Internet Engineering Task Force (IETF)
Request for Comments: 6190
Category: Standards Track
ISSN: 2070-1721

S. Wenger
Independent
Y.-K. Wang
Huawei Technologies
T. Schierl
Fraunhofer HHI
A. Eleftheriadis
Vidyo

May 2011

RTP Payload Format for Scalable Video Coding

Abstract

This memo describes an RTP payload format for Scalable Video Coding (SVC) as defined in Annex G of ITU-T Recommendation H.264, which is technically identical to Amendment 3 of ISO/IEC International Standard 14496-10. 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 in multiple RTP packets. Furthermore, it supports transmission of an SVC stream over a single as well as multiple RTP sessions. The payload format defines a new media subtype name "H264-SVC", but is still backward compatible to RFC 6184 since the base layer, when encapsulated in its own RTP stream, must use the H.264 media subtype name ("H264") and the packetization method specified in RFC 6184. The payload format has wide applicability in videoconferencing, Internet video streaming, and high-bitrate entertainment-quality video, among others.

Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc6190.

Table of Contents

1. Introduction

This memo specifies an RTP [RFC3550] payload format for the Scalable Video Coding (SVC) extension of the H.264/AVC video coding standard. SVC is specified in Amendment 3 to ISO/IEC 14496 Part 10 [ISO/IEC14496-10] and equivalently in Annex G of ITU-T Rec. H.264 [H.264]. In this memo, unless explicitly stated otherwise, "H.264/AVC" refers to the specification of [H.264] excluding Annex G.

SVC covers the entire application range of H.264/AVC, from low-bitrate mobile applications, to High-Definition Television (HDTV) broadcasting, and even Digital Cinema that requires nearly lossless coding and hundreds of megabits per second. The scalability features that SVC adds to H.264/AVC enable several system-level functionalities related to the ability of a system to adapt the signal to different system conditions with no or minimal processing. The adaptation relates both to the capabilities of potentially heterogeneous receivers (differing in screen resolution, processing speed, etc.), and to differing or time-varying network conditions. The adaptation can be performed at the source, the destination, or in intermediate media-aware network elements (MANEs). The payload format specified in this memo exposes these system-level functionalities so that system designers can take direct advantage of these features.

> Informative note: Since SVC streams contain, by design, a sub-stream that is compliant with H.264/AVC, it is trivial for a MANE to filter the stream so that all SVC-specific information is removed. This memo, in fact, defines a media type parameter (sprop-avc-ready, Section 7.2) that indicates whether or not the stream can be converted to one compliant with [RFC6184] by eliminating RTP packets, and rewriting RTP Control Protocol (RTCP) to match the changes to the RTP packet stream as specified in Section 7 of [RFC3550].

This memo defines two basic modes for transmission of SVC data, single-session transmission (SST) and multi-session transmission (MST). In SST, a single RTP session is used for the transmission of all scalability layers comprising an SVC bitstream; in MST, the scalability layers are transported on different RTP sessions. In SST, packetization is a straightforward extension of [RFC6184]. For MST, four different modes are defined in this memo. They differ on whether or not they allow interleaving, i.e., transmitting Network Abstraction Layer (NAL) units in an order different than the decoding order, and by the technique used to effect inter-session NAL unit decoding order recovery. Decoding order recovery is performed using either inter-session timestamp alignment [RFC3550] or cross-session decoding order numbers (CS-DONs). One of the MST modes supports both

decoding order recovery techniques, so that receivers can select their preferred technique. More details can be found in Section 1.2.2.

This memo further defines three new NAL unit types. The first type is the payload content scalability information (PACSI) NAL unit, which is used to provide an informative summary of the scalability information of the data contained in an RTP packet, as well as ancillary data (e.g., CS-DON values). The second and third new NAL unit types are the empty NAL unit and the non-interleaved multi-time aggregation packet (NI-MTAP) NAL unit. The empty NAL unit is used to ensure inter-session timestamp alignment required for decoding order recovery in MST. The NI-MTAP is used as a new payload structure allowing the grouping of NAL units of different time instances in decoding order. More details about the new packet structures can be found in Section 1.2.3.

This memo also defines the signaling support for SVC transport over RTP, including a new media subtype name (H264-SVC).

A non-normative overview of the SVC codec and the payload is given in the remainder of this section.

1.1. The SVC Codec

1.1.1. Overview

SVC defines a coded video representation in which a given bitstream offers representations of the source material at different levels of fidelity (hence the term "scalable"). Scalable video coding bitstreams, or scalable bitstreams, are constructed in a pyramidal fashion: the coding process creates bitstream components that improve the fidelity of hierarchically lower components.

The fidelity dimensions offered by SVC are spatial (picture size), quality (or Signal-to-Noise Ratio (SNR)), and temporal (pictures per second). Bitstream components associated with a given level of spatial, quality, and temporal fidelity are identified using corresponding parameters in the bitstream: dependency_id, quality_id, and temporal_id (see also Section 1.1.3). The fidelity identifiers have integer values, where higher values designate components that are higher in the hierarchy. It is noted that SVC offers significant flexibility in terms of how an encoder may choose to structure the dependencies between the various components. Decoding of a particular component requires the availability of all the components it depends upon, either directly, or indirectly. An operation point

of an SVC bitstream consists of the bitstream components required to be able to decode a particular dependency_id, quality_id, and temporal_id combination.

The term "layer" is used in various contexts in this memo. For example, in the terms "Video Coding Layer" and "Network Abstraction Layer" it refers to conceptual organization levels. When referring to bitstream syntax elements such as block layer or macroblock layer, it refers to hierarchical bitstream structure levels. When used in the context of bitstream scalability, e.g., "AVC base layer", it refers to a level of representation fidelity of the source signal with a specific set of NAL units included. The correct interpretation is supported by providing the appropriate context.

SVC maintains the bitstream organization introduced in H.264/AVC. Specifically, all bitstream components are encapsulated in Network Abstraction Layer (NAL) units, which are organized as Access Units (AUs). An AU is associated with a single sampling instance in time. A subset of the NAL unit types correspond to the Video Coding Layer (VCL), and contain the coded picture data associated with the source content. Non-VCL NAL units carry ancillary data that may be necessary for decoding (e.g., parameter sets as explained below) or that facilitate certain system operations but are not needed by the decoding process itself. Coded picture data at the various fidelity dimensions are organized in slices. Within one AU, a coded picture of an operation point consists of all the coded slices required for decoding up to the particular combination of dependency_id and quality_id values at the time instance corresponding to the AU.

It is noted that the concept of temporal scalability is already present in H.264/AVC, as profiles defined in Annex A of [H.264] already support it. Specifically, in H.264/AVC, the concept of sub-sequences has been introduced to allow optional use of temporal layers through Supplemental Enhancement Information (SEI) messages. SVC extends this approach by exposing the temporal scalability information using the temporal_id parameter, alongside (and unified with) the dependency_id and quality_id values that are used for spatial and quality scalability, respectively. For coded picture data defined in Annex G of [H.264], this is accomplished by using a new type of NAL unit, namely, coded slice in scalable extension NAL unit (type 20), where the fidelity parameters are part of its header. For coded picture data that follow H.264/AVC, and to ensure compatibility with existing H.264/AVC decoders, another new type of NAL unit, namely, prefix NAL unit (type 14), has been defined to carry this header information. SVC additionally specifies a third new type of NAL unit, namely, subset sequence parameter set NAL unit (type 15), to contain sequence parameter set information for quality and spatial enhancement layers. All these three newly specified NAL

unit types (14, 15, and 20) are among those reserved in H.264/AVC and are to be ignored by decoders conforming to one or more of the profiles specified in Annex A of [H.264].

Within an AU, the VCL NAL units associated with a given dependency_id and quality_id are referred to as a "layer representation". The layer representation corresponding to the lowest values of dependency_id and quality_id (i.e., zero for both) is compliant by design to H.264/AVC. The set of VCL and associated non-VCL NAL units across all AUs in a bitstream associated with a particular combination of values of dependency_id and quality_id, and regardless of the value of temporal_id, is conceptually a scalable layer. For backward compatibility with H.264/AVC, it is important to differentiate, however, whether or not SVC-specific NAL units are present in a given bitstream. This is particularly important for the lowest fidelity values in terms of dependency_id and quality_id (zero for both), as the corresponding VCL data are compliant with H.264/AVC, and may or may not be accompanied by associated prefix NAL units. This memo therefore uses the term "AVC base layer" to designate the layer that does not contain SVC-specific NAL units, and "SVC base layer" to designate the same layer but with the addition of the associated SVC prefix NAL units. Note that the SVC specification uses the term "base layer" for what in this memo will be referred to as "AVC base layer". Similarly, it is also important to be able to differentiate, within a layer, the temporal fidelity components it contains. This memo uses the term "T0" to indicate, within a particular layer, the subset that contains the NAL units associated with temporal_id equal to 0.

SNR scalability in SVC is offered in two different ways. In what is called coarse-grain scalability (CGS), scalability is provided by including or excluding a complete layer when decoding a particular bitstream. In contrast, in medium-grain scalability (MGS), scalability is provided by selectively omitting the decoding of specific NAL units belonging to MGS layers. The selection of the NAL units to omit can be based on fixed-length fields present in the NAL unit header (see also Sections 1.1.3 and 4.2).

1.1.2. Parameter Sets

SVC maintains the parameter sets concept in H.264/AVC and introduces a new type of sequence parameter set, referred to as the subset sequence parameter set [H.264]. Subset sequence parameter sets have NAL unit type equal to 15, which is different from the NAL unit type value (7) of sequence parameter sets. VCL NAL units of NAL unit type 1 to 5 must only (indirectly) refer to sequence parameter sets, while VCL NAL units of NAL unit type 20 must only (indirectly) refer to subset sequence parameter sets. The references are indirect because

VCL NAL units refer to picture parameter sets (in their slice header), which in turn refer to regular or subset sequence parameter sets. Subset sequence parameter sets use a separate identifier value space than sequence parameter sets.

In SVC, coded picture data from different layers may use the same or different sequence and picture parameter sets. Let the variable DQId be equal to dependency_id * 16 + quality_id. At any time instant during the decoding process there is one active sequence parameter set for the layer representation with the highest value of DQId and one or more active layer SVC sequence parameter set(s) for layer representations with lower values of DQId. The active sequence parameter set or an active layer SVC sequence parameter set remains unchanged throughout a coded video sequence in the scalable layer in which the active sequence parameter set or active layer SVC sequence parameter set is referred to. This means that the referred sequence parameter set or subset sequence parameter set can only change at instantaneous decoding refresh (IDR) access units for any layer. At any time instant during the decoding process there may be one active picture parameter set (for the layer representation with the highest value of DQId) and one or more active layer picture parameter set(s) (for layer representations with lower values of DQId). The active picture parameter set or an active layer picture parameter set remains unchanged throughout a layer representation in which the active picture parameter set or active layer picture parameter set is referred to, but may change from one AU to the next.

1.1.3. NAL Unit Header

SVC extends the one-byte H.264/AVC NAL unit header by three additional octets for NAL units of types 14 and 20. The header indicates the type of the NAL unit, the (potential) presence of bit errors or syntax violations in the NAL unit payload, information regarding the relative importance of the NAL unit for the decoding process, the layer identification information, and other fields as discussed below.

The syntax and semantics of the NAL unit header are specified in [H.264], but the essential properties of the NAL unit header are summarized below for convenience.

The first byte of the NAL unit header has the following format (the bit fields are the same as defined for the one-byte H.264/AVC NAL unit header, while the semantics of some fields have changed slightly, in a backward-compatible way):

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

The semantics of the components of the NAL unit type octet, as specified in [H.264], are described briefly below. In addition to the name and size of each field, the corresponding syntax element name in [H.264] is also provided.

F: 1 bit
forbidden_zero_bit. H.264/AVC declares a value of 1 as a syntax violation.

NRI: 2 bits
nal_ref_idc. A value of "00" (in binary form) indicates that the content of the NAL unit is not used to reconstruct reference pictures for future prediction. Such NAL units can be discarded without risking the integrity of the reference pictures in the same layer. A value greater than "00" indicates that the decoding of the NAL unit is required to maintain the integrity of reference pictures in the same layer or that the NAL unit contains parameter sets.

Type: 5 bits
nal_unit_type. This component specifies the NAL unit type as defined in Table 7-1 of [H.264], and later within this memo. For a reference of all currently defined NAL unit types and their semantics, please refer to Section 7.4.1 in [H.264].

In H.264/AVC, NAL unit types 14, 15, and 20 are reserved for future extensions. SVC uses these three NAL unit types as follows: NAL unit type 14 is used for prefix NAL unit, NAL unit type 15 is used for subset sequence parameter set, and NAL unit type 20 is used for coded slice in scalable extension (see Section 7.4.1 in [H.264]). NAL unit types 14 and 20 indicate the presence of three additional octets in the NAL unit header, as shown below.

```
   +---------------+---------------+---------------+
   |0|1|2|3|4|5|6|7|0|1|2|3|4|5|6|7|0|1|2|3|4|5|6|7|
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |R|I|   PRID    |N| DID |  QID  | TID |U|D|O| RR|
   +---------------+---------------+---------------+
```

R: 1 bit
reserved_one_bit. Reserved bit for future extension. R must be equal to 1. The value of R must be ignored by decoders.

I: 1 bit
idr_flag. This component specifies whether the layer representation is an instantaneous decoding refresh (IDR) layer representation (when equal to 1) or not (when equal to 0).

PRID: 6 bits
priority_id. This flag specifies a priority identifier for the NAL unit. A lower value of PRID indicates a higher priority.

N: 1 bit
no_inter_layer_pred_flag. This flag specifies, when present in a coded slice NAL unit, whether inter-layer prediction may be used for decoding the coded slice (when equal to 1) or not (when equal to 0).

DID: 3 bits
dependency_id. This component indicates the inter-layer coding dependency level of a layer representation. At any access unit, a layer representation with a given dependency_id may be used for inter-layer prediction for coding of a layer representation with a higher dependency_id, while a layer representation with a given dependency_id shall not be used for inter-layer prediction for coding of a layer representation with a lower dependency_id.

QID: 4 bits
quality_id. This component indicates the quality level of an MGS layer representation. At any access unit and for identical dependency_id values, a layer representation with quality_id equal to ql uses a layer representation with quality_id equal to ql-1 for inter-layer prediction.

TID: 3 bits
temporal_id. This component indicates the temporal level of a layer representation. The temporal_id is associated with the frame rate, with lower values of _temporal_id corresponding to lower frame rates. A layer representation at a given temporal_id typically depends on layer representations with lower temporal_id values, but it never depends on layer representations with higher temporal_id values.

U: 1 bit
use_ref_base_pic_flag. A value of 1 indicates that only reference base pictures are used during the inter prediction process. A value of 0 indicates that the reference base pictures are not used during the inter prediction process.

D: 1 bit
discardable_flag. A value of 1 indicates that the current NAL unit is not used for decoding NAL units with values of dependency_id higher than the one of the current NAL unit, in the current and all subsequent access units. Such NAL units can be discarded without risking the integrity of layers with higher dependency_id values. discardable_flag equal to 0 indicates that the decoding of the NAL unit is required to maintain the integrity of layers with higher dependency_id.

O: 1 bit
output_flag: Affects the decoded picture output process as defined in Annex C of [H.264].

RR: 2 bits
reserved_three_2bits. Reserved bits for future extension. RR MUST be equal to "11" (in binary form). The value of RR must be ignored by decoders.

This memo extends the semantics of F, NRI, I, PRID, DID, QID, TID, U, and D per Annex G of [H.264] as described in Section 4.2.

1.2. Overview of the Payload Format

Similar to [RFC6184], this payload format can only be used to carry the raw NAL unit stream over RTP and not the bytestream format specified in Annex B of [H.264].

The design principles, transmission modes, and packetization modes as well as new payload structures are summarized in this section. It is assumed that the reader is familiar with the terminology and concepts defined in [RFC6184].

1.2.1. Design Principles

The following design principles have been observed for this payload format:

o Backward compatibility with [RFC6184] wherever possible.

- The SVC base layer or any H.264/AVC compatible subset of the SVC base layer, when transmitted in its own RTP stream, must be encapsulated using [RFC6184]. This ensures that such an RTP stream can be understood by [RFC6184] receivers.

- Media-aware network elements (MANEs) as defined in [RFC6184] are signaling-aware, rely on signaling information, and have state.

- MANEs can aggregate multiple RTP streams, possibly from multiple RTP sessions.

- MANEs can perform media-aware stream thinning (selective elimination of packets or portions thereof). By using the payload header information identifying layers within an RTP session, MANEs are able to remove packets or portions thereof from the incoming RTP packet stream. This implies rewriting the RTP headers of the outgoing packet stream, and rewriting of RTCP packets as specified in Section 7 of [RFC3550].

1.2.2. Transmission Modes and Packetization Modes

This memo allows the packetization of SVC data for both single-session transmission (SST) and multi-session transmission (MST). In the case of SST all SVC data are carried in a single RTP session. In the case of MST two or more RTP sessions are used to carry the SVC data, in accordance with the MST-specific packetization modes defined in this memo, which are based on the packetization modes defined in [RFC6184]. In MST, each RTP session is associated with one RTP stream, which may carry one or more layers.

The base layer is, by design, compatible to H.264/AVC. During transmission, the associated prefix NAL units, which are introduced by SVC and, when present, are ignored by H.264/AVC decoders, may be encapsulated within the same RTP packet stream as the H.264/AVC VCL NAL units or in a different RTP packet stream (when MST is used). For convenience, the term "AVC base layer" is used to refer to the base layer without prefix NAL units, while the term "SVC base layer" is used to refer to the base layer with prefix NAL units.

Furthermore, the base layer may have multiple temporal components (i.e., supporting different frame rates). As a result, the lowest temporal component ("T0") of the AVC or SVC base layer is used as the starting point of the SVC bitstream hierarchy.

This memo allows encapsulating in a given RTP stream any of the following three alternatives of layer combinations:

1. the T0 AVC base layer or the T0 SVC base layer only;
2. one or more enhancement layers only; or
3. the T0 SVC base layer, and one or more enhancement layers.

SST should be used in point-to-point unicast applications and, in general, whenever the potential benefit of using multiple RTP sessions does not justify the added complexity. When SST is used, the layer combination cases 1 and 3 above can be used. When an H.264/AVC compatible subset of the SVC base layer is transmitted using SST, the packetization of [RFC6184] must be used, thus ensuring compatibility with [RFC6184] receivers. When, however, one or more SVC quality or spatial enhancement layers are transmitted using SST, the packetization defined in this memo must be used. In SST, any of the three [RFC6184] packetization modes, namely, single NAL unit mode, non-interleaved mode, and interleaved mode, can be used.

MST should be used in a multicast session when different receivers may request different layers of the scalable bitstream. An operation point for an SVC bitstream, as defined in this memo, corresponds to a set of layers that together conform to one of the profiles defined in Annex A or G of [H.264] and, when decoded, offer a representation of the original video at a certain fidelity. The number of streams used in MST should be at least equal to the number of operation points that may be requested by the receivers. Depending on the application, this may result in each layer being carried in its own RTP session, or in having multiple layers encapsulated within one RTP session.

> Informative note: Layered multicast is a term commonly used to describe the application where multicast is used to transmit layered or scalable data that has been encapsulated into more than one RTP session. This application allows different receivers in the multicast session to receive different operation points of the scalable bitstream. Layered multicast, among other application examples, is discussed in more detail in Section 11.2.

When MST is used, any of the three layer combinations above can be used for each of the sessions. When an H.264/AVC compatible subset of the SVC base layer is transmitted in its own session in MST, the packetization of [RFC6184] must be used, such that [RFC6184] receivers can be part of the MST and receive only this session. For MST, this memo defines four different MST-specific packetization modes, namely, non-interleaved timestamp (NI-T) based mode, non-interleaved CS-DON (NI-C) based mode, non-interleaved combined timestamp and CS-DON mode (NI-TC), and interleaved CS-DON (I-C) based mode (detailed in Section 4.5.2). The modes differ depending on whether the SVC data are allowed to be interleaved, i.e., to be transmitted in an order different than the intended decoding order,

and they also differ in the mechanisms provided in order to recover the correct decoding order of the NAL units across the multiple RTP sessions. These four MST modes reuse the packetization modes introduced in [RFC6184] for the packetization of NAL units in each of their individual RTP sessions.

As the names of the MST packetization modes imply, the NI-T, NI-C, and NI-TC modes do not allow interleaved transmission, while the I-C mode allows interleaved transmission. With any of the three non-interleaved MST packetization modes, legacy [RFC6184] receivers with implementation of the non-interleaved mode specified in [RFC6184] can join a multi-session transmission of SVC, to receive the base RTP session encapsulated according to [RFC6184].

1.2.3. New Payload Structures

[RFC6184] specifies three basic payload structures, namely, single NAL unit packet, aggregation packet, and fragmentation unit. Depending on the basic payload structure, an RTP packet may contain a NAL unit not aggregating other NAL units, one or more NAL units aggregated in another NAL unit, or a fragment of a NAL unit not aggregating other NAL units. Each NAL unit of a type specified in [H.264] (i.e., 1 to 23, inclusive) may be carried in its entirety in a single NAL unit packet, may be aggregated in an aggregation packet, or may be fragmented and carried in a number of fragmentation unit packets. To enable aggregation or fragmentation of NAL units while still ensuring that the RTP packet payload is only composed of NAL units, [RFC6184] introduced six new NAL unit types (24-29) to be used as payload structures, selected from the NAL unit types left unspecified in [H.264].

This memo reuses all the payload structures used in [RFC6184]. Furthermore, three new types of NAL units are defined: payload content scalability information (PACSI) NAL unit, empty NAL unit, and non-interleaved multi-time aggregation packet (NI-MTAP) (specified in Sections 4.9, 4.10, and 4.7.1, respectively).

PACSI NAL units may be used for the following purposes:

- To enable MANEs to decide whether to forward, process, or discard aggregation packets, by checking in PACSI NAL units the scalability information and other characteristics of the aggregated NAL units, rather than looking into the aggregated NAL units themselves, which are defined by the video coding specification.

- To enable correct decoding order recovery in MST using the NI-C or NI-TC mode, with the help of the CS-DON information included in PACSI NAL units.

- To improve resilience to packet losses, e.g., by utilizing the following data or information included in PACSI NAL units: repeated Supplemental Enhancement Information (SEI) messages, information regarding the start and end of layer representations, and the indices to layer representations of the lowest temporal subset.

Empty NAL units may be used to enable correct decoding order recovery in MST using the NI-T or NI-TC mode. NI-MTAP NAL units may be used to aggregate NAL units from multiple access units but without interleaving.

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 [H.264]. Section 3.1.1 lists relevant definitions copied from [H.264] for convenience.

When there is discrepancy, the definitions in [H.264] take precedence. Section 3.1.2 gives definitions specific to this memo. Some of the definitions in Section 3.1.2 are also present in [RFC6184] and copied here with slight adaptations as needed.

3.1.1. Definitions from the SVC Specification

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

base layer: A bitstream subset that contains all the NAL units with the nal_unit_type syntax element equal to 1 or 5 of the bitstream and does not contain any NAL unit with the nal_unit_type syntax element equal to 14, 15, or 20 and conforms to one or more of the profiles specified in Annex A of [H.264].

base quality layer representation: The layer representation of the target dependency representation of an access unit that is associated with the quality_id syntax element equal to 0.

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

dependency representation: A subset of Video Coding Layer (VCL) NAL units within an access unit that are associated with the same value of the dependency_id syntax element, which is provided as part of the NAL unit header or by an associated prefix NAL unit. A dependency representation consists of one or more layer representations.

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

IDR picture: Instantaneous decoding refresh picture. A coded picture in which all slices of the target dependency representation within the access unit are I or EI slices that causes the decoding process to mark all reference pictures as "unused for reference" immediately after decoding the IDR picture. After the decoding of an IDR picture all following coded pictures in decoding order can be decoded without inter prediction from any picture decoded prior to the IDR picture. The first picture of each coded video sequence is an IDR picture.

layer representation: A subset of VCL NAL units within an access unit that are associated with the same values of the dependency_id and quality_id syntax elements, which are provided as part of the VCL NAL unit header or by an associated prefix NAL unit. One or more layer representations represent a dependency representation.

prefix NAL unit: A NAL unit with nal_unit_type equal to 14 that immediately precedes in decoding order a NAL unit with nal_unit_type equal to 1, 5, or 12. The NAL unit that immediately succeeds in decoding order the prefix NAL unit is referred to as the associated NAL unit. The prefix NAL unit contains data associated with the associated NAL unit, which are considered to be part of the associated NAL unit.

reference base picture: A reference picture that is obtained by decoding a base quality layer representation with the nal_ref_idc syntax element not equal to 0 and the store_ref_base_pic_flag syntax element equal to 1 of an access unit and all layer representations of the access unit that are referred to by inter-layer prediction of the base quality layer representation. A reference base picture is not an output of the decoding process, but the samples of a reference base picture may be used for inter prediction in the decoding process of subsequent pictures in decoding order. Reference base picture is a collective term for a reference base field or a reference base frame.

scalable bitstream: A bitstream with the property that one or more bitstream subsets that are not identical to the scalable bitstream form another bitstream that conforms to the SVC specification [H.264].

target dependency representation: The dependency representation of an access unit that is associated with the largest value of the dependency_id syntax element for all dependency representations of the access unit.

target layer representation: The layer representation of the target dependency representation of an access unit that is associated with the largest value of the quality_id syntax element for all layer representations of the target dependency representation of the access unit.

3.1.2. Definitions Specific to This Memo

anchor layer representation: An anchor layer representation is such a layer representation that, if decoding of the operation point corresponding to the layer starts from the access unit containing this layer representation, all the following layer representations of the layer, in output order, can be correctly decoded. The output order is defined in [H.264] as the order in which decoded pictures are output from the decoded picture buffer of the decoder. As H.264 does not specify the picture display process, this more general term is used instead of display order. An anchor layer representation is a random access point to the layer the anchor layer representation belongs. However, some layer representations, succeeding an anchor layer representation in decoding order but preceding the anchor layer representation in output order, may refer to earlier layer representations for inter prediction, and hence the decoding may be incorrect if random access is performed at the anchor layer representation.

AVC base layer: The subset of the SVC base layer in which all prefix NAL units (type 14) are removed. Note that this is equivalent to the term "base layer" as defined in Annex G of [H.264].

base RTP session: When multi-session transmission is used, the RTP session that carries the RTP stream containing the T0 AVC base layer or the T0 SVC base layer, and zero or more enhancement layers. This RTP session does not depend on any other RTP session as indicated by mechanisms defined in Section 7.2.3. The base RTP session may carry NAL units of NAL unit type equal to 14 and 15.

decoding order number (DON): A field in the payload structure or a derived variable indicating NAL unit decoding order. Values of DON are in the range of 0 to 65535, inclusive. After reaching the maximum value, the value of DON wraps around to 0. Note that this definition also exists in [RFC6184] in exactly the same form.

Empty NAL unit: A NAL unit with NAL unit type equal to 31 and sub-type equal to 1. An empty NAL unit consists of only the two-byte NAL unit header with an empty payload.

enhancement RTP session: When multi-session transmission is used, an RTP session that is not the base RTP session. An enhancement RTP session typically contains an RTP stream that depends on at least one other RTP session as indicated by mechanisms defined in Section 7.2.3. A lower RTP session to an enhancement RTP session is an RTP session on which the enhancement RTP session depends. The lowest RTP session for a receiver is the RTP session that does not depend on any other RTP session received by the receiver. The highest RTP session for a receiver is the RTP session on which no other RTP session received by the receiver depends.

cross-session decoding order number (CS-DON): A derived variable indicating NAL unit decoding order number over all NAL units within all the session-multiplexed RTP sessions that carry the same SVC bitstream.

default level: The level indicated by the profile-level-id parameter. In Session Description Protocol (SDP) Offer/Answer, the level is downgradable, i.e., the answer may either use the default level or a lower level. Note that this definition also exists in [RFC6184] in a slightly different form.

default sub-profile: The subset of coding tools, which may be all coding tools of one profile or the common subset of coding tools of more than one profile, indicated by the profile-level-id parameter. In SDP Offer/Answer, the default sub-profile must be used in a

symmetric manner, i.e., the answer must either use the same sub-profile as the offer or reject the offer. Note that this definition also exists in [RFC6184] in a slightly different form.

enhancement layer: A layer in which at least one of the values of dependency_id or quality_id is higher than 0, or a layer in which none of the NAL units is associated with the value of temporal_id equal to 0. An operation point constructed using the maximum temporal_id, dependency_id, and quality_id values associated with an enhancement layer may or may not conform to one or more of the profiles specified in Annex A of [H.264].

H.264/AVC compatible: The property of a bitstream subset of conforming to one or more of the profiles specified in Annex A of [H.264].

intra layer representation: A layer representation that contains only slices that use intra prediction, and hence do not refer to any earlier layer representation in decoding order in the same layer. Note that in SVC intra prediction includes intra-layer intra prediction as well as inter-layer intra prediction.

layer: A bitstream subset in which all NAL units of type 1, 5, 12, 14, or 20 have the same values of dependency_id and quality_id, either directly through their NAL unit header (for NAL units of type 14 or 20) or through association to a prefix (type 14) NAL unit (for NAL unit type 1, 5, or 12). A layer may contain NAL units associated with more than one values of temporal_id.

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. Note that this definition also exists in [RFC6184] in exactly the same form.

> 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 Secure Real-time Transport Protocol (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].

multi-session transmission: The transmission mode in which the SVC stream is transmitted over multiple RTP sessions. Dependency between RTP sessions MUST be signaled according to Section 7.2.3 of this memo.

NAL unit decoding order: A NAL unit order that conforms to the constraints on NAL unit order given in Section G.7.4.1.2 in [H.264]. Note that this definition also exists in [RFC6184] in a slightly different form.

NALU-time: The value that the RTP timestamp would have if the NAL unit would be transported in its own RTP packet. Note that this definition also exists in [RFC6184] in exactly the same form.

operation point: An operation point is identified by a set of values of temporal_id, dependency_id, and quality_id. A bitstream corresponding to an operation point can be constructed by removing all NAL units associated with a higher value of dependency_id, and all NAL units associated with the same value of dependency_id but higher values of quality_id or temporal_id. An operation point bitstream conforms to at least one of the profiles defined in Annex A or G of [H.264], and offers a representation of the original video signal at a certain fidelity.

> Informative note: Additional NAL units may be removed (with lower dependency_id or same dependency_id but lower quality_id) if they are not required for decoding the bitstream at the particular operation point. The resulting bitstream, however, may no longer conform to any of the profiles defined in Annex A or G of [H.264].

operation point representation: The set of all NAL units of an operation point within the same access unit.

RTP packet stream: A sequence of RTP packets with increasing sequence numbers (except for wrap-around), identical payload type 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 layers.

single-session transmission: The transmission mode in which the SVC bitstream is transmitted over a single RTP session.

SVC base layer: The layer that includes all NAL units associated with dependency_id and quality_id values both equal to 0, including prefix NAL units (NAL unit type 14).

SVC enhancement layer: A layer in which at least one of the values of dependency_id or quality_id is higher than 0. An operation point constructed using the maximum dependency_id and quality_id values and any temporal_id value associated with an SVC enhancement layer does not conform to any of the profiles specified in Annex A of [H.264].

SVC NAL unit: A NAL unit of NAL unit type 14, 15, or 20 as specified in Annex G of [H.264].

SVC NAL unit header: A four-byte header resulting from the addition of a three-byte SVC-specific header extension added in NAL unit types 14 and 20.

SVC RTP session: Either the base RTP session or an enhancement RTP session.

T0 AVC base layer: A subset of the AVC base layer constructed by removing all VCL NAL units associated with temporal_id values higher than 0 and non-VCL NAL units and SEI messages associated only with the VCL NAL units being removed.

T0 SVC base layer: A subset of the SVC base layer constructed by removing all VCL NAL units associated with temporal_id values higher than 0 as well as prefix NAL units, non-VCL NAL units, and SEI messages associated only with the VCL NAL units being removed.

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. Note that this definition also exists in [RFC6184] in exactly the same form.

3.2. Abbreviations

In addition to the abbreviations defined in [RFC6184], the following abbreviations are used in this memo.

CGS:	Coarse-Grain Scalability
CS-DON:	Cross-Session Decoding Order Number
MGS:	Medium-Grain Scalability
MST:	Multi-Session Transmission
PACSI:	Payload Content Scalability Information
SST:	Single-Session Transmission
SNR:	Signal-to-Noise Ratio
SVC:	Scalable Video Coding

4. RTP Payload Format

4.1. RTP Header Usage

In addition to Section 5.1 of [RFC6184], the following rules apply.

o Setting of the M bit:

The M bit of an RTP packet for which the packet payload is an NI-MTAP MUST be equal to 1 if the last NAL unit, in decoding order, of the access unit associated with the RTP timestamp is contained in the packet.

o Setting of the RTP timestamp:

For an RTP packet for which the packet payload is an empty NAL unit, the RTP timestamp must be set according to Section 4.10.

For an RTP packet for which the packet payload is a PACSI NAL unit, the RTP timestamp MUST be equal to the NALU-time of the next non-PACSI NAL unit in transmission order. Recall that the NALU-time of a NAL unit in an MTAP is defined in [RFC6184] as the value that the RTP timestamp would have if that NAL unit would be transported in its own RTP packet.

o Setting of the SSRC:

For both SST and MST, the SSRC values MUST be set according to [RFC3550].

4.2. NAL Unit Extension and Header Usage

4.2.1. NAL Unit Extension

This memo specifies a NAL unit extension mechanism to allow for introduction of new types of NAL units, beyond the three NAL unit types left undefined in [RFC6184] (i.e., 0, 30, and 31). The extension mechanism utilizes the NAL unit type value 31 and is specified as follows. When the NAL unit type value is equal to 31, the one-byte NAL unit header consisting of the F, NRI, and Type fields as specified in Section 1.1.3 is extended by one additional octet, which consists of a 5-bit field named Subtype and three 1-bit fields named J, K, and L, respectively. The additional octet is shown in the following figure.

```
+---------------+
|0|1|2|3|4|5|6|7|
+-+-+-+-+-+-+-+-+
| Subtype |J|K|L|
+---------------+
```

The Subtype value determines the (extended) NAL unit type of this NAL unit. The interpretation of the fields J, K, and L depends on the Subtype. The semantics of the fields are as follows.

When Subtype is equal to 1, the NAL unit is an empty NAL unit as specified in Section 4.10. When Subtype is equal to 2, the NAL unit is an NI-MTAP NAL unit as specified in Section 4.7.1. All other values of Subtype (0, 3-31) are reserved for future extensions, and receivers MUST ignore the entire NAL unit when Subtype is equal to any of these reserved values.

4.2.2. NAL Unit Header Usage

The structure and semantics of the NAL unit header according to the H.264 specification [H.264] were introduced in Section 1.1.3. This section specifies the extended semantics of the NAL unit header fields F, NRI, I, PRID, DID, QID, TID, U, and D, according to this memo. When the Type field is equal to 31, the semantics of the fields in the extension NAL unit header were specified in Section 4.2.1.

The semantics of F specified in Section 5.3 of [RFC6184] also apply in this memo. That is, a value of 0 for F indicates that the NAL unit type octet and payload should not contain bit errors or other syntax violations, whereas a value of 1 for F indicates that the NAL unit type octet and payload may contain bit errors or other syntax violations. MANEs SHOULD set the F bit to indicate bit errors in the NAL unit.

For NRI, for a bitstream conforming to one of the profiles defined in Annex A of [H.264] and transported using [RFC6184], the semantics specified in Section 5.3 of [RFC6184] apply, i.e., NRI also indicates the relative importance of NAL units. For a bitstream conforming to one of the profiles defined in Annex G of [H.264] and transported using this memo, in addition to the semantics specified in Annex G of [H.264], NRI also indicates the relative importance of NAL units within a layer.

For I, in addition to the semantics specified in Annex G of [H.264], according to this memo, MANEs MAY use this information to protect NAL units with I equal to 1 better than NAL units with I equal to 0. MANEs MAY also utilize information of NAL units with I equal to 1 to

decide when to forward more packets for an RTP packet stream. For example, when it is detected that spatial layer switching has happened such that the operation point has changed to a higher value of DID, MANEs MAY start to forward NAL units with the higher value of DID only after forwarding a NAL unit with I equal to 1 with the higher value of DID.

Note that, in the context of this section, "protecting a NAL unit" means any RTP or network transport mechanism that could improve the probability of successful delivery of the packet conveying the NAL unit, including applying a Quality of Service (QoS) enabled network, Forward Error Correction (FEC), retransmissions, and advanced scheduling behavior, whenever possible.

For PRID, the semantics specified in Annex G of [H.264] apply. Note that MANEs implementing unequal error protection MAY use this information to protect NAL units with smaller PRID values better than those with larger PRID values, for example, by including only the more important NAL units in a FEC protection mechanism. The importance for the decoding process decreases as the PRID value increases.

For DID, QID, or TID, in addition to the semantics specified in Annex G of [H.264], according to this memo, values of DID, QID, or TID indicate the relative importance in their respective dimension. A lower value of DID, QID, or TID indicates a higher importance if the other two components are identical. MANEs MAY use this information to protect more important NAL units better than less important NAL units.

For U, in addition to the semantics specified in Annex G of [H.264], according to this memo, MANEs MAY use this information to protect NAL units with U equal to 1 better than NAL units with U equal to 0.

For D, in addition to the semantics specified in Annex G of [H.264], according to this memo, MANEs MAY use this information to determine whether a given NAL unit is required for successfully decoding a certain Operation Point of the SVC bitstream, hence to decide whether to forward the NAL unit.

4.3. Payload Structures

The NAL unit structure is central to H.264/AVC, [RFC6184], as well as SVC and this memo. In H.264/AVC and SVC, all coded bits for representing a video signal are encapsulated in NAL units. In [RFC6184], each RTP packet payload is structured as a NAL unit, which contains one or a part of one NAL unit specified in H.264/AVC, or aggregates one or more NAL units specified in H.264/AVC.

[RFC6184] specifies three basic payload structures (in Section 5.2 of [RFC6184]): single NAL unit packet, aggregation packet, fragmentation unit, and six new types (24 to 29) of NAL units. The value of the Type field of the RTP packet payload header (i.e., the first byte of the payload) may be equal to any value from 1 to 23 for a single NAL unit packet, any value from 24 to 27 for an aggregation packet, and 28 or 29 for a fragmentation unit.

In addition to the NAL unit types defined originally for H.264/AVC, SVC defines three new NAL unit types specifically for SVC: coded slice in scalable extension NAL units (type 20), prefix NAL units (type 14), and subset sequence parameter set NAL units (type 15), as described in Section 1.1.

This memo further introduces three new types of NAL units, PACSI NAL unit (NAL unit type 30) as specified in Section 4.9, empty NAL unit (type 31, subtype 1) as specified in Section 4.10, and NI-MTAP NAL unit (type 31, subtype 2) as specified in Section 4.7.1.

The RTP packet payload structure in [RFC6184] is maintained with slight extensions in this memo, as follows. Each RTP packet payload is still structured as a NAL unit, which contains one or a part of one NAL unit specified in H.264/AVC and SVC, or contains one PACSI NAL unit or one empty NAL unit, or aggregates zero or more NAL units specified in H.264/AVC and SVC, zero or one PACSI NAL unit, and zero or more empty NAL units.

In this memo, one of the three basic payload structures, fragmentation unit, remains the same as in [RFC6184], and the other two, single NAL unit packet and aggregation packet, are extended as follows. The value of the Type field of the payload header may be equal to any value from 1 to 23, inclusive, and 30 to 31, inclusive, for a single NAL unit packet, and any value from 24 to 27, inclusive, and 31, for an aggregation packet. When the Type field of the payload header is equal to 31 and the Subtype field of the payload header is equal to 2, the packet is an aggregation packet (containing an NI-MTAP NAL unit). When the Type field of the payload header is equal to 31 and the Subtype field of the payload header is equal to 1, the packet is a single NAL unit packet (containing an empty NAL unit).

Note that, in this memo, the length of the payload header varies depending on the value of the Type field in the first byte of the RTP packet payload. If the value is equal to 14, 20, or 30, the first four bytes of the packet payload form the payload header; otherwise, if the value is equal to 31, the first two bytes of the payload form the payload header; otherwise, the payload header is the first byte of the packet payload.

Table 1 lists the NAL unit types introduced in SVC and this memo and where they are described in this memo. Table 2 summarizes the basic payload structure types for all NAL unit types when they are directly used as RTP packet payloads according to this memo. Table 3 summarizes the NAL unit types allowed to be aggregated (i.e., used as aggregation units in aggregation packets) or fragmented (i.e., carried in fragmentation units) according to this memo.

Table 1. NAL unit types introduced in SVC and this memo

Type	Subtype	NAL Unit Name	Section Numbers
14	-	Prefix NAL unit	1.1
15	-	Subset sequence parameter set	1.1
20	-	Coded slice in scalable extension	1.1
30	-	PACSI NAL unit	4.9
31	0	reserved	4.2.1
31	1	Empty NAL unit	4.10
31	2	NI-MTAP	4.7.1
31	3-31	reserved	4.2.1

Table 2. Basic payload structure types for all NAL unit types when they are directly used as RTP packet payloads

Type	Subtype	Basic Payload Structure
0	-	reserved
1-23	-	Single NAL Unit Packet
24-27	-	Aggregation Packet
28-29	-	Fragmentation Unit
30	-	Single NAL Unit Packet
31	0	reserved
31	1	Single NAL Unit Packet
31	2	Aggregation Packet
31	3-31	reserved

Table 3. Summary of the NAL unit types allowed to be aggregated or fragmented (yes = allowed, no = disallowed, - = not applicable/not specified)

Type	Subtype	STAP-A	STAP-B	MTAP16	MTAP24	FU-A	FU-B	NI-MTAP
0	-	-	-	-	-	-	-	-
1-23	-	yes	yes	yes	yes	yes	yes	yes
24-29	-	no	no	no	no	no	no	no
30	-	yes	yes	yes	yes	no	no	yes
31	0	-	-	-	-	-	-	-
31	1	yes	no	no	no	no	no	yes
31	2	no	no	no	no	no	no	no
31	3-31	-	-	-	-	-	-	-

4.4. Transmission Modes

This memo enables transmission of an SVC bitstream over one or more RTP sessions. If only one RTP session is used for transmission of the SVC bitstream, the transmission mode is referred to as single-session transmission (SST); otherwise (more than one RTP session is used for transmission of the SVC bitstream), the transmission mode is referred to as multi-session transmission (MST).

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

If the OPTIONAL mst-mode media type parameter (see Section 7.1) is not present, SST MUST be used; otherwise (mst-mode is present), MST MUST be used.

4.5. Packetization Modes

4.5.1. Packetization Modes for Single-Session Transmission

When SST is in use, Section 5.4 of [RFC6184] applies with the following extensions.

The packetization modes specified in Section 5.4 of [RFC6184], namely, single NAL unit mode, non-interleaved mode, and interleaved mode, are also referred to as session packetization modes. Table 4 summarizes the allowed session packetization modes for SST.

Table 4. Summary of allowed session packetization modes (denoted as "Session Mode" for simplicity) for SST (yes = allowed, no = disallowed)

Session Mode	Allowed
Single NAL Unit Mode	yes
Non-Interleaved Mode	yes
Interleaved Mode	yes

For NAL unit types in the range of 0 to 29, inclusive, the NAL unit types allowed to be directly used as packet payloads for each session packetization mode are the same as specified in Section 5.4 of [RFC6184]. For other NAL unit types, which are newly introduced in this memo, the NAL unit types allowed to be directly used as packet payloads for each session packetization mode are summarized in Table 5.

Table 5. New NAL unit types allowed to be directly used as packet payloads for each session packetization mode (yes = allowed, no = disallowed, - = not applicable/not specified)

Type	Subtype	Single NAL Unit Mode	Non-Interleaved Mode	Interleaved Mode
30	-	yes	no	no
31	0	-	-	-
31	1	yes	yes	no
31	2	no	yes	no
31	3-31	-	-	-

4.5.2. Packetization Modes for Multi-Session Transmission

For MST, this memo specifies four MST packetization modes:

- Non-interleaved timestamp based mode (NI-T);
- Non-interleaved cross-session decoding order number (CS-DON) based mode (NI-C);
- Non-interleaved combined timestamp and CS-DON mode (NI-TC); and
- Interleaved CS-DON (I-C) mode.

These four modes differ in two ways. First, they differ in terms of whether NAL units are required to be transmitted within each RTP session in decoding order (i.e., non-interleaved), or they are allowed to be transmitted in a different order (i.e., interleaved).

Second, they differ in the mechanisms they provide in order to recover the correct decoding order of the NAL units across all RTP sessions involved.

The NI-T, NI-C, and NI-TC modes do not allow interleaving, and are thus targeted for systems that require relatively low end-to-end latency, e.g., conversational systems. The I-C mode allows interleaving and is thus targeted for systems that do not require very low end-to-end latency. The benefits of interleaving are the same as that of the interleaved mode specified in [RFC6184].

The NI-T mode uses timestamps to recover the decoding order of NAL units, whereas the NI-C and I-C modes both use the CS-DON mechanism (explained later) to do so. The NI-TC mode provides both timestamps and the CS-DON method; receivers in this case may choose to use either method for performing decoding order recovery. The MST packetization mode in use MUST be signaled by the value of the OPTIONAL mst-mode media type parameter. The used MST packetization mode governs which session packetization modes are allowed in the associated RTP sessions, which in turn govern which NAL unit types are allowed to be directly used as RTP packet payloads.

Table 6 summarizes the allowed session packetization modes for NI-T, NI-C, and NI-TC. Table 7 summarizes the allowed session packetization modes for I-C.

Table 6. Summary of allowed session packetization modes (denoted as "Session Mode" for simplicity) for NI-T, NI-C, and NI-TC (yes = allowed, no = disallowed)

Session Mode	Base Session	Enhancement Session
Single NAL Unit Mode	yes	no
Non-Interleaved Mode	yes	yes
Interleaved Mode	no	no

Table 7. Summary of allowed session packetization modes (denoted as "Session Mode" for simplicity) for I-C (yes = allowed, no = disallowed)

Session Mode	Base Session	Enhancement Session
Single NAL Unit Mode	no	no
Non-Interleaved Mode	no	no
Interleaved Mode	yes	yes

For NAL unit types in the range of 0 to 29, inclusive, the NAL unit types allowed to be directly used as packet payloads for each session packetization mode are the same as specified in Section 5.4 of [RFC6184]. For other NAL unit types, which are newly introduced in this memo, the NAL unit types allowed to be directly used as packet payloads for each allowed session packetization mode for NI-T, NI-C, NI-TC, and I-C are summarized in Tables 8, 9, 10, and 11, respectively.

Table 8. New NAL unit types allowed to be directly used as packet payloads for each allowed session packetization mode when NI-T is in use (yes = allowed, no = disallowed, - = not applicable/not specified)

Type	Subtype	Single NAL Unit Mode	Non-Interleaved Mode
30	-	yes	no
31	0	-	-
31	1	yes	yes
31	2	no	yes
31	3-31	-	-

Table 9. New NAL unit types allowed to be directly used as packet payloads for each allowed session packetization mode when NI-C is in use (yes = allowed, no = disallowed, - = not applicable/not specified)

Type	Subtype	Single NAL Unit Mode	Non-Interleaved Mode
30	-	yes	yes
31	0	-	-
31	1	no	no
31	2	no	yes
31	3-31	-	-

Table 10. New NAL unit types allowed to be directly used as packet payloads for each allowed session packetization mode when NI-TC is in use (yes = allowed, no = disallowed, - = not applicable/not specified)

Type	Subtype	Single NAL Unit Mode	Non-Interleaved Mode
30	-	yes	yes
31	0	-	-
31	1	yes	yes
31	2	no	yes
31	3-31	-	-

Table 11. New NAL unit types allowed to be directly used as packet payloads for the allowed session packetization mode when I-C is in use (yes = allowed, no = disallowed, - = not applicable/not specified)

Type	Subtype	Interleaved Mode
30	-	no
31	0	-
31	1	no
31	2	no
31	3-31	-

When MST is in use and the MST packetization mode in use is NI-C, empty NAL units (type 31, subtype 1) MUST NOT be used, i.e., no RTP packet is allowed to contain one or more empty NAL units.

When MST is in use and the MST packetization mode in use is I-C, both empty NAL units (type 31, subtype 1) and NI-MTAP NAL units (type 31, subtype 2) MUST NOT be used, i.e., no RTP packet is allowed to contain one or more empty NAL units or an NI-MTAP NAL unit.

4.6. Single NAL Unit Packets

Section 5.6 of [RFC6184] applies with the following extensions.

The payload of a single NAL unit packet MAY be a PACSI NAL unit (Type 30) or an empty NAL unit (Type 31 and Subtype 1), in addition to a NAL unit with NAL unit type equal to any value from 1 to 23, inclusive.

If the Type field of the first byte of the payload is not equal to 31, the payload header is the first byte of the payload. Otherwise, (the Type field of the first byte of the payload is equal to 31), the payload header is the first two bytes of the payload.

4.7. Aggregation Packets

In addition to Section 5.7 of [RFC6184], the following applies in this memo.

4.7.1. Non-Interleaved Multi-Time Aggregation Packets (NI-MTAPs)

One new NAL unit type introduced in this memo is the non-interleaved multi-time aggregation packet (NI-MTAP). An NI-MTAP consists of one or more non-interleaved multi-time aggregation units.

The NAL units contained in NI-MTAPs MUST be aggregated in decoding order.

A non-interleaved multi-time aggregation unit for the NI-MTAP consists of 16 bits of unsigned size information of the following NAL unit (in network byte order), and 16 bits (in network byte order) of timestamp offset (TS offset) for the NAL unit. The structure is presented in Figure 1. The starting or ending position of an aggregation unit within a packet may or may not be on a 32-bit word boundary. The NAL units in the NI-MTAP are ordered in NAL unit decoding order.

The Type field of the NI-MTAP MUST be set equal to "31".

The F bit MUST be set to 0 if all the F bits of the aggregated NAL units are zero; otherwise, it MUST be set to 1.

The value of NRI MUST be the maximum value of NRI across all NAL units carried in the NI-MTAP packet.

The field Subtype MUST be equal to 2.

If the field J is equal to 1, the optional DON field MUST be present for each of the non-interleaved multi-time aggregation units. For SST, the J field MUST be equal to 0. For MST, in the NI-T mode the J field MUST be equal to 0, whereas in the NI-C or NI-TC mode the J field MUST be equal to 1. When the NI-C or NI-TC mode is in use, the DON field, when present, MUST represent the CS-DON value for the particular NAL unit as defined in Section 6.2.2.

The fields K and L MUST be both equal to 0.

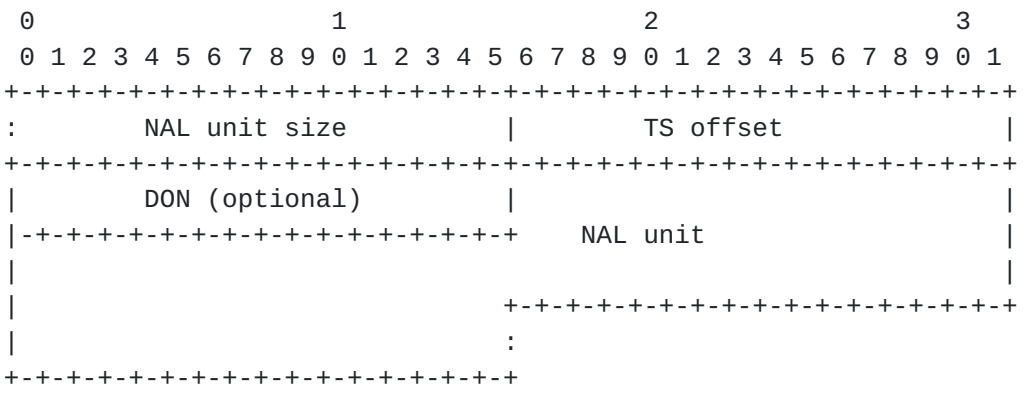

Figure 1. Non-interleaved multi-time aggregation unit for NI-MTAP

Let TS be the RTP timestamp of the packet carrying the NAL unit. Recall that the NALU-time of a NAL unit in an MTAP is defined in [RFC6184] as the value that the RTP timestamp would have if that NAL unit would be transported in its own RTP packet. The timestamp offset field MUST be set to a value equal to the value of the following formula:

```
if NALU-time >= TS, TS offset = NALU-time - TS
else, TS offset = NALU-time + (2^32 - TS)
```

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

> Informative note: The "earliest" multi-time aggregation unit is the one that would have the smallest extended RTP timestamp among all the aggregation units of an NI-MTAP if the aggregation units were encapsulated in single NAL unit packets. An extended timestamp is a timestamp that has more than 32 bits and is capable of counting the wraparound of the timestamp field, thus enabling one to determine the smallest value if the timestamp wraps. Such an "earliest" aggregation unit may or may not be the first one in the order in which the aggregation units are encapsulated in an NI-MTAP. The "earliest" NAL unit need not be the same as the first NAL unit in the NAL unit decoding order either.

Figure 2 presents an example of an RTP packet that contains an NI-MTAP that contains two non-interleaved multi-time aggregation units, labeled as 1 and 2 in the figure.

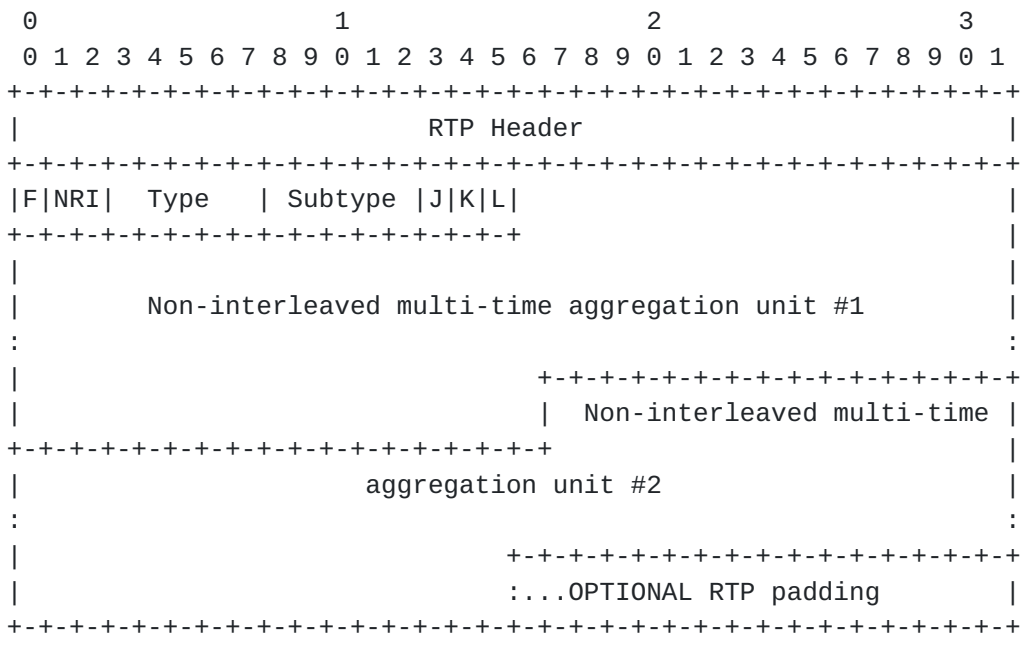

Figure 2. An RTP packet including an NI-MTAP containing two non-interleaved multi-time aggregation units

4.8. Fragmentation Units (FUs)

Section 5.8 of [RFC6184] applies.

> Informative note: In case a NAL unit with the four-byte SVC NAL unit header is fragmented, the three-byte SVC-specific header extension is considered as part of the NAL unit payload. That is, the three-byte SVC-specific header extension is only available in the first fragment of the fragmented NAL unit.

4.9. Payload Content Scalability Information (PACSI) NAL Unit

Another new type of NAL unit specified in this memo is the payload content scalability information (PACSI) NAL unit. The Type field of PACSI NAL units MUST be equal to 30 (a NAL unit type value left unspecified in [H.264] and [RFC6184]). A PACSI NAL unit MAY be carried in a single NAL unit packet or an aggregation packet, and MUST NOT be fragmented.

PACSI NAL units may be used for the following purposes:

- To enable MANEs to decide whether to forward, process, or discard aggregation packets, by checking in PACSI NAL units the scalability information and other characteristics of the

aggregated NAL units, rather than looking into the aggregated NAL units themselves, which are defined by the video coding specification;

- To enable correct decoding order recovery in MST using the NI-C or NI-TC mode, with the help of the CS-DON information included in PACSI NAL units; and

- To improve resilience to packet losses, e.g., by utilizing the following data or information included in PACSI NAL units: repeated Supplemental Enhancement Information (SEI) messages, information regarding the start and end of layer representations, and the indices to layer representations of the lowest temporal subset.

PACSI NAL units MAY be ignored in the NI-T mode without affecting the decoding order recovery process.

When a PACSI NAL unit is present in an aggregation packet, the following applies.

- The PACSI NAL unit MUST be the first aggregated NAL unit in the aggregation packet.

- There MUST be at least one additional aggregated NAL unit in the aggregation packet.

- The RTP header fields and the payload header fields of the aggregation packet are set as if the PACSI NAL unit was not included in the aggregation packet.

- If the aggregation packet is an MTAP16, MTAP24, or NI-MTAP with the J field equal to 1, the decoding order number (DON) for the PACSI NAL unit MUST be set to indicate that the PACSI NAL unit has an identical DON to the first NAL unit in decoding order among the remaining NAL units in the aggregation packet.

When a PACSI NAL unit is included in a single NAL unit packet, it is associated with the next non-PACSI NAL unit in transmission order, and the RTP header fields of the packet are set as if the next non-PACSI NAL unit in transmission order was included in a single NAL unit packet.

The PACSI NAL unit structure is as follows. The first four octets are exactly the same as the four-byte SVC NAL unit header discussed in Section 1.1.3. They are followed by one octet containing several flags, then five optional octets, and finally zero or more SEI NAL units. Each SEI NAL unit is preceded by a 16-bit unsigned size field

(in network byte order) that indicates the size of the following NAL unit in bytes (excluding these two octets, but including the NAL unit header octet of the SEI NAL unit). Figure 3 illustrates the PACSI NAL unit structure and an example of a PACSI NAL unit containing two SEI NAL units.

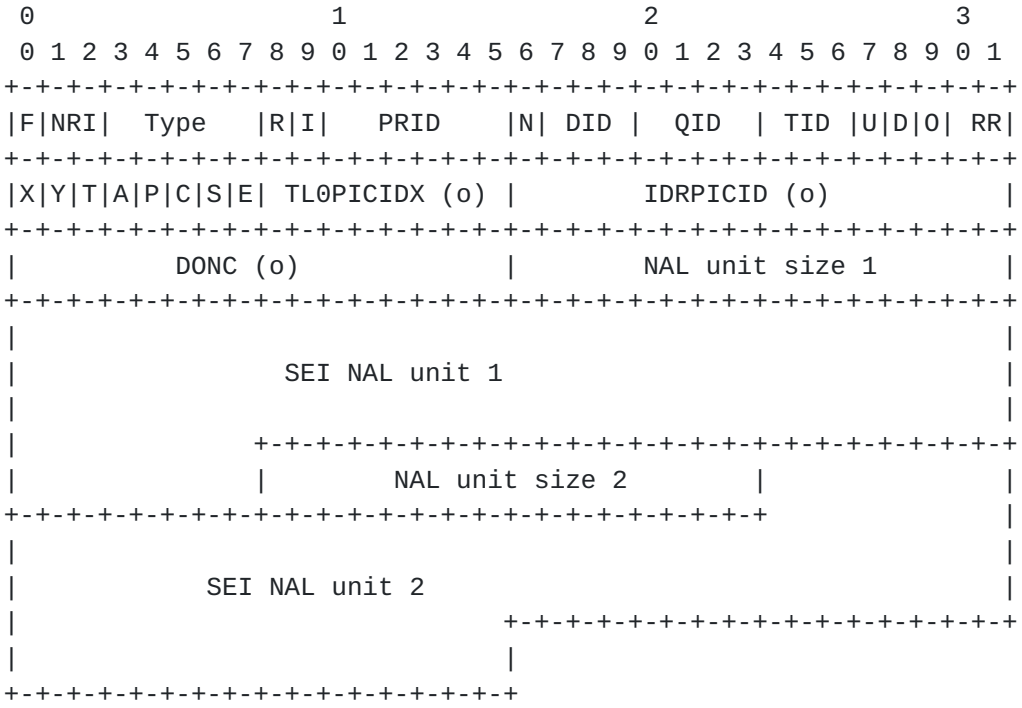

Figure 3. PACSI NAL unit structure. Fields suffixed by "(o)" are OPTIONAL.

The bits A, P, and C are specified only if the bit X is equal to 1. The bits S and E are specified, and the fields TL0PICIDX and IDRPICID are present, only if the bit Y is equal to 1. The field DONC is present only if the bit T is equal to 1. The field T MUST be equal to 0 if the PACSI NAL unit is contained in an STAP-B, MTAP16, MTAP24, or NI-MTAP with the J field equal to 1.

The values of the fields in PACSI NAL unit MUST be set as follows.

o The F bit MUST be set to 1 if the F bit in at least one of the remaining NAL units in the aggregation packet is equal to 1 (when the PACSI NAL unit is included in an aggregation packet) or if the next non-PACSI NAL unit in transmission order has the F bit equal to 1 (when the PACSI NAL unit is included in a single NAL unit packet). Otherwise, the F bit MUST be set to 0.

- The NRI field MUST be set to the highest value of NRI field among all the remaining NAL units in the aggregation packet (when the PACSI NAL unit is included in an aggregation packet) or the value of the NRI field of the next non-PACSI NAL unit in transmission order (when the PACSI NAL unit is included in a single NAL unit packet).

- The Type field MUST be set to 30.

- The R bit MUST be set to 1. Receivers MUST ignore the value of R.

- The I bit MUST be set to 1 if the I bit of at least one of the remaining NAL units in the aggregation packet is equal to 1 (when the PACSI NAL unit is included in an aggregation packet) or if the I bit of the next non-PACSI NAL unit in transmission order is equal to 1 (when the PACSI NAL unit is included in a single NAL unit packet). Otherwise, the I bit MUST be set to 0.

- The PRID field MUST be set to the lowest value of the PRID values of the remaining NAL units in the aggregation packet (when the PACSI NAL unit is included in an aggregation packet) or the PRID value of the next non-PACSI NAL unit in transmission order (when the PACSI NAL unit is included in a single NAL unit packet).

- The N bit MUST be set to 1 if the N bit of all the remaining NAL units in the aggregation packet is equal to 1 (when the PACSI NAL unit is included in an aggregation packet) or if the N bit of the next non-PACSI NAL unit in transmission order is equal to 1 (when the PACSI NAL unit is included in a single NAL unit packet). Otherwise, the N bit MUST be set to 0.

- The DID field MUST be set to the lowest value of the DID values of the remaining NAL units in the aggregation packet (when the PACSI NAL unit is included in an aggregation packet) or the DID value of the next non-PACSI NAL unit in transmission order (when the PACSI NAL unit is included in a single NAL unit packet).

- The QID field MUST be set to the lowest value of the QID values of the remaining NAL units with the lowest value of DID in the aggregation packet (when the PACSI NAL unit is included in an aggregation packet) or the QID value of the next non-PACSI NAL unit in transmission order (when the PACSI NAL unit is included in a single NAL unit packet).

- The TID field MUST be set to the lowest value of the TID values of the remaining NAL units with the lowest value of DID in the aggregation packet (when the PACSI NAL unit is included in an

aggregation packet) or the TID value of the next non-PACSI NAL unit in transmission order (when the PACSI NAL unit is included in a single NAL unit packet).

- The U bit MUST be set to 1 if the U bit of at least one of the remaining NAL units in the aggregation packet is equal to 1 (when the PACSI NAL unit is included in an aggregation packet) or if the U bit of the next non-PACSI NAL unit in transmission order is equal to 1 (when the PACSI NAL unit is included in a single NAL unit packet). Otherwise, the U bit MUST be set to 0.

- The D bit MUST be set to 1 if the D value of all the remaining NAL units in the aggregation packet is equal to 1 (when the PACSI NAL unit is included in an aggregation packet) or if the D bit of the next non-PACSI NAL unit in transmission order is equal to 1 (when the PACSI NAL unit is included in a single NAL unit packet). Otherwise, the D bit MUST be set to 0.

- The O bit MUST be set to 1 if the O bit of at least one of the remaining NAL units in the aggregation packet is equal to 1 (when the PACSI NAL unit is included in an aggregation packet) or if the O bit of the next non-PACSI NAL unit in transmission order is equal to 1 (when the PACSI NAL unit is included in a single NAL unit packet). Otherwise, the O bit MUST be set to 0.

- The RR field MUST be set to "11" (in binary form). Receivers MUST ignore the value of RR.

- If the X bit is equal to 1, the bits A, P, and C are specified as below. Otherwise, the bits A, P, and C are unspecified, and receivers MUST ignore the values of these bits. The X bit SHOULD be identical for all the PACSI NAL units in all the RTP sessions carrying the same SVC bitstream.

- If the Y bit is equal to 1, the OPTIONAL fields TL0PICIDX and IDRPICID MUST be present and specified as below, and the bits S and E are also specified as below. Otherwise, the fields TL0PICIDX and IDRPICID MUST NOT be present, while the S and E bits are unspecified and receivers MUST ignore the values of these bits. The Y bit MUST be identical for all the PACSI NAL units in all the RTP sessions carrying the same SVC bitstream. The Y bit MUST be equal to 0 when the parameter packetization-mode is equal to 2.

- If the T bit is equal to 1, the OPTIONAL field DONC MUST be present and specified as below. Otherwise, the field DONC MUST NOT be present. The field T MUST be equal to 0 if the PACSI NAL unit is contained in an STAP-B, MTAP16, MTAP24, or NI-MTAP.

- The A bit MUST be set to 1 if at least one of the remaining NAL units in the aggregation packet belongs to an anchor layer representation (when the PACSI NAL unit is included in an aggregation packet) or if the next non-PACSI NAL unit in transmission order belongs to an anchor layer representation (when the PACSI NAL unit is included in a single NAL unit packet). Otherwise, the A bit MUST be set to 0.

  Informative note: The A bit indicates whether CGS or spatial layer switching at a non-IDR layer representation (a layer representation with nal_unit_type not equal to 5 and idr_flag not equal to 1) can be performed. With some picture coding structures a non-IDR intra layer representation can be used for random access. Compared to using only IDR layer representations, higher coding efficiency can be achieved. The H.264/AVC or SVC solution to indicate the random accessibility of a non-IDR intra layer representation is using a recovery point SEI message. The A bit offers direct access to this information, without having to parse the recovery point SEI message, which may be buried deeply in an SEI NAL unit. Furthermore, the SEI message may or may not be present in the bitstream.

- The P bit MUST be set to 1 if all the remaining NAL units in the aggregation packet have redundant_pic_cnt greater than 0 (when the PACSI NAL unit is included in an aggregation packet) or the next non-PACSI NAL unit in transmission order has redundant_pic_cnt greater than 0 (when the PACSI NAL unit is included in a single NAL unit packet). Otherwise, the P bit MUST be set to 0.

  Informative note: The P bit indicates whether a packet can be discarded because it contains only redundant slice NAL units. Without this bit, the corresponding information can be obtained from the syntax element redundant_pic_cnt, which is contained in the variable-length coded slice header.

- The C bit MUST be set to 1 if at least one of the remaining NAL units in the aggregation packet belongs to an intra layer representation (when the PACSI NAL unit is included in an aggregation packet) or if the next non-PACSI NAL unit in transmission order belongs to an intra layer representation (when the PACSI NAL unit is included in a single NAL unit packet). Otherwise, the C bit MUST be set to 0.

  Informative note: The C bit indicates whether a packet contains intra slices, which may be the only packets to be forwarded, e.g., when the network conditions are particularly adverse.

- The S bit MUST be set to 1, if the first NAL unit following the PACSI NAL unit in an aggregation packet is the first VCL NAL unit, in decoding order, of a layer representation (when the PACSI NAL unit is included in an aggregation packet) or if the next non-PACSI NAL unit in transmission order is the first VCL NAL unit, in decoding order, of a layer representation(when the PACSI NAL unit is included in a single NAL unit packet). Otherwise, the S bit MUST be set to 0.

- The E bit MUST be set to 1, if the last NAL unit following the PACSI NAL unit in an aggregation packet is the last VCL NAL unit, in decoding order, of a layer representation (when the PACSI NAL unit is included in an aggregation packet) or if the next non-PACSI NAL unit in transmission order is the last VCL NAL unit, in decoding order, of a layer representation (when the PACSI NAL unit is included in a single NAL unit packet). Otherwise, the E bit MUST be set to 0.

  Informative note: In an aggregation packet it is always possible to detect the beginning or end of a layer representation by detecting changes in the values of dependency_id, quality_id, and temporal_id in NAL unit headers, except from the first and last NAL units of a packet. The S or E bits are used to provide this information, for both single NAL unit and aggregation packets, so that previous or following packets do not have to be examined. This enables MANEs to detect slice loss and take proper action such as requesting a retransmission as soon as possible, as well as to allow efficient playout buffer handling similarly to the M bit present in the RTP header. The M bit in the RTP header still indicates the end of an access unit, not the end of a layer representation.

- When present, the TL0PICIDX field MUST be set to equal to tl0_dep_rep_idx as specified in Annex G of [H.264] for the layer representation containing the first NAL unit following the PACSI NAL unit in the aggregation packet (when the PACSI NAL unit is included in an aggregation packet) or containing the next non-PACSI NAL unit in transmission order (when the PACSI NAL unit is included in a single NAL unit packet).

- When present, the IDRPICID field MUST be set to equal to effective_idr_pic_id as specified in Annex G of [H.264] for the layer representation containing the first NAL unit following the PACSI NAL unit in the aggregation packet (when the PACSI NAL unit is included in an aggregation packet) or containing the next non-PACSI NAL unit in transmission order (when the PACSI NAL unit is included in a single NAL unit packet).

Informative note: The TL0PICIDX and IDRPICID fields enable the detection of the loss of layer representations in the most important temporal layer (with temporal_id equal to 0) by receivers as well as MANEs. SVC provides a solution that uses SEI messages, which are harder to parse and may or may not be present in the bitstream. When the PACSI NAL unit is part of an NI-MTAP packet, it is possible to infer the correct values of tl0_dep_rep_idx and idr_pic_id for all layer representations contained in the NI-MTAP by following the rules that specify how these parameters are set as given in Annex G of [H.264] and by detecting the different layer representations contained in the NI-MTAP packet by detecting changes in the values of dependency_id_, quality_id, and temporal_id in the NAL unit headers as well as using the S and E flags. The only exception is if NAL units of an IDR picture are present in the NI-MTAP in a position other than the first NAL unit following the PACSI NAL unit, in which case the value of idr_pic_id cannot be inferred. In this case the NAL unit has to be partially parsed to obtain the idr_pic_id. Note that, due to the large size of IDR pictures, their inclusion in an NI-MTAP, and especially in a position other than the first NAL unit following the PACSI NAL unit, may be neither practical nor useful.

- When present, the field DONC indicates the cross-session decoding order number (CS-DON) for the first of the remaining NAL units in the aggregation packet (when the PACSI NAL unit is included in an aggregation packet) or the CS-DON of the next non-PACSI NAL unit in transmission order (when the PACSI NAL unit is included in a single NAL unit packet). CS-DON is further discussed in Section 4.11.

The PACSI NAL unit MAY include a subset of the SEI NAL units associated with the access unit to which the first non-PACSI NAL unit in the aggregation packet belongs, and MUST NOT contain SEI NAL units associated with any other access unit.

Informative note: In H.264/AVC and SVC, within each access unit, SEI NAL units must appear before any VCL NAL unit in decoding order. Therefore, without using PACSI NAL units, SEI messages are typically only conveyed in the first of the packets carrying an access unit. Senders may repeat SEI NAL units in PACSI NAL units, so that they are repeated in more than one packet and thus increase robustness against packet losses. Receivers may use the repeated SEI messages in place of missing SEI messages.

For a PACSI NAL unit included in an aggregation packet, an SEI message SHOULD NOT be included in the PACSI NAL unit and also included in one of the remaining NAL units contained in the same aggregation packet.

4.10. Empty NAL unit

An empty NAL unit MAY be included in a single NAL unit packet, an STAP-A or an NI-MTAP packet. Empty NAL units MUST have an RTP timestamp (when transported in a single NAL unit packet) or NALU-time (when transported in an aggregation packet) that is associated with an access unit for which there exists at least one NAL unit of type 1, 5, or 20. When MST is used, the type 1, 5, or 20 NAL unit may be in a different RTP session. Empty NAL units may be used in the decoding order recovery process of the NI-T mode as described in Section 5.2.1.

The packet structure is shown in the following figure.

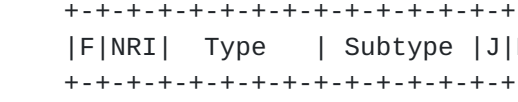

Figure 4. Empty NAL unit structure.

The fields MUST be set as follows:

F MUST be equal to 0
NRI MUST be equal to 3
Type MUST be equal to 31
Subtype MUST be equal to 1
J MUST be equal to 0
K MUST be equal to 0
L MUST be equal to 0

4.11. Decoding Order Number (DON)

The DON concept is introduced in [RFC6184] and is used to recover the decoding order when interleaving is used within a single session. Section 5.5 of [RFC6184] applies when using SST.

When using MST, it is necessary to recover the decoding order across the various RTP sessions regardless if interleaving is used or not. In addition to the timestamp mechanism described later, the CS-DON mechanism is an extension of the DON facility that can be used for this purpose, and is defined in the following section.

4.11.1. Cross-Session DON (CS-DON) for Multi-Session Transmission

The cross-session decoding order number (CS-DON) is a number that indicates the decoding order of NAL units across all RTP sessions involved in MST. It is similar to the DON concept in [RFC6184], but contrary to [RFC6184] where the DON was used only for interleaved

packetization, in this memo it is used not only in the interleaved MST mode (I-C) but also in two of the non-interleaved MST modes (NI-C and NI-TC).

When the NI-C or NI-TC MST modes are in use, the packetization of each session MUST be as specified in Section 5.2.2. In PACSI NAL units the CS-DON value is explicitly coded in the field DONC. For non-PACSI NAL units the CS-DON value is derived as follows. Let SN indicate the RTP sequence number of a packet.

- For each non-PACSI NAL unit carried in a session using the single NAL unit session packetization mode, the CS-DON value of the NAL unit is equal to (DONC_prev_PACSI + SN_diff - 1) % 65536, wherein "%" is the modulo operation, DONC_prev_PACSI is the DONC value of the previous PACSI NAL unit with the same NALU-time as the current NAL unit, and SN_diff is calculated as follows:

  ```
  if SN1 > SN2, SN_diff = SN1 - SN2
  else SN_diff = SN2 + 65536 - SN1
  ```

  where SN1 and SN2 are the SNs of the current NAL unit and the previous PACSI NAL unit with the same NALU-time, respectively.

- For non-PACSI NAL units carried in a session using the non-interleaved session packetization mode, the CS-DON value of each non-PACSI NAL unit is derived as follows.

  For a non-PACSI NAL unit in a single NAL unit packet, the following applies.

  If the previous PACSI NAL unit is contained in a single NAL unit packet, the CS-DON value of the NAL unit is calculated as above;

  otherwise (the previous PACSI NAL unit is contained in an STAP-A packet), the CS-DON value of the NAL unit is calculated as above, with DONC_prev_PACSI being replaced by the CS-DON value of the previous non-PACSI NAL unit in decoding order (i.e., the CS-DON value of the last NAL unit of the STAP-A packet).

  For a non-PACSI NAL unit in an STAP-A packet, the following applies.

  If the non-PACSI NAL unit is the first non-PACSI NAL unit in the STAP-A packet, the CS-DON value of the NAL unit is equal to DONC of the PACSI NAL unit in the STAP-A packet;

otherwise (the non-PACSI NAL unit is not the first non-PACSI NAL unit in the STAP-A packet), the CS-DON value of the NAL unit is equal to: (the CS-DON value of the previous non-PACSI NAL unit in decoding order + 1) % 65536, wherein "%" is the modulo operation.

For a non-PACSI NAL unit in a number of FU-A packets, the CS-DON value of the NAL unit is calculated the same way as when the single NAL unit session packetization mode is in use, with SN1 being the SN value of the first FU-A packet.

For a non-PACSI NAL unit in an NI-MTAP packet, the CS-DON value is equal to the value of the DON field of the non-interleaved multi-time aggregation unit.

When the I-C MST packetization mode is in use, the DON values derived according to [RFC6184] for all the NAL units in each of the RTP sessions MUST indicate CS-DON values.

5. Packetization Rules

Section 6 of [RFC6184] applies in this memo, with the following additions.

5.1. Packetization Rules for Single-Session Transmission

All receivers MUST support the single NAL unit packetization mode to provide backward compatibility to endpoints supporting only the single NAL unit mode of [RFC6184]. However, the use of single NAL unit packetization mode (packetization-mode equal to 0) SHOULD be avoided whenever possible, because encapsulating NAL units of small sizes in their own packets (e.g., small NAL units containing parameter sets, prefix NAL units, or SEI messages) is less efficient due to the packet header overhead.

All receivers MUST support the non-interleaved mode.

Informative note: The non-interleaved mode of [RFC6184] does allow an application to encapsulate a single NAL unit in a single RTP packet. Historically, the single NAL unit mode has been included in [RFC6184] only for compatibility with ITU-T Rec. H.241 Annex A [H.241]. There is no point in carrying this historic ballast towards a new application space such as the one provided with SVC. The implementation complexity increase for supporting the additional mechanisms of the non-interleaved mode (namely, STAP-A and FU-A) is minor, whereas the benefits are significant. As a result, the support of STAP-A and FU-A is required. Additionally,

support for two of the three NAL unit types defined in this memo, namely, empty NAL units and NI-MTAP is needed, as specified in Section 4.5.1.

A NAL unit of small size SHOULD be encapsulated in an aggregation packet together with one or more other NAL units. For example, non-VCL NAL units such as access unit delimiters, parameter sets, or SEI NAL units are typically small.

A prefix NAL unit and the NAL unit with which it is associated, and which follows the prefix NAL unit in decoding order, SHOULD be included in the same aggregation packet whenever an aggregation packet is used for the associated NAL unit, unless this would violate session MTU constraints or if fragmentation units are used for the associated NAL unit.

Informative note: Although the prefix NAL unit is ignored by an H.264/AVC decoder, it is necessary in the SVC decoding process.

Given the small size of the prefix NAL unit, it is best if it is transported in the same RTP packet as its associated NAL unit.

When only an H.264/AVC compatible subset of the SVC base layer is transmitted in an RTP session, the subset MUST be encapsulated according to [RFC6184]. This way, an [RFC6184] receiver will be able to receive the H.264/AVC compatible bitstream subset.

When a set of layers including one or more SVC enhancement layers is transmitted in an RTP session, the set SHOULD be carried in one RTP stream that SHOULD be encapsulated according to this memo.

5.2. Packetization Rules for Multi-Session Transmission

When MST is used, the packetization rules specified in Section 5.1 still apply. In addition, the following packetization rules MUST be followed, to ensure that decoding order of NAL units carried in the sessions can be correctly recovered for each of the MST packetization modes using the de-packetization process specified in Section 6.2.

The NI-T and NI-TC modes both use timestamps to recover the decoding order. In order to be able to do so, it is necessary for the RTP packet stream to contain data for all sampling instances of a given RTP session in all enhancement RTP sessions that depend on the given RTP session. The NI-C and I-C modes do not have this limitation, and use the CS-DON values as a means to explicitly indicate decoding order, either directly coded in PACSI NAL units, or inferred from

them using the packetization rules. It is noted that the NI-TC mode offers both alternatives and it is up to the receiver to select which one to use.

5.2.1. NI-T/NI-TC Packetization Rules

When using the NI-T mode and a PACSI NAL unit is present, the T bit MUST be equal to 0, i.e., the DONC field MUST NOT be present.

When using the NI-T mode, the optional parameters sprop-mst-remux-buf-size, sprop-remux-buf-req, remux-buf-cap, sprop-remux-init-buf-time, sprop-mst-max-don-diff MUST NOT be present.

When the NI-T or NI-TC MST mode is in use, the following applies.

If one or more NAL units of an access unit of sampling time instance t is present in RTP session A, then one or more NAL units of the same access unit MUST be present in any enhancement RTP session that depends on RTP session A.

> Informative note: The mapping between RTP and NTP format timestamps is conveyed in RTCP SR packets. In addition, the mechanisms for faster media timestamp synchronization discussed in [RFC6051] may be used to speed up the acquisition of the RTP-to-wall-clock mapping.
>
> Informative note: The rule above may require the insertion of NAL units, typically when temporal scalability is used, i.e., an enhancement RTP session does not contain any NAL units for an access unit with a particular NTP timestamp (media timestamp), which, however, is present in a lower enhancement RTP session or the base RTP session. There are two ways to insert additional NAL units in order to satisfy this rule:
>
> - One option for adding additional NAL units is to use empty NAL units (defined in Section 4.10), which can be used by the process described in Section 6.2.1 for the access unit reordering process.
>
> - Additional NAL units may also be added by the encoder itself, for example, by transmitting coded data that simply instruct the decoder to repeat the previous picture. This option, however, may be difficult to use with pre-encoded content.

If a packet must be inserted in order to satisfy the above rule, e.g., in case of a MANE generating multiple RTP streams out of a single RTP stream, the inserted packet must have an RTP timestamp that maps to the same wall-clock time (in NTP format) as the one of

the RTP timestamp of any packet of the access unit present in any lower enhancement RTP session or the base RTP session. This is easy to accomplish if the NAL unit or the packet can be inserted at the time of the RTP stream generation, since the media timestamp (NTP timestamp) must be the same for the inserted packet and the packet of the corresponding access unit. If there is no knowledge of the media time at RTP stream generation or if the RTP streams are not generated at the same instance, this can be also applied later in the transmission process. In this case the NTP timestamp of the inserted packet can be calculated as follows.

Assume that a packet A2 of an access unit with RTP timestamp TS_A2 is present in base RTP session A, and that no packet of that access unit is present in enhancement RTP session B, as shown in Figure 5. Thus, a packet B2 must be inserted into session B following the rule above. The most recent RTCP sender report in session A carries NTP timestamp NTP_A and the RTP timestamp TS_A. The sender report in session B with a lower NTP timestamp than NTP_A is NTP_B, and carries the RTP timestamp TS_B.

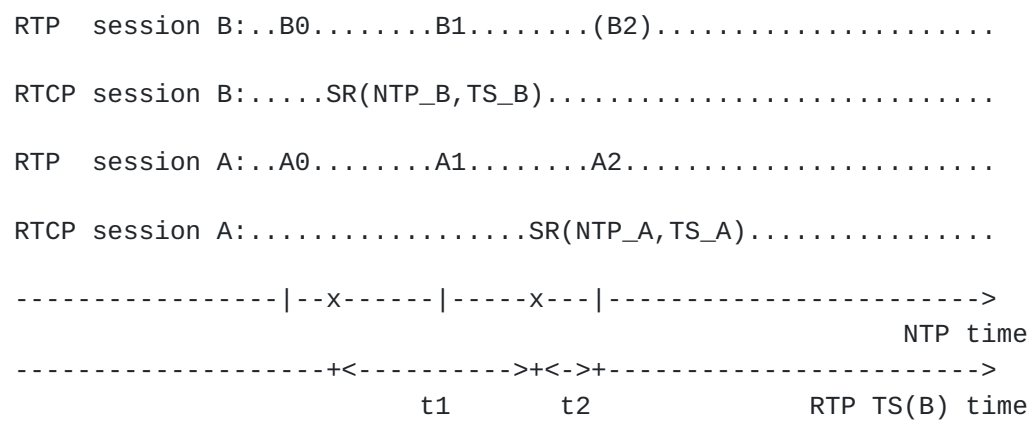

Figure 5. Example calculation of RTP timestamp for packet insertion in an enhancement layer RTP session

The vertical bars ("|")in the NTP time line in the figure above indicate that access unit data is present in at least one of the sessions. The "x" marks indicate the times of the sender reports. The RTP timestamp time line for session B, shown right below the NTP time line, indicates two time segments, t1 and t2. t1 is the time difference between the sender reports between the two sessions, expressed in RTP timestamp clock ticks, and t2 is the time difference from the session A sender report to the A2 packet, again expressed in RTP timestamp clock ticks. The sum of these differences is added to

the RTP timestamp of the session report from session B in order to derive the correct RTP timestamp for the inserted packet B2. In other words:

```
TS_B2 = TS_B + t1 + t2
```

Let toRTP() be a function that calculates the RTP time difference (in clock ticks of the used clock) given an NTP timestamp difference, and effRTPdiff() be a function that calculates the effective difference between two timestamps, including wraparounds:

```
effRTPdiff( ts1, ts2 ):

    if( ts1 <= ts2 ) then
        effRTPdiff := ts1-ts2
    else
        effRTPDiff := (4294967296 + ts2) - ts1
```

We have:

```
t1 = toRTP(NTP_A - NTP_B) and t2 = effRTPdiff(TS_A2, TS_A)
```

Hence in order to generate the RTP timestamp TS_B2 for the inserted packet B2, the RTP timestamp for packet B2 TS_B2 can be calculated as follows.

```
TS_B2 =  TS_B + toRTP(NTP_A - NTP_B) +  effRTPdiff(TS_A2, TS_A)
```

5.2.2. NI-C/NI-TC Packetization Rules

When the NI-C or NI-TC MST mode is in use, the following applies for each of the RTP sessions.

- For each single NAL unit packet containing a non-PACSI NAL unit, the previous packet, if present, MUST have the same RTP timestamp as the single NAL unit packet, and the following applies.
  - If the NALU-time of the non-PACSI NAL unit is not equal to the NALU-time of the previous non-PACSI NAL unit in decoding order, the previous packet MUST contain a PACSI NAL unit containing the DONC field.
- In an STAP-A packet the first NAL unit in the STAP-A packet MUST be a PACSI NAL unit containing the DONC field.
- For an FU-A packet the previous packet MUST have the same RTP timestamp as the FU-A packet, and the following applies.

- If the FU-A packet is the start of the fragmented NAL unit, the following applies.
  - If the NALU-time of the fragmented NAL unit is not equal to the NALU-time of the previous non-PACSI NAL unit in decoding order, the previous packet MUST contain a PACSI NAL unit containing the DONC field;
  - Otherwise, (the NALU-time of the fragmented NAL unit is equal to the NALU-time of the previous non-PACSI NAL unit in decoding order), the previous packet MAY contain a PACSI NAL unit containing the DONC field.
- Otherwise, if the FU-A packet is the end of the fragmented NAL unit, the following applies.
  - If the next non-PACSI NAL unit in decoding order has NALU-time equal to the NALU-time of the fragmented NAL unit, and is carried in a number of FU-A packets or a single NAL unit packet, the next packet MUST be a single NAL unit packet containing a PACSI NAL unit containing the DONC field.
  - Otherwise (the FU-A packet is neither the start nor the end of the fragmented NAL unit), the previous packet MUST be a FU-A packet.

- For each single NAL unit packet containing a PACSI NAL unit, if present, the PACSI NAL unit MUST contain the DONC field.
- When the optional media type parameter sprop-mst-csdon-always-present is equal to 1, the session packetization mode in use MUST be the non-interleaved mode, and only STAP-A and NI-MTAP packets can be used.

5.2.3. I-C Packetization Rules

When the I-C MST packetization mode is in use, the following applies.

- When a PACSI NAL unit is present, the T bit MUST be equal to 0, i.e., the DONC field is not present, and the Y bit MUST be equal to 0, i.e., the TL0PICIDX and IDRPICID are not present.

5.2.4. Packetization Rules for Non-VCL NAL Units

NAL units that do not directly encode video slices are known in H.264 as non-VCL NAL units. Non-VCL units that are only used by, or only relevant to, enhancement RTP sessions SHOULD be sent in the lowest session to which they are relevant.

Some senders, however, such as those sending pre-encoded data, may be unable to easily determine which non-VCL units are relevant to which session. Thus, non-VCL NAL units MAY, instead, be sent in a session on which the session using these non-VCL NAL units depends (e.g., the base RTP session).

If a non-VCL unit is relevant to more than one RTP session, neither of which depends on the other(s), the NAL unit MAY be sent in another session on which all these sessions depend.

5.2.5. Packetization Rules for Prefix NAL Units

Section 5.1 of this memo applies, with the following addition. If the base layer is sent in a base RTP session using [RFC6184], prefix NAL units MAY be sent in the lowest enhancement RTP session rather than in the base RTP session.

6. De-Packetization Process

6.1. De-Packetization Process for Single-Session Transmission

For single-session transmission, where a single RTP session is used, the de-packetization process specified in Section 7 of [RFC6184] applies.

6.2. De-Packetization Process for Multi-Session Transmission

For multi-session transmission, where more than one RTP session is used to receive data from the same SVC bitstream, the de-packetization process is specified as follows.

As for a single RTP session, the general concept behind the de-packetization process is to reorder NAL units from transmission order to the NAL unit decoding order.

The sessions to be received MUST be identified by mechanisms specified in Section 7.2.3. An enhancement RTP session typically contains an RTP stream that depends on at least one other RTP session, as indicated by mechanisms defined in Section 7.2.3. A lower RTP session to an enhancement RTP session is an RTP session on which the enhancement RTP session depends. The lowest RTP session for a receiver is the base RTP session, which does not depend on any other RTP session received by the receiver. The highest RTP session for a receiver is the RTP session on which no other RTP session received by the receiver depends.

For each of the RTP sessions, the RTP reception process as specified in RFC 3550 is applied. Then the received packets are passed into the payload de-packetization process as defined in this memo.

The decoding order of the NAL units carried in all the associated RTP sessions is then recovered by applying one of the following subsections, depending on which of the MST packetization modes is in use.

6.2.1. Decoding Order Recovery for the NI-T and NI-TC Modes

The following process MUST be applied when the NI-T packetization mode is in use. The following process MAY be applied when the NI-TC packetization mode is in use.

The process is based on RTP session dependency signaling, RTP sequence numbers, and timestamps.

The decoding order of NAL units within an RTP packet stream in RTP session is given by the ordering of sequence numbers SN of the RTP packets that contain the NAL units, and the order of appearance of NAL units within a packet.

Timing information according to the media timestamp TS, i.e., the NTP timestamp as derived from the RTP timestamp of an RTP packet, is associated with all NAL units contained in the same RTP packet received in an RTP session.

For NI-MTAP packets the NALU-time is derived for each contained NAL unit by using the "TS offset" value in the NI-MTAP packet as defined in Section 4.10, and is used instead of the RTP packet timestamp to derive the media timestamp, e.g., using the NTP wall clock as provided via RTCP sender reports. NAL units contained in fragmentation packets are handled as defragmented, entire NAL units with their own media timestamps. All NAL units associated with the same value of media timestamp TS are part of the same access unit AU(TS). Any empty NAL units SHOULD be kept as, effectively, access unit indicators in the reordering process. Empty NAL units and PACSI NAL units SHOULD be removed before passing access unit data to the decoder.

> Informative note: These empty NAL units are used to associate NAL units present in other RTP sessions with RTP sessions not containing any data for an access unit of a particular time instance. They act as access unit indicators in sessions that would otherwise contain no data for the particular access unit. The presence of these NAL units is ensured by the packetization rules in Section 5.2.1.

It is assumed that the receiver has established an operation point (DID, QID, and TID values), and has identified the highest enhancement RTP session for this operation point. The decoding order of NAL units from multiple RTP streams in multiple RTP sessions MUST be recovered into a single sequence of NAL units, grouped into access units, by performing any process equivalent to the following steps. The general process is described in Section 4.2 of [RFC6051]. For convenience the instructions of [RFC6051] are repeated and applied to NAL units rather than to full RTP packets. Additionally, SVC-specific extensions to the procedure in Section 4.2. of [RFC6051] are presented in the following list:

- The process should be started with the NAL units received in the highest RTP session with the first media timestamp TS (in NTP format) available in the session's (de-jittering) buffer. It is assumed that packets in the de-jittering buffer are already stored in RTP sequence number order.

- Collect all NAL units associated with the same value of media timestamp TS, starting from the highest RTP session, from all the (de-jittering) buffers of the received RTP sessions. The collected NAL units will be those associated with the access unit AU(TS).

- Place the collected NAL units in the order of session dependency as derived by the dependency indication as specified in Section 7.2.3, starting from the lowest RTP session.

- Place the session ordered NAL units in decoding order within the particular access unit by satisfying the NAL unit ordering rules for SVC access units, as described in the informative algorithm provided in Section 6.2.1.1.

- Remove NI-MTAP and any PACSI NAL units from the access unit AU(TS).

- The access units can then be transferred to the decoder. Access units AU(TS) are transferred to the decoder in the order of appearance (given by the order of RTP sequence numbers) of media timestamp values TS in the highest RTP session associated with access unit AU(TS).

  Informative note: Due to packet loss it is possible that not all sessions may have NAL units present for the media timestamp value TS present in the highest RTP session. In such a case, an algorithm may: a) proceed to the next complete access unit with NAL units present in all the received RTP sessions; or b) consider a new highest RTP

session, the highest RTP session for which the access unit is complete, and apply the process above. The algorithm may return to the original highest RTP session when a complete and error-free access unit that contains NAL units in all the sessions is received.

The following gives an informative example.

The example shown in Figure 6 refers to three RTP sessions A, B, and C containing an SVC bitstream transmitted as 3 sources. In the example, the dependency signaling (described in Section 7.2.3) indicates that session A is the base RTP session, B is the first enhancement RTP session and depends on A, and C is the second enhancement RTP session and depends on A and B. A hierarchical picture coding prediction structure is used, in which session A has the lowest frame rate and sessions B and C have the same but higher frame rate.

The figure shows NAL units contained in RTP packets that are stored in the de-jittering buffer at the receiver for session de-packetization. The NAL units are already reordered according to their RTP sequence number order and, if within an aggregation packet, according to the order of their appearance within the aggregation packet. The figure indicates for the received NAL units the decoding order within the sessions, as well as the associated media (NTP) timestamps ("TS[..]"). NAL units of the same access unit within a session are grouped by "(.,.)" and share the same media timestamp TS, which is shown at the bottom of the figure. Note that the timestamps are not in increasing order since, in this example, the decoding order is different from the output/display order.

The process first proceeds to the NAL units associated with the first media timestamp TS[1] present in the highest session C and removes/ignores all preceding (in decoding order) NAL units to NAL units with TS[1] in each of the de-jittering buffers of RTP sessions A, B, and C. Then, starting from session C, the first media timestamp available in decoding order (TS[1]) is selected and NAL units starting from RTP session A, and sessions B and C are placed in order of the RTP session dependency as required by Section 7.2.3 of this memo (in the example for TS[1]: first session B and then session C) into the access unit AU(TS[1]) associated with media timestamp TS[1]. Then the next media timestamp TS[3] in order of appearance in the highest RTP session C is processed and the process described above is repeated. Note that there may be access units with no NAL units present, e.g., in the lowest RTP session A (see, e.g., TS[1]). With TS[8], the first access unit with NAL units present in all the RTP sessions appears in the buffers.

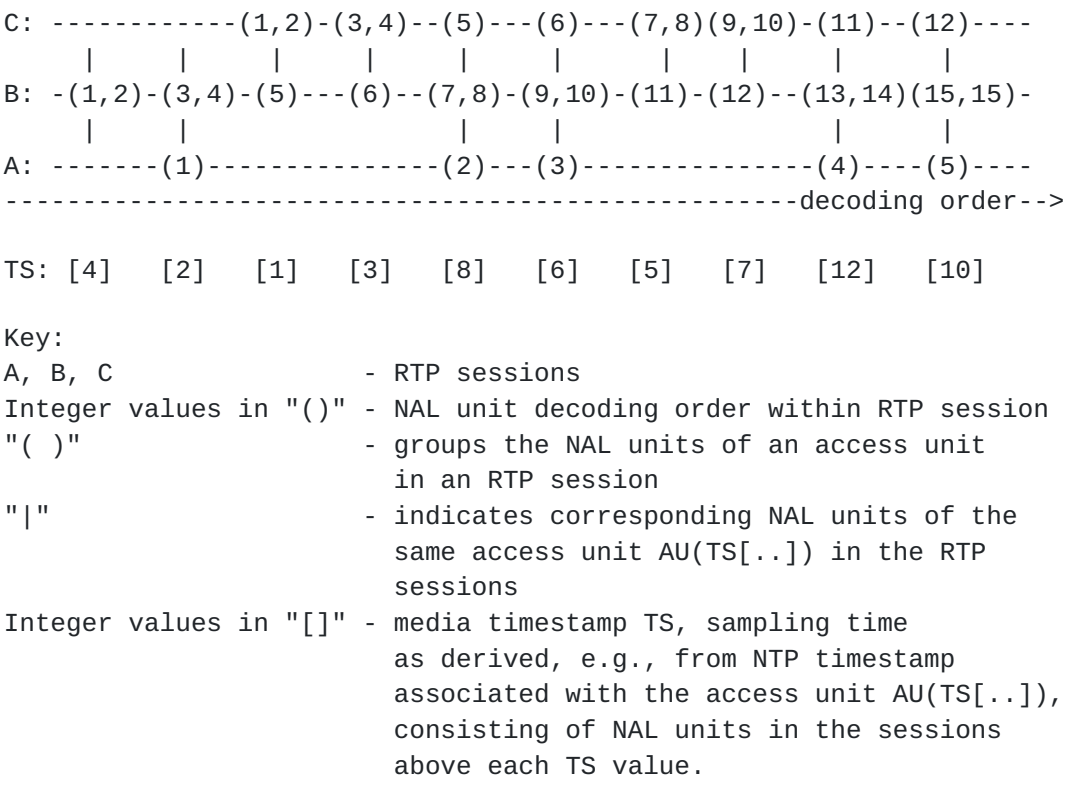

Figure 6. Example of decoding order recovery in multi-source transmission.

6.2.1.1. Informative Algorithm for NI-T Decoding Order Recovery within an Access Unit

Within an access unit, the [H.264] specification (Sections 7.4.1.2.3 and G.7.4.1.2.3) constrains the valid decoding order of NAL units.

These constraints make it possible to reconstruct a valid decoding order for the NAL units of an access unit based only on the order of NAL units in each session, the NAL unit headers, and Supplemental Enhancement Information message headers.

This section specifies an informative algorithm to reconstruct a valid decoding order for NAL units within an access unit. Other NAL unit orderings may also be valid; however, any compliant NAL unit ordering will describe the same video stream and ancillary data as the one produced by this algorithm.

An actual implementation, of course, needs only to behave "as if" this reordering is done. In particular, NAL units that are discarded by an implementation's decoding process do not need to be reordered.

In this algorithm, NAL units within an access unit are first ordered by NAL unit type, in the order specified in Table 12 below, except from NAL unit type 14, which is handled specially as described in the table. NAL units of the same type are then ordered as specified for the type, if necessary.

For the purposes of this algorithm, "session order" is the order of NAL units implied by their transmission order within an RTP session. For the non-interleaved and single NAL unit modes, this is the RTP sequence number order coupled with the order of NAL units within an aggregation unit.

Table 12. Ordering of NAL unit types within an Access Unit

Type	Description / Comments
9	Access unit delimiter
7	Sequence parameter set
13	Sequence parameter set extension
15	Subset sequence parameter set
8	Picture parameter set
16-18	Reserved
6	Supplemental enhancement information (SEI) If an SEI message with a first payload of 0 (Buffering Period) is present, it must be the first SEI message. If SEI messages with a Scalable Nesting (30) payload and a nested payload of 0 (Buffering Period) are present, these then follow the first SEI message. Such an SEI message with the all_layer_representations_in_au_flag equal to 1 is placed first, followed by any others, sorted in increasing order of DQId. All other SEI messages follow in any order.
14	Prefix NAL unit in scalable extension
1	Coded slice of a non-IDR picture
5	Coded slice of an IDR picture

NAL units of type 1 or 5 will be sent within only a single session for any given access unit. They are placed in session order. (Note: Any given access unit will contain only NAL units of type 1 or type 5, not both.)

If NAL units of type 14 are present, every NAL unit of type 1 or 5 is prefixed by a NAL unit of type 14. (Note: Within an access unit, every NAL unit of type 14 is identical, so correlation of type 14 NAL units with the other NAL units is not necessary.)

12 Filler data

The only restriction of filler data NAL units within an access unit is that they shall not precede the first VCL NAL unit with the same access unit.

19 Coded slice of an auxiliary coded picture without partitioning

These NAL units will be sent within only a single session for any given access unit, and are placed in session order.

20 Coded slice in scalable extension
21-23 Reserved

Type 20 NAL units are placed in increasing order of DQId. Within each DQId value, they are placed in session order.

(Note: SVC slices with a given DQId value will be sent within only a single session for any given access unit.)

Type 21-23 NAL units are placed immediately following the non-reserved-type VCL NAL unit they follow in session order.

10 End of sequence

11 End of stream

6.2.2. Decoding Order Recovery for the NI-C, NI-TC, and I-C Modes

The following process MUST be used when either the NI-C or I-C MST packetization mode is in use. The following process MAY be applied when the NI-TC MST packetization mode is in use.

The RTP packets output from the RTP-level reception processing for each session are placed into a re-multiplexing buffer.

It is RECOMMENDED to set the size of the re-multiplexing buffer (in bytes) equal to or greater than the value of the sprop-remux-buf-req media type parameter of the highest RTP session the receiver receives.

The CS-DON value is calculated and stored for each NAL unit.

> Informative note: The CS-DON value of a NAL unit may rely on information carried in another packet than the packet containing the NAL unit. This happens, e.g., when the CS-DON values need to be derived for non-PACSI NAL units contained in single NAL unit packets, as the single NAL unit packets themselves do not contain CS-DON information. In this case, when no packet containing required CS-DON information is received for a NAL unit, this NAL unit has to be discarded by the receiver as it cannot be fed to the decoder in the correct order. When the optional media type parameter sprop-mst-csdon-always-present is equal to 1, no such dependency exists, i.e., the CS-DON value of any particular NAL unit can be derived solely according to information in the packet containing the NAL unit, and therefore, the receiver does not need to discard any received NAL units.

The receiver operation is described below with the help of the following functions and constants:

- Function AbsDON is specified in Section 8.1 of [RFC6184].
- Function don_diff is specified in Section 5.5 of [RFC6184].
- Constant N is the value of the OPTIONAL sprop-mst-remux-buf-size media type parameter of the highest RTP session incremented by 1.

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

- There are N or more VCL NAL units in the re-multiplexing buffer.
- If sprop-mst-max-don-diff of the highest RTP session is present, don_diff(m,n) is greater than the value of sprop-mst-max-don-diff of the highest RTP session, where n corresponds to the NAL unit having the greatest value of AbsDON among the received NAL units and m corresponds to the NAL unit having the smallest value of AbsDON among the received NAL units.

- Initial buffering has lasted for the duration equal to or greater than the value of the OPTIONAL sprop-remux-init-buf-time media type parameter of the highest RTP session.

The NAL units to be removed from the re-multiplexing buffer are determined as follows:

- If the re-multiplexing buffer contains at least N VCL NAL units, NAL units are removed from the re-multiplexing buffer and passed to the decoder in the order specified below until the buffer contains N-1 VCL NAL units.

- If sprop-mst-max-don-diff of the highest RTP session is present, all NAL units m for which don_diff(m,n) is greater than sprop-max-don-diff of the highest RTP session are removed from the re-multiplexing buffer and passed to the decoder in the order specified below. Herein, n corresponds to the NAL unit having the greatest value of AbsDON among the NAL units in the re-multiplexing buffer.

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

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

- For each NAL unit associated with a value of CS-DON, a CS-DON distance is calculated as follows. If the value of CS-DON of the NAL unit is larger than the value of PDON, the CS-DON distance is equal to CS-DON - PDON. Otherwise, the CS-DON distance is equal to 65535 - PDON + CS-DON + 1.

- NAL units are delivered to the decoder in increasing order of CS-DON distance. If several NAL units share the same value of CS-DON distance, they can be passed to the decoder in any order.

- When a desired number of NAL units have been passed to the decoder, the value of PDON is set to the value of CS-DON for the last NAL unit passed to the decoder.

7. Payload Format Parameters

This section specifies the parameters that MAY be used to select optional features of the payload format and certain features of the bitstream. The parameters are specified here as part of the media type registration for the SVC 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.

Some parameters provide a receiver with the properties of the stream that will be sent. The names of all these parameters start with "sprop" for stream properties. Some of these "sprop" parameters are limited by other payload or codec configuration parameters. For example, the sprop-parameter-sets parameter is constrained by the profile-level-id parameter. The media sender selects all "sprop" parameters rather than the receiver. This uncommon characteristic of the "sprop" parameters may be incompatible with some signaling protocol concepts, in which case the use of these parameters SHOULD be avoided.

7.1. Media Type Registration

The media subtype for the SVC codec has been allocated from the IETF tree.

The receiver MUST ignore any unspecified parameter.

> Informative note: Requiring that the receiver ignore unspecified parameters allows for backward compatibility of future extensions. For example, if a future specification that is backward compatible to this specification specifies some new parameters, then a receiver according to this specification is capable of receiving data per the new payload but ignoring those parameters newly specified in the new payload specification. This provision is also present in [RFC6184].

Media Type name: video

Media subtype name: H264-SVC

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-level-id:
A base16 [RFC4648] (hexadecimal) representation of the following three bytes in the sequence parameter set or subset sequence parameter set NAL unit specified in [H.264]: 1) profile_idc; 2) a byte herein referred to as profile-iop, composed of the values of constraint_set0_flag, constraint_set1_flag, constraint_set2_flag, constraint_set3_flag, constraint_set4_flag, constraint_set5_flag, and reserved_zero_2bits, in bit-significance order, starting from the most-significant bit, and 3) level_idc. Note that reserved_zero_2bits is required to be equal to 0 in [H.264], but other values for it may be specified in the future by ITU-T or ISO/IEC.

The profile-level-id parameter indicates the default sub-profile, i.e., the subset of coding tools that may have been used to generate the stream or that the receiver supports, and the default level of the stream or the one that the receiver supports.

The default sub-profile is indicated collectively by the profile_idc byte and some fields in the profile-iop byte. Depending on the values of the fields in the profile-iop byte, the default sub-profile may be the same set of coding tools supported by one profile, or a common subset of coding tools of multiple profiles, as specified in Subsection G.7.4.2.1.1 of [H.264]. The default level is indicated by the level_idc byte, and, when profile_idc is equal to 66, 77, or 88 (the Baseline, Main, or Extended profile) and level_idc is equal to 11, additionally by bit 4 (constraint_set3_flag) of the profile-iop byte. When profile_idc is equal to 66, 77, or 88 (the Baseline, Main, or Extended profile) and level_idc is equal to 11, and bit 4 (constraint_set3_flag) of the profile-iop byte is equal to 1, the default level is Level 1b.

Table 13 lists all profiles defined in Annexes A and G of [H.264] and, for each of the profiles, the possible combinations of profile_idc and profile-iop that represent the same sub-profile.

Table 13. Combinations of profile_idc and profile-iop representing the same sub-profile corresponding to the full set of coding tools supported by one profile. In the following, x may be either 0 or 1, while the profile names are indicated as follows. CB: Constrained Baseline profile, B: Baseline profile, M: Main profile, E: Extended profile, H: High profile, H10: High 10 profile, H42: High 4:2:2 profile, H44: High 4:4:4 Predictive profile, H10I: High 10 Intra profile, H42I: High

4:2:2 Intra profile, H44I: High 4:4:4 Intra profile, C44I: CAVLC 4:4:4 Intra profile, SB: Scalable Baseline profile, SH: Scalable High profile, and SHI: Scalable High Intra profile.

Profile	profile_idc (hexadecimal)	profile-iop (binary)
CB	42 (B)	x1xx0000
same as:	4D (M)	1xxx0000
same as:	58 (E)	11xx0000
B	42 (B)	x0xx0000
same as:	58 (E)	10xx0000
M	4D (M)	0x0x0000
E	58	00xx0000
H	64	00000000
H10	6E	00000000
H42	7A	00000000
H44	F4	00000000
H10I	6E	00010000
H42I	7A	00010000
H44I	F4	00010000
C44I	2C	00010000
SB	53	x0000000
SH	56	0x000000
SHI	56	0x010000

For example, in the table above, profile_idc equal to 58 (Extended) with profile-iop equal to 11xx0000 indicates the same sub-profile corresponding to profile_idc equal to 42 (Baseline) with profile-iop equal to x1xx0000. Note that other combinations of profile_idc and profile-iop (not listed in Table 13) may represent a sub-profile equivalent to the common subset of coding tools for more than one profile. Note also that a decoder conforming to a certain profile may be able to decode bitstreams conforming to other profiles.

If profile-level-id is used to indicate stream properties, it indicates that, to decode the stream, the minimum subset of coding tools a decoder has to support is the default sub-profile, and the lowest level the decoder has to support is the default level.

If the profile-level-id parameter is used for capability exchange or session setup, it indicates the subset of coding tools, which is equal to the default sub-profile, that the codec supports for both receiving and sending. If max-recv-level is not present, the default level from profile-level-id indicates the highest level the codec wishes to support. If

max-recv-level is present, it indicates 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.

Informative note: Capability exchange and session setup procedures should provide means to list the capabilities for each supported sub-profile separately. For example, the one-of-N codec selection procedure of the SDP Offer/Answer model can be used (Section 10.2 of [RFC3264]). The one-of-N codec selection procedure may also be used to provide different combinations of profile_idc and profile-iop that represent the same sub-profile. When there are many different combinations of profile_idc and profile-iop that represent the same sub-profile, using the one-of-N codec selection procedure may result in a fairly large SDP message. Therefore, a receiver should understand the different equivalent combinations of profile_idc and profile-iop that represent the same sub-profile, and be ready to accept an offer using any of the equivalent combinations.

If no profile-level-id is present, the Baseline Profile without additional constraints at Level 1 MUST be implied.

max-recv-level:
This parameter MAY be used to indicate the highest level a receiver supports when the highest level is higher than the default level (the level indicated by profile-level-id). The value of max-recv-level is a base16 (hexadecimal) representation of the two bytes after the syntax element profile_idc in the sequence parameter set NAL unit specified in [H.264]: profile-iop (as defined above) and level_idc. If (the level_idc byte of max-recv-level is equal to 11 and bit 4 of the profile-iop byte of max-recv-level is equal to 1) or (the level_idc byte of max-recv-level is equal to 9 and bit 4 of the profile-iop byte of max-recv-level is equal to 0), the highest level the receiver supports is Level 1b. Otherwise, the highest level the receiver supports is equal to the level_idc byte of max-recv-level divided by 10.

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

max-recv-base-level:
This parameter MAY be used to indicate the highest level a receiver supports for the base layer when negotiating an SVC stream. The value of max-recv-base-level is a base16

(hexadecimal) representation of the two bytes after the syntax element profile_idc in the sequence parameter set NAL unit specified in [H.264]: profile-iop (as defined above) and level_idc. If (the level_idc byte of max-recv-level is equal to 11 and bit 4 of the profile-iop byte of max-recv-level is equal to 1) or (the level_idc byte of max-recv-level is equal to 9 and bit 4 of the profile-iop byte of max-recv-level is equal to 0), the highest level the receiver supports for the base layer is Level 1b. Otherwise, the highest level the receiver supports for the base layer is equal to the level_idc byte of max-recv-level divided by 10.

max-mbps, max-fs, max-cpb, max-dpb, and max-br:
The common properties of these parameters are specified in [RFC6184].

max-mbps: This parameter is as specified in [RFC6184].

max-fs: This parameter is as specified in [RFC6184].

max-cpb: The value of max-cpb is an integer indicating the maximum coded picture buffer size in units of 1000 bits for the VCL HRD parameters and in units of 1200 bits for the NAL HRD parameters. Note that this parameter does not use units of cpbBrVclFactor and cpbBrNALFactor (see Table A-1 of [H.264]). 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 conveyed in the value of the profile-level-id parameter or the max-recv-level parameter. 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 [H.264] for the signaled highest level is replaced with the value of max-cpb (after taking cpbBrVclFactor and cpbBrNALFactor into consideration when needed). The value of max-cpb (after taking cpbBrVclFactor and cpbBrNALFactor into consideration when needed) MUST be greater than or equal to the value of MaxCPB given in Table A-1 of [H.264] 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 [H.264].

Informative note: The coded picture buffer is used in the Hypothetical Reference Decoder (HRD, Annex C) of [H.264]. The use of the HRD is recommended in SVC 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-interleaving, re-multiplexing, and de-jitter buffers. The coded picture buffer need not be implemented in decoders as specified in Annex C of [H.264]; standard-compliant decoders can have any buffering arrangements provided that they can decode standard-compliant bitstreams. Thus, in practice, the input buffer for video decoder can be integrated with the de-interleaving, re-multiplexing, and de-jitter buffers of the receiver.

max-dpb: This parameter is as specified in [RFC6184].

max-br: The value of max-br is an integer indicating the maximum video bitrate in units of 1000 bits per second for the VCL HRD parameters and in units of 1200 bits per second for the NAL HRD parameters. Note that this parameter does not use units of cpbBrVclFactor and cpbBrNALFactor (see Table A-1 of [H.264]).

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 conveyed in the value of the profile-level-id parameter or the max-recv-level parameter.

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 (after taking cpbBrVclFactor and cpbBrNALFactor into consideration when needed) replaces the MaxBR value in Table A-1 of [H.264] for the highest level.

- When the max-cpb parameter is not present, the result of the following formula replaces the value of MaxCPB in Table A-1 of [H.264]: (MaxCPB of the signaled level) * max-br / (MaxBR of the signaled highest level).

For example, if a receiver signals capability for Main profile Level 1.2 with max-br equal to 1550, this indicates a maximum video bitrate of 1550 kbits/sec for VCL HRD parameters, a

maximum video bitrate of 1860 kbits/sec for NAL HRD parameters, and a CPB size of 4036458 bits (1550000 / 384000 * 1000 * 1000).

The value of max-br (after taking cpbBrVclFactor and cpbBrNALFactor into consideration when needed) MUST be greater than or equal to the value MaxBR given in Table A-1 of [H.264] for the signaled highest level.

Senders MAY use this knowledge to send higher-bitrate video as allowed in the level definition of SVC, 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. No assumption can be made from the value of this parameter that the network is capable of handling such bitrates at any given time. In particular, no conclusion can be drawn that the signaled bitrate is possible under congestion control constraints.

redundant-pic-cap:
This parameter is as specified in [RFC6184].

sprop-parameter-sets:
This parameter MAY be used to convey any sequence parameter set, subset sequence parameter set, and picture parameter set NAL units (herein referred to as the initial parameter set NAL units) that can be placed in the NAL unit stream to precede any other NAL units in decoding order and that are associated with the default level of profile-level-id. The parameter MUST NOT be used to indicate codec capability in any capability exchange procedure. The value of the parameter is a comma (',') separated list of base64 [RFC4648] representations of the parameter set NAL units as specified in Sections 7.3.2.1, 7.3.2.2, and G.7.3.2.1 of [H.264]. Note that the number of bytes in a parameter set NAL unit is typically less than 10, but a picture parameter set NAL unit can contain several hundreds of bytes.

Informative note: When several payload types are offered in the SDP Offer/Answer model, each with its own sprop-parameter-sets parameter, then the receiver cannot assume that those parameter sets do not use conflicting storage locations (i.e., identical values of parameter set

identifiers). Therefore, a receiver should buffer all sprop-parameter-sets and make them available to the decoder instance that decodes a certain payload type.

sprop-level-parameter-sets:
This parameter MAY be used to convey any sequence, subset sequence, and picture parameter set NAL units (herein referred to as the initial parameter set NAL units) that can be placed in the NAL unit stream to precede any other NAL units in decoding order and that are associated with one or more levels different than the default level of profile-level-id. The parameter MUST NOT be used to indicate codec capability in any capability exchange procedure.

The sprop-level-parameter-sets parameter contains parameter sets for one or more levels that are different than the default level. All parameter sets targeted for use when one level of the default sub-profile is accepted by a receiver are clustered and prefixed with a three-byte field that has the same syntax as profile-level-id. This enables the receiver to install the parameter sets for the accepted level and discard the rest. The three-byte field is named PLId, and all parameter sets associated with one level are named PSL, which has the same syntax as sprop-parameter-sets. Parameter sets for each level are represented in the form of PLId:PSL, i.e., PLId followed by a colon (':') and the base64 [RFC4648] representation of the initial parameter set NAL units for the level. Each pair of PLId:PSL is also separated by a colon. Note that a PSL can contain multiple parameter sets for that level, separated with commas (',').

The subset of coding tools indicated by each PLId field MUST be equal to the default sub-profile, and the level indicated by each PLId field MUST be different than the default level.

Informative note: This parameter allows for efficient level downgrade or upgrade in SDP Offer/Answer and out-of-band transport of parameter sets, simultaneously.

in-band-parameter-sets:
This parameter MAY be used to indicate a receiver capability. The value MAY be equal to either 0 or 1. The value 1 indicates that the receiver discards out-of-band parameter sets in sprop-parameter-sets and sprop-level-parameter-sets, therefore the sender MUST transmit all parameter sets in-band. The value 0 indicates that the receiver utilizes out-of-band parameter sets included in sprop-parameter-sets and/or sprop-level-parameter-sets. However, in this case, the sender MAY still choose to

send parameter sets in-band. When the parameter is not present, this receiver capability is not specified, and therefore the sender MAY send out-of-band parameter sets only, or it MAY send in-band-parameter-sets only, or it MAY send both.

packetization-mode:
This parameter is as specified in [RFC6184]. When the mst-mode parameter is present, the value of this parameter is additionally constrained as follows. If mst-mode is equal to "NI-T", "NI-C", or "NI-TC", packetization-mode MUST NOT be equal to 2. Otherwise, (mst-mode is equal to "I-C"), packetization-mode MUST be equal to 2.

sprop-interleaving-depth:
This parameter is as specified in [RFC6184].

sprop-deint-buf-req:
This parameter is as specified in [RFC6184].

deint-buf-cap:
This parameter is as specified in [RFC6184].

sprop-init-buf-time:
This parameter is as specified in [RFC6184].

sprop-max-don-diff:
This parameter is as specified in [RFC6184].

max-rcmd-nalu-size:
This parameter is as specified in [RFC6184].

mst-mode:
This parameter MAY be used to signal the properties of a NAL unit stream or the capabilities of a receiver implementation. If this parameter is present, multi-session transmission MUST be used. Otherwise (this parameter is not present), single-session transmission MUST be used. When this parameter is present, the following applies. When the value of mst-mode is equal to "NI-T", the NI-T mode MUST be used. When the value of mst-mode is equal to "NI-C", the NI-C mode MUST be used. When the value of mst-mode is equal to "NI-TC", the NI-TC mode MUST be used. When the value of mst-mode is equal to "I-C", the I-C mode MUST be used. The value of mst-mode MUST have one of the following tokens: "NI-T", "NI-C", "NI-TC", or "I-C".

All RTP sessions in an MST MUST have the same value of mst-mode.

sprop-mst-csdon-always-present:
This parameter MUST NOT be present when mst-mode is not present or the value of mst-mode is equal to "NI-T" or "I-C". This parameter signals the properties of the NAL unit stream. When sprop-mst-csdon-always-present is present and the value is equal to 1, packetization-mode MUST be equal to 1, and all the RTP packets carrying the NAL unit stream MUST be STAP-A packets containing a PACSI NAL unit that further contains the DONC field or NI-MTAP packets with the J field equal to 1. When sprop-mst-csdon-always-present is present and the value is equal to 1, the CS-DON value of any particular NAL unit can be derived solely according to information in the packet containing the NAL unit.

When sprop-mst-csdon-always-present is present in the current RTP session, it MUST be present also in all the RTP sessions the current RTP session depends on and the value of sprop-mst-csdon-always-present is identical for the current RTP session and all the RTP sessions on which the current RTP session depends.

sprop-mst-remux-buf-size:
This parameter MUST NOT be present when mst-mode is not present or the value of mst-mode is equal to "NI-T". This parameter MUST be present when mst-mode is present and the value of mst-mode is equal to "NI-C", "NI-TC", or "I-C".

This parameter signals the properties of the NAL unit stream. It MUST be set to a value one less than the minimum re-multiplexing buffer size (in NAL units), so that it is guaranteed that receivers can reconstruct NAL unit decoding order as specified in Subsection 6.2.2.

The value of sprop-mst-remux-buf-size MUST be an integer in the range of 0 to 32767, inclusive.

sprop-remux-buf-req:
This parameter MUST NOT be present when mst-mode is not present or the value of mst-mode is equal to "NI-T". It MUST be present when mst-mode is present and the value of mst-mode is equal to "NI-C", "NI-TC", or "I-C".

sprop-remux-buf-req signals the required size of the re-multiplexing buffer for the NAL unit stream. It is guaranteed that receivers can recover the decoding order of the received NAL units from the current RTP session and the RTP sessions the

current RTP session depends on as specified in Section 6.2.2, when the re-multiplexing buffer size is of at least the value of sprop-remux-buf-req in units of bytes.

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

remux-buf-cap:
This parameter MUST NOT be present when mst-mode is not present or the value of mst-mode is equal to "NI-T". This parameter MAY be used to signal the capabilities of a receiver implementation and indicates the amount of re-multiplexing buffer space in units of bytes that the receiver has available for recovering the NAL unit decoding order as specified in Section 6.2.2. A receiver is able to handle any NAL unit stream for which the value of the sprop-remux-buf-req parameter is smaller than or equal to this parameter.

If the parameter is not present, then a value of 0 MUST be used for remux-buf-cap. The value of remux-buf-cap MUST be an integer in the range of 0 to 4294967295, inclusive.

sprop-remux-init-buf-time:
This parameter MAY be used to signal the properties of the NAL unit stream. The parameter MUST NOT be present if mst-mode is not present or the value of mst-mode is equal to "NI-T".

The parameter signals the initial buffering time that a receiver MUST wait before starting to recover the NAL unit decoding order as specified in Section 6.2.2 of this memo.

The parameter is coded as a non-negative base10 integer representation in clock ticks of a 90-kHz clock. If the parameter is not present, then no initial buffering time value is defined. Otherwise, the value of sprop-remux-init-buf-time MUST be an integer in the range of 0 to 4294967295, inclusive.

sprop-mst-max-don-diff:
This parameter MAY be used to signal the properties of the NAL unit stream. It MUST NOT be used to signal transmitter or receiver or codec capabilities. The parameter MUST NOT be present if mst-mode is not present or the value of mst-mode is equal to "NI-T". sprop-mst-max-don-diff is an integer in the range of 0 to 32767, inclusive. If sprop-mst-max-don-diff is not present, the value of the parameter is unspecified. sprop-mst-max-don-diff is calculated same as sprop-max-don-diff as specified in [RFC6184], with decoding order number being replaced by cross-session decoding order number.

sprop-scalability-info:
This parameter MAY be used to convey the NAL unit containing the scalability information SEI message as specified in Annex G of [H.264]. This parameter MAY be used to signal the contained layers of an SVC bitstream. The parameter MUST NOT be used to indicate codec capability in any capability exchange procedure. The value of the parameter is the base64 [RFC4648] representation of the NAL unit containing the scalability information SEI message. If present, the NAL unit MUST contain only one SEI message that is a scalability information SEI message.

This parameter MAY be used in an offering or declarative SDP message to indicate what layers (operation points) can be provided. A receiver MAY indicate its choice of one layer using the optional media type parameter scalable-layer-id.

scalable-layer-id:
This parameter MAY be used to signal a receiver's choice of the offers or declared operation points or layers using sprop-scalability-info or sprop-operation-point-info. The value of scalable-layer-id is a base16 representation of the layer_id[i] syntax element in the scalability information SEI message as specified in Annex G of [H.264] or layer-ID contained in sprop-operation-point-info.

sprop-operation-point-info:
This parameter MAY be used to describe the operation points of an RTP session. The value of this parameter consists of a comma-separated list of operation-point-description vectors. The values given by the operation-point-description vectors are the same as, or are derived from, the values that would be given for a scalable layer in the scalability information SEI message as specified in Annex G of [H.264], where the term scalable layer in the scalability information SEI message refers to all NAL units associated with the same values of temporal_id, dependency_id, and quality_id. In this memo, such a set of NAL units is called an operation point.

Each operation-point-description vector has ten elements, provided as a comma-separated list of values as defined below. The first value of the operation-point-description vector is preceded by a '<', and the last value of the operation-point-description vector is followed by a '>'. If the sprop-operation-point-info is followed by exactly one operation-point-description vector, this describes the highest operation point contained in the RTP session. If there are two or more

operation-point-description vectors, the first describes the lowest and the last describes the highest operation point contained in the RTP session.

The values given by the operation-point-description vector are as follows, in the order listed:

- layer-ID: This value specifies the layer identifier of the operation point, which is identical to the layer_id that would be indicated (for the same values of dependency_id, quality_id, and temporal_id) in the scalability information SEI message. This field MAY be empty, indicating that the value is unspecified. When there are multiple operation-point-description vectors with layer-ID, the values of layer-ID do not need to be consecutive.

- temporal-ID: This value specifies the temporal_id of the operation point. This field MUST NOT be empty.

- dependency-ID: This values specifies the dependency_id of the operation point. This field MUST NOT be empty.

- quality-ID: This values specifies the quality_id of the operation point. This field MUST NOT be empty.

- profile-level-ID: This value specifies the profile-level-idc of the operation point in the base16 format. The default sub-profile or default level indicated by the parameter profile-level-ID in the sprop-operation-point-info vector SHALL be equal to or lower than the default sub-profile or default level indicated by profile-level-id, which may be either present or the default value is taken. This field MAY be empty, indicating that the value is unspecified.

- avg-framerate: This value specifies the average frame rate of the operation point. This value is given as an integer in frames per 256 seconds. The field MAY be empty, indicating that the value is unspecified.

- width: This value specifies the width dimension in pixels of decoded frames for the operation point. This parameter is not directly given in the scalability information SEI message. This field MAY be empty, indicating that the value is unspecified.

- height: This value gives the height dimension in pixels of decoded frames for the operation point. This parameter is not directly given in the scalability information SEI. This field MAY be empty, indicating that the value is unspecified.

- avg-bitrate: This value specifies the average bitrate of the operation point. This parameter is given as an integer in kbits per second over the entire stream. Note that this parameter is provided in the scalability information SEI message in bits per second and calculated over a variable time window. This field MAY be empty, indicating that the value is unspecified.

- max-bitrate: This value specifies the maximum bitrate of the operation point. This parameter is given as an integer in kbits per second and describes the maximum bitrate per each one-second window. Note that this parameter is provided in the scalability information SEI message in bits per second and is calculated over a variable time window. This field MAY be empty, indicating that the value is unspecified.

  Similarly to sprop-scalability-info, this parameter MAY be used in an offering or declarative SDP message to indicate what layers (operation points) can be provided. A receiver MAY indicate its choice of the highest layer it wants to send and/or receive using the optional media type parameter scalable-layer-id.

sprop-no-NAL-reordering-required:
This parameter MAY be used to signal the properties of the NAL unit stream. This parameter MUST NOT be present when mst-mode is not present or the value of mst-mode is not equal to "NI-T". The presence of this parameter indicates that no reordering of non-VCL or VCL NAL units is required for the decoding order recovery process.

sprop-avc-ready:
This parameter MAY be used to indicate the properties of the NAL unit stream. The presence of this parameter indicates that the RTP session, if used in SST, or used in MST combined with other RTP sessions also with this parameter present, can be processed by a [RFC6184] receiver. This parameter MAY be used with RTP sessions with media subtype H264-SVC.

Encoding considerations:
This media type is framed and binary; see Section 4.8 of RFC 4288 [RFC4288].

Security considerations:
   See Section 8 of RFC 6190.

Published specification:
   Please refer to RFC 6190 and its Section 13.

Additional information:
   none

File extensions: none

Macintosh file type code: none

Object identifier or OID: none

Person & email address to contact for further information:

   Ye-Kui Wang, yekui.wang@huawei.com

Intended usage: COMMON

Restrictions on usage:
   This media type depends on RTP framing, and hence is only
   defined for transfer via RTP [RFC3550]. Transport within other
   framing protocols is not defined at this time.

Interoperability considerations:
   The media subtype name contains "SVC" to avoid potential
   conflict with RFC 3984 and its potential future replacement RTP
   payload format for H.264 non-SVC profiles.

Applications that use this media type:
   Real-time video applications like video streaming, video
   telephony, and video conferencing.

Author:

   Ye-Kui Wang, yekui.wang@huawei.com

Change controller:
   IETF Audio/Video Transport working group delegated from the
   IESG.

7.2. SDP Parameters

7.2.1. Mapping of Payload Type Parameters to SDP

The media type video/H264-SVC string is mapped to fields in the Session Description Protocol (SDP) 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 H264-SVC (the media subtype).

- The clock rate in the "a=rtpmap" line MUST be 90000.

- The OPTIONAL parameters profile-level-id, max-recv-level, max-recv-base-level, max-mbps, max-fs, max-cpb, max-dpb, max-br, redundant-pic-cap, in-band-parameter-sets, packetization-mode, sprop-interleaving-depth, deint-buf-cap, sprop-deint-buf-req, sprop-init-buf-time, sprop-max-don-diff, max-rcmd-nalu-size, mst-mode, sprop-mst-csdon-always-present, sprop-mst-remux-buf-size, sprop-remux-buf-req, remux-buf-cap, sprop-remux-init-buf-time, sprop-mst-max-don-diff, and scalable-layer-id, when present, MUST be 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.

- The OPTIONAL parameters sprop-parameter-sets, sprop-level-parameter-sets, sprop-scalability-info, sprop-operation-point-info, sprop-no-NAL-reordering-required, and sprop-avc-ready, 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), a sprop-parameter-sets or sprop-level-parameter-sets 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-parameter-sets and sprop-level-parameter-sets using the "fmtp" source attribute allows for out-of-band transport of parameter sets in topologies like Topo-Video-switch-MCU [RFC5117].

7.2.2. Usage with the SDP Offer/Answer Model

When an SVC stream (with media subtype H264-SVC) 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 SVC are profile-level-id, packetization-mode, and mst-mode. These media configuration parameters (except for the level part of profile-level-id) MUST be used symmetrically when the answerer does not include scalable-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. Note that the level part of profile-level-id includes level_idc, and, for indication of level 1b when profile_idc is equal to 66, 77, or 88, bit 4 (constraint_set3_flag) of profile-iop. The level part of profile-level-id is changeable.

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

  Informative note: In [H.264], all the levels except for Level 1b are equal to the value of level_idc divided by 10. Level 1b is a level higher than Level 1.0 but lower than Level 1.1, and is signaled in an ad hoc manner. For the Baseline, Main, and Extended profiles (with profile_idc equal to 66, 77, and 88, respectively), Level 1b is indicated by level_idc equal to 11 (i.e., the same as level 1.1) and constraint_set3_flag equal to 1. For other profiles, Level 1b is indicated by level_idc equal to 9 (but note that Level 1b for these profiles is still higher than Level 1, which has level_idc equal to 10, and lower than Level 1.1). In SDP Offer/Answer, an answer may indicate a level equal to or lower than the level indicated in the offer. Due to the ad hoc indication of Level 1b, offerers and answerers must check the value of bit 4 (constraint_set3_flag) of the middle octet of the parameter profile-level-id, when profile_idc is equal to 66, 77, or 88 and level_idc is equal to 11.

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 scalable-layer-id. When the answer does not include scalable-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 level lower than the default level offered.

> Informative note: When an offerer receives an answer that does not include scalable-layer-id it has to compare payload types not declared in the offer based on the media type (i.e., video/H264-SVC) 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.

Since an SVC stream may contain multiple operation points, a facility is provided so that an answerer can select a different operation point than the entire SVC stream. Specifically, different operation points MAY be described using the sprop-scalability-info or sprop-operation-point-info parameters. The first one carries the entire scalability information SEI message defined in Annex G of [H.264], whereas the second one may be derived, e.g., as a subset of this SEI message that only contains key information about an operation point. Operation points, in both cases, are associated with a layer identifier.

If such information (sprop-operation-point-info or sprop-scalability-info) is provided in an offer, an answerer MAY select from the various operation points offered in the sprop-scalability-information or sprop-operation-point-info parameters by including scalable-layer-id in the answer. By this, the answerer indicates its selection of a particular operation point in the received and/or in the sent stream. When such operation point selection takes place, i.e., the answerer includes scalable-layer-id 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: The ability to perform operation point selection enables a receiver to utilize the scalable nature of an SVC stream.

- The parameter max-recv-level, when present, declares the highest level supported for receiving. In case max-recv-level is not present, the highest level supported for receiving is equal to the

default level indicated by the level part of profile-level-id. max-recv-level, when present, MUST be higher than the default level.

- The parameter max-recv-base-level, when present, declares the highest level of the base layer supported for receiving. When max-recv-base-level is not present, the highest level supported for the base layer is not constrained separately from the SVC stream containing the base layer. The endpoint at the other side MUST NOT send a scalable stream for which the base layer is of a level higher than max-recv-base-level. Parameters declaring receiver capabilities above the default level (max-mbps, max-smbps, max-fs, max-cpb, max-dpb, max-br, and max-recv-level) do not apply to the base layer when max-recv-base-level is present.

- The parameters sprop-deint-buf-req, sprop-interleaving-depth, sprop-max-don-diff, sprop-init-buf-time, sprop-mst-csdon-always-present, sprop-remux-buf-req, sprop-mst-remux-buf-size, sprop-remux-init-buf-time, sprop-mst-max-don-diff, sprop-scalability-information, sprop-operation-point-info, sprop-no-NAL-reordering-required, and sprop-avc-ready describe the properties of the NAL unit stream that the offerer or 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 SVC, 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 the 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-mbps, max-fs, max-cpb, max-dpb, max-br, redundant-pic-cap, and max-rcmd-nalu-size 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 the parameters describe the limitations of what the offerer or answerer accepts for receiving streams.

- When mst-mode is not present and packetization-mode is equal to 2, the following applies.

- An offerer has to include the size of the de-interleaving buffer, sprop-deint-buf-req, in the offer. To enable the offerer and answerer to inform each other about their capabilities for de-interleaving buffering, both parties are RECOMMENDED to include deint-buf-cap. It is also RECOMMENDED to consider offering multiple payload types with different buffering requirements when the capabilities of the receiver are unknown.

- When mst-mode is present and equal to "NI-C", "NI-TC", or "I-C", the following applies.

  - An offerer has to include sprop-remux-buf-req in the offer. To enable the offerer and answerer to inform each other about their capabilities for re-multiplexing buffering, both parties are RECOMMENDED to include remux-buf-cap. It is also RECOMMENDED to consider offering multiple payload types with different buffering requirements when the capabilities of the receiver are unknown.

- The sprop-parameter-sets or sprop-level-parameter-sets parameter, when present (included in the "a=fmtp" line of SDP or conveyed using the "fmtp" source attribute as specified in Section 6.3 of [RFC5576]), is used for out-of-band transport of parameter sets. However, when out-of-band transport of parameter sets is used, parameter sets MAY still be additionally transported in-band.

  The answerer MAY use either out-of-band or in-band transport of parameter sets for the stream it is sending, regardless of whether out-of-band parameter sets transport has been used in the offerer-to-answerer direction. Parameter sets included in an answer are independent of those parameter sets included in the offer, as they are used for decoding two different video streams, one from the answerer to the offerer, and the other in the opposite direction.

  The following rules apply to transport of parameter sets in the offerer-to-answerer direction.

  - An offer MAY include either or both of sprop-parameter- sets and sprop-level-parameter-sets. If neither sprop-parameter-sets nor sprop-level-parameter-sets is present in the offer, then only in-band transport of parameter sets is used.

  - If the answer includes in-band-parameter-sets equal to 1, then the offerer MUST transmit parameter sets in-band. Otherwise, the following applies.

- If the level to use in the offerer-to-answerer direction is equal to the default level in the offer, the following applies.

  The answerer MUST be prepared to use the parameter sets included in sprop-parameter-sets, when present, for decoding the incoming NAL unit stream, and ignore sprop-level-parameter-sets, when present.

  When sprop-parameter-sets is not present in the offer, in-band transport of parameter sets MUST be used.

- Otherwise (the level to use in the offerer-to-answerer direction is not equal to the default level in the offer), the following applies.

  The answerer MUST be prepared to use the parameter sets that are included in sprop-level-parameter-sets for the accepted level (i.e., the default level in the answer, which is also the level to use in the offerer-to-answerer direction), when present, for decoding the incoming NAL unit stream, and ignore all other parameter sets included in sprop-level-parameter-sets and sprop-parameter-sets, when present.

  When no parameter sets for the accepted level are present in the sprop-level-parameter-sets, in-band transport of parameter sets MUST be used.

The following rules apply to transport of parameter sets in the answerer-to-offerer direction.

- An answer MAY include either sprop-parameter-sets or sprop-level-parameter-sets, but MUST NOT include both of the two. If neither sprop-parameter-sets nor sprop-level-parameter-sets is present in the answer, then only in-band transport of parameter sets is used.

- If the offer includes in-band-parameter-sets equal to 1, then the answerer MUST NOT include sprop-parameter-sets or sprop-level-parameter-sets in the answer and MUST transmit parameter sets in-band. Otherwise, the following applies.

  - If the level to use in the answerer-to-offerer direction is equal to the default level in the answer, the following applies.

The offerer MUST be prepared to use the parameter sets included in sprop-parameter-sets, when present, for decoding the incoming NAL unit stream, and ignore sprop-level-parameter-sets, when present.

When sprop-parameter-sets is not present in the answer, the answerer MUST transmit parameter sets in-band.

- Otherwise (the level to use in the answerer-to-offerer direction is not equal to the default level in the answer), the following applies.

  The offerer MUST be prepared to use the parameter sets that are included in sprop-level-parameter-sets for the level to use in the answerer-to-offerer direction, when present in the answer, for decoding the incoming NAL unit stream, and ignore all other parameter sets included in sprop-level-parameter-sets and sprop-parameter-sets, when present in the answer.

  When no parameter sets for the level to use in the answerer-to-offerer direction are present in sprop-level-parameter-sets in the answer, the answerer MUST transmit parameter sets in-band.

When sprop-parameter-sets or sprop-level-parameter-sets is conveyed using the "fmtp" source attribute as specified in Section 6.3 of [RFC5576], the receiver of the parameters MUST store the parameter sets included in the sprop-parameter-sets or sprop-level-parameter-sets for the accepted level and associate them to the source given as a part of the "fmtp" source attribute. Parameter sets associated with one source MUST only be used to decode NAL units conveyed in RTP packets from the same source. When this mechanism is in use, SSRC collision detection and resolution MUST be performed as specified in [RFC5576].

Informative note: Conveyance of sprop-parameter-sets and sprop-level-parameter-sets using the "fmtp" source attribute may be used in topologies like Topo-Video-switch-MCU [RFC5117] to enable out-of-band transport of parameter sets.

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

- The media format configuration is identified by profile-level- id, including the level part, packetization-mode, and mst-mode. These media format configuration parameters (including the level part of profile-level-id) MUST be used symmetrically; i.e., the answerer

MUST either maintain all configuration parameters or remove the media format (payload type) completely. Note that this implies that the level part of profile-level-id for Offer/Answer in multicast is not changeable.

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]. An answer MUST NOT contain a payload type number used in the offer unless the configuration is the same as in the offer.

- Parameter sets received MUST be associated with the originating source, and MUST be only used in decoding the incoming NAL unit stream from the same source.

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

Table 14 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 scalable-layer-id parameter is used only apply to answers, whereas the other columns apply to both offers and answers.

Table 14. Interpretation of parameters for various combinations of offers, answers, direction attributes, with and without scalable-layer-id. Columns that do not indicate offer or answer apply to both.

	sendonly --+			
	answer: recvonly,scalable-layer-id --+			
	recvonly w/o scalable-layer-id --+			
	answer: sendrecv, scalable-layer-id --+			
	sendrecv w/o scalable-layer-id --+			
	|	|	|	|
profile-level-id	C	X	C	X	P
max-recv-level	R	R	R	R	-
max-recv-base-level	R	R	R	R	-
packetization-mode	C	X	C	X	P
mst-mode	C	X	C	X	P
sprop-avc-ready	P	P	-	-	P
sprop-deint-buf-req	P	P	-	-	P
sprop-init-buf-time	P	P	-	-	P
sprop-interleaving-depth	P	P	-	-	P
sprop-max-don-diff	P	P	-	-	P
sprop-mst-csdon-always-present	P	P	-	-	P
sprop-mst-max-don-diff	P	P	-	-	P
sprop-mst-remux-buf-size	P	P	-	-	P
sprop-no-NAL-reordering-required	P	P	-	-	P
sprop-operation-point-info	P	P	-	-	P
sprop-remux-buf-req	P	P	-	-	P
sprop-remux-init-buf-time	P	P	-	-	P
sprop-scalability-info	P	P	-	-	P
deint-buf-cap	R	R	R	R	-
max-br	R	R	R	R	-
max-cpb	R	R	R	R	-
max-dpb	R	R	R	R	-
max-fs	R	R	R	R	-
max-mbps	R	R	R	R	-
max-rcmd-nalu-size	R	R	R	R	-
redundant-pic-cap	R	R	R	R	-
remux-buf-cap	R	R	R	R	-
in-band-parameter-sets	R	R	R	R	-
sprop-parameter-sets	S	S	-	-	S
sprop-level-parameter-sets	S	S	-	-	S
scalable-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
S: out-of-band parameter sets
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 part of the profile-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 downgrading (for profile-level-id) is used, an answerer MUST NOT include the scalable-layer-id parameter.

When a sender's capabilities are declared, and non-downgradable parameters are used in this declaration, then 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, e.g., for the packetization mode. 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.

If an offerer wishes to have non-symmetric capabilities between sending and receiving, the offerer can allow asymmetric levels via level-asymmetry-allowed equal to 1. Alternatively, the offerer can offer different RTP sessions, i.e., different media lines declared as "recvonly" and "sendonly", respectively. This may have further implications on the system, and may require additional external semantics to associate the two media lines.

7.2.3. 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. Additionally, the order of dependencies of the RTP sessions indicated by the "a=depend" attribute as defined in [RFC5583] MUST represent the decoding order of the VC) NAL units in an access unit, i.e., the order of session dependency is given from the base or the lowest enhancement RTP session (the most important) to the highest enhancement RTP session (the least important).

7.2.4. Usage in Declarative Session Descriptions

When SVC 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-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-level-id
  - packetization-mode
  - mst-mode
  - sprop-deint-buf-req
  - sprop-interleaving-depth
  - sprop-max-don-diff
  - sprop-init-buf-time
  - sprop-mst-csdon-always-present
  - sprop-mst-remux-buf-size
  - sprop-remux-buf-req
  - sprop-remux-init-buf-time
  - sprop-mst-max-don-diff
  - sprop-scalability-info
  - sprop-operation-point-info
  - sprop-no-NAL-reordering-required
  - sprop-avc-ready

  Out-of-band transporting of parameter sets:

  - sprop-parameter-sets
  - sprop-level-parameter-sets

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

- max-mbps
- max-fs
- max-cpb
- max-dpb
- max-br
- max-recv-level
- max-recv-base-level
- redundant-pic-cap
- max-rcmd-nalu-size
- deint-buf-cap
- remux-buf-cap
- scalable-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.

7.3. Examples

In the following examples, "{data}" is used to indicate a data string encoded as base64.

7.3.1. Example for Offering a Single SVC Session

Example 1: The offerer offers one video media description including two RTP payload types. The first payload type offers H264, and the second offers H264-SVC. Both payload types have different fmtp parameters as profile-level-id, packetization-mode, and sprop-parameter-sets.

Offerer -> Answerer SDP message:

```
m=video 20000 RTP/AVP 97 96
a=rtpmap:96 H264/90000
a=fmtp:96 profile-level-id=4de00a; packetization-mode=0;
 sprop-parameter-sets={sps0},{pps0};
a=rtpmap:97 H264-SVC/90000
a=fmtp:97 profile-level-id=53000c; packetization-mode=1;
 sprop-parameter-sets={sps0},{pps0},{sps1},{pps1};
```

If the answerer does not support media subtype H264-SVC, it can issue an answer accepting only the base layer offer (payload type 96). In the following example, the receiver supports H264-SVC, so it lists payload type 97 first as the preferred option.

```
Answerer -> Offerer SDP message:

m=video 40000 RTP/AVP 97 96
a=rtpmap:96 H264/90000
a=fmtp:96 profile-level-id=4de00a; packetization-mode=0;
 sprop-parameter-sets={sps2},{pps2};
a=rtpmap:97 H264-SVC/90000
a=fmtp:97 profile-level-id=53000c; packetization-mode=1;
 sprop-parameter-sets={sps2},{pps2},{sps3},{pps3};
```

7.3.2. Example for Offering a Single SVC Session Using scalable-layer-id

Example 2: Offerer offers the same media configurations as shown in the example above for receiving and sending the stream, but using a single RTP payload type and including sprop-operation-point-info.

```
Offerer -> Answerer SDP message:

m=video 20000 RTP/AVP 97
a=rtpmap:97 H264-SVC/90000
a=fmtp:97 profile-level-id=53000c; packetization-mode=1;
 sprop-parameter-sets={sps0},{sps1},{pps0},{pps1};
 sprop-operation-point-info=<1,0,0,0,4de00a,3200,176,144,128,
256>,<2,1,1,0,53000c,6400,352,288,256,512>;
```

In this example, the receiver supports H264-SVC and chooses the lower operation point offered in the RTP payload type for sending and receiving the stream.

```
Answerer -> Offerer SDP message:

m=video 40000 RTP/AVP 97
a=rtpmap:97 H264-SVC/90000
a=fmtp:97 sprop-parameter-sets={sps2},{sps3},{pps2},{pps3};
 scalable-layer-id=1;
```

In an equivalent example showing the use of sprop-scalability-info instead using the sprop-operation-point-info, the sprop-operation-point-info would be exchanged by the sprop-scalability-info followed by the binary (base16) representation of the Scalability Information SEI message.

7.3.3. Example for Offering Multiple Sessions in MST

Example 3: In this example, the offerer offers a multi-session transmission with up to three sessions. The base session media description includes payload types that are backward compatible with

[RFC6184], and three different payload types are offered. The other two media are using payload types with media subtype H264-SVC. In each media description, different values of profile-level-id, packetization-mode, mst-mode, and sprop-parameter-sets are offered.

Offerer -> Answerer SDP message:

```
a=group:DDP L1 L2 L3
m=video 20000 RTP/AVP 96 97 98
a=rtpmap:96 H264/90000
a=fmtp:96 profile-level-id=4de00a; packetization-mode=0;
 mst-mode=NI-T; sprop-parameter-sets={sps0},{pps0};
a=rtpmap:97 H264/90000
a=fmtp:97 profile-level-id=4de00a; packetization-mode=1;
 mst-mode=NI-TC; sprop-parameter-sets={sps0},{pps0};
a=rtpmap:98 H264/90000
a=fmtp:98 profile-level-id=4de00a; packetization-mode=2;
 mst-mode=I-C; init-buf-time=156320;
 sprop-parameter-sets={sps0},{pps0};
a=mid:L1
m=video 20002 RTP/AVP 99 100
a=rtpmap:99 H264-SVC/90000
a=fmtp:99 profile-level-id=53000c; packetization-mode=1;
 mst-mode=NI-T; sprop-parameter-sets={sps1},{pps1};
a=rtpmap:100 H264-SVC/90000
a=fmtp:100 profile-level-id=53000c; packetization-mode=2;
 mst-mode=I-C; sprop-parameter-sets={sps1},{pps1};
a=mid:L2
a=depend:99 lay L1:96,97; 100 lay L1:98
m=video 20004 RTP/AVP 101
a=rtpmap:101 H264-SVC/90000
a=fmtp:101 profile-level-id=53001F; packetization-mode=1;
 mst-mode=NI-T; sprop-parameter-sets={sps2},{pps2};
a=mid:L3
a=depend:101 lay L1:96,97 L2:99
```

It is assumed that in this example the answerer only supports the NI-T mode for multi-session transmission. For this reason, it chooses the corresponding payload type (96) for the base RTP session. For the two enhancement RTP sessions, the answerer also chooses the payload types that use the NI-T mode (99 and 101).

```
Answerer -> Offerer SDP message:

a=group:DDP L1 L2 L3
m=video 40000 RTP/AVP 96
a=rtpmap:96 H264/90000
a=fmtp:96 profile-level-id=4de00a; packetization-mode=0;
 mst-mode=NI-T; sprop-parameter-sets={sps3},{pps3};
a=mid:L1
m=video 40002 RTP/AVP 99
a=rtpmap:99 H264-SVC/90000
a=fmtp:99 profile-level-id=53000c; packetization-mode=1;
 mst-mode=NI-T; sprop-parameter-sets={sps4},{pps4};
a=mid:L2
a=depend:99 lay L1:96
m=video 40004 RTP/AVP 101
a=rtpmap:101 H264-SVC/90000
a=fmtp:101 profile-level-id=53001F; packetization-mode=1;
 mst-mode=NI-T; sprop-parameter-sets={sps5},{pps5};
a=mid:L3
a=depend:101 lay L1:96 L2:99
```

7.3.4. Example for Offering Multiple Sessions in MST Including Operation with Answerer Using scalable-layer-id

Example 4: In this example, the offerer offers a multi-session transmission of three layers with up to two sessions. The base session media description has a payload type that is backward compatible with [RFC6184]. Note that no parameter sets are provided, in which case in-band transport must be used. The other media description contains two enhancement layers and uses the media subtype H264-SVC. It includes two operation point definitions.

```
Offerer -> Answerer SDP message:

a=group:DDP L1 L2
m=video 20000 RTP/AVP 96
a=rtpmap:96 H264/90000
a=fmtp:96 profile-level-id=4de00a; packetization-mode=0;
 mst-mode=NI-T;
a=mid:L1
m=video 20002 RTP/AVP 97
a=rtpmap:97 H264-SVC/90000
a=fmtp:97 profile-level-id=53001F; packetization-mode=1;
 mst-mode=NI-TC; sprop-operation-point-info=<2,0,1,0,53000c,
3200,352,288,384,512>,<3,1,2,0,53001F,6400,704,576,768,1024>;
a=mid:L2
a=depend:97 lay L1:96
```

It is assumed that the answerer wants to send and receive the base layer (payload type 96), but it only wants to send and receive the lower enhancement layer, i.e., the one with layer id equal to 2. For this reason, the response will include the selection of the desired layer by setting scalable-layer-id equal to 2. Note that the answer only includes the scalable-layer-id information. The answer could include sprop-parameter-sets in the response.

Answerer -> Offerer SDP message:

```
a=group:DDP L1 L2
m=video 40000 RTP/AVP 96
a=rtpmap:96 H264/90000
a=fmtp:96 profile-level-id=4de00a; packetization-mode=0;
 mst-mode=NI-T;
a=mid:L1
m=video 40002 RTP/AVP 97
a=rtpmap:97 H264-SVC/90000
a=fmtp:97 scalable-layer-id=2;
a=mid:L2
a=depend:97 lay L1:96
```

7.3.5. Example for Negotiating an SVC Stream with a Constrained Base Layer in SST

Example 5: The offerer (Alice) offers one video description including two RTP payload types with differing levels and packetization modes.

Offerer -> Answerer SDP message:

```
m=video 20000 RTP/AVP 97 96
a=rtpmap:96 H264-SVC/90000
a=fmtp:96 profile-level-id=53001e; packetization-mode=0;
a=rtpmap:97 H264-SVC/90000
a=fmtp:97 profile-level-id=53001f; packetization-mode=1;
```

The answerer (Bridge) chooses packetization mode 1, and indicates that it would receive an SVC stream with the base layer being constrained.

Answerer -> Offerer SDP message:

```
m=video 40000 RTP/AVP 97
a=rtpmap:97 H264-SVC/90000
a=fmtp:97 profile-level-id=53001f; packetization-mode=1;
  max-recv-base-level=000d
```

The answering endpoint must send an SVC stream at Level 3.1. Since the offering endpoint did not declare max-recv-base-level, the base layer of the SVC stream the answering endpoint must send is not specifically constrained. The offering endpoint (Alice) must send an SVC stream at Level 3.1, for which the base layer must be of a level not higher than Level 1.3.

7.4. Parameter Set Considerations

Section 8.4 of [RFC6184] applies in this memo, with the following applies additionally for multi-session transmission (MST).

In MST, regardless of out-of-band or in-band transport of parameter sets are in use, parameter sets required for decoding NAL units carried in one particular RTP session SHOULD be carried in the same session, MAY be carried in a session that the particular RTP session depends on, and MUST NOT be carried in a session that the particular RTP session does not depend on.

8. Security Considerations

The security considerations of the RTP Payload Format for H.264 Video specification [RFC6184] apply. Additionally, the following applies.

Decoders MUST exercise caution with respect to the handling of reserved NAL unit types and reserved SEI messages, particularly if they contain active elements, and MUST restrict their domain of applicability to the presentation containing the stream. The safest way is to simply discard these NAL units and SEI messages.

When integrity protection is applied to a stream, care MUST be taken that the stream being transported may be scalable; hence a receiver may be able to access only part of the entire stream.

End-to-end security with either authentication, integrity, or confidentiality protection will prevent a MANE from performing media-aware operations other than discarding complete packets. And in the case of confidentiality protection it will even be prevented from performing discarding of packets in a media-aware way. To allow any MANE to perform its operations, it will be required to be a trusted entity that is included in the security context establishment. This applies both for the media path and for the RTCP path, if RTCP packets need to be rewritten.

9. Congestion Control

Within any given RTP session carrying payload according to this specification, the provisions of Section 10 of [RFC6184] apply. Reducing the session bitrate is possible by one or more of the following means:

a) Within the highest layer identified by the DID field remove any NAL units with QID higher than a certain value.

b) Remove all NAL units with TID higher than a certain value.

c) Remove all NAL units associated with a DID higher than a certain value.

   Informative note: Removal of all coded slice NAL units associated with DIDs higher than a certain value in the entire stream is required in order to preserve conformance of the resulting SVC stream.

d) Utilize the PRID field to indicate the relative importance of NAL units, and remove all NAL units associated with a PRID higher than a certain value. Note that the use of the PRID is application-specific.

e) Remove NAL units or entire packets according to application-specific rules. The result will depend on the particular coding structure used as well as any additional application-specific functionality (e.g., concealment performed at the receiving decoder). In general, this will result in the reception of a non-conforming bitstream and hence the decoder behavior is not specified by [H.264]. Significant artifacts may therefore appear in the decoded output if the particular decoder implementation does not take appropriate action in response to congestion control.

   Informative note: The discussion above is centered on NAL units rather than packets, primarily because that is the level where senders can meaningfully manipulate the scalable bitstream. The mapping of NAL units to RTP packets is fairly flexible when using aggregation packets. Depending on the nature of the congestion control algorithm, the "dimension" of congestion measurement (packet count or bitrate) and reaction to it (reducing packet count or bitrate or both) can be adjusted accordingly.

All aforementioned means are available to the RTP sender, regardless of whether that sender is located in the sending endpoint or in a mixer-based MANE.

When a translator-based MANE is employed, then the MANE MAY manipulate the session only on the MANE's outgoing path, so that the sensed end-to-end congestion falls within the permissible envelope. As with all translators, in this case, the MANE needs to rewrite RTCP RRs to reflect the manipulations it has performed on the session.

> Informative note: Applications MAY also implement, in addition or separately, other congestion control mechanisms, e.g., as described in [RFC5775] and [Yan].

10. IANA Considerations

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

11. Informative Appendix: Application Examples

11.1. Introduction

Scalable video coding is a concept that has been around since at least MPEG-2 [MPEG2], which goes back as early as 1993. Nevertheless, it has never gained wide acceptance, perhaps partly because applications didn't materialize in the form envisioned during standardization.

ISO/IEC MPEG and ITU-T VCEG, respectively, performed a requirement analysis for the SVC project. The MPEG and VCEG requirement documents are available in [JVT-N026] and [JVT-N027], respectively.

The following introduces four main application scenarios that the authors consider relevant and that are implementable with this specification.

11.2. Layered Multicast

This well-understood form of the use of layered coding [McCanne] implies that all layers are individually conveyed in their own RTP packet streams, each carried in its own RTP session using the IP (multicast) address and port number as the single demultiplexing point. Receivers "tune" into the layers by subscribing to the IP multicast, normally by using IGMP [IGMP]. Depending on the application scenario, it is also possible to convey a number of layers in one RTP session, when finer operation points within the subset of layers are not needed.

Layered multicast has the great advantage of simplicity and easy implementation. However, it has also the great disadvantage of utilizing many different transport addresses. While the authors

consider this not to be a major problem for a professionally maintained content server, receiving client endpoints need to open many ports to IP multicast addresses in their firewalls. This is a practical problem from a firewall and network address translation (NAT) viewpoint. Furthermore, even today IP multicast is not as widely deployed as many wish.

The authors consider layered multicast an important application scenario for the following reasons. First, it is well understood and the implementation constraints are well known. Second, there may well be large-scale IP networks outside the immediate Internet context that may wish to employ layered multicast in the future. One possible example could be a combination of content creation and core-network distribution for the various mobile TV services, e.g., those being developed by 3GPP (MBMS) [MBMS] and DVB (DVB-H) [DVB-H].

11.3. Streaming

In this scenario, a streaming server has a repository of stored SVC coded layers for a given content. At the time of streaming, and according to the capabilities, connectivity, and congestion situation of the client(s), the streaming server generates and serves a scalable stream. Both unicast and multicast serving is possible. At the same time, the streaming server may use the same repository of stored layers to compose different streams (with a different set of layers) intended for other audiences.

As every endpoint receives only a single SVC RTP session, the number of firewall pinholes can be optimized to one.

The main difference between this scenario and straightforward simulcasting lies in the architecture and the requirements of the streaming server, and is therefore out of the scope of IETF standardization. However, compelling arguments can be made why such a streaming server design makes sense. One possible argument is related to storage space and channel bandwidth. Another is bandwidth adaptability without transcoding -- a considerable advantage in a congestion controlled network. When the streaming server learns about congestion, it can reduce the sending bitrate by choosing fewer layers when composing the layered stream; see Section 9. SVC is designed to gracefully support both bandwidth ramp-down and bandwidth ramp-up with a considerable dynamic range. This payload format is designed to allow for bandwidth flexibility in the mentioned sense. While, in theory, a transcoding step could achieve a similar dynamic range, the computational demands are impractically high and video quality is typically lowered -- therefore, few (if any) streaming servers implement full transcoding.

11.4. Videoconferencing (Unicast to MANE, Unicast to Endpoints)

Videoconferencing has traditionally relied on Multipoint Control Units (MCUs). These units connect endpoints in a star configuration and operate as follows. Coded video is transmitted from each endpoint to the MCU, where it is decoded, scaled, and composited to construct output frames, which are then re-encoded and transmitted to the endpoint(s). In systems supporting personalized layout (each user is allowed to select the layout of his/her screen), the compositing and encoding process is performed for each of the receiving endpoints. Even without personalized layout, rate matching still requires that the encoding process at the MCU is performed separately for each endpoint. As a result, MCUs have considerable complexity and introduce significant delay. The cascaded encodings also reduce the video quality. Particularly for multipoint connections, interactive communication is cumbersome as the end-to-end delay is very high [G.114]. A simpler architecture is the switching MCU, in which one of the incoming video streams is redirected to the receiving endpoints. Obviously, only one user at a time can be seen and rate matching cannot be performed, thus forcing all transmitting endpoints to transmit at the lowest bit rate available in the MCU-to-endpoint connections.

With scalable video coding the MCU can be replaced with an application-level router (ALR): this unit simply selects which incoming packets should be transmitted to which of the receiving endpoints [Eleft]. In such a system, each endpoint performs its own composition of the incoming video streams. Assuming, for example, a system that uses spatial scalability with two layers, personalized layout is equivalent to instructing the ALR to only send the required packets for the corresponding resolution to the particular endpoint. Similarly, rate matching at the ALR for a particular endpoint can be performed by selecting an appropriate subset of the incoming video packets to transmit to the particular endpoint. Personalized layout and rate matching thus become routing decisions, and require no signal processing. Note that scalability also allows participants to enjoy the best video quality afforded by their links, i.e., users no longer have to be forced to operate at the quality supported by the weakest endpoint. Most importantly, the ALR has an insignificant contribution to the end-to-end delay, typically an order of magnitude less than an MCU. This makes it possible to have fully interactive multipoint conferences with even a very large number of participants. There are significant advantages as well in terms of error resilience and, in fact, error tolerance can be increased by nearly an order of magnitude here as well (e.g., using unequal error protection). Finally, the very low delay of an ALR allows these systems to be

cascaded, with significant benefits in terms of system design and deployment. Cascading of traditional MCUs is impossible due to the very high delay that even a single MCU introduces.

Scalable video coding enables a very significant paradigm shift in videoconferencing systems, bringing the complexity of video communication systems (particularly the servers residing within the network) in line with other types of network applications.

11.5. Mobile TV (Multicast to MANE, Unicast to Endpoint)

This scenario is a bit more complex, and designed to optimize the network traffic in a core network, while still requiring only a single pinhole in the endpoint's firewall. One of its key applications is the mobile TV market.

Consider a large private IP network, e.g., the core network of the Third Generation Partnership Project (3GPP). Streaming servers within this core network can be assumed to be professionally maintained. It is assumed that these servers can have many ports open to the network and that layered multicast is a real option. Therefore, the streaming server multicasts SVC scalable layers, instead of simulcasting different representations of the same content at different bitrates.

Also consider many endpoints of different classes. Some of these endpoints may lack the processing power or the display size to meaningfully decode all layers; others may have these capabilities. Users of some endpoints may wish not to pay for high quality and are happy with a base service, which may be cheaper or even free. Other users are willing to pay for high quality. Finally, some connected users may have a bandwidth problem in that they can't receive the bandwidth they would want to receive -- be it through congestion, connectivity, change of service quality, or for whatever other reasons. However, all these users have in common that they don't want to be exposed too much, and therefore the number of firewall pinholes needs to be small.

This situation can be handled best by introducing middleboxes close to the edge of the core network, which receive the layered multicast streams and compose the single SVC scalable bitstream according to the needs of the endpoint connected. These middleboxes are called MANEs throughout this specification. In practice, the authors envision the MANE to be part of (or at least physically and topologically close to) the base station of a mobile network, where all the signaling and media traffic necessarily are multiplexed on the same physical link.

MANEs necessarily need to be fairly complex devices. They certainly need to understand the signaling, so, for example, to associate the payload type octet in the RTP header with the SVC payload type.

A MANE may aggregate multiple RTP streams, possibly from multiple RTP sessions, thus to reduce the number of firewall pinholes required at the endpoints, or may optimize the outgoing RTP stream to the MTU size of the outgoing path by utilizing the aggregation and fragmentation mechanisms of this memo. This type of MANE is conceptually easy to implement and can offer powerful features, primarily because it necessarily can "see" the payload (including the RTP payload headers), utilize the wealth of layering information available therein, and manipulate it.

A MANE can also perform stream thinning, in order to adhere to congestion control principles as discussed in Section 9. While the implementation of the forward (media) channel of such a MANE appears to be comparatively simple, the need to rewrite RTCP RRs makes even such a MANE a complex device.

While the implementation complexity of either case of a MANE, as discussed above, is fairly high, the computational demands are comparatively low.

12. Acknowledgements

Miska Hannuksela contributed significantly to the designs of the PACSI NAL unit and the NI-C mode for decoding order recovery. Roni Even organized and coordinated the design team for the development of this memo, and provided valuable comments. Jonathan Lennox contributed to the NAL unit reordering algorithm for MST and provided input on several parts of this memo. Peter Amon, Sam Ganesan, Mike Nilsson, Colin Perkins, and Thomas Wiegand were members of the design team and provided valuable contributions. Magnus Westerlund has also made valuable comments. Charles Eckel and Stuart Taylor provided valuable comments after the first WGLC for this document. Xiaohui (Joanne) Wei helped improving Table 13 and the SDP examples.

The work of Thomas Schierl has been supported by the European Commission under contract number FP7-ICT-248036, project COAST.

13. References

13.1. Normative References

[H.264] ITU-T Recommendation H.264, "Advanced video coding for generic audiovisual services", March 2010.

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

[ISO/IEC14496-10] ISO/IEC International Standard 14496-10:2005.

[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.

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

[RFC4288] Freed, N. and J. Klensin, "Media Type Specifications and Registration Procedures", BCP 13, RFC 4288, December 2005.

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

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

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

[RFC5583] Schierl, T. and S. Wenger, "Signaling Media Decoding Dependency in the Session Description Protocol (SDP)", RFC 5583, July 2009.

[RFC6051] Perkins, C. and T. Schierl, "Rapid Synchronisation of RTP Flows", RFC 6051, November 2010.

13.2. Informative References

[DVB-H] DVB - Digital Video Broadcasting (DVB); DVB-H Implementation Guidelines, ETSI TR 102 377, 2005.

[Eleft] Eleftheriadis, A., R. Civanlar, and O. Shapiro, "Multipoint Videoconferencing with Scalable Video Coding", Journal of Zhejiang University SCIENCE A, Vol. 7, Nr. 5, April 2006, pp. 696-705. (Proceedings of the Packet Video 2006 Workshop.)

[G.114] ITU-T Rec. G.114, "One-way transmission time", May 2003.

[H.241] ITU-T Rec. H.241, "Extended video procedures and control signals for H.300-series terminals", May 2006.

[IGMP] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, October 2002.

[JVT-N026] Ohm J.-R., Koenen, R., and Chiariglione, L. (ed.), "SVC requirements specified by MPEG (ISO/IEC JTC1 SC29 WG11)", JVT-N026, available from http://ftp3.itu.ch/av-arch/jvt-site/2005_01_HongKong/JVT-N026.doc, Hong Kong, China, January 2005.

[JVT-N027] Sullivan, G. and Wiegand, T. (ed.), "SVC requirements specified by VCEG (ITU-T SG16 Q.6)", JVT-N027, available from http://ftp3.itu.int/av-arch/jvt-site/2005_01_HongKong/JVT-N027.doc, Hong Kong, China, January 2005.

[McCanne] McCanne, S., Jacobson, V., and Vetterli, M., "Receiver-driven layered multicast", in Proc. of ACM SIGCOMM'96, pages 117-130, Stanford, CA, August 1996.

[MBMS] 3GPP - Technical Specification Group Services and System Aspects; Multimedia Broadcast/Multicast Service (MBMS); Protocols and codecs (Release 6), December 2005.

[MPEG2] ISO/IEC International Standard 13818-2:1993.

[RFC2326] Schulzrinne, H., Rao, A., and R. Lanphier, "Real Time Streaming Protocol (RTSP)", RFC 2326, April 1998.

[RFC2974] Handley, M., Perkins, C., and E. Whelan, "Session Announcement Protocol", RFC 2974, October 2000.

[RFC5117] Westerlund, M. and S. Wenger, "RTP Topologies", RFC 5117, January 2008.

[RFC5775] Luby, M., Watson, M., and L. Vicisano, "Asynchronous Layered Coding (ALC) Protocol Instantiation", RFC 5775, April 2010.

[Yan] Yan, J., Katrinis, K., May, M., and Plattner, R., "Media- and TCP-friendly congestion control for scalable video streams", in IEEE Trans. Multimedia, pages 196-206, April 2006.

Authors' Addresses

Stephan Wenger
2400 Skyfarm Dr.
Hillsborough, CA 94010
USA

Phone: +1-415-713-5473
EMail: stewe@stewe.org

Ye-Kui Wang
Huawei Technologies
400 Crossing Blvd, 2nd Floor
Bridgewater, NJ 08807
USA

Phone: +1-908-541-3518
EMail: yekui.wang@huawei.com

Thomas Schierl
Fraunhofer HHI
Einsteinufer 37
D-10587 Berlin
Germany

Phone: +49-30-31002-227
EMail: ts@thomas-schierl.de

Alex Eleftheriadis
Vidyo, Inc.
433 Hackensack Ave.
Hackensack, NJ 07601
USA

Phone: +1-201-467-5135
EMail: alex@vidyo.com