indicating the maximum processing rate in units of luma samples per second. The max-ls parameter signals that the receiver is capable of decoding video at a higher rate than is required by the signaled highest level.

When max-ls is signaled, the receiver MUST be able to decode NAL unit streams that conform to the signaled highest level, with the exception that the MaxLumaSR value in Table A-2 of [HEVC] for the signaled highest level is replaced with the value of max-ls. The value of max-ls MUST be greater than or equal to the value of MaxLumaSR given in Table A-2 of [HEVC] for the highest level. Senders MAY use this knowledge to send pictures of a given size at a higher picture rate than is indicated in the signaled highest level.

max-lps:
The value of max-lps is an integer indicating the maximum picture size in units of luma samples. The max-lps parameter signals that the receiver is capable of decoding larger picture sizes than are required by the signaled highest level. When max-lps is signaled, the receiver MUST be able to decode NAL unit streams that conform to the signaled highest level, with the exception that the MaxLumaPS value in Table A-1 of [HEVC] for the signaled highest level is replaced with the value of max-lps. The value of max-lps MUST be greater than or equal to the value of MaxLumaPS given in Table A-1 of [HEVC] for the highest level. Senders MAY use this knowledge to send larger pictures at a proportionally lower frame rate than is indicated in the signaled highest level.

max-cpb:
The value of max-cpb is an integer indicating the maximum coded picture buffer size in units of CpbBrVclFactor bits for the VCL HRD parameters and in units of CpbBrNalFactor bits for the NAL HRD parameters, where CpbBrVclFactor and CpbBrNalFactor are defined in Section A.4 of [HEVC]. The max-cpb parameter signals that the receiver has more memory than the minimum amount of coded picture buffer memory required by the signaled highest level. When max-cpb is signaled, the receiver MUST be able to decode NAL unit streams that conform to the signaled highest level, with the exception that the MaxCPB value in Table A-1 of [HEVC] for the signaled highest level is replaced with the value of max-cpb. The value of max-cpb MUST be greater than or equal to the value of MaxCPB given in Table A-1 of [HEVC] for the highest level. Senders MAY use this knowledge to construct coded video streams with greater

variation of bitrate than can be achieved with the MaxCPB value in Table A-1 of [HEVC].

Informative note: The coded picture buffer is used in the hypothetical reference decoder (Annex C of HEVC). The use of the hypothetical reference decoder is recommended in HEVC encoders to verify that the produced bitstream conforms to the standard and to control the output bitrate. Thus, the coded picture buffer is conceptually independent of any other potential buffers in the receiver, including de-packetization and de-jitter buffers. The coded picture buffer need not be implemented in decoders as specified in Annex C of HEVC, but rather standard-compliant decoders can have any buffering arrangements provided that they can decode standard-compliant bitstreams. Thus, in practice, the input buffer for a video decoder can be integrated with de-packetization and de-jitter buffers of the receiver.

max-dpb:

The value of max-dpb is an integer indicating the maximum decoded picture buffer size in units decoded pictures at the MaxLumaPS for the highest level, i.e. number of decoded pictures at the maximum picture size defined by the highest level. The value of max-dpb MUST be smaller than or equal to 16. The max-dpb parameter signals that the receiver has more memory than the minimum amount of decoded picture buffer memory required by default, which is MaxDpbPicBuf as defined in [HEVC](equal to 6). When max-dpb is signaled, the receiver MUST be able to decode NAL unit streams that conform to the signaled highest level, with the exception that the MaxDpbPicBuff value defined in [HEVC] as 6 is replaced with the value of max-dpb. Consequently, a receiver that signals max-dpb MUST be capable of storing the following number of decoded frames (MaxDpbSize) in its decoded picture buffer:

```
                    if( PicSizeInSamplesY <= ( MaxLumaPS >> 2 ) )
       MaxDpbSize = Min( 4 * max-dpb, 16 )
     else if ( PicSizeInSamplesY <= ( MaxLumaPS >> 1 ) )
       MaxDpbSize = Min( 2 * max-dpb, 16 )
     else if ( PicSizeInSamplesY <= ( ( 3 * MaxLumaPS ) >> 2 ) )
       MaxDpbSize = Min( (4 * max-dpb) / 3, 16 )
```

```
            else
               MaxDpbSize = max-dpb
```

Wherein MaxLumaPS given in Table A-1 of [HEVC] for the highest level and PicSizeInSamplesY is the current size of each decoded picture in units of luma samples as defined in [HEVC].

The value of max-dpb MUST be greater than or equal to the value of MaxDpbPicBuf (i.e. 6) as defined in [HEVC]. Senders MAY use this knowledge to construct coded video streams with improved compression.

Informative note: This parameter was added primarily to complement a similar codepoint in the ITU-T Recommendation H.245, so as to facilitate signaling gateway designs. The decoded picture buffer stores reconstructed samples. There is no relationship between the size of the decoded picture buffer and the buffers used in RTP, especially de-packetization and de-jitter buffers.

max-br:
The value of max-br is an integer indicating the maximum video bitrate in units of CpbBrVclFactor bits per second for the VCL HRD parameters and in units of CpbBrNalFactor bits per second for the NAL HRD parameters, where CpbBrVclFactor and CpbBrNalFactor are defined in Section A.4 of [HEVC].

The max-br parameter signals that the video decoder of the receiver is capable of decoding video at a higher bitrate than is required by the signaled highest level.

When max-br is signaled, the video codec of the receiver MUST be able to decode NAL unit streams that conform to the signaled highest level, with the following exceptions in the limits specified by the highest level:

- The value of max-br replaces the MaxBR value in Table A-2 of [HEVC] for the highest level.

o When the max-cpb parameter is not present, the result of the following formula replaces the value of MaxCPB in Table A-1 of [HEVC]:

(MaxCPB of the signaled level) * max-br / (MaxBR of the signaled highest level).

For example, if a receiver signals capability for Main profile Level 2 with max-br equal to 2000, this indicates a maximum video bitrate of 2000 kbits/sec for VCL HRD parameters, a maximum video bitrate of 2200 kbits/sec for NAL HRD parameters, and a CPB size of 2000000 bits (2000000 / 1500000 * 1500000).

The value of max-br MUST be greater than or equal to the value MaxBR given in Table A-2 of [HEVC] for the signaled highest level.

Senders MAY use this knowledge to send higher bitrate video as allowed in the level definition of Annex A of HEVC to achieve improved video quality.

Informative note: This parameter was added primarily to complement a similar codepoint in the ITU-T Recommendation H.245, so as to facilitate signaling gateway designs. The assumption that the network is capable of handling such bitrates at any given time cannot be made from the value of this parameter. In particular, no conclusion can be drawn that the signaled bitrate is possible under congestion control constraints.

tx-mode:

This parameter indicates whether the transmission mode is SST or MST.

The value of tx-mode MUST be equal to either "MST" or "SST". When not present, the value of tx-mode is inferred to be equal to "SST".

If the value is equal to "MST", MST MUST be in use. Otherwise (the value is equal to "SST"), SST MUST be in use.

The value of tx-mode MUST be equal to "MST" for all RTP sessions in an MST.

sprop-depack-buf-nalus:

This parameter specifies the maximum number of NAL units that precede a NAL unit in the de-packetization buffer in reception order and follow the NAL unit in decoding order.

The value of sprop-depack-buf-nalus MUST be an integer in the range of 0 to 32767, inclusive.

When not present, the value of sprop-depack-buf-nalus is inferred to be equal to 0.

When the RTP session depends on one or more other RTP sessions (in this case tx-mode MUST be equal to "MST"), this parameter MUST be present and the value of sprop-depack-buf-nalus MUST be greater than 0.

sprop-depack-buf-bytes:

This parameter signals the required size of the de-packetization buffer in units of bytes. The value of the parameter MUST be greater than or equal to the maximum buffer occupancy (in units of bytes) of the de-packetization buffer as specified in section 6.

The value of sprop-depack-buf-bytes MUST be an integer in the range of 0 to 4294967295, inclusive.

When the RTP session depends on one or more other RTP sessions (in this case tx-mode MUST be equal to "MST") or sprop-depack-buf-nalus is present and is greater than 0, this parameter MUST be present and the value of sprop-depack-buf-bytes MUST be greater than 0.

Informative note: sprop-depack-buf-bytes indicates the required size of the de-packetization buffer only. When

network jitter can occur, an appropriately sized jitter buffer has to be available as well.

depack-buf-cap:

This parameter signals the capabilities of a receiver implementation and indicates the amount of de-packetization buffer space in units of bytes that the receiver has available for reconstructing the NAL unit decoding order. A receiver is able to handle any stream for which the value of the sprop-depack-buf-bytes parameter is smaller than or equal to this parameter.

When not present, the value of depack-buf-req is inferred to be equal to 0. The value of depack-buf-cap MUST be an integer in the range of 0 to 4294967295, inclusive.

Informative note: depack-buf-cap indicates the maximum possible size of the de-packetization buffer of the receiver only. When network jitter can occur, an appropriately sized jitter buffer has to be available as well.

segmentation-id:

This parameter MAY be used to signal the segmentation tools present in the stream and that can be used for parallelization. The value of segmentation-id MUST be an integer in the range of 0 to 3, inclusive. When not present, the value of segmentation-id is inferred to be equal to 0.

When segmentation-id is equal to 0, no information about the segmentation tools is provided. When segmentation-id is equal to 1, it indicates that slices are present in the stream. When segmentation-id is equal to 2, it indicates that tiles are present in the stream. When segmentation-id is equal to 3, it indicates that WPP is used in the stream.

spatial-segmentation-idc:

A base16 [RFC4648] representation of the syntax element min_spatial_segmentation_idc as specified in [HEVC]. This parameter MAY be used to describe parallelization capabilities of the stream.

Encoding considerations:

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

Security considerations:

See Section 9 of RFC XXXX.

Public specification:

Please refer to Section 13 of RFC XXXX.

Additional information: None

File extensions: none

Macintosh file type code: none

Object identifier or OID: none

Person & email address to contact for further information:

Intended usage: COMMON

Author: See Section 14 of RFC XXXX.

Change controller:

IETF Audio/Video Transport Payloads working group delegated from the IESG.

7.2 SDP Parameters

The receiver MUST ignore any parameter unspecified in this memo.

7.2.1 Mapping of Payload Type Parameters to SDP

The media type video/H265 string is mapped to fields in the Session Description Protocol (SDP) [RFC4566] as follows:

- The media name in the "m=" line of SDP MUST be video.
- The encoding name in the "a=rtpmap" line of SDP MUST be H265 (the media subtype).
- The clock rate in the "a=rtpmap" line MUST be 90000.
- The OPTIONAL parameters "profile-space", "profile-id", "tier-flag", "level-id", "interop-constraints", "profile-compatibility-indicator", "sub-layer-id", "recv-sub-layer-id", "max-recv-level-id", "max-ls", "max-lps", "max-cpb", "max-dpb", "max-br", "tx-mode", "sprop-depack-buf-nalus", "sprop-depack-buf-bytes", "depack-buf-cap", "segmentation-id", and "spatial-segmentation-idc", when present, MUST be included in the "a=fmtp" line of SDP. This parameter is expressed as a media type string, in the form of a semicolon separated list of parameter=value pairs.
- The OPTIONAL parameters "sprop-vps", "sprop-sps", and "sprop-pps", when present, MUST be included in the "a=fmtp" line of SDP or conveyed using the "fmtp" source attribute as specified in section 6.3 of [RFC5576]. For a particular media format (i.e., RTP payload type), "sprop-vps" "sprop-sps", or "sprop-pps" MUST NOT be both included in the "a=fmtp" line of SDP and conveyed using the "fmtp" source attribute. When included in the "a=fmtp" line of SDP, these parameters are expressed as a media type string, in the form of a semicolon separated list of parameter=value pairs. When conveyed using the "fmtp" source attribute, these parameters are only associated with the given source and payload type as parts of the "fmtp" source attribute.

  Informative note: Conveyance of "sprop-vps", "sprop-sps", and "sprop-pps" using the "fmtp" source attribute allows for out-of-band transport of parameter sets in topologies like Topo-Video-switch-MCU as specified in [RFC5117].

An example of media representation in SDP is as follows:

```
m=video 49170 RTP/AVP 98
a=rtpmap:98 H265/90000
a=fmtp:98 profile-id=ST;
          sprop-vps=<video parameter sets data>
```

7.2.2 Usage with SDP Offer/Answer Model

When HEVC is offered over RTP using SDP in an Offer/Answer model [RFC3264] for negotiation for unicast usage, the following limitations and rules apply:

- The parameters identifying a media format configuration for HEVC are profile-space, profile-id, tier-flag, level-id, interop-constraints, tx-mode, and sprop-depack-buf-nalus. These media configuration parameters, except for level-id, MUST be used symmetrically when the answerer does not include recv-sub-layer-id in the answer; i.e., the answerer MUST either maintain all configuration parameters or remove the media format (payload type) completely, if one or more of the parameter values are not supported. The value of level-id) is changeable.

  Informative note: The requirement for symmetric use does not apply for level-id, and does not apply for the other stream properties and capability parameters.

To simplify handling and matching of these configurations, the same RTP payload type number used in the offer SHOULD also be used in the answer, as specified in [RFC3264]. The same RTP payload type number used in the offer MUST also be used in the answer when the answer includes recv-sub-layer-id. When the answer does not include recv-sub-layer-id, the answer MUST NOT contain a payload type number used in the offer unless the configuration is exactly the same as in the offer or the configuration in the answer only differs from that in the offer with a different value of level-id. The answer MAY contain the recv-sub-layer-id parameter if an HEVC stream contains multiple operation points (using temporal scalability and sub-layers) and sprop-vps is included in the offer where sub-layers are present in the video parameter set. If the sprop-vps is provided in an offer, an answerer MAY select a particular operation point in the received and/or in the sent stream. When recv-sub-layer-id is present in the answer, the media configuration parameters MUST NOT

be present in the answer. Rather, the media configuration that the answerer will use for receiving and/or sending is the one used for the selected operation point as indicated in the offer.

> Informative note: When an offerer receives an answer that does not include recv-sub-layer-id, it has to compare payload types not declared in the offer based on the media type (i.e., video/H265) and the above media configuration parameters with any payload types it has already declared. This will enable it to determine whether the configuration in question is new or if it is equivalent to configuration already offered, since a different payload type number may be used in the answer. The ability to perform operation point selection enables a receiver to utilize the temporal scalable nature of an HEVC stream.

- The parameters sprop-depack-buf-nalus and sprop-depack-buf-bytes describe the properties of the RTP packet stream that the offerer or the answerer is sending for the media format configuration. This differs from the normal usage of the Offer/Answer parameters: normally such parameters declare the properties of the stream that the offerer or the answerer is able to receive. When dealing with HEVC, the offerer assumes that the answerer will be able to receive media encoded using the configuration being offered.

  > Informative note: The above parameters apply for any stream sent by a declaring entity with the same configuration; i.e., they are dependent on their source. Rather than being bound to the payload type, the values may have to be applied to another payload type when being sent, as they apply for the configuration.

- The capability parameters max-ls, max-lps, max-cpb, max-dpb, and max-br MAY be used to declare further capabilities of the offerer or answerer for receiving. These parameters MUST NOT be present when the direction attribute is "sendonly" and when the parameters describe the limitations of what the offerer or answerer accepts for receiving streams.

- An offerer has to include the size of the de-packetization buffer, sprop-depack-buf-bytes, and sprop-depack-buf-nalus, in the offer for an interleaved HEVC stream or for the MST transmission mode. To enable the offerer and answerer to inform each other about their capabilities for de-packetization buffering in receiving streams, both parties are RECOMMENDED to include depack-buf-cap. For interleaved streams or in MST, it is also RECOMMENDED to consider offering multiple payload types with different buffering requirements when the capabilities of the receiver are unknown.

For streams being delivered over multicast, the following rules apply:

- The media format configuration is identified by profile-space, profile-id, tier-flag, level-id, interop-constraints, tx-mode and sprop-depack-buf-nalus. These media format configuration parameters, including level-id, MUST be used symmetrically; that is, the answerer MUST either maintain all configuration parameters or remove the media format (payload type) completely. Note that this implies that the level-id for Offer/Answer in multicast is not changeable.

To simplify the handling and matching of these configurations, the same RTP payload type number used in the offer SHOULD also be used in the answer, as specified in [RFC3264]. An answer MUST NOT contain a payload type number used in the offer unless the configuration is the same as in the offer.

- The rules for other parameters are the same as above for unicast as long as the above rules are obeyed.

Table 1 lists the interpretation of all the parameters that MUST be used for the various combinations of offer, answer, and direction attributes. Note that the two columns wherein the recv-sub-layer-id parameter is used only apply to answers, whereas the other columns apply to both offers and answers.

Table 1. Interpretation of parameters for various combinations of offers, answers, direction attributes, with and without recv-sub-

layer-id. Columns that do not indicate offer or answer apply to both.

				sendonly --+	
			answer: recvonly,recv-sub-layer-id --+		
			recvonly w/o recv-sub-layer-id --+			
		answer: sendrecv, recv-sub-layer-id --+				
		sendrecv w/o recv-sub-layer-id --+				
profile-space		C	X	C	X	P
profile-id		C	X	C	X	P
tier-flag		C	X	C	X	P
level-id		C	X	C	X	P
interop-constraints		C	X	C	X	P
profile-compatibility-indicator		C	X	C	X	P
max-recv-level-id		R	R	R	R	-
tx-mode		C	X	C	X	P
sprop-depack-buf-nalus		P	P	-	-	P
sprop-depack-buf-bytes		P	P	-	-	P
depack-buf-cap		R	R	R	R	-
segmentation-id		P	P	P	P	P
spatial-segmentation-idc		P	P	P	P	P
max-br		R	R	R	R	-
max-cpb		R	R	R	R	-
max-dpb		R	R	R	R	-
max-ls		R	R	R	R	-
max-lps		R	R	R	R	-
sprop-parameter-sets		P	P	-	-	P
recv-sub-layer-id		X	O	X	O	-

Legend:

C: configuration for sending and receiving streams
P: properties of the stream to be sent
R: receiver capabilities
O: operation point selection
X: MUST NOT be present
-: not usable, when present SHOULD be ignored

Parameters used for declaring receiver capabilities are in general downgradable; i.e., they express the upper limit for a sender's

possible behavior. Thus, a sender MAY select to set its encoder using only lower/lesser or equal values of these parameters.

Parameters declaring a configuration point are not changeable, with the exception of the level-id parameter for unicast usage. This expresses values a receiver expects to be used and MUST be used verbatim on the sender side. If level-id is changed, an answerer MUST NOT include the recv-sub-layer-id parameter.

When a sender's capabilities are declared, and non-changeable parameters are used in this declaration, these parameters express a configuration that is acceptable for the sender to receive streams. In order to achieve high interoperability levels, it is often advisable to offer multiple alternative configurations. It is impossible to offer multiple configurations in a single payload type. Thus, when multiple configuration offers are made, each offer requires its own RTP payload type associated with the offer.

A receiver SHOULD understand all media type parameters, even if it only supports a subset of the payload format's functionality. This ensures that a receiver is capable of understanding when an offer to receive media can be downgraded to what is supported by the receiver of the offer.

An answerer MAY extend the offer with additional media format configurations. However, to enable their usage, in most cases a second offer is required from the offerer to provide the stream property parameters that the media sender will use. This also has the effect that the offerer has to be able to receive this media format configuration, not only to send it.

7.2.3 Usage in Declarative Session Descriptions

When HEVC over RTP is offered with SDP in a declarative style, as in Real Time Streaming Protocol (RTSP) [RFC2326] or Session Announcement Protocol (SAP) [RFC2974], the following considerations are necessary.

o All parameters capable of indicating both stream properties and receiver capabilities are used to indicate only stream properties. For example, in this case, the parameter profile-tier-level-id declares the values used by the stream, not the capabilities for receiving streams. This results in that the following interpretation of the parameters MUST be used:

Declaring actual configuration or stream properties:

- profile-space
- profile-id
- tier-flag
- level-id
- interop-constraints
- tx-mode
- sprop-parameter-sets
- sprop-depack-buf-nalus
- sprop-depack-buf-bytes
- segmentation-id
- spatial-segmentation-idc

Not usable (when present, they SHOULD be ignored):

- max-lps
- max-ls
- max-cpb
- max-dpb
- max-br
- max-recv-level-id
- depack-buf-cap
- sub-layer-id

o A receiver of the SDP is required to support all parameters and values of the parameters provided; otherwise, the receiver MUST reject (RTSP) or not participate in (SAP) the session. It falls on the creator of the session to use values that are expected to be supported by the receiving application.

7.2.4 Dependency Signaling in Multi-Session Transmission

If MST is used, the rules on signaling media decoding dependency in SDP as defined in [RFC5583] apply. The rules on "hierarchical or layered encoding" with multicast in Section 5.7 of [RFC4566] do not apply, i.e., the notation for Connection Data "c=" SHALL NOT be used with more than one address. The order of session dependency is given from the RTP session containing the lowest temporal sub-layer to the RTP session containing the highest temporal sub-layer.

8. Use with Feedback Messages

As specified in section 6.1 of RFC 4585 [RFC4585], payload Specific Feedback messages are identified by the RTCP packet type value PSFB (206). AVPF [RFC4585] defines three payload-specific feedback messages and one application layer feedback message, and CCM [RFC5104] specifies four payload-specific feedback messages.

In addition, this memo defined two payload-specific feedback messages. These feedback messages are identified by means of the feedback message type (FMT) parameter as follows:

Assigned in [RFC4585]:

```
   1:     Picture Loss Indication (PLI)
   2:     Slice Lost Indication (SLI)
   3:     Reference Picture Selection Indication (RPSI)
   15:    Application layer FB message
   31:    reserved for future expansion of the number space
```

Assigned in [RFC5104]:

```
   4:     Full Intra Request (FIR) Command
   5:     Temporal-Spatial Trade-off Request (TSTR)
   6:     Temporal-Spatial Trade-off Notification (TSTN)
   7:     Video Back Channel Message (VBCM)
```

Assigned in this memo:

```
   8:     Specific Picture Loss Indication (SPLI)
```

Unassigned:

```
0:      unassigned
9-14:   unassigned
16-30:  unassigned
```

The following subsections define the Feedback Control Information (FCI) format for the new payload-specific feedback message and how to use HEVC with the RPSI and SPLI messages, both for the purpose of feedback based reference picture selection for improved error resilience in real-time conversational video applications such as video telephone and video conferencing.

Feedback based reference picture selection has been shown as a powerful tool to stop temporal error propagation for improved error resilience [Girod99][Wang05]. In one approach, the decoder side tracks errors in the decoded pictures and informs to the encoder side that a particular picture that has been decoded relatively earlier is correct and still present in the decoded picture buffer and requests the encoder to use that correct picture for reference when encoding the next picture, so to stop further temporal error propagation. For this approach, the decoder side should use the RPSI feedback message. In another approach, the decoder side only reports, to the encoder side, which pictures has been entirely or partially lost, and the encoder tracks errors in the decoded pictures at the decoder side based on the feedback messages, and if it infers that an earlier decoded picture is correct at the decoder side and is still in the decoded picture buffer of the decoder, it encodes the next picture using that correct picture for reference. The SPLI message defined below is for use with the second approach described above.

Encoders can encode some long-term reference pictures as specified in H.264 or HEVC for purposes described in the previous paragraph without the need of a huge decoded picture buffer. As shown in [Wang05], with a flexible reference picture management scheme as in H.264 and HEVC, even a decoded picture buffer size of two would work for both the approaches described in the previous paragraph.

8.1 Definition of the SPLI Feedback Message

The SPLI feedback message is identified by PT=PSFB and FMT=8. There MUST be exactly one RPSI contained in the FCI field.

Informative note: The SPLI message defined in this memo also applies to other codecs, and may later be moved to another extension of RFC 4585.

The FCI format of the SPLI message is exactly the same as that of the RPSI message, with the name of the field "Native RPSI bit string defined per codec" being replaced with "Native SPLI bit string defined per codec", as shown in Figure 11.

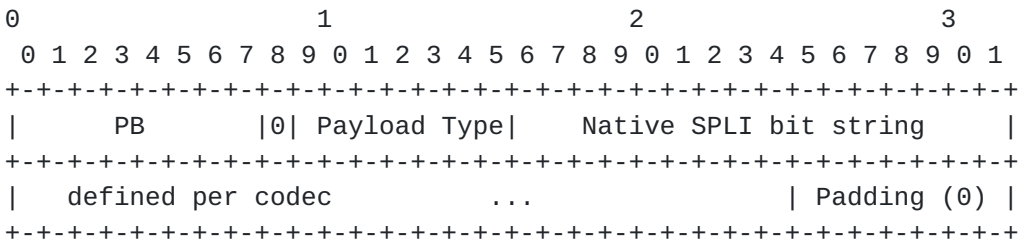

Figure 11 The PCI format of the SPLI

PB: 8 bits

The number of unused bits required to pad the length of the SPLI message to a multiple of 32 bits.

0: 1 bit

MUST be set to zero upon transmission and ignored upon reception.

Payload Type: 7 bits

Indicates the RTP payload type in the context of which the native SPLI bit string MUST be interpreted.

Native SPLI bit string: variable length

Indicates the SPLI information as natively defined by the video codec.

Padding: #PB bits

A number of bits set to zero to fill up the contents of the SPLI message to the next 32-bit boundary. The number of padding bits MUST be indicated by the PB field.

The same timing rules as for the RPSI message, as defined in [RFC4585], apply for the SPLI message.

8.2 Use of HEVC with the RPSI Feedback Message

The field "Native RPSI bit string defined per codec" is a base16 [RFC4648] representation of the 8 bits consisting of 2 most significant bits equal to 0 and 6 bits of nuh_layer_id, as defined in [HEVC], followed by the 32 bits representing the value of the PicOrderCntVal (in network byte order), as defined in [HEVC], for the picture that is requested to be used for reference when encoding the next picture.

Use of the RPSI feedback message as positive acknowledgement is deprecated. In other words, the RPSI feedback message MUST only be used as a reference picture selection request, such that it can also be used in multicast.

8.3 Use of HEVC with the SPLI Feedback Message

The field "Native SPLI bit string defined per codec" is a base16 [RFC4648] representation of the 8 bits consisting of 2 most significant bits equal to 0 and 6 bits of nuh_layer_id, as defined in [HEVC], followed by the 32 bits representing the value of the PicOrderCntVal, as defined in [HEVC], for the picture that is indicated as entirely or partially lost.

9. Security Considerations

RTP packets using the payload format defined in this specification are subject to the security considerations discussed in the RTP specification [RFC3550], and in any applicable RTP profile such as RTP/AVP [RFC3551], RTP/AVPF [RFC4585], RTP/SAVP [RFC3711] or

RTP/SAVPF [RFC5124]. However, as "Securing the RTP Protocol Framework: Why RTP Does Not Mandate a Single Media Security Solution" [I-D.ietf-avt-srtp-not-mandatory] discusses it is not an RTP payload format's responsibility to discuss or mandate what solutions are used to meet the basic security goals like confidentiality, integrity, and source authenticity for RTP in general. This responsibility lays on anyone using RTP in an application. They can find guidance on available security mechanisms and important considerations as discussed in "Options for Securing RTP Sessions" [I-D.ietf-avtcore-rtp-security-options].

The rest of this section discusses the security impacting properties of the payload format itself.

Because the data compression used with this payload format is applied end-to-end, any encryption needs to be performed after compression. 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 pathological datagrams into the stream that are complex to decode and that cause the receiver to be overloaded. H.265 is particularly vulnerable to such attacks, as it is extremely simple to generate datagrams containing NAL units that affect the decoding process of many future NAL units. Therefore, the usage of data origin authentication and data integrity protection of at least the RTP packet is RECOMMENDED, for example, with SRTP [RFC 3711].

Note that the appropriate mechanism to ensure confidentiality and integrity of RTP packets and their payloads is very dependent on the application and on the transport and signaling protocols employed. Thus, although SRTP is given as an example above, other possible choices exist.

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.

End-to-end security with authentication, integrity, or confidentiality protection will prevent a MANE from performing media-aware operations other than discarding complete packets. In

the case of confidentiality protection, it will even be prevented from discarding packets in a media-aware way. To be allowed to perform such operations, a MANE is required to be a trusted entity that is included in the security context establishment.

10. Congestion Control

Congestion control for RTP SHALL be used in accordance with RTP [RFC3550] and with any applicable RTP profile, e.g., AVP [RFC 3551]. If best-effort service is being used, an additional requirement is that users of this payload format MUST monitor packet loss to ensure that the packet loss rate is within an acceptable range. Packet loss is considered acceptable if a TCP flow across the same network path, and experiencing the same network conditions, would achieve an average throughput, measured on a reasonable timescale, that is not less than the RTP flow is achieving. This condition can be satisfied by implementing congestion control mechanisms to adapt the transmission rate, the number of layers subscribed for a layered multicast session, or by arranging for a receiver to leave the session if the loss rate is unacceptably high.

The bitrate adaptation necessary for obeying the congestion control principle is easily achievable when real-time encoding is used, for example by adequately tuning the quantization parameter.

However, when pre-encoded content is being transmitted, bandwidth adaptation requires the pre-coded bitstream to be tailored for such adaptivity. The key mechanism available in HEVC is temporal scalability. A media sender can remove NAL units belonging to higher temporal sub-layers (i.e. those NAL units with a high value of TID) until the sending bitrate drops to an acceptable range. HEVC contains mechanisms that allow the lightweight identification of switching points in temporal enhancement layers, as discussed in Section 1.1.2 of this memo. An HEVC media sender can send packets belonging to NAL units of temporal enhancement layers starting from these switching points to probe for available bandwidth and to utilized bandwidth that has been shown to be available.

Above mechanisms generally work within a defined profile and level and, therefore, no renegotiation of the channel is required. Only when non-downgradable parameters (such as profile) are required to

be changed does it become necessary to terminate and restart the media stream. This may be accomplished by using a different RTP payload type.

MANEs MAY remove certain unusable packets from the packet stream when that stream was damaged due to previous packet losses. This can help reduce the network load in certain special cases. For example, MANES can remove those FUs where the leading FUs belonging to the same NAL unit have been lost or those dependent slice segments when the leading slice segments belonging to the same slice have been lost, because the trailing FUs or dependent slice segments are meaningless to most decoders. MANES can also remove higher temporal scalable layers if the outbound transmission (from the MANE's viewpoint) experiences congestion.

11. IANA Consideration

A new media type, as specified in Section 7.1 of this memo, should be registered with IANA.

12. Acknowledgements

TBD

This document was prepared using 2-Word-v2.0.template.dot.

13. References

13.1 Normative References

[HEVC] JCT-VC, "High Efficiency Video Coding (HEVC) text specification draft 10 (for FDIS & Last Call)", JCTVC-L1003v34, March 2013.

[H.264] ITU-T Recommendation H.264, "Advanced video coding for generic audiovisual services", January 2012.

[RFC6184] Wang, Y.-K., Even, R., Kristensen, T., and R. Jesup, "RTP Payload Format for H.264 Video", RFC 6184, May 2011.

[RFC6190] Wenger, S., Wang, Y.-K., Schierl, T., and A. Eleftheriadis, "RTP Payload Format for Scalable Video Coding", RFC 6190, May 2011.

[RFC6051] C. Perkins and T. Schierl, "Rapid Synchronisation of RTP Flows"

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC3264] Rosenberg, J. and H. Schulzrinne, "An Offer/Answer Model with Session Description Protocol (SDP)", RFC 3264, June 2002.

[RFC4648] Josefsson, S., "The Base16, Base32, and Base64 Data Encodings", RFC 4648, October 2006.

[RFC3550] Schulzrinne, H., Casner, S., Frederick, R., and Jacobson, V., "RTP: A Transport Protocol for Real-Time Applications", STD 64, RFC 3550, July 2003.

[RFC4566] Handley, M., Jacobson, V., and Perkins, C., "SDP: Session Description Protocol", RFC 4566, July 2006.

[RFC5576] Lennox, J., Ott, J., and Schierl, T., "Source-Specific Media Attributes in the Session Description Protocol", RFC 5576, June 2009.

[RFC4585] Ott, J., Wenger, S., Sato, N., Burmeister, C., and Rey, J., "Extended RTP Profile for Real-time Transport Control Protocol (RTCP)-Based Feedback (RTP/AVPF)", RFC 4585, July 2006.

[RFC5104] Wenger, S., Chandra, U., Westerlund, M., and Burman, B., "Codec Control Messages in the RTP Audio-Visual Profile with Feedback (AVPF)", RFC 5104, February 2008.

13.2 Informative References

[Ed. (YK): Details for some of the following references are to be added.]

[3GPDASH] 3GPP TS 26.247.

[3GPPFF] 3GPP TS 26.244.

[Girod99] Girod, B. and Faerber, F., "Feedback-based error control for mobile video transmission", Proceedings IEEE, Vol. 87, No. 10, pp. 1707-1723, October 1999.

[ISOBMFF] IS0/IEC 14496-12.

[MPEG2S] IS0/IEC 13818-2.

[MPEGDASH] IS0/IEC 23009-1.

[RFC5109] Li, A., "RTP Payload Format for Generic Forward Error Correction", RFC 5109, December 2007.

[Wang05] Wang, Y.-K., Zhu, C., and Li, H., "Error resilient video coding using flexible reference fames", Visual Communications and Image Processing 2005 (VCIP 2005), July 2005, Beijing, China.

14. Authors' Addresses

Thomas Schierl
Fraunhofer HHI
Einsteinufer 37
D-10587 Berlin
Germany
Phone: +49-30-31002-227
Email: ts@thomas-schierl.de

Stephan Wenger
Vidyo, Inc. th 433 Hackensack Ave., 7 floor
Hackensack, N.J. 07601
USA
Phone: +1-415-713-5473
EMail: stewe@stewe.org

Ye-Kui Wang
Qualcomm Incorporated
5775 Morehouse Drive
San Diego, CA 92121
USA
Phone: +1-858-651-8345
EMail: yekuiw@qti.qualcomm.com

Miska M. Hannuksela
Nokia Corporation
P.O. Box 1000
33721 Tampere
Finland
Phone: +358-7180-08000
EMail: miska.hannuksela@nokia.com

Yago Sanchez
Fraunhofer HHI
Einsteinufer 37
D-10587 Berlin
Germany
Phone: +49-30-31002-227
Email: yago.sanchez@hhi.fraunhofer.de