RTP Payload Format for SVC Video
draft-ietf-avt-rtp-svc-08.txt

Status of this Memo

By submitting this Internet-Draft, each author represents that any applicable patent or other IPR claims of which he or she is aware have been or will be disclosed, and any of which he or she becomes aware will be disclosed, in accordance with Section 6 of BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt.

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

This Internet-Draft will expire on August 24, 2008.

Copyright Notice

Copyright (C) The IETF Trust (2008).

Abstract

This memo describes an RTP payload format for scalable video coding (SVC) defined in Annex G of the ITU-T Recommendation H.264 video codec 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, produced by the video encoder, in each RTP packet payload. The payload format has wide applicability, such as low bit-rate conversational, Internet video streaming, or high bit-rate entertainment quality video.
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. Formally, SVC takes the form of Amendment 3 to ISO/IEC 14496 Part 10 [MPEG4-10], and Annex G of ITU-T Rec. H.264/AVC [H.264]. The specification of SVC is available in [SVC].

SVC covers the whole application ranges of H.264/AVC, starting with low bit-rate Internet streaming applications to HDTV broadcast and Digital Cinema with nearly lossless coding and requiring dozens or hundreds of MBit/s.

This memo defines a backwards compatible enhancement to the H264/AVC payload format [RFC3984], in which the specific features introduced by SVC are taken into account. [Edt. Note (AE): Review backwards compatibility assertion, and qualify, when memo is completed.] Specifically, it documents the enhancements relevant from an RTP transport viewpoint, and defines signaling support for SVC, including a new media subtype name.

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. The SVC Codec

3.1. 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), as well as temporal...
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 3.3). 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 all the bitstream components required to be able to decode a particular dependency_id, quality_id, and temporal_id combination.

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 (AU). 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. 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. The NAL encapsulates each slice generated by the VCL into one or more NAL units. Please consult RFC 3984 for a more in-depth discussion of the NAL unit concept. SVC specifies the decoding order of NAL units.

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] sub-sequences have been introduced in order to allow optional use of temporal layers. SVC extends this approach by exposing the temporal scalability information using the temporal_id parameter, alongside the dependency_id and quality_id values that are used for spatial and quality scalability. For coded picture data defined in Annex G of [SVC] this is accomplished by using a new type of NAL unit 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 receivers, a new type of "prefix" NAL unit has been defined to carry this header information. This prefix NAL unit type is among those ignored by H.264/AVC receivers as explained in [RFC3984].

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) is the base layer representation and 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. Due to the backwards compatibility with H.264/AVC, it is important to differentiate however whether or not SVC-specific NAL units are present in a given bitstream or not. 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 to 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 contains only H.264/AVC VCL 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 [SVC] 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.

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.

SNR scalability in SVC is offered in two different ways. In what is called Coarse-Grained Scalability (CGS), scalability is provided by including or excluding a complete layer when decoding a particular bitstream. In contrast, in Medium-Grained 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 in the NAL unit header.

3.2. Parameter Set Concept
The parameter set concept is inherited from [H.264]. Please refer to section 1.2 of RFC 3984 for more details.

SVC introduced a new type of sequence parameter set, referred to as a subset sequence parameter set. Subset sequence parameter sets have NAL unit type equal to 15, which is different from the NAL unit type value (7) of sequence parameter set. 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. 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. At any time instant during the decoding process there may be one active sequence parameter set (for the layer representation with the highest value of (dependency_id * 16 + quality_id)) and one or more active layer SVC sequence parameter set(s) (for layer representations with lower values of (dependency_id * 16 + quality_id)). 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 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 (dependency_id * 16 + quality_id)) and one or more active layer picture parameter set(s) (for layer representations with lower values of (dependency_id * 16 + quality_id)). 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.

### 3.3. Network Abstraction Layer Unit Header

SVC NAL units of type 20 encapsulate VCL data as defined in Annex G of [SVC]. A special type of an SVC NAL unit is the prefix NAL unit (type 14) that includes descriptive information of the associated H.264/AVC VCL NAL unit (type 1 or 5) that immediately follows the prefix NAL unit.

SVC extends the one-byte H.264/AVC NAL unit header by three additional octets. 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 formally
specified in [SVC], but the essential properties of the NAL unit
header are summarized below.

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   |
+---------------+

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 [SVC], 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 [SVC].

In H.264/AVC, NAL unit types 14, 15 and 20 are reserved for future
extensions. SVC uses these three NAL unit types. 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 [SVC]). 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: 1 bit
reserved_one_bit. Reserved bit for future extension. R MUST be equal to 1. Receivers SHOULD ignore the value of R.

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 [SVC].

RR: 2 bits
reserved_three_2bits. Reserved bits for future extension. RR MUST be equal to ‘11’ (in binary form). Receivers SHOULD ignore the value of RR.

This memo reuses the same additional NAL unit types introduced in RFC 3984, which are presented in section 6.3. In addition, this memo introduces one OPTIONAL NAL unit type, 30, as specified in section 6.9. These NAL unit types are marked as unspecified in [SVC] and intentionally reserved for use in systems specifications like this memo. Moreover, this specification extends the semantics of F, NRI, I, PRID, DID, QID, TID, U, and D as described in section 6.4.

4. Scope

This payload specification can only be used to carry the "naked" NAL unit stream over RTP, and not the byte stream format according to Annex B of [SVC]. The likely applications of this specification will be in the IP based multimedia communications fields including conversational multimedia, video telephony or video conferencing, Internet streaming and TV over IP.

This specification allows, in a given RTP stream, to encapsulate NAL units belonging to
- the T0 AVC base layer or the T0 SVC base layer only, as detailed in [RFC3984], or
- one or more enhancement layers, or
Session multiplexing SHOULD be used when different receivers in the multicast session may request different operation points of the scalable bitstream. In session multiplexing, layers are carried in multiple RTP sessions, and each RTP session is associated with one RTP stream. The RTP stream in each RTP session MAY carry one or more layers, which can be any of the above three. When each operation point corresponding to a layer may be required by some receivers, then each Layer SHOULD be carried in its own RTP stream and its own RTP session. When fewer operation points are required by the receivers, then multiple layers MAY be encapsulated within one RTP stream in one RTP session.

Informative note: Layered multicast is a term commonly used to describe the application where multicast is used to transmit data that has been encapsulated into more than one RTP session using session multiplexing. 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 the informative Section 13.2.

When session multiplexing is not used, the following applies.

- When an H.264/AVC compatible subset of the SVC base layer is transmitted, the subset SHOULD be carried in one RTP stream that MUST be encapsulated according to RFC 3984. This way, a legacy RFC 3984 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, the set SHOULD be carried in one RTP stream that SHOULD be encapsulated according to this memo.

This RTP payload specification is designed to be unaware of the octet string in the NAL unit payload defined in [SVC]. The NAL unit header defined in [SVC] co-serves as the payload header of this RTP payload format, when single NAL unit packetization is used, i.e. one NAL unit per RTP packet. In this case, the payload of a NAL unit follows immediately. Additionally to [RFC3984], this memo locally defines a NAL unit type in the unspecified NAL unit type space of [SVC]. If other than the single NAL unit packetization mode is used as defined in [RFC3984] or this memo, locally defined NAL unit types may be additionally present in the RTP packets, together with one or more NAL unit types as specified in [SVC].
5. Definitions and Abbreviations

5.1. Definitions

5.1.1. Definitions per SVC specification

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

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 and 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 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 consist of one or more layer representations.

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

IDR 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[SVC].

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.

5.1.2. Definitions local 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. An anchor layer representation is a random access point to the layer the anchor layer representation belongs to. 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 may not be correctly decoded 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 [SVC].

base RTP session: The RTP session, among all the RTP sessions using session multiplexing, 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 [I-D.ietf-mmusic-decoding-dependency]. The base RTP session may carry NAL units of NAL unit type equal to 14 and 15.

enhancement RTP session: An RTP session, among all the RTP sessions using session multiplexing, 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 [I-D.ietf-mmusic-decoding-dependency]. A lower RTP session to an enhancement RTP session is an RTP session which the enhancement RTP session depends on.

cross-layer decoding order number (CL-DON): A derived variable indicating NAL unit decoding order number over all NAL units within all layers of an SVC bitstream.

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: A bitstream subset that conforms 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 types 1, 5, or 12) whether these prefix NAL units are present or not. A layer may contain NAL units associated with more than one values of temporal_id.

operation point: An operation point is identified by a set of values of temporal_id, dependency_id, and quality_id and is a bitstream subset 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 conforms to at least one of the profiles defined in Annex A or Annex G of [SVC], and offers a representation of the original video signal at a certain fidelity. [Edt. Note (YkW): The definition implies that all the non-VCL NAL units that are not directly associated with temporal_id, dependency_id, and quality_id are included any operation point. Let’s see whether this is always OK or any improvement is needed.]

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 PT and identical SSRC (Synchronization Source), carried in one RTP session. Within the scope of this memo, one RTP packet stream is utilized to transport one or more layers.

session multiplexing: The scalable SVC bitstream is distributed onto different RTP sessions, whereby each RTP session carries a single RTP packet stream. Each RTP session requires a separate signaling and has a separate Timestamp, Sequence Number, and SSRC space. Timestamps for the RTP sessions SHALL be derived from the same clock instance. Dependency between sessions MUST be signaled according to [I-D.ietf-mmusic-decoding-dependency] and this memo.

SVC base layer: The layer that includes all NAL units associated with dependency_id and quality_id values both equal to 0.
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 temporal_id, dependency_id, and quality_id values 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 [SVC].

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. The lowest SVC RTP session is the base RTP session, and the highest RTP session is the enhancement RTP session which no other RTP session depends on, or the base RTP session if no enhancement RTP session exists.

[Edt. Note (YkW): There may be multiple RTP sessions that no other RTP session depends on. We may limit the scope of lower or higher RTP sessions to be within a given receiver, which either receive one session or multiple sessions with at most one session no other session depends on. However, in that case, the lowest session may not be the base 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.

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 their associated prefix NAL units.

5.2. Abbreviations

In addition to the abbreviations defined in [RFC3984], the following ones are defined.

CGS: Coarse-Grain Scalability
CL-DON: Cross-Layer Decoding Order Number
MGS: Medium-Grain Scalability
PACSI: Payload Content Scalability Information
SVC: Scalable Video Coding

6. RTP Payload Format
6.1. Design Principles

The following design principles have been observed:

- Backward compatibility with [RFC3984] wherever possible.

- The SVC base layer or any H.264/AVC compatible subset containing
  the T0 SVC base layer and one or more temporal enhancement layers,
  when transmitted in its own session, MUST be
  encapsulated using [RFC3984]. Requiring this has the desirable
  side effect that it can be used by [RFC3984] legacy devices.

- MANEs are signaling aware and rely on signaling information.
  MANEs have state.

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

- MANEs can perform media-aware stream thinning. By using the
  payload
  header information identifying Layers within an RTP session,
  MANEs are able to remove packets from the incoming RTP packet
  stream. This implies rewriting
  the RTP headers of the outgoing packet stream and rewriting of
  RTCP Receiver Reports.

6.2. RTP Header Usage

Please see section 5.1 of [RFC3984].

6.3. Common Structure of the RTP Payload Format

Please see section 5.2 of [RFC3984].

6.4. NAL Unit Header Usage

The structure and semantics of the NAL unit header were introduced
in section 3.3. This section specifies the semantics of F, NRI, I,
PRID, DID, QID, TID, U, and D according to this specification.

The semantics of F specified in section 5.3 of [RFC3984] also
applies herein.

For NRI, for the bitstream conforming to one of the profiles defined
in Annex A of [H.264] and transported using [RFC3984], the semantics
specified in section 5.3 of [RFC3984] are applicable, i.e., NRI also
indicates the relative importance of NAL units. In an SVC context,
in addition to the semantics specified in Annex G of [SVC], NRI also indicates the relative importance of NAL units within a layer.

[Edt. Note (YkJW): "SVC context" to be clearly specified.]

For I, in addition to the semantics specified in Annex G of [SVC], 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 sensed 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 success delivery of the packet conveying the NAL unit, including applying a QoS-enabled network, Forward Error Correction (FEC), retransmissions, and advanced scheduling behavior, whenever possible.

For PRID, the semantics specified in Annex G of [SVC] applies. 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 an FEC protection mechanism. The importance for the decoding process decreases as the PRID value increases.

For DID, QID, TID, in addition to the semantics specified in Annex G of [SVC], 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 [SVC], 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 [SVC], 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.

6.5. Packetization Modes

Please see section 5.4 of [RFC3984].

6.6. Decoding Order Number (DON)

Please see section 5.5 of [RFC3984]. The following applies in addition.

If different layers of a SVC bitstream are transported in more than one RTP session, the DON values derived according to RFC 3984 of all the NAL units in the RTP sessions using interleaved mode MUST indicate CL-DON values.

When the CL-DON decoding order recovery mode is used with session multiplexing as described in section 7.1 and at least one STAP-A packet is present in any of the RTP sessions, the following applies.
- A PACSI NAL unit MUST be present in each STAP-A packet.
- A DONC field MUST be present in the PACSI NAL unit included in each STAP-A.
- The DON values for the NAL units in each STAP-A packet MUST be derived as follows and MUST indicate CL-DON values.
  o The DONC field in the PACSI NAL unit specifies the value of DON for the first NAL unit in the STAP-A in transmission order. For each successive NAL unit in appearance order in the STAP-A, the value of DON is equal to (the value of DON of the previous NAL unit in the STAP-A + 1) % 65536, wherein ‘%’ stands for modulo operation.

6.7. Aggregation Packets

Please see section 5.7 of [RFC3984].

6.8. Fragmentation Units (FUs)

Please see section 5.8 of [RFC3984].

6.9. Payload Content Scalability Information (PACSI) NAL Unit

A new NAL unit type is specified in this memo, and referred to as payload content scalability information (PACSI) NAL unit. The OPTIONAL PACSI NAL unit, if present, MUST be the first NAL unit in an aggregation packet, and it MUST NOT be present in other types of packets. The PACSI NAL unit indicates scalability information and
other characteristics that are common for all the remaining NAL units in the payload of the aggregation packet. Furthermore, a PACSI NAL unit MAY contain zero or more SEI NAL units. PACSI NAL unit makes it easier for MANEs to decide whether to forward/process/discard the aggregation packet containing the PACSI NAL unit. Other reasons to use PACSI NAL units are explained later when specifying the semantics of the fields. Senders MAY create PACSI NAL units and receivers MAY ignore them, or use them as hints to enable efficient aggregation packet processing. Note that the NAL unit type for the PACSI NAL unit is selected among those values that are unspecified in [SVC] and [RFC3984].

When the first aggregation unit of an aggregation packet contains a PACSI NAL unit, there MUST be at least one additional aggregation unit present in the same packet. The RTP header and payload header fields of the aggregation packet are set according to the remaining NAL units in the aggregation packet.

When a PACSI NAL unit is included in a multi-time aggregation packet (MTAP), 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.

The structure of a PACSI NAL unit is as follows. The first four octets are exactly the same as the four-byte SVC NAL unit header as discussed in section 3.3. They are followed by one always present octet, five optional octets, and zero or more SEI NAL units, each SEI NAL unit 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 type octet of the SEI NAL unit). Figure 1 illustrates the PACSI NAL unit structure and an example of a PACSI NAL unit containing two SEI NAL units.

The bits A, P, C, S, and E are specified only if the bit X is equal to 1. The fields TL0PICIDX and IDR PICID are present only if the bit Y is equal to 1. The fields TL0PICIDX and IDR PICID MUST NOT be present if the bit Y is equal to 0. The field DONC is present only if the bit T is equal to 1. The field T MUST be equal to 0 if the aggregation packet containing the PACSI NAL unit is not an STAP-A packet.

```
0                   1                   2                   3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|F|NRI|  Type   |R|I|   PRID    |N| DID |  QID  | TID |U|D|O| RR|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Wenger, Wang, Schierl Expires August 24, 2008 [Page 21]
The values of the fields in PACSI NAL unit MUST be set as follows. The term "target NAL units" are used in the semantics of some fields. The target NAL units are such NAL units contained in the aggregation packet, but not included in the PACSI NAL unit, that are within the access unit to which the first NAL unit following the PACSI NAL unit in the aggregation packet belongs.

- The F bit MUST be set to 1 if the F bit in at least one of the remaining NAL units in the payload of the aggregation packet is equal to 1. 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 payload of the aggregation packet.

- The Type field MUST be set to 30.

- The R bit MUST be set to 1. Receivers SHOULD 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 payload of the aggregation packet is equal to 1. Otherwise, the I bit MUST be set to 0.

- The PRID field MUST be set to the lowest value of the PRID values of all the remaining NAL units in the payload of the aggregation packet.
o The N bit MUST be set to 1 if the N bit of all the remaining NAL units in the payload is equal to 1. Otherwise, the N bit MUST be set to 0.

o The DID field MUST be set to the lowest value of the DID values of all the remaining NAL units in the payload of the aggregation packet.

o The QID field MUST be set to the lowest value of the QID values of all the remaining NAL units with the lowest value of DID in the payload.

o The TID field MUST be set to the lowest value of the TID values of all the remaining NAL units with the lowest value of DID in the payload.

o The U bit MUST be set to 1 if the U bit of at least one of the remaining NAL units in the payload of the aggregation packet is equal to 1. Otherwise, the U bit MUST be set to 0.

o The D bit MUST be set to 1 if the D value of all the remaining NAL unit in the payload is equal to 1. Otherwise, the D bit MUST be set to 0.

o The O bit MUST be set to 1 if the O bit of at least one of the remaining NAL units in the payload of the aggregation packet is equal to 1. Otherwise, the O bit MUST be set to 0.

o The RR field MUST be set to '11' (in binary form). Receivers SHOULD ignore the value of RR.

o If the X bit is equal to 1, the bits A, P, C, S, and E are specified as in below. Otherwise, the bits A, P, C, S, and E are unspecified, and receivers MUST ignore these bits. The X bit SHOULD be identical for all the PACSI NAL units in all the RTP sessions conveying an SVC bitstream.

o If the Y bit is equal to 1, the OPTIONAL fields TL0PICIDX and IDRPLICID MUST be present and specified as in below. Otherwise, the fields TL0PICIDX and IDRPLICID MUST NOT be present. The Y bit SHOULD be identical for all the PACSI NAL units involved in all the RTP sessions conveying an SVC bitstream.
o 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.

o The A bit MUST be set to 1 if all the target NAL units belong to anchor layer representations. Otherwise, the A bit MUST be set to 0. The A bit SHOULD be identical for all the PACSI NAL units for which the target NAL units belong to the same access unit.

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. When the coded pattern like IBBP is in use, 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 recovery point SEI message. However, with this A bit it is much easier to parse than to parse the recovery point SEI message, which may even be buried deeply in an SEI NAL unit. Furthermore, the SEI message may not be present in the bitstream.

o The P bit MUST be set to 1 if all the remaining NAL units in the payload of the aggregation packet are with redundant_pic_cnt greater than 0, i.e. the slices are redundant slices. Otherwise, the P bit MUST be set to 0.

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

o The C bit MUST be set to 1 if the target NAL units (as defined above) belong to an access unit for which the layer representation having the greatest value of dependency_id among all the layer representations containing the target NAL units is an intra layer representation. Otherwise, the C bit MUST be set to 0. The C bit SHOULD be identical for all the PACSI NAL units for which the target NAL units belong to the same access unit.

Informative note: The C bit indicates whether the packet contains intra slices which may be the only packets to be forwarded for a fast forward playback, e.g. when the network condition is extremely bad.
o The S bit MUST be set to 1, if the first VCL NAL unit, in decoding order, of the layer representation containing the first NAL unit following the PACSI NAL unit in the aggregation packet is present in the payload. Otherwise, the S bit MUST be set to 0.

o The E bit MUST be set to 1, if the last VCL NAL unit, in decoding order, of the layer representation containing the first NAL unit following the PACSI NAL unit in the aggregation packet is present in the payload. Otherwise, the E field MUST be set to 0.

Informative note: The S or E bit indicates whether the first or last slice, in decoding order, of a layer representation is in the packet, to enable a MANE to detect slice loss and take proper action such as requesting a retransmission as soon as possible, as well as to allow an efficient playout buffer handling similarly as the M bit 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.

o When present, the TL0PICIDX field MUST be set to equal to t10_dep_rep_idx as specified in Annex G of [SVC] for the layer representation containing the first NAL unit following the PACSI NAL unit in the aggregation packet.

o When present, the IDRPICID field MUST be set to equal to effective_idr_pic_id as specified in Annex G of [SVC] for the layer representation containing the first NAL unit following the PACSI NAL unit in the aggregation packet.

Informative note: The TL0PICIDX and IDRPICID fields enable the detection of the loss of layer representations in the most important temporal layer, by receivers as well as MANEs. SVC includes a solution by using SEI messages, which are harder to parse and may not be present in the bitstream at all.

o When present, the field DONC indicates the cross-layer decoding order number for the first NAL unit in the STAP-A in transmission order.

The PACSI NAL unit SHALL include a subset (zero to all) of the SEI NAL units associated with the access unit to which the target NAL units belong, and SHALL NOT contain SEI NAL units associated with any other access unit.

Informative note: Senders may repeat such SEI NAL units in the PACSI NAL unit the presence of which in more than one packet is essential for packet loss robustness. Receivers may use the
repeated SEI messages in place of missing SEI messages. 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 packet of those packets conveying an access unit.

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

7. Packetization Rules

Please see section 6 of [RFC3984]. The following rules apply in addition.

All receivers MUST support the single NAL unit packetization mode to provide backward compatibility to endpoints supporting only the single NAL unit mode of RFC 3984. 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, e.g. small NAL units containing parameter sets or SEI messages, in their own packets is typically less efficient because of the relatively big overhead.

All receivers MUST support the non-interleaved mode of [RFC3984].

Informative note: The non-interleaved mode allows an application to encapsulate a single NAL unit in a single RTP packet. Historically, the single NAL unit mode has been included into [RFC3984] 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. More technically speaking, the implementation complexity increase for providing the additional mechanisms of the non-interleaved mode (namely STAP-A and FU-A) is minor, and the benefits are great, that STAP-A and FU-A implementation is required.

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 delimiter, parameter set, or SEI NAL unit 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.

Informative note: When either the prefix NAL unit or the associated NAL unit containing an H.264/AVC coded slice is lost, the remaining one would be hardly useful in SVC context, wherein the prefix NAL unit must be available for decoded picture buffer management operations of the decoding process.

When the first aggregation unit of an aggregation packet contains a PACSI NAL unit, there MUST be at least one additional aggregation unit present in the same packet.

7.1. Packetization Rules for session multiplexing

When session multiplexing is used, decoding order recovery for NAL units carried in all the RTP sessions is needed. Two alternative decoding order recovery modes are provided for session multiplexing. The first is referred to as the classical RTP decoding order recovery mode, where CL-DON MUST NOT be used. The second is referred to as the CL-DON decoding order recovery mode, where CL-DON MUST be used.

[Edt. Note (TS): Definition of the CL-DON as in section 6.6 in this draft allows for INTERLEAVED transmission of NAL units using the non-INTERLEAVED packetization mode. This is really dangerous for interoperability and backward compatibility. Thus the whole specification of the (CL-)DON basically allows for the INTERLEAVING feature, the only way out may be to connect the use of CL-DON to the interleaved mode (packetization-mode equal to 2).]

If the classical RTP decoding order recovery mode is in use, either the single NAL unit packetization mode, the non-interleaved or the interleaved packetization mode can be used. Different RTP sessions MAY still use different packetization modes.

If the CL-DON decoding order recovery mode is in use, either the non-interleaved packetization mode, restricted to STAP-A packets only, i.e. FU-A and Single NAL unit packets MUST NOT be used, or the interleaved packetization mode MAY be used. As CL-DON MUST be used, the CL-DON value must be derivable from the payload structure for this decoding order recovery mode. Different RTP sessions MAY still use different packetization modes.

The respective packetization rules for the two decoding order recovery modes in session multiplexing are as follows.
I. The classical RTP decoding order recovery mode

a. If an access unit of sampling time instance X is present in RTP session A, this access unit MUST be also present in any RTP session, which depends on RTP session A.

b. When a PACSI NAL unit is present, the T bit MUST be equal to 0, i.e. the DONC field MUST NOT be present.

c. The sprop-cl-don parameter MUST NOT be present in the session description.

   Informative note: Restriction a. may be achieved for pre-encoded content by inserting filler data NAL units (NAL unit type 12) or filler payload SEI messages (NAL unit type 6, SEI message payload type equal to 3) as defined in [SVC]. This insertion can be achieved by encoders, servers, as well as by MANEs.

II. The CL-DON decoding order recovery mode

a. For each RTP session, the non-interleaved packetization mode or the interleaved packetization mode MUST be used.

b. For any RTP session that uses the interleaved packetization mode, the DON values derived as specified in RFC 3984 MUST indicate CL-DON.

c. For any RTP session that uses the non-interleaved packetization mode, the following applies.

   i. STAP-A MUST be used, and any other type of packets allowed (i.e. single NAL unit packet or FU-A packets) MUST NOT be used.

   ii. Each STAP-A MUST contain a PACSI NAL unit and the DONC field MUST be present in the PACSI NAL unit.

d. The sprop-cl-don parameter MUST be present in the session description.

8. De-Packetization Process (Informative)

   For a single RTP session, the de-packetization process specified in section 7 of [RFC3984] applies [Edt. Note: with some fixes to
For receiving more than one of multiple RTP sessions conveying a scalable bitstream, the de-packetization process is specified in section 7.3 (Additional De-Packetization Guidelines) of RFC 3984. 

8.1. De-Packetization Process for NAL Units Conveyed using Session Multiplexing

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

In this section, "the RTP sessions" refer to the RTP sessions for which the NAL units are de-packetized.

The sessions to be received SHALL be identified by mechanisms specified in [I-D.ietf-mmusic-decoding-dependency].

For each of the RTP sessions, the RTP reception process as specified in RFC 3550 is applied, such that the received packets are passed in increasing order of timestamp, and, for those RTP packets with identical RTP timestamp, in increasing order of sequence number, into the payload de-packetization to NAL units as defined in this memo.

The decoding order of the NAL units carried in all the RTP sessions is then recovered by applying section 8.1.1 or 8.1.2, depending on the presence of the parameter sprop-cl-don in the session description.

8.1.1. The Classical RTP Decoding Order Recovery Mode

This process SHALL be used when the parameter sprop-cl-don is not present in the session description.

In this section, the NAL unit decoding order recovery process is described for the constraints in section 7 using the non-interleaved, interleaved mode or Single NAL unit mode for all RTP sessions, i.e. CL-DON SHALL NOT be present in any of these sessions. The process is based on RTP session dependency signaling, RTP sequence numbers, and timestamps.

1. Within each RTP stream, the decoding order of NAL units SHALL be recovered according to the following rules:
When using the single NAL unit mode (packetization-mode equal to 0) or the non-interleaved mode (packetization mode equal to 1), the RTP header sequence number SHALL give the decoding order as specified in [RFC3984].

When the interleaved mode (packetization-mode equal to 2) is used, the Decoding Order Number SHALL give the decoding order as specified in [RFC3984].

Informative note:
The decoding order recovery process cannot rely on timestamp increase as indicator for decoding order.

2. The decoding order of NAL units from multiple RTP streams in multiple RTP sessions SHALL be recovered into a single sequence of NAL units, grouped into access units, by performing the following rules:

- NAL units with the same (RTP or NTP) timestamp are grouped in decoding order to operation point representations in each RTP stream.

- Operation point representations with the same (NTP) timestamp SHALL be grouped to access units in order of the SVC RTP session dependency, from lowest to highest.
  
  Note: There may be operation point representations which do not have any corresponding operation point representations in RTP streams of lower SVC RTP sessions.

- [Edt. Note (TS): When SEI messages are not part of the base RTP session, they are not transported in decoding order. Only in that case this paragraph is required. See open issue section.] SEI messages (NAL unit type equal to 6), when present in any RTP session not equal to the base RTP session, SHALL be re-ordered so that the resulting access unit order conforms to [SVC], i.e. NAL units with nal_unit_type equal to 6 present in any of the RTP streams shall be grouped and precede directly any NAL units of type 1, 5, 14, 15 and 20 in the access unit.

- The access units SHALL be passed in order of decoding order appearance of timestamps in the highest SVC RTP session to the decoder.

Informative example:
The example shown in Figure 2. refers to three SVC RTP sessions A, B and C. In the example, the dependency signaling as described in 9.2.3, indicates, that Session A does not depend on any other of the sessions; B depends on A; C depends on A and B as restricted in section 7. Session A has the lowest frame rate and Session B and C have the same, but a higher frame rate. Figure 2. shows an example for buffering with different jitters in the sessions, i.e. at buffering startup not all packets of the same time instance are available.

The process first proceeds to TS [8] and remove/ignore all preceding NAL units in each of the buffers of RTP session A, B, and C. Then starting from session C, the first timestamp available in decoding order (TS [1]) is selected and all operation point representations in lower RTP sessions A and B are moved in order of the RTP session dependency (in the example form session A -> B -> C) into the decoder. Then the next timestamp in the highest RTP session C is selected and the process described above is repeated. Note, that there may be no matching operation point representations at the lowest RTP session A, e.g. for TS[5].

In case of "real" packet loss at TS[4] and TS[2], a similar processing as described above may be applied to get synchronized with the timestamp order as given by the highest RTP session (in the example - session "C"). During an interval where losses are present in the highest RTP session, reordering may be in some cases only possible up to the highest loss-free received RTP session. Such cases may occur, when multiple losses in the highest RTP session are present and the correct order of timestamps over all the sessions cannot be recovered by any other than the highest RTP session.

Decoding order and dependency of NAL units per received RTP session with different jitters in sessions at buffering startup time:

C: 6---(1,2)---(3,4)---(5)---(6,7,8)---(9,10)---(11,12)---
   |    |    |    |    |    |    |    |    |    |
B: 6---(1,2)---(3,4)---(5)---(6,7,8)---(9,10)---(11,12)---(13,14)---(15,15)---
   |    |    |    |    |    |    |    |    |
A: 6---(1)---(2)---(3)---(4)---(5)---
   |    |    |    |    |    |    |

Key:
A, B, C - RTP sessions
8.1.2. The CL-DON Decoding Order Recovery Mode

This process SHALL be used when the parameter sprop-cl-don is present in the session description.

In this mode, for each NAL unit a CL-DON value can be derived. This enables NAL unit decoding order recovery for all the RTP sessions without requiring separate decoding order recovery for each RTP session beforehand.

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

The NAL unit decoding order recovery process as described below is then exactly the same as the single session decoding order recovery process for the interleaved packetization mode as specified in subsection 7.2 of RFC 3984, with deinterleaving buffer replaced by de-session-multiplexing buffer and DON replaced by CL-DON.

It is RECOMMENDED to set the size of the de-session-multiplexing buffer, in terms of number of bytes, equal to or greater than the value of the sprop-deint-buf-req media type parameter of the RTP session conveying the SVC Layer for which the decoding requires the presence of the SVC Layers conveyed in all the other RTP sessions, referred to the highest RTP session.

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

Regardless of the buffering state, the receiver stores incoming NAL units. The CL-DON value is calculated and stored for each NAL unit.

The receiver operation is described below with the help of the following functions and constants:
o Function AbsDON is specified in section 9.1 of this specification.

o Function don_diff is specified in section 5.5 of RFC 3984.

o Constant N is the value of the OPTIONAL sprop-interleaving-depth media type parameter of the highest RTP session incremented by 1.

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

o There are N or more VCL NAL units in the de-session-multiplexing buffer.

o If sprop-max-don-diff of the highest SVC RTP session is present, don_diff(m,n) is greater than the value of sprop-max-don-diff of the highest RTP session, in which n corresponds to the NAL unit having the greatest value of AbsDON among the received NAL units and m corresponds to the NAL unit having the smallest value of AbsDON among the received NAL units.

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

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

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

o If sprop-max-don-diff of the highest SVC 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 de-session-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 de-session-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 CL-DON, a CL-DON distance is calculated as follows. If the value of CL-DON of the NAL unit is larger than the value of PDON, the CL-DON distance is equal to CL-DON - PDON. Otherwise, the CL-DON distance is equal to 65535 - PDON + CL-DON + 1.

NAL units are delivered to the decoder in ascending order of CL-DON distance. If several NAL units share the same value of CL-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 CL-DON for the last NAL unit passed to the decoder.

9. 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 not be compatible with some signaling protocol concepts, in which case the use of these parameters SHOULD be avoided.

9.1. Media Type Registration

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

The receiver MUST ignore any unspecified parameter.

Informative note: Requiring that the receiver ignores 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 RFC 3984.

Media Type name: video

Media subtype name: H264-SVC or H264

The media subtype "H264" MUST be used for RTP streams using RFC 3984, i.e. not using any of the new features introduced by this specification compared to RFC 3984. [Edt. Note: The new features are to be listed herein.] For RTP streams using any of the new features introduced by this specification compared to RFC 3984, the media subtype "H264-SVC" SHOULD be used, and the media subtype "H264" MAY be used. Use of the media subtype "H264" for RTP streams using the new features allows for RFC 3984 receivers to negotiate and receive H.264/AVC or SVC streams packetized according to this specification, but to ignore media parameters and NAL unit types it does not recognize.

Required parameters: none

OPTIONAL parameters:

profile-level-id:

A base16 [RFC3548] (hexadecimal) representation of the following three bytes in the sequence parameter set NAL unit specified in [SVC]: 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, and reserved_zero_4bits in bit-significance order, starting from the most significant bit, and 3) level_idc. Note that reserved_zero_4bits is required to be equal to 0 in [SVC], but other values for it may be specified in the future by ITU-T or ISO/IEC.

If the profile-level-id parameter is used to indicate properties of a NAL unit stream, it indicates the profile and level that a decoder has to support in order to comply with [SVC] when it decodes the NAL unit stream. The profile-iop byte indicates whether the NAL unit stream also
obeys all the constraints as specified in subsection 7.4.2.1.1 of [SVC]. Herein the NAL unit stream refers to the one consisting of all NAL units conveyed in the current RTP session, and all NAL units conveyed in other RTP sessions, if present, the current RTP session depends on. The current RTP session MAY depend on other RTP sessions when a scalable bitstream is transported with more than one RTP session and the current session is not an independent RTP session.

If the profile-level-id parameter is used for capability exchange or session setup procedure, it indicates the profile that the codec supports and the highest level supported for the signaled profile. The profile-iop byte indicates whether the codec has additional limitations whereby only the common subset of the algorithmic features and limitations signaled with the profile-iop byte is supported by the codec. For example, if a codec supports only the common subset of the coding tools of the Baseline profile and the Main profile at level 2.1 and below, the profile-level-id becomes 42E015, in which 42 stands for the Baseline profile, E0 indicates that only the common subset for all profiles is supported, and 15 indicates level 2.1.

Informative note: Capability exchange and session setup procedures should provide means to list the capabilities for each supported codec profile separately. For example, the one-of-N codec selection procedure of the SDP Offer/Answer model can be used (section 10.2 of [RFC4566]).

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

max-mbps, max-fs, max-cpb, max-dpb, and max-br:
These parameters MAY be used to signal the capabilities of a receiver or a sender implementation.
These parameters MUST NOT be used for any other purpose. The profile-level-id parameter MUST be present in the same receiver capability description that contains any of these parameters. The level conveyed in the value of
the profile-level-id parameter MUST be such that the receiver is fully capable of supporting. max-mbps, max-fs, max-cpb, max-dpb, and max-br MAY be used to indicate capabilities of the receiver that extend the required capabilities of the signaled level, as specified below.

When more than one parameter from the set (max-mbps, max-fs, max-cpb, max-dpb, max-br) is present, the receiver MUST support all signaled capabilities simultaneously. For example, if both max-mbps and max-br are present, the signaled level with the extension of both the frame rate and bit rate is supported. That is, the receiver is able to decode NAL unit streams in which the macroblock processing rate is up to max-mbps (inclusive), the bit rate is up to max-br (inclusive), the coded picture buffer size is derived as specified in the semantics of the max-br parameter below, and other properties comply with the level specified in the value of the profile-level-id parameter.

A receiver MUST NOT signal values of max-mbps, max-fs, max-cpb, max-dpb, and max-br that meet the requirements of a higher level, referred to as level A herein, compared to the level specified in the value of the profile-level-id parameter, if the receiver can support all the properties of level A.

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

max-mbps: The value of max-mbps is an integer indicating the maximum macroblock processing rate in units of macroblocks per second. The max-mbps parameter signals that the receiver is capable of decoding video at a higher rate than is required by the signaled level conveyed in the value of the profile-level-id parameter. When max-mbps is signaled, the receiver MUST be able to decode NAL unit streams that conform to the
signaled level, with the exception that the MaxMBPS value in Table A-1 or Table G-n of [SVC] for the signaled level is replaced with the value of max-mbps. The value of max-mbps MUST be greater than or equal to the value of MaxMBPS for the level given in Table A-1 or Table G-n of [SVC]. Senders MAY use this knowledge to send pictures of a given size at a higher picture rate than is indicated in the signaled level.

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

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 (see A.3.1 item i or G.n item m of [SVC]) and in units of 1200 bits for the NAL HRD parameters (see A.3.1 item j or G.n item m of [SVC]). The max-cpb parameter signals that the receiver has more memory than the minimum amount of coded picture buffer memory required by the signaled level conveyed in the value of the profile-level-id parameter. When max-cpb is signaled, the receiver MUST be able to decode NAL unit streams that conform to the signaled level, with the exception that the MaxCPB value in Table A-1 or Table G-n of [SVC] for the signaled level is replaced with the value of max-cpb. The value of max-cpb MUST be greater
than or equal to the value of MaxCPB for the
level given in Table A-1 or Table G-n of [SVC].
Senders MAY
use this knowledge to construct coded video
streams with greater variation of bit rate
than can be achieved with the
MaxCPB value in Table A-1 or Table G-n of [SVC].

Informative note: The coded picture buffer
is used in the hypothetical reference
decoder (Annex C) of SVC. The use of the
hypothetical reference decoder 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 and de-jitter
buffers. The coded picture buffer need not
be implemented in decoders as specified in
Annex C of SVC, but rather standard-
compliant decoders can have any buffering
arrangements provided that they can decode
standard-compliant bitstreams. Thus, in
practice, the input buffer for video
decoder can be integrated with de-
interleaving and de-jitter buffers of the
receiver.

max-dpb: The value of max-dpb is an integer indicating
the maximum decoded picture buffer size in
units of 1024 bytes. The max-dpb parameter
signals that the receiver has more memory than
the minimum amount of decoded picture buffer
memory required by the signaled level conveyed
in the value of the profile-level-id parameter.
When max-dpb is signaled, the receiver MUST be
able to decode NAL unit streams that conform to
the signaled level, with the exception that the
MaxDPB value in Table A-1 or Table G-n of [SVC]
for the
signaled level is replaced with the value of
max-dpb. Consequently, a receiver that signals
max-dpb MUST be capable of storing the
following number of decoded frames,
complementary field pairs, and non-paired
fields in its decoded picture buffer:

\[
\text{Min}(1024 \times \text{max-dpb} / \{ \text{PicWidthInMbs} \times \\
\text{FrameHeightInMbs} \times 256 \times \text{ChromaFormatFactor} \})
\]
PicWidthInMbs, FrameHeightInMbs, and ChromaFormatFactor are defined in [SVC].

The value of max-dpb MUST be greater than or equal to the value of MaxDPB for the level given in Table A-1 or Table G-n of [SVC]. Senders MAY use this knowledge to construct coded video streams with improved compression.

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

**max-br:**

The value of max-br is an integer indicating the maximum video bit rate in units of 1000 bits per second for the VCL HRD parameters (see A.3.1 item i or G.n item m of [SVC]) and in units of 1200 bits per second for the NAL HRD parameters (see A.3.1 item j or G.n item m of [SVC]).

The max-br parameter signals that the video decoder of the receiver is capable of decoding video at a higher bit rate than is required by the signaled level conveyed in the value of the profile-level-id 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 level, conveyed in the profile-level-id parameter, with the following exceptions in the limits specified by the level:

- The value of max-br replaces the MaxBR value of the signaled level (in Table A-1 of or Table G-n of [SVC]).
- When the max-cpb parameter is not present, the result of the following formula replaces the value of MaxCPB in Table A-1 or Table G-n of [SVC]:
  \[(\text{MaxCPB of the signaled level}) \times \text{max-br} / \]
For example, if a receiver signals capability for 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 MUST be greater than or equal to the value MaxBR for the signaled level given in Table A-1 or Table G-n of [SVC].

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 bit rates at any given time. In particular, no conclusion can be drawn that the signaled bit rate is possible under congestion control constraints.

**redundant-pic-cap:**

This parameter signals the capabilities of a receiver implementation. When equal to 0, the parameter indicates that the receiver makes no attempt to use redundant coded pictures to correct incorrectly decoded primary coded pictures. When equal to 0, the receiver is not capable of using redundant slices; therefore, a sender SHOULD avoid sending redundant slices to save bandwidth. When equal to 1, the receiver is capable of decoding any such redundant slice that covers a corrupted area in a primary decoded picture (at least partly), and therefore a sender MAY send redundant slices. When the parameter is not present, then a value of 0 MUST be used for redundant-pic-cap. When present, the value of redundant-pic-cap MUST be either 0 or 1.

When the profile-level-id parameter is present
in the same capability signaling as the redundant-pic-cap parameter, and the profile indicated in profile-level-id is such that it disallows the use of redundant coded pictures (e.g., Main Profile), the value of redundant-pic-cap MUST be equal to 0. When a receiver indicates redundant-pic-cap equal to 0, the received stream SHOULD NOT contain redundant coded pictures.

Informative note: Even if redundant-pic-cap is equal to 0, the decoder is able to ignore redundant codec pictures provided that the decoder supports such a profile (Baseline, Extended) in which redundant coded pictures are allowed.

Informative note: Even if redundant-pic-cap is equal to 1, the receiver may also choose other error concealment strategies to replace or complement decoding of redundant slices.

sprop-parameter-sets:
This parameter MAY be used to convey any sequence and picture parameter set NAL units (herein referred to as the initial parameter set NAL units) that MUST be placed in the NAL unit stream to precede any other NAL units in decoding order by the receiver. The parameter MUST NOT be used to indicate codec capability in any capability exchange procedure. The value of the parameter is the base64 [RFC3548] representation of the initial parameter set NAL units as specified in sections 7.3.2.1, 7.3.2.2 and G.7.3.2.1.3 of [SVC]. The parameter sets are conveyed in decoding order, and no framing of the parameter set NAL units takes place. A comma is used to separate any pair of parameter sets in the list. Note that the number of bytes in a parameter set NAL unit is typically less than 10, 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 double-buffer all sprop-parameter-sets and make them available to the decoder instance that decodes a certain payload type.

parameter-add:
This parameter MAY be used to signal whether the receiver of this parameter is allowed to add parameter sets in its signaling response using the sprop-parameter-sets media parameter. The value of this parameter is either 0 or 1. 0 is equal to false; i.e., it is not allowed to add parameter sets. 1 is equal to true; i.e., it is allowed to add parameter sets. If the parameter is not present, its value MUST be 1.

packetization-mode:
This parameter signals the properties of an RTP payload type or the capabilities of a receiver implementation. Only a single configuration point can be indicated; thus, when capabilities to support more than one packetization-mode are declared, multiple configuration points (RTP payload types) must be used.

When the value of packetization-mode is equal to 0 or packetization-mode is not present, the single NAL mode, as defined in section 6.2 of RFC 3984, MUST be used. This mode is in use in standards using ITU-T Recommendation H.241 [H.241] (see section 12.1 of RFC 3984). When the value of packetization-mode is equal to 1, the non-interleaved mode, as defined in section 6.3 of RFC 3984, MUST be used. When the value of packetization-mode is equal to 2, the interleaved mode, as defined in section 6.4 of RFC 3984, MUST be used. The value of packetization mode MUST be an integer in the range of 0 to 2, inclusive.

sprop-interleaving-depth:
This parameter MUST NOT be present when the current RTP session does not depend on any other RTP session, and packetization-mode is not present or the value of packetization-mode is equal to 0 or 1. This parameter MUST be present when sprop-cl-don value is present or the value of
packetization-mode is equal to 2.

This parameter signals the properties of a NAL unit stream. It specifies the maximum number of VCL NAL units that precede any VCL NAL unit in the NAL unit stream in transmission order and follow the VCL NAL unit in decoding order. Consequently, it is guaranteed that receivers can reconstruct NAL unit decoding order when the buffer size for NAL unit decoding order recovery is at least the value of sprop-interleaving-depth + 1 in terms of VCL NAL units. Herein the NAL unit stream refers to the one consisting of all NAL units conveyed in the current RTP session, and all NAL units conveyed in other RTP sessions, if present, the current RTP session depends on.

The value of sprop-interleaving-depth MUST be an integer in the range of 0 to 32767, inclusive.

sprop-deint-buf-req:

This parameter MUST NOT be present when the current RTP session does not depend on any other RTP session, and packetization-mode is not present or the value of packetization-mode is equal to 0 or 1. This parameter MUST be present when the sprop-cl-don value is present or the value of packetization-mode is equal to 2.

sprop-deint-buf-req signals the required size of the deinterleaving buffer for the NAL unit stream. The value of the parameter MUST be greater than or equal to the maximum buffer occupancy (in units of bytes) required in such a deinterleaving buffer that is specified in section 8 of this specification. It is guaranteed that receivers can perform the deinterleaving of interleaved NAL units into NAL unit decoding order, when the deinterleaving buffer size is at least the value of sprop-deint-buf-req in terms of bytes. Herein the NAL unit stream refers to the one consisting of all NAL units conveyed in the current RTP session, and all NAL units conveyed in other RTP sessions, if present, the current RTP session depends on.

The value of sprop-deint-buf-req MUST be an integer in the range of 0 to 4294967295,
Informative note: sprop-deint-buf-req indicates the required size of the deinterleaving buffer only. When network jitter can occur, an appropriately sized jitter buffer has to be provisioned for as well. When a scalable bitstream is conveyed in more than one RTP session, and the sessions initiates at different time, the session initiation variation has also to be compensated by an appropriately sized buffer.

**deint-buf-cap:**

This parameter signals the capabilities of a receiver implementation and indicates the amount of deinterleaving buffer space in units of bytes that the receiver has available for reconstructing the NAL unit decoding order, and that the receiver is able to handle any stream for which the value of the sprop-deint-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 deint-buf-cap. The value of deint-buf-cap MUST be an integer in the range of 0 to 4294967295, inclusive.

Informative note: deint-buf-cap indicates the maximum possible size of the deinterleaving buffer of the receiver only. When network jitter can occur, an appropriately sized jitter buffer has to be provisioned for as well.

**sprop-init-buf-time:**

This parameter MAY be used to signal the properties of a NAL unit stream. Herein the NAL unit stream refers to the one consisting of all NAL units conveyed in the current RTP session, and all NAL units conveyed in other RTP sessions, if present, the current RTP session depends on and sprop-cl-don value is present.

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

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

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

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

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

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

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

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

The parameter is coded as a non-negative base-10 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-init-buf-time MUST be an integer in the range of 0 to 4294967295, inclusive.

In addition to the signaled sprop-init-buf-time, receivers SHOULD take into account the transmission delay jitter buffering, including buffering for the delay jitter caused by mixers, translators, gateways, proxies, traffic-shapers, and other network elements. Yet another aspect receivers SHOULD take into account is the session initiation variation when a scalable bitstream is conveyed in more than one session, including buffering the variation.
This parameter MAY be used to signal the properties of a NAL unit stream. It MUST NOT be used to signal transmitter or receiver or codec capabilities. sprop-max-don-diff is an integer in the range of 0 to 32767, inclusive. If sprop-max-don-diff is not present, the value of the parameter is unspecified. Herein the NAL unit stream refers to the one consisting of all NAL units conveyed in the current RTP session, and all NAL units conveyed in other RTP sessions, if present, the current RTP session depends on.

sprop-max-don-diff is calculated as follows:

\[
\text{sprop-max-don-diff} = \max\{\text{AbsDON}(i) - \text{AbsDON}(j)\},
\]
for any \(i\) and any \(j > i\),

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

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

where DON(i) is the decoding order number of the NAL unit having index \(i\) in the transmission
order. The decoding order number is specified in section 6.6 of this specification.

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

max-rcmd-nalu-size:
This parameter MAY be used to signal the capabilities of a receiver. The parameter MUST NOT be used for any other purposes. The value of the parameter indicates the largest NALU size in bytes that the receiver can handle efficiently. The parameter value is a recommendation, not a strict upper boundary. The sender MAY create larger NALUs but must be aware that the handling of these may come at a higher cost than NALUs conforming to the limitation.

The value of max-rcmd-nalu-size MUST be an integer in the range of 0 to 4294967295, inclusive. If this parameter is not specified, no known limitation to the NALU size exists. Senders still have to consider the MTU size available between the sender and the receiver and SHOULD run MTU discovery for this purpose.

This parameter is motivated by, for example, an IP to H.223 video telephony gateway, where NALUs smaller than the H.223 transport data unit will be more efficient. A gateway may terminate IP; thus, MTU discovery will normally not work beyond the gateway.

Informative note: Setting this parameter to a lower than necessary value may have a negative impact.

sprop-cl-don:
When present in an RTP session description, the DONC field MUST be present in the PACSI NAL unit contained in any STAP-A packet in the current RTP session, and, if the interleaved packetization mode is in use, the DON values derived as specified in RFC 3984 MUST indicate CL-DON values.

sprop-prebuf-size:
This parameter MAY be present when the current RTP session depends on any other RTP session. This parameter MUST NOT be present when
sprop-cl-don is present. sprop-prebuf-size MAY signal the required size of the receiver buffer for the NAL unit stream per RTP session. This parameter may be useful to compensate the impact of inter-RTP session jitter, when the receiver buffer size is at least the value of sprop-prebuf-size in terms of bytes. Herein the NAL unit stream refers to the one consisting of all NAL units conveyed in the current RTP session.

The value of sprop-prebuf-size MUST be an integer in the range of 0 to 4294967295, inclusive.

Informative note: sprop-prebuf-size indicates the required size of the prebuffering receiver buffer only. When network jitter can occur, an appropriately sized jitter buffer has to be provisioned for as well. When a scalable bitstream is conveyed in more than one RTP session, and the sessions initiates at different time, the session initiation variation has also to be compensated by an appropriately sized buffer.

sprop-prebuf-time:
This parameter MAY be used to signal the properties of a NAL unit stream within a session multiplexing. Herein the NAL unit stream refers to the one consisting of all NAL units conveyed in the current RTP session. This parameter MUST NOT be present when sprop-cl-don is present.

The parameter signals the initial buffering time is used for a receiver before starting to recover the NAL unit decoding order for more than one RTP session. The parameter is the maximum value of (transmission time of a NAL unit – decoding time of the NAL unit), assuming reliable and instantaneous transmission, the same timeline for transmission and decoding, and that decoding starts when the first packet arrives.

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-prebuf-time MUST be an integer in the range of 0 to 4294967295, inclusive.

In addition to the signaled sprop-prebuf-time, receivers SHOULD take into account the transmission delay jitter buffering, including buffering for the delay jitter caused by mixers, translators, gateways, proxies, traffic-shapers, and other network elements. Yet another aspect receivers SHOULD take into account is the session initiation variation when a scalable bitstream is conveyed in more than one session, including buffering the variation.
sprop-scalability-info:
The parameter MAY be used to convey the NAL unit containing the scalability information SEI message as specified in Annex G of [SVC]. 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 representation of the NAL unit containing the scalability information SEI message. If present, the NAL unit MUST contain only a scalability information SEI message.

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

sprop-layer-range:
This parameter MAY be used to signal two sets of the layer identification values of the lowest and highest operation points conveyed in the RTP session. Each set is a base16 representation of a three-character value, with the first character representing DID, the second character representing QID, and the third character representing TID. The two sets are comma separated. Let DIDl and DIDh be the least DID value and the greatest DID value, respectively, among all the NAL units conveyed in the RTP session. Let QIDl and TIDl be the least QID value and the least TID value, respectively, among all the NAL units that are conveyed in the RTP session and that have DID equal to DIDl. Let QIDh and TIDh be the greatest QID value and the great TID value, respectively, among all the NAL units that are conveyed in the RTP session and that have DID equal to DIDh. The first set indicates the DID, QID and TID values of the lowest operation point, for which the DID, QID and TID values are equal to DIDl, QIDl, and TIDl, respectively. The second set indicates the DID, QID and TID values of the highest operation point, for which the DID, QID and TID values are equal to DIDh, QIDh, and TIDh, respectively.

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. 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 [SVC].

[Edt. Note (TS): That is, a SDP capable receiver/middle-box must decode the sprop-scalability-info syntax, which is not specified in this memo, to select a scalable-layer-id. This is currently not addressed in the offer answer section!]

Wenger, Wang, Schierl Expires August 24, 2008
sprop-spatial-resolution: [Edt. Note: I know that framerate and bitrate SDP parameters are already available, but failed to find a spatial resolution SDP parameter. It would be good if this is already defined. Otherwise, it would be better to be defined somewhere else because it is a generic parameter.]

This parameter MAY be used to indicate the property of a stream or the capability of a receiver or sender implementation. The value is a base16 of the width and height of the spatial resolution, in pixels, separated by a comma.
[Edt. Note (TS): Shouldn’t this be a generic SDP parameter?]

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

Security considerations:
See section 10 of RFC XXXX.

Public specification:
Please refer to RFC XXXX and its section 14.

Additional information:
None

File extensions: none
Macintosh file type code: none
Object identifier or OID: none
Person & email address to contact for further information:
Intended usage: COMMON
Author:
Change controller: IETF Audio/Video Transport working group
delegated from the IESG.

9.2. SDP Parameters

[Edt. Note: For agreeing on a Layer or OP in unicast, an SDP can contain multiple m lines with bitrate, framerate and spatial resolution parameters available, in addition to sprop-scalability-info. The receive can select one of the m lines, or, for operation points that are not included in the m lines, one of the "scalable layers" specified by sprop-scalabilité-info using scalable-layer-id.

For layered multicast, then the grouping signaling in I-D.ietf-mmusic-decoding-dependency is needed.
The above would conveniently support also the normal ROI use cases (with a few ROIs each indicated as a "scalable layer") but not the interactive ROI use cases. The quality layer using priority_id use cases are not supported either. That would need one more optional media type parameter, to identify a quality layer. The lightweight transcoding use cases should be supported well by using (multiple) normal AVC SDP offering messages.

9.2.1. 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.

9.2.2. Usage with the SDP Offer/Answer Model

When H.264 or 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 H.264 or SVC are "profile-level-id", "packetization-mode", and, if required by "packetization-mode", "sprop-deint-buf-req". These three parameters MUST be used symmetrically; 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.

Informative note: The requirement for symmetric use applies only for the above three parameters and not for the other stream properties and capability parameters.

To simplify handling and matching of these configurations, the same RTP payload type number used in the offer SHOULD also be used in the answer, as specified in [RFC3264]. An answer MUST NOT contain a payload type number used in the offer unless the configuration ("profile-level-id", "packetization-mode", and, if present, "sprop-deint-buf-req") is the same as in the offer.

Informative note: An offerer, when receiving the answer, has to compare payload types not declared in the offer based on media type (i.e., video/H264-SVC) and the above three parameters with any payload types it has already declared, in order to determine whether the configuration in question is new or equivalent to a configuration already offered.

An answerer MAY select from the layers offered in the "sprop-scalability-information" parameter by including "scalable-layer-id" or "sprop-layer-range" in the answer. [Edt. Note: do we need to additionally define behavior with snd/rcvonly parameter?]

- The parameters "sprop-parameter-sets", "sprop-deint-buf-req", "sprop-interleaving-depth", "sprop-max-don-diff", "sprop-init-buf-time", "sprop-prebuf-size", "sprop-prebuf-time", "sprop-scalability-information", "sprop-layer-range" describe the properties of the NAL unit stream that the offerer or answerer is sending for this 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 answerer is able to receive. When dealing with H.264 or 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", ","max rcmd- nalu-size") MAY be used to declare further capabilities. Their interpretation depends on the direction attribute. When the direction attribute is sendonly, then the parameters describe the limits of the RTP packets and the NAL unit stream that the sender is capable of producing. When the direction attribute is sendrecv or recvonly, then the parameters describe the limitations of what the receiver accepts.

As specified above, an offerer has to include the size of the deinterleaving buffer in the offer for an interleaved H.264 or SVC stream. To enable the offerer and answerer to inform each other about their capabilities for deinterleaving buffering, both parties are RECOMMENDED to include "deint-buf-cap". This information MAY be used when the value for "sprop-deint-buf-req" is selected in a second round of offer and answer. For interleaved streams, 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" parameter is used as described above. In addition, an answerer MUST maintain all parameter sets received in the offer in its answer. Depending on the value of the "parameter-add" parameter, different rules apply: If "parameter-add" is false (0), the answer MUST NOT add any additional parameter sets. If "parameter-add" is true (1), the answerer, in its answer, MAY add additional parameter sets to the "sprop-parameter-sets" parameter. The answerer MUST also, independent of the value of "parameter-add", accept to receive a video stream using the sprop-parameter-sets it declared in the answer.

Informative note: care must be taken when parameter sets are added not to cause overwriting of already transmitted parameter sets by using conflicting parameter set identifiers.

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

The stream properties parameters ("sprop-parameter-sets", "sprop-deint-buf-req", "sprop-interleaving-depth", "sprop-max-don-diff", "sprop-init-buf-time", "sprop-prebuf-size", "sprop-prebuf-time", "sprop-scalability-information", and "sprop-layer-range") MUST NOT be changed by the answerer. Thus, a payload type can either be accepted unaltered or removed.

The receiver capability parameters "max- mbps", "max- fs", "max- cpb", "max- dpb", "max- br", and "max rcmd- nalu-size" MUST be
supported by the answerer for all streams declared as sendrecv or recvonly; otherwise, one of the following actions MUST be performed: the media format is removed, or the session rejected.

- The receiver capability parameter redundant-pic-cap SHOULD be supported by the answerer for all streams declared as sendrecv or recvonly as follows: The answerer SHOULD NOT include redundant coded pictures in the transmitted stream if the offerer indicated redundant-pic-cap equal to 0. Otherwise (when redundant_pic_cap is equal to 1), it is beyond the scope of this memo to recommend how the answerer should use redundant coded pictures.

Below are the complete lists of how the different parameters shall be interpreted in the different combinations of offer or answer and direction attribute.

- In offers and answers for which "a=sendrecv" or no direction attribute is used, or in offers and answers for which "a=recvonly" is used, the following interpretation of the parameters MUST be used.

Declaring actual configuration or properties for receiving:

- profile-level-id
- packetization-mode

Declaring actual properties of the stream to be sent (applicable only when "a=sendrecv" or no direction attribute is used):

- sprop-deint-buf-req
- sprop-interleaving-depth
- sprop-parameter-sets
- sprop-max-don-diff
- sprop-init-buf-time
- sprop-prebuf-size
- sprop-prebuf-time
- sprop-scalability-information
- sprop-layer-range
- scalable-layer-id
- sprop-cl-don

Declaring receiver implementation capabilities:

- max-mbps
- max-fs
- max-cpb
- max-dpb
Declaring how Offer/Answer negotiation shall be performed:

- parameter-add

- In an offer or answer for which the direction attribute "a=sendonly" is included for the media stream, the following interpretation of the parameters MUST be used:

Declaring actual configuration and properties of stream proposed to be sent:

- profile-level-id
- packetization-mode
- sprop-deint-buf-req
- sprop-max-don-diff
- sprop-init-buf-time
- sprop-parameter-sets
- sprop-interleaving-depth
- sprop-prebuf-size
- sprop-prebuf-time
- sprop-scalability-information
- sprop-layer-range
- sprop-spatial-resolution
- sprop-cl-don

Declaring how Offer/Answer negotiation shall be performed:

- parameter-add

Furthermore, the following considerations are necessary:

- 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. "sprop-parameter-sets" MUST NOT be used in a sender’s declaration of its capabilities, as the limits of the values that are carried inside the parameter sets are implicit with the profile and level used.

- Parameters declaring a configuration point are not downgradable, with the exception of the level part of the "profile-level-id"
parameter. This expresses values a receiver expects to be used and must be used verbatim on the sender side.

- 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. 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 MIME 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 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 properties 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 has to offer different RTP sessions; i.e., different media lines declared as "recvonly" and "sendonly", respectively. This may have further implications on the system.

9.2.3. Usage with Session Multiplexing

If Session multiplexing is used, the rules on signaling media decoding dependency in SDP as defined in [I-D.ietf-mmusic-decoding-dependency] apply.

[Ed. Note (TS): We may want to connect mid-value with e.g. lowest TDQ value.]

9.2.4. Usage in Declarative Session Descriptions

When H.264 or SVC over RTP is offered with SDP in a declarative style, as in RTSP [RFC2326] or SAP [RFC2974], the following considerations are necessary.
All parameters capable of indicating the properties of both a NAL unit stream and a receiver are used to indicate the properties of a NAL unit stream. For example, in this case, the parameter "profile-level-id" declares the values used by the stream, instead of the capabilities of the sender. This results in that the following interpretation of the parameters MUST be used:

Declaring actual configuration or properties:

- profile-level-id
- sprop-parameter-sets
- packetization-mode
- sprop-interleaving-depth
- sprop-deint-buf-req
- sprop-max-don-diff
- sprop-init-buf-time
- sprop-prebuf-size
- sprop-prebuf-time
- sprop-layer-range
- sprop-spatial-resolution
- sprop-scalability-info
- sprop-cl-don

Not usable:

- max-mbps
- max-fs
- max-cpb
- max-dpb
- max-br
- redundant-pic-cap
- max-rcmd-nalu-size
- parameter-add
- deint-buf-cap
- scalable-layer-id

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.

9.3. Examples

9.3.1. Example for offering a single SVC session

Offerer -> Answerer SDP message:
9.3.2. Example for offering session multiplexing

Offerer -> Answerer SDP message:

```
Offerer
m = video 20000 RTP/AVP 96 97 98
a = rtpmap:96 H264/90000
a = rtpmap:97 H264-SVC/90000
a = rtpmap:98 H264-SVC/90000
a = mid:1

Answerer
m = video 20000 RTP/AVP 96 97 98
a = rtpmap:96 H264/90000
a = rtpmap:97 H264/90000
a = rtpmap:98 H264-SVC/90000
a = mid:2
a = depend:99 lay 1:97,98;

m = video 20002 RTP/AVP 99
a = rtpmap:99 H264-SVC/90000
a = rtpmap:99 H264-SVC/90000
a = mid:3
a = depend:100 lay 1:97,98 2:99;
```

9.4. Parameter Set Considerations

Please see section 8.4 of [RFC3984].

10. Security Considerations

Section 9 of [RFC3984] applies. 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, 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.

Informative note: Other security aspects, including confidentiality, authentication, and denial-of-service threat, for SVC are similar as H.264/AVC, as discussed in section 9 of [RFC3984].

11. Congestion Control

Within any given RTP session carrying payload according to this specification, the provisions of section 12 of [RFC3984] apply. Reducing the session bandwidth is possible by one or more of the following means, listed in an order that, in most cases, will assure the least negative impact to the user experience:

a) within the highest Layer identified by the DID field, utilize the TID and/or QID fields in the NAL unit header to drop NAL units with lower importance for the decoding process or human perception.
b) drop all NAL units belonging to the highest enhancement Layer as identified by the highest DID value.
c) dropping NAL units according to their importance for the decoding process, as indicated by the fields in the NAL unit header of the NAL units or in the prefix NAL units.
d) dropping NAL units or entire packets not according to the aforementioned rules (media-unaware stream thinning). This
results in the reception of a non-compliant bitstream and, most likely, in very annoying artifacts

Informative note: The discussion above is centered on NAL units and not on 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 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 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 [RFC3450] and [Yan].

12. IANA Consideration

[Edt. Note: A new media type should be registered from IANA.]

13. Informative Appendix: Application Examples

13.1. Introduction

Scalable video coding is a concept that has been around at least since 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. Dozens of scenarios have been studied. While some of the scenarios appear not to follow the most basic design principles of the Internet, e.g. as discussed in section 13.5, -- and are therefore not appropriate for IETF standardization -- others are clearly in the scope of IETF work. Of
these, this draft chooses the following subset for immediate consideration. The MPEG and VCEG requirement documents are available in [JVT-N026] and [JVT-N027], respectively.

With these remarks, we now introduce three main application scenarios that we consider as relevant, and that are implementable with this specification.

13.2. 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 we 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.

We 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].

13.3. Streaming of an SVC scalable stream

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 sending bitrate by choosing fewer layers, when composing the layered stream; see section 11. SVC is designed to gracefully support both bandwidth rampdown and bandwidth rampup 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.

13.4. Multicast to MANE, SVC scalable stream 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 3GPP. Streaming servers within this core network can be assumed to be professionally maintained. We assume that these servers can have many ports open to the network and that layered multicast is a real option. Therefore, we assume that the streaming server multicasts SVC scalable layers, instead of simulcasting different representations of the same content at different bit rates.

Also consider many endpoints of different classes. Some of these endpoints may not have the processing power or the display size to meaningfully decode all layers; others may have these capabilities. Users of some endpoints may not wish 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 need 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 bit stream according to the needs of the endpoint connected. These middleboxes are called MANEs throughout this specification. In practice, we 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. This is why we do not worry too much about decomposition aspects of the MANE as such.

MANEs necessarily need to be fairly complex devices. They certainly need to understand the signaling, so, for example, to associate the PT 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. This type of MANEs 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.

While such an MANE operation in its most trivial form (combining multiple RTP packet streams into a single one) can be implemented comparatively simply -- reordering the incoming packets according to the DON and sending them in the appropriate order -- more complex forms can also be envisioned. For example, a MANE can be optimizing the outgoing RTP stream to the MTU size of the outgoing path by utilizing the aggregation and fragmentation mechanisms of this memo.

A MANE can also perform stream thinning, so to adhere to congestion control principles as discussed in section 11. 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. In particular, SVC and/or this specification contain means to easily generate the correct inter-layer decoding
order of NAL units. No serious bit-oriented processing is required and no significant state information (beyond that of the signaling and perhaps the SVC sequence parameter sets) need to be kept.

13.5. Scenarios currently not considered for being unaligned with IP philosophy

Remarks have been made that the current draft does not take into consideration at least one application scenario which some JVT folks considered important. In particular, their idea was to make the RTP payload format (or the media stream itself) self-contained enough that a stateless, non-signaling-aware device can "thin" an RTP session to meet the bandwidth demands of the endpoint. They called this device a "Router" or "Gateway", and sometimes a MANE. Obviously, it’s not a Router or Gateway in the IETF sense. To distinguish it from a MANE as defined in RFC 3984 and in this specification, let’s call it an MDfH (Magic Device from Heaven).

To simplify discussions, let’s assume point-to-point traffic only. The endpoint has a signaling relationship with the streaming server, but it is known that the MDfH is somewhere in the media path (e.g. because the physical network topology ensures this). It has been requested, at least implicitly through MPEG’s and JVT’s requirements document, that the MDfH should be capable to intercept the SVC scalable bit stream, modify it by dropping packets or parts thereof, and forwarding the resulting packet stream to the receiving endpoint. It has been requested that this payload specification contains protocol elements facilitating such an operation, and the argument has been made that the NRI field of RFC 3984 serves exactly the same purpose.

The authors of this I-D do not consider the scenario above to be aligned with the most basic design philosophies the IETF follows, and therefore have not addressed the comments made (except through this section). In particular, we see the following problems with the MDfH approach:

- As the very minimum, the MDfH would need to know which RTP streams are carrying SVC. We don’t see how this could be accomplished but by using a static payload type. None of the IETF defined RTP profiles envision static payload types for SVC, and even the de-facto profiles developed by some application standard organizations (3GPP for example) do not use this outdated concept. Therefore, the MDfH necessarily needs to be at least "listening" to the signaling.
- If the RTP packet payload were encrypted, it would be impossible to interpret the payload header and/or the first bytes of the
media stream. We understand that there are crypto schemes under discussion that encrypt only the last n bytes of an RTP payload, but we are more than unsure that this is fully in line with the IETF’s security vision.

Even if the above two problems would have been overcome through standardization outside of the IETF, we still foresee serious design flaws:

- An MDfH can’t simply dump RTP packets it doesn’t want to forward. It either needs to act as a full RTP Translator (implying that it rewrites RTCP RRs and such), or it needs to patch the RTP sequence numbers to fulfill the RTP specification. Not doing either would, for the receiver, look like the gaps in the sequence numbers occurred due to unintentional erasures, which has interesting effects on congestion control (if implemented), will break pretty much every meta-payload ever developed, and so on. (Many more points could be made here).

In summary, based on our current knowledge we are not willing to specify protocol mechanisms that support an operation point that has so little in common with classic RTP use.

13.6. SSRC Multiplexing

The authors have played with the idea of introducing SSRC multiplexing, i.e. allowing sending multiple RTP packet streams containing layers in the same RTP session, differentiated by SSRC values. Our intention was to minimize the number of firewall pinholes in an endpoint to one, by using MANEs to aggregate multiple outgoing sessions stemming from a server into a single session (with SSRC multiplexed packet streams). We were hoping that would be feasible even with encrypted packets in an SRTP context.

While an implementation along these lines indeed appears to be feasible for the forward media path, the RTCP RR rewrite cannot be implemented in the way necessary for this scheme to work. This relates to the need to authenticate the RTCP RRs as per SRTP [RFC3711]. While the RTCP RR itself does not need to be rewritten by the scheme we envisioned, its transport addresses needs to be manipulated. This, in turn, is incompatible with the mandatory authentication of RTCP RRs. As a result, there would be a requirement that a MANE needs to be in the RTCP security context of the sessions, which was not envisioned in our use case.
As the envisioned use case cannot be implemented, we refrained to add the considerable document complexity to support SSRC multiplexing herein.

14. References

14.1. Normative References


14.2. Informative References

Stephan Wenger  
Nokia  
955 Page Mill Road  
Palo Alto, CA 94304  
USA  
Phone: +1-650-862-7368  
Email: stewe@stewe.org

Ye-Kui Wang  
Nokia Research Center  
P.O. Box 100  
FIN-33721 Tampere  
Finland  
Phone: +358-50-486-7004  
Email: ye-kui.wang@nokia.com

Thomas Schierl  
Fraunhofer HHI  
Einsteinufer 37  
D-10587 Berlin  
Germany
Copyright Statement

Full Copyright Statement

Copyright (C) The IETF Trust (2008).

This document is subject to the rights, licenses and restrictions contained in BCP 78, and except as set forth therein, the authors retain all their rights.

Disclaimer of Validity

This document and the information contained herein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY, THE IETF TRUST AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Intellectual Property Statement

Intellectual Property

The IETF takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights. Information on the procedures with respect to rights in RFC documents can be found in BCP 78 and BCP 79.

Copies of IPR disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at http://www.ietf.org/ipr.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at ietf-ipr@ietf.org.
19. Acknowledgement

Funding for the RFC Editor function is provided by the IETF Administrative Support Activity (IASA).

20. RFC Editor Considerations

none

21. Open Issues

1) Cross layer decoding order dependency - two suggested solutions on the table. Need to agree if use one or both. In the case of both how to resolve interoperability. Initial step is to update text explaining the usage.

2) Backward compatibility to H.264, enabling H.264 (RFC 3984 single NAL unit mode) to interoperate with SVC using base layer. Need more definition.

3) Clarify the PACSI packet since there were changes between the draft revision

4) Review the SDP parameters.

5) Changed semantics between RFC 3984 and svc like sprop-deint-buf-req - probably will need new parameters.

6) What to do with bugs in RFC 3984.

7) Clarify the usage of the new parameters like sprop-scalability-info, relation to SEI and usage in offer/answer.

8) The text should be clear enough to allow an implementer to use it for creating the payload without having to read the H.264 SVC document.

9) Non-VCL NAL units, e.g. SEI messages and parameter sets, may be needed by an enhancement layer but not the base layer. However, according to SVC, within an access unit, these non-VCL NAL units must precede VCL NAL units in decoding order. In session multiplexing, should non-VCL NAL units be transported in the same session as the layer that requires the non-VCL NAL unit, or should they be always transported in the base session? It may be impossible to find out without parsing details which session respectively SPS/subset SPS a picture parameter set belongs to. It may make sense for simplicity to allow a MANE to include all of the non-VCL NAL units within all the sessions.

10) sprop-spatial-resolution: in this draft or a more generic draft?

11) Shall we allow NAL unit types 14 and 15 to be present in the RTP stream carrying the T0 base Layer, when RFC3984 encapsulated?

12) Further to the comments 9 and 11 above, if different sessions carry different temporal enhancement layers, then who should get, e.g., the subset SPS. It is actually possible that none does, if transmitted out-of-band. We should enumerate the possibilities and leave no doubt about how it is supposed to work. This can be done in the definition of the RTP sessions in 5.1.2, but it’s
even better if specific text is added (after
discussion/approval).

13) Do we need to describe the filler NAL unit insertion?
14) Current draft allows for interleaving capabilities within non-
interleaved packetization mode (packetization-mode==1) via PACSIs
with CL-DON. There is no need for mixing the capabilities of
decoding order recovery for session multiplexing and
interleaving. Shouldn’t these two features (decoding order
recovery and interleaving) be clearly separated, e.g. via the
explicit indication of packetization-mode==2 as specified by
RFC3984?

22. Changes Log

Version 00
- 29.08.2005, YkW: Initial version
- 29.09.2005, Miska: Reviewed and commented throughout the document
- 05.10.2006, StW: Editorial changes through the document, and
formatted the document in RFC payload format style

From -00 to -01
- 04.02.2006, StW: Added details to scope
- 04.02.2006, StW: Added short subsection 6.1 "Design Principles"
- 04.02.2006, StW: Added section 15, "Application Examples"
- 06.02 - 03.03.2006, YkW: Various modifications throughout the
document
- 13.02.2006 - 03.03.2006 , ThS: Added definitions and additional
information to section 3.3, 5.1, 7 and 8, parameters in section 9.1 and
added section 14 for NAL unit re-ordering for layered multicast.
Further modifications throughout the document

From -01 to -02
- 06.03.2006, StW: Editorial improvements
- 26.05.2006, YkW: Updated NAL unit header syntax and semantics
according to the latest draft SVC spec
- 20.06.2006, Miska/YkW: Added section 6.10 "Payload Content
Scalability Information (PACSI) NAL Unit"
- 20.06.2006, YkW: Updated the NAL unit reordering process for layered
multicast (removed the old section 14 "Informative Appendix: NAL Unit
Re-ordering for Layered Multicast" and added the new section 13 "NAL
Unit Reordering for Layered Multicast")

From -02 to -03
- 05.09.2006, YkW: Updated the NAL unit header syntax, definitions, etc., according to the foreseen July JVT output. Updated possible MANE adaptation operations according to SPID, TL, DID and QL. Clarified the removal of single NAL unit packetization mode. Added the support of SSRC multiplexing in layered multicast.
- 08.09.2006, StW: Editorial changes throughout the document.
- 08.09.2006, YkW: Added the packetization rule for suffix NAL unit.
- 19.09.2006, YkW: Moved/updated SSRC multiplexing support to section 6.2 ‘‘RTP header usage’’. Moved/updated the cross layer DON constraint to Section 6.6 ‘‘Decoding order number’’. Moved/updated the packetization rule when a SVC bistream is transported over more than one RTP session to Section 7 ‘‘Packetization rules’’. Removed Section 13 "Support of layered multicast".
- 17.10, StW: Fixed many editorials, clarified MANE, mixer, translator and RTP packet stream throughout doc (hopefully consistently) 18.10., removed comments, clarified B-Bit, changed definition of base-layer (do not need to be of the lowest temporal resolution).

From -03 to draft-ietf-avt-rtp-svc-00

- 23.11.06, StW: Editorials throughout the memo
- 23.11.06, StW: removed all occurrences of the security discussions, as they are incorrect. When using SRTP, the RTCP is authenticated, implying that a translator cannot rewrite RTCP RRs, implying that RRs would be incorrect as soon as the session is modified (i.e. packets are being removed), implying that SSRC-mux does not work in multicast.
- 23.11.06, StW: rewrote congestion control
- 23.11.06, StW: removed application scenario related to SRTP, as this does not work (see above
- 23.11.06, StW: added informative reference to H.241
- 27/29.11.06, YkW: editorial changes throughout the document
- 27/29.11.06, YkW: alignment with the SVC specification
- 19.12.06, TS:
  TS: [SVC] is now the complete Joint Draft of H.264
  TS: Removed SSRC Multiplexing
  TS: Changed use cases for MANE as a translator
  TS: Editorials throughout the document, alignment with SVC spec.
- 20-28.12.06, StW/TS/YkW: editorial changes throughout the document

From draft-ietf-avt-rtp-svc-00 to draft-ietf-avt-rtp-svc-01
- 23.02.07, YkW/Miska Hannuksela: Added enhancements to PACSI NAL unit
- 01.03.07, Jonathan Lennox/YkW: Added recommendatory packetization rules for SEI messages and non-VCL NAL units
- 05.03.07, Thomas Wiegand/YkW: Added the fields of picture start, picture end, and T10PicIdx to PACSI NAL unit
- 05.03.07, TS: Draft conforms to new I-D style

From draft-ietf-avt-rtp-svc-01 to draft-ietf-avt-rtp-svc-02
25-June-2007: TS
Clarified definitions Layer, Operation Points, Removed FGS
Use of DON in de-packetization
Congestion control
25-June-2007: YkW
Edit throughout the spec, aligned with JVT-X201 SVC spec
09-July-2007: TS
Further modifications and alignments with JVT-X201.
05-Dec-2007: TS
Formatting corrected, ref to signaling draft corrected

From draft-ietf-avt-rtp-svc-02 to draft-ietf-avt-rtp-svc-03
- 21-Aug-2007 to 24-Sep-2007: YkW
  1) Resolved most of the comments sent to the AVT reflector and to the editors
  2) Updated the intro text for parameter sets
  3) Reordered the definitions according to alphabetical order and added some definitions
  4) Added the NAL unit order recovery process for layered multicast using CL-DON in the PACSI NAL unit, thus to allow for layered multicast without requiring the non-interleaved packetization mode. The detailed NAL unit order recovery process added to section 8.
  5) Added some packetization rules. Some of these were to resolve the "single NAL unit mode deprecation" issue.
  6) Added semantics of the media type parameters inherited from RFC 3984, and added a couple of new parameters for negotiation of operation point.

Wenger, Wang, Schierl  Expires August 24, 2008
7) Other edits throughout the document.
   - 16 to 18 November 2007: TS
     1) Added the NAL unit order recovery process for layered multicast
        without using CL-DON, thus to allow for layered multicast without
        requiring the non-interleaved packetization mode.
     2) Added the usages of the media type parameters, including SDP
        usage with offer/answer model, declarative usage, and examples.
   - 08 to 19 November 2007: YkW
     1) Aligned the spec with the final version of the SVC spec.
     2) Updated the congestion control part according to Colin Perkins’
        comment.
     3) Checked the parameter set considerations and confirmed that the
        text in RFC 3984 is OK.
     4) Updated the security considerations part.
     5) Added justifications for some fields in the PACSI NAL units.

From draft-ietf-avt-rtp-svc-03 to draft-ietf-avt-rtp-svc-04
   - 18 December 2007: TS
     1) Updated formatting in the Media Type Registration section
     2) Updated the semantics of sprop-layer-range
     3) Updated Open issues according to Roni’s email
     4) Corrected usage of "depend" in SDP example

From draft-ietf-avt-rtp-svc-04 to draft-ietf-avt-rtp-svc-05
   08 January 2008: TS
     1) Back to original word source document for draft-ietf-avt-rtp-
        svc-03.
     2) Changed/corrected formatting of document:
        a. Word source document margins
        b. Footer/Header adjustment
        c. Formatting of RFC default sections
        d. Formatting of Author’s section
        e. Formatting of Reference sections
        f. Corrected formatting of Media Type section

After TXT document generation, document should be readable by
Internet Draft submission tool.
   3) Fixed SDP example (fmtp:97 to 96) [Mike’s comment]
   4) Changed "sprop-layer-range" definition according to Ye-Kui’s
      internal proposal.
   5) Corrected usage of "depend" in SDP example

From draft-ietf-avt-rtp-svc-05 to draft-ietf-avt-rtp-svc-06
   16-17 January 2008: YkW
   1) Updated Sections 13.1 and 13.2 with some corrections, and added
      two informative references on SVC requirements.
   2) Added text (in Section 4) to explain
      a. When layered multicast (i.e. session multiplexing) should be
         used.
      b. Whether one or more layers should be carried in one RTP
         stream when layered multicast is used or not.
3) Added new Subsection 7.1 to contain packetization rules for
layered multicast, and within the new subsection,
a. Tried to improve the readability
b. Added text to explain which packetization modes can be used
   in each decoder order recovery mode for layered multicast
4) Aligned the definitions with the SVC specification, including
   the term base layer, and used the term T0 base layer for the
   minimum subset of the base layer.
5) Systematically checked throughout the document the places that
   use the terms base layer or T0 base layer.
6) Corrected the semantics of the reserved fields in the SVC NAL
   unit header extension.
7) Renamed the CL-DON field with DONC, and systematically updated
   throughout the document texts containing instances of "CL-DON".
8) Clarified numerous instances of "the remaining NAL units in the
   payload" in the PACSI NAL unit description by adding "of the
   aggregation packet".
9) Moved the definition of "target NAL units" to be beginning of
   the semantics of PACSI NAL unit fields.
10) Removed the obsolete semantics of the T bit in the PACSI NAL
    unit.
11) Updated the semantics of the P bit in the PACSI NAL unit, to
    indicate that all the remaining NAL units in the payload of the
    aggregation packet are redundant slices.
12) Removed some obsolete text
    a. Two paragraphs in Subsection 3.1
    b. The editor note on max-mbps, max-fs, max-cpb, max-dpb, and
       max-br semantics.
21 January 2008: TS
1) New parameter sprop-cl-don indicating use of CL DON for decoding
   order recovery
2) Non-CL DON mode is now referred to as "classical RTP decoding
   order recovery".
3) Extended SVC definition section by target dependency and target
   layer representation
4) Extended Skope section.
5) Clarified rules for CL-DON mode and classical RTP mode.
6) Extended usage of classical RTP mode for interleaved mode
7) General depacketization description in section 8.1 for classical
   RTP and CL-DON mode.
8) Removed rule for presence of SEI messages for enhancement
    layers.

21 January 2008: YkW
1) Updated sub-section 8.1.2.

From draft-ietf-avt-rtp-svc-06 to draft-ietf-avt-rtp-svc-07
28-29 January 2008: TS
1) Removed/changed constraints on session multiplexing in section 4. (scope)
2) Removed constraint on non-VCL NAL units to be in the same session as related VCL NAL units in section 7.
3) Removed some of the old constraints in section 7 on process I. (re-ordering without CL-DON).
4) Removed decoding order constraints of section 8.1 and rewriting of 8.1.
5) Rewriting of section 8.1.1.
6) New local definitions: Operation Point representation, Base RTP session, Enhancement RTP session
7) Changed meaning of sprop-prebuf-size and sprop-prebuf-time

31 January 2008: YkW
1) Clarified "subset of the base layer" per Roni’s comment, in sections 3.1, 4, and 6.1.
2) Updated Section 4 (Scope) per Roni’s and Mike’s comments. Session multiplexing is now considered a distinct thing as layered multicast, which uses session multiplexing.
3) Commented text in section 8.1 and suggested an alternative text.
4) Updated sub-section 8.1.2 (CL-DON decoding order recovery mode).
5) Corrected the semantics of sprop-cl-don.
6) Updated open issues.

1 February 2008: TS
1) Updated text in section 8.1
2) Updated open issues

From draft-ietf-avt-rtp-svc-07 to draft-ietf-avt-rtp-svc-08
11 February 2008: AE (Alex Eleftheriadis)
1) Updated text in sections 1, 3.1, 3.2, based on comments posted by the author.

13 February 2008: YkW
1) Made some corrections and improvements to the newly updated text in sections 1, 3.1, 3.2, removed the last paragraph in section 1 that was used to point out the cross-layer decoding order recovery issue, and added clarifications regarding picture parameter set usage in SVC.

18 February 2008: AE
1) Revised all definitions (5.1.2) based on extensive discussions with editors to ensure that the definitions are logically consistent, are inline with the SVC specification, and avoid confusion (to the extent possible given existing SVC terminology).
2) Major changes include: new 'layer' definition that is not temporal_id specific; new definitions of 'AVC base layer' and 'SVC base layer' to distinguish with/without prefix NALs; T0 used for all temporal subcomponents; new definitions of 'enhancement layer'; new 'base RTP session' and 'enhancement RTP session'; introduction of 'SVC RTP session' to provide a reference to either base or enhancement; change of 'SVC NAL unit' to include type 15.

3) Reviewed entire document to ensure consistent use of new terminology.

4) Changed capitalization to be consistent throughout the document, and removed smart quotes where present.

5) Changed definitions of 'S' and 'E' bits to indicate transmission order, rather than decoding order, per Mike's posting and this author's response.

6) Identified as open issue the placement of non-VLC NAL units in the various RTP sessions. The draft is currently more or less silent about who should get what (e.g., if temporal_id 0 and 1 sessions exist, who should get the SPS, both?). Current definitions of layers etc. are on purpose silent about this as various options exist - the definitions should not be locked to any particular choice unless we want them to be.

18 February 2008: YkW
1) Made a couple corrections in sections 3.1 and 3.2, added back the SVC usage of sequence parameter sets in section 3.2.
2) Updated the definitions of base RTP session, enhancement RTP session, and cross-layer decoding order number.
3) Added a few editing comments regarding the definitions of operation point and SVC RTP session and the use of "SVC context".
4) Other various editorial changes throughout sections 1-6.

19 February 2008: AE
1) Corrected minor typos and some remaining instances of 'Layer' (wrong capitalization).
2) Reworded 6.4, specifically the discussion about NRI use.
3) Reworded 9.1, informative note about ignoring unspecified parameters.
4) Ensured consistent change of the word 'greater' to 'higher' in the definitions section (5.1.2), when referring to values of D, T, or Q.

20 February 2008: TS
1) RTP base session definition modified
2) Timestamps are derived from same clock instance for session multiplexing
3) Added FU-A to text in section 7.
4) Corrected Offer examples
5) Added various [Edt. Note] - comments

25 February 2008: TS
1) Integrated/addressed Mike’s comments sent by email on
   02/08/2008: 6), 7), 9), 10), 16), 17), 18), 19), 20), 21), 22),
   23), 24), 25), 26), 27). Other comments have been addressed by
   the authors earlier.
   Open comments: 8), 11)/12), 13), 15),
2) Updated open issues.

25 February 2008: YkW
1) Various minor editorial changes.