

BS ISO/IEC 14496-15:2014

Incorporating corrigendum December 2014



BSI Standards Publication

Information technology — Coding of audio-visual objects

Part 15: Carriage of network abstraction
layer (NAL) unit structured video in ISO base
media file format

bsi.

...making excellence a habit.
AVA Documents™
Document Sharing Hub

National foreword

This British Standard is the UK implementation of ISO/IEC 14496-15:2014, incorporating corrigendum December 2014. It supersedes BS ISO/IEC 14496-15:2004 which is withdrawn.

The start and finish of text introduced or altered by corrigendum is indicated in the text by tags. Text altered by ISO/IEC corrigendum December 2014 is indicated in the text by AC₁ AC₁.

The UK participation in its preparation was entrusted to Technical Committee IST/37, Coding of picture, audio, multimedia and hypermedia information.

A list of organizations represented on this committee can be obtained on request to its secretary.

This publication does not purport to include all the necessary provisions of a contract. Users are responsible for its correct application.

© The British Standards Institution 2015.
Published by BSI Standards Limited 2015

ISBN 978 0 580 90312 0
ICS 35.040

Compliance with a British Standard cannot confer immunity from legal obligations.

This British Standard was published under the authority of the Standards Policy and Strategy Committee on 31 July 2014.

Amendments/corrigenda issued since publication

Date	Text affected
31 May 2015	Implementation of ISO/IEC corrigendum December 2014

**Information technology — Coding of audio-
visual objects —**

**Part 15:
Carriage of network abstraction layer
(NAL) unit structured video in ISO base
media file format**

Technologies de l'information — Codage des objets audiovisuels —

*Partie 15: Transport de vidéo structuré en unités NAL au format ISO de base
pour les fichiers médias*



Reference number
ISO/IEC 14496-15:2014(E)

AVA Documents
Document Sharing Hub
© ISO/IEC 2014



COPYRIGHT PROTECTED DOCUMENT

© ISO/IEC 2014

All rights reserved. Unless otherwise specified, no part of this publication may be reproduced or utilized otherwise in any form or by any means, electronic or mechanical, including photocopying, or posting on the internet or an intranet, without prior written permission. Permission can be requested from either ISO at the address below or ISO's member body in the country of the requester.

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents	Page
Foreword.....	v
Introduction.....	vii
1 Scope	1
2 Normative references	1
3 Terms, definitions and abbreviated terms.....	1
3.1 Terms and definitions.....	1
3.2 Abbreviated terms	5
4 General Definitions.....	6
4.1 Introduction.....	6
4.2 Elementary stream structure	6
4.3 Sample and Configuration definition	6
4.4 Video Track Structure.....	8
4.5 Template fields used	8
4.6 Visual width and height.....	9
4.7 Decoding time (DTS) and composition time (CTS).....	9
4.8 Sync sample (IDR)	9
4.9 Shadow sync	10
4.10 Sample groups on random access recovery points and random access points	10
4.11 Hinting.....	10
5 AVC elementary streams and sample definitions	11
5.1 Introduction.....	11
5.2 Elementary stream structure.....	11
5.3 Sample and Configuration definition	14
5.4 Derivation from ISO Base Media File Format	18
6 SVC elementary stream and sample definitions.....	29
6.1 Introduction.....	29
6.2 Elementary stream structure.....	30
6.3 Use of the plain AVC file format	31
6.4 Sample and configuration definition.....	31
6.5 Derivation from the ISO base media file format.....	33
7 MVC elementary stream and sample definitions	39
7.1 Introduction.....	39

7.2	Overview of MVC Storage	40
7.3	MVC Track Structure	41
7.4	Use of the plain AVC File Format.....	42
7.5	Sample and configuration definition	42
7.6	Derivation from the ISO base media file format	45
7.7	MVC specific information boxes.....	54
8	HEVC elementary streams and sample definitions.....	63
8.1	Introduction	63
8.2	Elementary Stream Structure	64
8.3	Sample and configuration definition	64
8.4	Derivation from ISO base media file format.....	69
Annex A (normative)	In-stream structures specific to SVC and MVC	76
Annex B (normative)	SVC and MVC sample group and sub-track definitions	81
Annex C (normative)	Temporal metadata support.....	102
Annex D (normative)	File format toolsets.....	110
Annex E (normative)	Sub-parameters for the MIME type ‘Codecs’ parameter	112
Annex F (Informative)	Patent Statements.....	114

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of the joint technical committee is to prepare International Standards. Draft International Standards adopted by the joint technical committee are circulated to national bodies for voting. Publication as an International Standard requires approval by at least 75 % of the national bodies casting a vote.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

ISO/IEC 14496-15 was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 29, *Coding of audio, picture, multimedia and hypermedia information*.

This third edition cancels and replaces the second edition (ISO/IEC 14496-15:2010), which has been technically revised. It also incorporates the Amendment ISO/IEC 14496-15:2010/Amd.1:2011 and the Technical Corrigenda ISO/IEC 14496-15:2010/Cor.1:2011 and ISO/IEC 14496-15:2010/Cor.2:2012.

ISO/IEC 14496 consists of the following parts, under the general title *Information technology — Coding of audio-visual objects*:

- *Part 1: Systems*
- *Part 2: Visual*
- *Part 3: Audio*
- *Part 4: Conformance testing*
- *Part 5: Reference software*
- *Part 6: Delivery Multimedia Integration Framework (DMIF)*
- *Part 7: Optimized reference software for coding of audio-visual objects* [Technical Report]
- *Part 8: Carriage of ISO/IEC 14496 contents over IP networks*
- *Part 9: Reference hardware description* [Technical Report]
- *Part 10: Advanced Video Coding*

- *Part 11: Scene description and application engine*
- *Part 12: ISO base media file format*
- *Part 13: Intellectual Property Management and Protection (IPMP) extensions*
- *Part 14: MP4 file format*
- *Part 15: Carriage of network abstraction layer (NAL) unit structured video in the ISO base media file format*
- *Part 16: Animation Framework eXtension (AFX)*
- *Part 17: Streaming text format*
- *Part 18: Font compression and streaming*
- *Part 19: Synthesized texture stream*
- *Part 20: Lightweight Application Scene Representation (LSeR) and Simple Aggregation Format (SAF)*
- *Part 21: MPEG-J Graphics Framework eXtension (GFX)*
- *Part 22: Open Font Format*
- *Part 23: Symbolic Music Representation*
- *Part 24: Audio and systems interaction*
- *Part 25: 3D Graphics Compression Model*
- *Part 26: Audio conformance*
- *Part 27: 3D Graphics conformance*
- *Part 28: Composite font representation*

Introduction

This part of ISO/IEC 14496 defines a storage format based on, and compatible with, the ISO Base Media File Format (ISO/IEC 14496-12 and ISO/IEC 15444-12), which is used by the MP4 file format (ISO/IEC 14496-14) and the Motion JPEG 2000 file format (ISO/IEC 15444-3) among others. This part of ISO/IEC 14496 enables video streams formatted as Network Adaptation Layer Units (NAL Units) to

- be used in conjunction with other media streams, such as audio,
- be used in an MPEG-4 systems environment, if desired,
- be formatted for delivery by a streaming server, using hint tracks, and
- inherit all the use cases and features of the ISO Base Media File Format on which MP4 and MJ2 are based.

This part of ISO/IEC 14496 may be used as a standalone specification; it specifies how NAL unit structured video content shall be stored in an ISO Base Media File Format compliant format. However, it is normally used in the context of a specification, such as the MP4 file format, derived from the ISO Base Media File Format, that permits the use of NAL unit structured video such as AVC (ISO/IEC 14496-10) and video and High Efficiency Video Coding (HEVC, ISO/IEC 23008-2) video.

The ISO Base Media File Format is becoming increasingly common as a general-purpose media container format for the exchange of digital media, and its use in this context should accelerate both adoption and interoperability.

The International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC) draw attention to the fact that it is claimed that compliance with this document may involve the use of a patent.

The ISO and IEC take no position concerning the evidence, validity and scope of this patent right.

The holder of this patent right has assured the ISO and IEC that he is willing to negotiate licences under reasonable and non-discriminatory terms and conditions with applicants throughout the world. In this respect, the statement of the holder of this patent right is registered with the ISO and IEC. Information may be obtained from the companies listed in Annex F.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights other than those identified in Annex F. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

Information technology — Coding of audio-visual objects —

Part 15:

Carriage of network abstraction layer (NAL) unit structured video in the ISO base media file format

1 Scope

This part of ISO/IEC 14496 specifies the storage format for streams of video that is structured as NAL Units, such as AVC (ISO/IEC 14496-10) and HEVC (ISO/IEC 23008-2) video streams.

2 Normative references

The following documents, in whole or in part, are normatively referenced in this document and are indispensable for its application. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 14496-10, *Information technology — Coding of audio-visual objects — Part 10: Advanced Video Coding*

ISO/IEC 14496-12, *Information technology — Coding of audio-visual objects — Part 12: ISO base media file format*¹⁾

ISO/IEC 23008-2, *Information technology — High efficiency coding and media delivery in heterogeneous environments — Part 2: High efficiency video coding*

3 Terms, definitions and abbreviated terms

3.1 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 14496-10 or ISO/IEC 23008-2, and the following apply.

3.1.1

aggregator

in-stream structure using a NAL unit header

NOTE Aggregators are used to group NAL units belonging to the same sample.

¹⁾ ISO/IEC 14496-12 is technically identical to ISO/IEC 15444-12.

3.1.2

AVC base layer

maximum subset of a bitstream that is AVC compatible (i.e. a bitstream not using any of the functionality of ISO/IEC 14496-10 Annex G or Annex H)

NOTE 1 The AVC base layer is represented by AVC VCL NAL units and associated non-VCL NAL units.

NOTE 2 The AVC base layer itself can be a temporal scalable bitstream.

3.1.3

AVC NAL unit

AVC VCL NAL unit and its associated non-VCL NAL units in a bitstream

3.1.4

AVC VCL NAL unit

NAL unit with type 1 to 5 (inclusive) as specified in ISO/IEC 14496-10

3.1.5

extraction path

set of operations on the original bitstream, each yielding a subset bitstream, ordered such that the complete bitstream is first in the set, and the base layer is last, and all the bitstreams are in decreasing complexity (along one of the scalability axes, such as resolution), and where every bitstream is a valid operating point

NOTE An extraction path may be represented by the values of `priority_id` in the NAL unit headers. Alternatively an extraction path can be represented by the run of tiers or by a set of hierarchically dependent tracks.

3.1.6

extractor

in-stream structure using a NAL unit header including a NAL unit header extension

NOTE Extractors contain instructions on how to extract data from other tracks. Logically an Extractor can be seen as a 'link'. While accessing a track containing Extractors, the Extractor is replaced by the data it is referencing.

3.1.7

in-stream structure

structure residing within sample data

3.1.8

MVC VCL NAL unit

NAL unit with type 20, and NAL units with type 14, as specified in ISO/IEC 14496-10, when the immediately following NAL units are AVC VCL NAL units.

NOTE MVC VCL NAL units do not affect the decoding process of a legacy AVC decoder.

3.1.9

operating point

subset of a scalable bitstream, representing in SVC a particular spatial resolution, temporal resolution, and quality, or in MVC a set of target output views

NOTE 1 Each operating point consists of all the data needed to decode this particular bitstream subset.

NOTE 2 In an SVC stream an operating point can be represented either by (i) specific values of DTQ (`dependency_id`, `temporal_id` and `quality_id`) or (ii) specific values of P (`priority_id`) or (iii) combinations of them (e.g. PDTQ). Note that the usage of `priority_id` is defined by the application. In an SVC file a track represents one or more operating points. Within a track tiers may be used to define multiple operating points.

NOTE 3 The bitstream subset of an MVC operating point represents a particular set of target output views at a particular temporal resolution, and consists of all the data needed to decode this particular bitstream subset.

NOTE 4 An operating point is referred to as an operation point in Annex H of ISO/IEC 14496-10 or in ISO/IEC 23008-2.

3.1.10

parameter set

video parameter set, sequence parameter set, or picture parameter set, as defined in the applicable video standard (e.g. ISO/IEC 14496-10 or ISO/IEC 23008-2)

NOTE This term is used to refer to all types of parameter sets.

3.1.11

parameter set elementary stream

elementary stream containing samples made up of only sequence and picture parameter set NAL units synchronized with the video elementary stream

3.1.12

prefix NAL unit

NAL units with type 14 as specified in ISO/IEC 14496-10

NOTE Prefix NAL units provide scalability information about AVC VCL NAL units and filler data NAL units. Prefix NAL units do not affect the decoding process of a legacy AVC decoder. The behaviour of a legacy AVC file reader as a response to prefix NAL units is undefined.

3.1.13

scalable layer; layer

set of VCL NAL units with the same values of *dependency_id*, *quality_id*, and *temporal_id*, and the associated non-VCL NAL units as specified in ISO/IEC 14496-10.

NOTE 1 A scalable layer with any of *dependency_id*, *quality_id*, and *temporal_id* not equal to 0 enhances the video by one or more scalability levels in at least one direction (temporal, quality or spatial resolution)

NOTE 2 SVC uses a “layered” encoder design which results in a bitstream representing “coding layers”. In some publications the ‘base layer’ is the first quality layer of a specific coding layer. In some publications the base layer is the scalable layer with the lowest priority. The SVC file format uses “scalable layer” or “layer” in a general way for describing nested bitstreams (using terms like AVC base layer or SVC enhancement layer).

3.1.14

scalable layer representation

bitstream subset that is required for decoding the scalable layer, consisting of the scalable layer itself and all the scalable layers on which the scalable layer depends

NOTE A scalable layer representation is also referred to as the representation of the scalable layer.

3.1.15

sub-picture

proper subset of coded slices of a layer representation

3.1.16

sub-picture tier

tier that consists of sub-pictures

NOTE Any coded slice that is not included in the tier representation of a sub-picture tier is not to be referred to in inter prediction or inter-layer prediction for decoding of the sub-picture tier.

3.1.17

SVC enhancement layer

layer that specifies a part of a scalable bitstream that enhances the video

NOTE 1 An SVC enhancement layer is represented by SVC VCL NAL units and the associated non-VCL NAL units and SEI messages.

NOTE 2 Usually an SVC enhancement layer represents a spatial or coarse-grain scalability (CGS) coding layer (identified by a specific value of dependency_id).

3.1.18

SVC NAL unit

SVC VCL NAL unit and its associated non-VCL NAL units in an SVC stream

3.1.19

SVC stream

bitstream represented by the operating point for which dependency_id is equal to mDid, temporal_id is the greatest temporal_id value among mOpSet, and quality_id is the greatest quality_id value among mOpSet, where the greatest value of dependency_id of all the operating points represented by DTQ (dependency_id, temporal_id and quality_id) combinations is equal to mDid, and the set of all the operating points with dependency_id equal to mDid is mOpSet.

NOTE The term "SVC stream" is referenced by 'decoding/accessing the entire stream' in this document. There may be NAL units which are not required for decoding this operating point.

3.1.20

SVC VCL NAL unit

NAL unit with type 20, and NAL units with type 14 when the immediately following NAL units are AVC VCL NAL units

NOTE SVC VCL NAL units do not affect the decoding process of a legacy AVC decoder.

3.1.21

temporal layer representation

representation of a temporal layer

temporal layer and all lower temporal layers

3.1.22

tier

set of operating points within a track, providing information about the operating points and instructions on how to access the corresponding bitstream portions (using maps and groups)

NOTE 1 A tier represents one or more scalable layers of an SVC bitstream.

NOTE 2 The term "tier" is used to avoid confusion with the frequently used term layer. A tier represents a subset of a track and represents an operating point of an SVC bitstream. Tiers in a track subset the entire track, no matter whether the track references another track by extractors.

NOTE 3 An MVC tier represents a particular set of temporal subsets of a particular set of views.

3.1.23

tier representation; representation of the tier

bitstream subset that is required for decoding the tier, consisting of the tier itself and all the tiers on which the tier depends

3.1.24

video elementary stream

elementary stream containing access units made up of NAL units for coded picture data

3.1.25

virtual base view

AVC compatible representation of an independently coded non-base view

NOTE The virtual base view of an independently coded non-base view is created according to the process specified in H.8.5.5 of ISO/IEC 14496-10. Samples containing data units of an independently coded non-base view and samples of the virtual base view are aligned by decoding times.

3.2 Abbreviated terms

AVC	Advanced Video Coding. Where contrasted with SVC or MVC in this International Standard, this term refers to the main part of ISO/IEC 14496-10, including neither Annex G (Scalable Video Coding) nor Annex H (Multiview Video Coding)
BLA	Broken Link Access
CRA	Clean Random Access
CTU	Coding Tree Unit
HEVC	High Efficiency Video Coding
FF	File Format
HRD	Hypothetical Reference Decoder
IDR	Instantaneous Decoding Refresh
MVC	Multiview Video Coding [refers to ISO/IEC 14496-10 when the techniques in Annex H (Multiview Video Coding) are in use]
NAL	Network Abstraction Layer
PPS	Picture Parameter Set
ROI	Region-Of-Interest
SEI	Supplementary Enhancement Information
SPS	Sequence Parameter Set
STSA	Step-wise Temporal Sub-layer Access
SVC	Scalable Video Coding [refers to ISO/IEC 14496-10 when the techniques in Annex G (Scalable Video Coding) are in use]
TSA	Temporal Sub-layer Access
VCL	Video Coding Layer
VPS	Video Parameter Set

4 General Definitions

4.1 Introduction

The specifications in this clause apply to all coding systems identified by chapters in this specification, unless specifically over-ridden by definitions in the clause for a specific coding system.

The following table summarizes the correspondences between the sets of terminology used in video specifications and the ISO Base Media File Format.

Table 1 – Correspondence of terms in video and ISO Base Media File Format

Video	ISO Base Media File Format
-	Movie
Bitstream	Track
Access Unit	Sample

4.2 Elementary stream structure

This specification concerns video coding systems that specify a set of Network Abstraction Layer (NAL) units, which contain different types of data. This subclause specifies the format of the elementary streams for storing such content.

4.3 Sample and Configuration definition

4.3.1 Introduction

Sample: A sample is an access unit as defined in the appropriate specification.

Parameter set sample: A parameter set sample is a sample in a parameter set stream which shall consist of those parameter set NAL units that are to be considered as if present in the video elementary stream at the same instant in time.

4.3.2 Canonical order and restrictions

The elementary stream is stored in the ISO Base Media File Format in a *canonical* format. The canonical format is as *neutral* as possible so that systems that need to customize the stream for delivery over different transport protocols — MPEG-2 Systems, RTP, and so on — should not have to *remove* information from the stream while being free to *add* to the stream. Furthermore, a canonical format allows such operations to be performed against a known initial state.

The canonical stream format is an elementary stream that satisfies the following conditions:

- **Video data NAL units:** All video data NAL units for a single picture shall be contained with the sample whose decoding time and composition time are those of the picture. Each sample shall contain at least one video data NAL unit of the primary picture.

- **SEI NAL units:** All SEI NAL units shall be contained in the parameter set arrays, or in the sample whose decoding time is at the time, or immediately precedes the time (with no intervening samples), when the SEI messages come into effect instantaneously. In general, SEI messages for a picture shall be included in the sample containing that picture and that SEI messages pertaining to a sequence of pictures shall be included in the sample containing the first picture of the sequence to which the SEI message pertains. The order of SEI messages within a sample is as defined in the applicable video coding standard.
- The sequence of NAL units in an elementary stream and within a single sample must be in a valid decoding order for those NAL units as specified in the applicable video coding standard.
- **All timing information is external to stream.** Picture Timing SEI messages that define presentation or composition timestamps may be included in the video elementary stream, as these messages contain other information than timing, and may be required for conformance checking. However, all timing information is provided by the information stored in the various sample metadata tables, and this information over-rides any timing provided in the video layer. Timing provided within the video stream in this file format should be ignored as it may contradict the timing provided by the file format and may not be correct or consistent within itself.

NOTE This constraint is imposed due to the fact that post-compression editing, combination, or re-timing of a stream at the file format level may invalidate or make inconsistent any embedded timing information present within the video stream.

- **No start codes.** The elementary streams shall not include start codes. As stored, each NAL unit is preceded by a length field as specified in 4.3.3; this enables easy scanning of the sample's NAL units. Systems that wish to deliver, from this file format, a stream using start codes will need to reformat the stream to insert those start codes.

4.3.3 Sample format

4.3.3.1 Definition

This subclause defines the structure of the samples. Samples are externally framed and have a size supplied by that external framing. The syntax of a sample is configured via the decoder specific configuration for the elementary stream. An example of the structure of a video sample is depicted in the following figure.

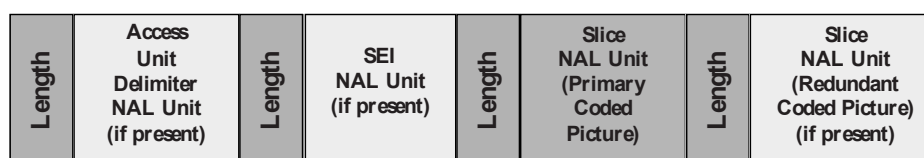


Figure 1 — Example structure of a sample

An access unit is made up of a set of NAL units. Each NAL unit is represented with a:

- *Length:* Indicates the length in bytes of the following NAL unit. The length field can be configured to be of 1, 2, or 4 bytes.
- *NAL Unit:* Contains the NAL unit data as specified in the applicable video coding standard.

4.3.3.2 Syntax

```
aligned(8) class NALUSample
{
    unsigned int PictureLength = sample_size; //Size of Sample from SampleSizeBox
    for (i=0; i<PictureLength; ) //to end of the picture
    {
        unsigned int((DecoderConfigurationRecord.LengthSizeMinusOne+1)*8)
            NALUnitLength;
        bit(NALUnitLength * 8) NALUnit;
        i += (DecoderConfigurationRecord.LengthSizeMinusOne+1) + NALUnitLength;
    }
}
```

4.3.3.3 Semantics

DecoderConfigurationRecord indicates the record in the matching sample entry (e.g. AVCDecoderConfigurationRecord in the case of AVC)

NALUnitLength indicates the size of a NAL unit measured in bytes. The length field includes the size of both the one byte NAL header and the RBSP payload but does not include the length field itself.

NALUnit contains a single NAL unit. The syntax of a NAL unit is defined in the appropriate specification (e.g. ISO/IEC 14496-10) and includes both the one byte NAL header and the variable length encapsulated byte stream payload.

4.4 Video Track Structure

In the terminology of ISO/IEC 14496-12, both video and parameter set tracks are video or visual tracks. They therefore use:

- a) a handler_type of 'vide' in the HandlerBox;
- b) a video media header 'vmhd';
- c) and, as defined below, a derivative of the VisualSampleEntry.

A video stream is represented by one or more video tracks in a file.

If there is more than one track representing scalable aspects of a single stream, then they form alternatives to each other, and the field 'alternate_group' should be used, or the composition system used should select one of them, as appropriate. See 8.10.3 "Track Selection Box" of ISO/IEC 14496-12 for informative labelling of why tracks are members of alternate groups.

4.5 Template fields used

The ISO Base Media File Format defines a number of fields which have default values but which may be defined for use by specific sub-systems. Tracks containing video data may use the following template fields:

- a) alternate_group in the TrackHeaderBox (see 5.4.6 on stream switching).
- b) template field 'depth' in the VisualSampleEntry to document the presence of alpha.

depth takes one of the following values

0x18 – the video sequence is in colour with no alpha

0x28 – the video sequence is in grayscale with no alpha

0x20 – the video sequence has alpha (gray or colour)

4.6 Visual width and height

The width and height fields in a `VisualSampleEntry` must correctly document the cropped frame dimensions (visual presentation size) of the video stream that is described by that entry. The width and height fields do *not* reflect any changes in size caused by SEI messages such as pan-scan. The visual handling of SEI messages such as pan-scan is both optional and terminal-dependent. If the width and height of the sequence changes, then a new sample entry is needed.

Note that the visual size in the SPS may be either frame or field size; in the sample entry, it is always the frame size.

The width and height fields in the track header may not be the same as the width and height fields in the one or more `VisualSampleEntry` in the video track. As specified in the ISO Base Media File Format, if normalized visual presentation is needed, all the sequences are normalized to the track width and height for presentation.

4.7 Decoding time (DTS) and composition time (CTS)

Samples are stored in the file format in decoding order. If picture reordering is not used and decoding and composition times are the same, then presentation is the same as decoding order and only the time-to-sample 'stts' table is used. Note that any kind of picture may be reordered, not only B-pictures.

If decoding time and composition time differ, the composition time-to-sample 'ctts' table is also used in conjunction with the 'stts' table.

4.8 Sync sample (IDR)

A sample is considered as a sync sample if ALL of the following conditions are met:

- The video data NAL units in the sample indicate that the primary picture contained in the sample is an instantaneous decoding refresh (IDR) picture.
- When the sample entry name is 'avc1' or 'avc2', all SPSs and PPSs needed to decode the video data NAL units in the sample of the IDR picture and the following samples in decode order are contained in the decoder configuration of the video elementary stream or in a separate parameter set elementary stream sample.
- When the sample entry name is 'avc3' or 'avc4', the following applies:
 1. If the sample is an IDR access unit, all parameter sets needed for decoding that sample shall be included either in the sample entry or in the sample itself.
 2. Otherwise (the sample is not an IDR access unit), all parameter sets needed for decoding the sample shall be included either in the sample entry or in any of the samples since the previous random access point to the sample itself, inclusive.

A parameter set elementary stream sample is a sync sample if and only if all parameter sets required by the associated video elementary stream from the time of the parameter set sample forward are supplied, in the parameter set stream, before they are required by the associated video elementary stream.

4.9 Shadow sync

The use of the shadow sync table to indicate alternate encodings of a sample for random access are supported as defined in the ISO Base Media File Format. A shadow sync shall indicate a sample that is a random access point as specified in the general requirements and for the specific coding format in the track.

While the use of shadow sync is supported for backward compatibility reasons, this use is deprecated and use of the mechanisms defined in 5.4.6 is recommended.

4.10 Sample groups on random access recovery points and random access points

The video coding system can include the concept of a 'gradual decoding refresh' or random access recovery point. This may be signalled in the bit-stream using a mechanism such as the recovery point SEI message. This message is found at the beginning of the random access, and indicates how much data must be decoded subsequent to the access unit at the position of the SEI message before the recovery is complete.

When all access units in output order starting from the access unit at the position of the SEI message can be successfully decoded after random access, i.e. when the `recovery_frame_cnt` syntax element of the recovery point SEI message is 0, the Random Access Point ('rap ') sample grouping should be used.

This concept of gradual recovery is supported in the file format also by using RollRecoveryEntry Groups [4.5]. In order that the group membership marks the sample containing the SEI message the '`roll-distance`' is constrained to being only positive (i.e. a post-roll). In other words, RollRecoveryEntry Groups can be used when the value of the `recovery_frame_cnt` syntax element of the recovery point SEI message is greater than 0.

Note – The roll-group counts samples in the file format; this may not match the way that the distances are represented in the SEI message.

Within a stream, it is necessary to mark the beginning of the pre-roll, so that a stream decoder may start decoding there. However, in a file, when performing random access, a deterministic search is desired for the closest preceding frame which can be decoded perfectly (either a sync sample, or the end of a pre-roll).

4.11 Hinting

Note that what the hint tracks call "B frames" are actually 'disposable' pictures or non-reference pictures, for example as defined in ISO/IEC 14496-10.

Care should be taken when the structures in Annex A (aggregators or extractors) are in use and the track is hinted. These structures are defined only for use in the file format and should not be transmitted. In particular, a hint track that points at an extractor in a video track would cause the extractor itself to be transmitted (which is probably both incorrect and not the desired behaviour), not the data the extractor references. Hint tracks should normally directly reference NAL units specified in the applicable video coding standard.

5 AVC elementary streams and sample definitions

5.1 Introduction

The Advanced Video Coding (AVC) standard, jointly developed by the ITU-T and ISO/IEC JTC 1/SC 29/WG 11 (MPEG), offers not only increased coding efficiency and enhanced robustness, but also many features for the systems that use it. To enable the best visibility of, and access to, those features, and to enhance the opportunities for the interchange and interoperability of media, this part of ISO/IEC 14496 defines a storage format for video streams compressed using AVC.

This clause defines the storage for plain AVC streams, where 'plain AVC' refers to the main part of ISO/IEC 14496-10, excluding Annex G (Scalable Video Coding) and Annex H (Multiview Video Coding).

This clause specifies the elementary stream and sample structure used to store AVC visual content.

The storage of AVC content uses the existing capabilities of the ISO base media file format but also defines extensions to support the following features of the AVC codec.

- Switching pictures:
to enable switching between different coded streams and substitution of pictures within the same stream.
- Sub-sequences and layers:
provides a structuring of the dependencies of a group of pictures to provide for a flexible stream structure (e.g. in terms of temporal scalability and layering).
- Parameter sets:
the sequence and picture parameter set mechanism decouples the transmission of infrequently changing information from the transmission of coded macroblock data. Each slice containing the coded macroblock data references the picture parameter set containing its decoding parameters. In turn, the picture parameter set references a sequence parameter set that contains sequence level decoding parameter information.

5.2 Elementary stream structure

Two types of elementary streams are defined for storing AVC content (see also Figure 2):

- **Video Elementary Streams** shall contain all video coding related NAL units (i.e. those NAL units containing video data or signalling video structure) and may contain non-video coding related NAL units such as SEI messages and access unit delimiter NAL units. Other NAL units that are not expressly prohibited may be present, and if they are unrecognized should be ignored (e.g. not placed in the output buffer while accessing the file).
- **Parameter set elementary streams** shall not contain video coding related NAL units (i.e. those NAL units containing video data or signalling video structure), and would normally contain only sequence parameter sets, picture parameter sets and sequence parameter set extension NAL units.

Using these stream types, AVC content shall be stored in one of these configurations:

- **Video elementary stream with no parameter sets:** In this case, sequence and picture parameter set NAL units shall be stored in the sample entries of this track. Sequence and picture parameter set NAL units shall not be part of AVC samples within the stream itself.

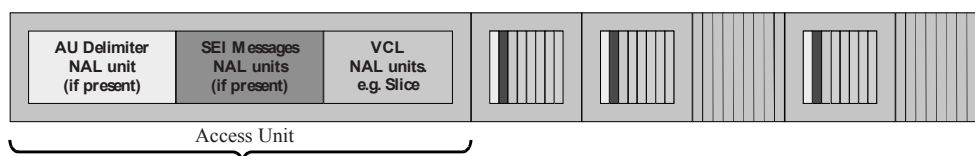
- **Video elementary stream possibly including parameter sets:** In this case, the sample entry indicates whether the stream may contain parameter sets of given types, in addition to other parameters provided in the sample entry. Sequence and picture parameter set NAL units may therefore be part of AVC samples within the stream itself.
- **Video elementary stream and parameter set elementary stream:** In this case, sequence and picture parameter set NAL units shall be transmitted only in the parameter set elementary stream and shall neither be present in the sample entries nor the AVC samples of the video elementary stream.

The types of NAL units that are allowed in each of the video and parameter set elementary streams are specified in the following table.

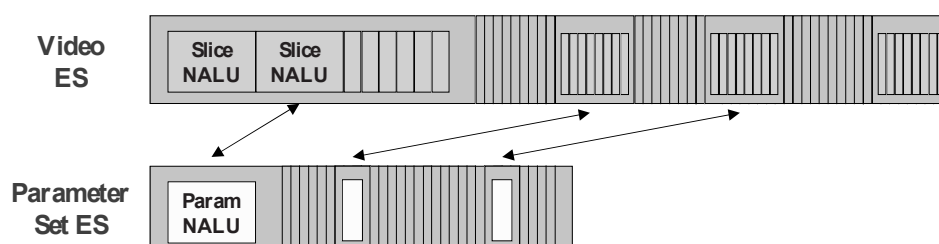
Table 2 – NAL Unit types in elementary Streams

Value of nal_unit_type	Description	Video elementary stream (sample entry 'avc1' or 'avc2')	Video elementary stream (sample entry 'avc3' or 'avc4')	Parameter set elementary stream
0	Unspecified	Not specified by this part of ISO/IEC 14496	Not specified by this part of ISO/IEC 14496	Not specified by this part of ISO/IEC 14496
1	Coded slice of a non-IDR picture slice_layer_without_partitioning_rbsp()	Yes	Yes	No
2	Coded slice data partition A slice_data_partition_a_layer_rbsp()	Yes	Yes	No
3	Coded slice data partition B slice_data_partition_b_layer_rbsp()	Yes	Yes	No
4	Coded slice data partition C slice_data_partition_c_layer_rbsp()	Yes	Yes	No
5	Coded slice of an IDR picture slice_layer_without_partitioning_rbsp()	Yes	Yes	No
6	Supplemental enhancement information(SEI) sei_rbsp()	Yes. Except for the Sub-sequence, layering or Filler SEI messages	Yes Except for the Sub-sequence, or layering SEI messages	Only 'declarative' SEIs should be present
7	Sequence parameter set (SPS) seq_parameter_set_rbsp()	No. If parameter set elementary stream is not used, SPS shall be stored in the Decoder Specific Information.	Yes Parameter set elementary stream shall not be used	Yes
8	Picture parameter set (PPS) pic_parameter_set_rbsp()	No. If parameter set elementary stream is not used, PPS shall be stored in the Decoder Specific Information.	Yes Parameter set elementary stream shall not be used	Yes
9	Access unit delimiter (AU Delimiter) access_unit_delimiter_rbsp()	Yes	Yes	No

10	End of sequence end_of_seq_rbsp()	Yes	Yes	No
11	End of stream end_of_stream_rbsp()	Yes	Yes	No
12	Filler data (FD) filler_data_rbsp()	No	Yes	No
13	Sequence parameter set extension seq_parameter_set_extension_rbsp()	No. If parameter set elementary stream is not used, Sequence Parameter Set Extension shall be stored in the Decoder Specific Information.	Yes Parameter set elementary stream shall not be used	Yes
14...18	Reserved	Not specified by this part of ISO/IEC 14496	Not specified by this part of ISO/IEC 14496	Not specified by this part of ISO/IEC 14496
19	Coded slice of an auxiliary coded picture without partitioning slice_layer_without_partitioning_rbsp()	Yes	Yes	No
20...23	Reserved	Not specified by this part of ISO/IEC 14496	Not specified by this part of ISO/IEC 14496	Not specified by this part of ISO/IEC 14496
24 – 31	Unspecified	Not specified by this part of ISO/IEC 14496	Not specified by this part of ISO/IEC 14496	Not specified by this part of ISO/IEC 14496



(a) Single video elementary stream containing NAL units



(b) Synchronized video and parameter sets with arrows denoting synchronization between streams

Figure 2 — AVC elementary stream structure

5.3 Sample and Configuration definition

5.3.1 Introduction

AVC sample: An AVC sample is an access unit as defined in ISO/IEC 14496-10, 7.4.1.2.

AVC parameter set sample: An AVC parameter set sample is a sample in a parameter set stream which shall consist of those parameter set NAL units that are to be considered as if present in the video elementary stream at the same instant in time.

5.3.2 Canonical order and restrictions

The canonical stream format is an AVC elementary stream that satisfies the following conditions in addition to the general conditions in 4.3.2:

- **Video data NAL units (Coded Slice, Coded Slice Data Partition A, Coded Slice Data Partition B, Coded Slice Data Partition C, Coded Slice IDR Pictures):** All slice and data partition NAL units for a single picture shall be contained with the sample whose decoding time and composition time are those of the picture.
- **Parameter sets:** If a parameter set elementary stream is used, then the sample in the parameter stream shall have a decoding time equal or prior to when the parameter set(s) comes into effect instantaneously. This means that for a parameter set to be used in a picture it must be sent prior to the sample containing that picture or in the sample for that picture.
- **NOTE –** When the sample entry name is 'avc1' or 'avc2', parameter sets are stored either in the sample entries of the video stream or in the parameter set stream, but never in both. This ensures that it is not necessary to examine every part of the video elementary stream to find relevant parameter sets. It also avoids dependencies of indefinite duration between the sample that contains the parameter set definition and the samples that use it. Storing parameter sets in the sample entries of a video stream provides a simple and static way to supply parameter sets. Parameter set elementary streams on the other hand are more complex but allow for more dynamism in the case of updates. Parameter sets may be inserted into the video elementary stream when the file is streamed over a transport that permits such parameter set updates. When the sample entry name is 'avc3' or 'avc4', then parameter sets may be present both the sample entries and as part of the samples.
- **Parameter set track:** A sync sample in a parameter set track indicates that all parameter sets needed from that (decoding) time forward in the video elementary stream are in that or succeeding parameter stream samples. Also there shall be a parameter set sample at each point a parameter set is updated. Each parameter set sample shall contain exactly the sequence and picture parameter sets needed to decode the relevant section of the video elementary stream.

NOTE The use of a parameter set track in the file format does not require that a system delivering video content use a separate elementary stream for parameter sets. Instead, implementations may choose to map parameter sets to in-band parameter set NAL units in the video elementary stream or use some out-of-band delivery mechanism defined by the transport layer.

- **SEI message NAL units:** The order of SEI messages within a sample is as defined in ISO/IEC 14496-10, 7.4.1.2.
- **Access unit delimiter NAL units:** The constraints obeyed by access unit delimiter NAL units are defined in ISO/IEC 14496-10, 7.4.1.2.3.

- **[AC1] Sub-sequence and layering SEI messages.** Sub-sequence or layering SEI messages should not occur in the AVC elementary stream. Specifically, the sub-sequence information, sub-sequence layer characteristics, and sub-sequence characteristics SEI messages should not occur in the stored AVC video elementary stream. Instead, all such information is stored as external metadata as described in 5.4.5. **[AC1]**
- **Redundant picture:** NAL units within a single access unit shall be ordered in non-decreasing order of redundant picture count (`redundant_pic_cnt`).
- **Slice groups:** NAL units within a primary coded picture or a redundant coded picture shall be ordered in non-decreasing order of slice group identifier. Within the same slice group, slices shall be ordered by their first Macroblock location (`first_mb_in_slice` in the slice header).

NOTE Slice groups are stored in a canonical order to ease hinting, and to make it easier to find a primary picture within a sample.

- **Filler data.** Video data is naturally represented as variable bit rate in the file format and should be filled for transmission if needed. Filler Data NAL units and Filler Data SEI messages shall not be present in the file format stored stream when the sample entry does not also permit parameter sets.

NOTE The removal or addition of Filler Data NAL units, start codes, SEI messages or Filler Data SEI messages may change the bit-stream characteristics with respect to conformance with the HRD when operating the HRD in CBR mode as specified in ISO/IEC 14496-10, Annex C.

5.3.3 Decoder configuration information

This subclause specifies the decoder configuration information for ISO/IEC 14496-10 video content.

5.3.3.1 AVC decoder configuration record

5.3.3.1.1 Definition

This record contains the size of the length field used in each sample to indicate the length of its contained NAL units as well as the initial parameter sets. This record is externally framed (its size must be supplied by the structure which contains it).

This record contains a version field. This version of the specification defines version 1 of this record. Incompatible changes to the record will be indicated by a change of version number. Readers must not attempt to decode this record or the streams to which it applies if the version number is unrecognised.

Compatible extensions to this record will extend it and will not change the configuration version code. Readers should be prepared to ignore unrecognised data beyond the definition of the data they understand (e.g. after the parameter sets in this specification).

When used to provide the configuration of

- a parameter set elementary stream,
- a video elementary stream used in conjunction with a parameter set elementary stream,

the configuration record shall contain no sequence or picture parameter sets (`numOfSequenceParameterSets` and `numOfPictureParameterSets` shall both have the value 0).

When used to provide the configuration of a video elementary stream used without a parameter set elementary stream, the configuration record may or may not contain sequence or picture parameter sets (`numOfSequenceParameterSets` or `numOfPictureParameterSets` may or may not have the value 0).

The values for `AVCProfileIndication`, `AVCLevelIndication`, and the flags which indicate profile compatibility must be valid for all parameter sets of the stream described by this record. The level indication must indicate a level of capability equal to or greater than the highest level indicated in the included parameter sets; each profile compatibility flag may only be set if all the included parameter sets set that flag. The profile indication must indicate a profile to which the entire stream associated with this configuration record conforms. If the sequence parameter sets are marked with different profiles, and the relevant profile compatibility flags are all zero, then the stream may need examination to determine which profile, if any, the entire stream conforms to. If the entire stream is not examined, or the examination reveals that there is no profile to which the entire stream conforms, then the stream must be split into two or more sub-streams with separate configuration records in which these rules can be met.

Explicit indication can be provided in the AVC Decoder Configuration Record about the chroma format and bit depth used by the avc video elementary stream. The parameter '`chroma_format_idc`' present in the sequence parameter set in AVC specifies the chroma sampling relative to the luma sampling. Similarly the parameters '`bit_depth_luma_minus8`' and '`bit_depth_chroma_minus8`' in the sequence parameter set specify the bit depth of the samples of the luma and chroma arrays. The values of `chroma_format_idc`, `bit_depth_luma_minus8` and '`bit_depth_chroma_minus8`' must be identical in all sequence parameter sets in a single AVC configuration record. If two sequences differ in any of these values, two different AVC configuration records will be needed. If the two sequences differ in color space indications in their VUI information, then two different configuration records are also required.

The array of sequence parameter sets, and the array of picture parameter sets, may contain SEI messages of a 'declarative' nature, that is, those that provide information about the stream as a whole. An example of such an SEI is a user-data SEI. Such SEIs may also be placed in a parameter set elementary stream. NAL unit types that are reserved in ISO/IEC 14496-10 and in this specification may acquire a definition in future, and readers should ignore NAL units with reserved values of NAL unit type when they are present in these arrays.

NOTE - this 'tolerant' behaviour is designed so that errors are not raised, allowing the possibility of backwards-compatible extensions to these arrays in future specifications.

When Sequence Parameter Set Extension NAL units occur in this record in profiles other than those indicated for the array specific to such NAL units (`profile_idc` not equal to any of 100, 110, 122, 144), they should be placed in the Sequence Parameter Set Array.

NOTE - the profile identified by `profile_idc` value 144 is deprecated in ISO/IEC 14496-10.

5.3.3.1.2 Syntax

```

[AC1] aligned(8) class AVCDecoderConfigurationRecord {
    unsigned int(8) configurationVersion = 1;
    unsigned int(8) AVCProfileIndication; [AC1]
    unsigned int(8) profile_compatibility;
    unsigned int(8) AVCLevelIndication;
    bit(6) reserved = '111111'b;
    unsigned int(2) lengthSizeMinusOne;
    bit(3) reserved = '111'b;
    unsigned int(5) numOfSequenceParameterSets;
    for (i=0; i< numOfSequenceParameterSets; i++) {
        unsigned int(16) sequenceParameterSetLength;
        bit(8*sequenceParameterSetLength) sequenceParameterSetNALUnit;
    }
    unsigned int(8) numOfPictureParameterSets;
    for (i=0; i< numOfPictureParameterSets; i++) {
        unsigned int(16) pictureParameterSetLength;
        bit(8*pictureParameterSetLength) pictureParameterSetNALUnit;
    }
    [AC1] if ( AVCProfileIndication == 100 || AVCProfileIndication == 110 ||
        AVCProfileIndication == 122 || AVCProfileIndication == 144 ) [AC1]
    {
        bit(6) reserved = '111111'b;
        unsigned int(2) chroma_format;
        bit(5) reserved = '11111'b;
        unsigned int(3) bit_depth_luma_minus8;
        bit(5) reserved = '11111'b;
        unsigned int(3) bit_depth_chroma_minus8;
        unsigned int(8) numOfSequenceParameterSetExt;
        for (i=0; i< numOfSequenceParameterSetExt; i++) {
            unsigned int(16) sequenceParameterSetExtLength;
            bit(8*sequenceParameterSetExtLength) sequenceParameterSetExtNALUnit;
        }
    }
}

```

5.3.3.1.3 Semantics

AVCProfileIndication contains the profile code as defined in ISO/IEC 14496-10.

profile_compatibility is a byte defined exactly the same as the byte which occurs between the profile_IDC and level_IDC in a sequence parameter set (SPS), as defined in ISO/IEC 14496-10.

AVCLevelIndication contains the level code as defined in ISO/IEC 14496-10.

lengthSizeMinusOne indicates the length in bytes of the NALUnitLength field in an AVC video sample or AVC parameter set sample of the associated stream minus one. For example, a size of one byte is indicated with a value of 0. The value of this field shall be one of 0, 1, or 3 corresponding to a length encoded with 1, 2, or 4 bytes, respectively.

numOfSequenceParameterSets indicates the number of SPSs that are used as the initial set of SPSs for decoding the AVC elementary stream.

sequenceParameterSetLength indicates the length in bytes of the SPS NAL unit as defined in ISO/IEC 14496-10.

sequenceParameterSetNALUnit contains a SPS NAL unit, as specified in ISO/IEC 14496-10. SPSs shall occur in order of ascending parameter set identifier with gaps being allowed.

numOfPictureParameterSets indicates the number of picture parameter sets (PPSs) that are used as the initial set of PPSs for decoding the AVC elementary stream.

pictureParameterSetLength indicates the length in bytes of the PPS NAL unit as defined in ISO/IEC 14496-10.

`pictureParameterSetNALUnit` contains a PPS NAL unit, as specified in ISO/IEC 14496-10. PPSs shall occur in order of ascending parameter set identifier with gaps being allowed.

`chroma_format` contains the `chroma_format` indicator as defined by the `chroma_format_idc` parameter in ISO/IEC 14496-10.

`bit_depth_luma_minus8` indicates the bit depth of the samples in the Luma arrays. For example, a bit depth of 8 is indicated with a value of zero ($\text{BitDepth} = 8 + \text{bit_depth_luma_minus8}$). The value of this field shall be in the range of 0 to 4, inclusive.

`bit_depth_chroma_minus8` indicates the bit depth of the samples in the Chroma arrays. For example, a bit depth of 8 is indicated with a value of zero ($\text{BitDepth} = 8 + \text{bit_depth_chroma_minus8}$). The value of this field shall be in the range of 0 to 4, inclusive.

`numOfSequenceParameterSetExt` indicates the number of Sequence Parameter Set Extensions that are used for decoding the AVC elementary stream.

`sequenceParameterSetExtLength` indicates the length in bytes of the SPS Extension NAL unit as defined in ISO/IEC 14496-10.

`sequenceParameterSetExtNALUnit` contains a SPS Extension NAL unit, as specified in ISO/IEC 14496-10.

5.4 Derivation from ISO Base Media File Format

5.4.1 AVC File type and identification

Conformance with this part of ISO/IEC 14496 is indicated by the presence of the brand of a specification that permits the inclusion of AVC content, in the compatible brands list of the `FileTypeBox` as defined in ISO/IEC 14496-12. The file extension normally matches the major brand.

AVC content may be used in an MPEG-4 context; in a file with extension “.mp4”, the major brand may be ‘avc1’.

Readers conformant to this part of ISO/IEC 14496 should read the file if a suitable brand occurs in the compatible-brands list. Other structures and/or track types, defined in specifications other than that identified by the brand, may be present, and these may be ignored by a reader conformant with the specification identified by the brand.

5.4.2 AVC Video Stream Definition

This subclause defines the sample entry and sample format for AVC video elementary streams.

5.4.2.1 Sample entry name and format

5.4.2.1.1 Definition

Box Types:	‘avc1’, ‘avc2’, ‘avc3’, ‘avc4’, ‘avcC’, ‘m4ds’, ‘btrt’
Container:	Sample Description Box (‘std’)
Mandatory:	An ‘avc1’, ‘avc2’, ‘avc3’ or ‘avc4’ sample entry is mandatory
Quantity:	One or more sample entries may be present

An AVC visual sample entry shall contain an AVC Configuration Box, as defined below. This includes an `AVCDecoderConfigurationRecord`, as defined in 5.3.3.1.

An optional `MPEG4BitRateBox` may be present in the AVC visual sample entry to signal the bit rate information of the AVC video stream. Extension descriptors that should be inserted into the Elementary Stream Descriptor, when used in MPEG-4, may also be present.

Multiple sample entries may be used, as permitted by the ISO Base Media File Format specification, to indicate sections of video that use different configurations or parameter sets.

The sample entry name 'avc1' or 'avc3' may only be used when the stream to which this sample entry applies is a compliant and usable AVC stream as viewed by an AVC decoder operating under the configuration (including profile and level) given in the `AVCConfigurationBox`. The file format specific structures that resemble NAL units (see Annex A) may be present but must not be used to access the AVC base data; that is, the AVC data must not be contained in Aggregators (though they may be included within the bytes referenced by the `additional_bytes` field) nor referenced by Extractors.

The sample entry name 'avc2' or 'avc4' may only be used when Extractors or Aggregators (Annex A) are required to be supported, and an appropriate Toolset is required (for example, as indicated by the file-type brands). This sample entry type indicates that, in order to form the intended AVC stream, Extractors must be replaced with the data they are referencing, and Aggregators must be examined for contained NAL Units. Tier grouping may be present.

5.4.2.1.2 Syntax

```
// Visual Sequences
class AVCConfigurationBox extends Box('avcC') {
    AVCDecoderConfigurationRecord() AVCCConfig;
}

class MPEG4BitRateBox extends Box('btrt'){
    unsigned int(32) bufferSizeDB;
    unsigned int(32) maxBitrate;
    unsigned int(32) avgBitrate;
}

class MPEG4ExtensionDescriptorsBox extends Box('m4ds') {
    Descriptor Descr[0 .. 255];
}

class AVCSampleEntry() extends VisualSampleEntry (type){
    // type is 'avc1' or 'avc3'
    AVCConfigurationBox config;
    MPEG4BitRateBox (); // optional
    MPEG4ExtensionDescriptorsBox (); // optional
}

class AVC2SampleEntry() extends VisualSampleEntry (type){
    // type is 'avc2' or 'avc4'
    AVCConfigurationBox avcconfig;
    MPEG4BitRateBox bitrate; // optional
    MPEG4ExtensionDescriptorsBox descr; // optional
    extra_boxes boxes; // optional
}
```

5.4.2.1.3 Semantics

Compressorname in the base class `VisualSampleEntry` indicates the name of the compressor used with the value "\012AVC Coding" being recommended; the first byte is a count of the remaining bytes, here represented by \012, which (being octal 12) is 10 (decimal), the number of bytes in the rest of the string

`config` is defined in 5.3.3. If a separate parameter set stream is used, `numOfSequenceParameterSets` and `numOfPictureParameterSets` must both be zero.

`Descr` is a descriptor which should be placed in the `ElementaryStreamDescriptor` when this stream is used in an MPEG-4 systems context. This does not include `SLConfigDescriptor` or `DecoderConfigDescriptor`, but includes the other descriptors in order to be placed after the `SLConfigDescriptor`.

`bufferSizeDB` gives the size of the decoding buffer for the elementary stream in bytes.

`maxBitrate` gives the maximum rate in bits/second over any window of one second.

`avgBitrate` gives the average rate in bits/second over the entire presentation.

5.4.3 AVC parameter set stream definition

5.4.3.1 Sample entry name and format

5.4.3.1.1 Definition

Box Types: 'avcp'
Container: Sample Description Box ('stsd')
Mandatory: Yes
Quantity: One or more sample entries may be present

An AVC parameter stream sample entry shall contain an AVC Parameter Stream Configuration Box, as defined below.

5.4.3.1.2 Syntax

```
class AVCPParameterSampleEntry() extends VisualSampleEntry ('avcp'){  
    AVCConfigurationBox config;  
}
```

5.4.3.1.3 Semantics

`Compressorname` in the base class `VisualSampleEntry` indicates the name of the compressor used with the value "\016AVC Parameters" being recommended (\016 is 14, the length of the string as a byte).

`config` is defined in 5.3.3. `numOfSequenceParameterSets` and `numOfPictureParameterSets` must both be zero.

5.4.3.2 Sample format

5.4.3.2.1 Definition

This subclause defines the sample format for AVC Parameter set streams. An AVC parameter set sample contains only one or more sequence, picture parameter set, or sequence parameter set extension NAL units.

5.4.3.2.2 Syntax

```

aligned(8) class AVCPParameterSample
{
    unsigned int PictureLength = sample_size;
    //Size of AVCPParameterSample from SampleSizeBox
    for (i=0; i<PictureLength; ) // to end of the picture
    {
        unsigned int((AVCDecoderConfigurationRecord.LengthSizeMinusOne+1)*8)
        NALUnitLength;
        bit(NALUnitLength * 8) NALUnit;
        i += (AVCDecoderConfigurationRecord.LengthSizeMinusOne+1) + NALUnitLength;
    }
}

```

5.4.3.2.3 Semantics

NALUnitLength indicates the size of a NAL unit measured in bytes. The length field includes the size of both the one byte NAL header and the EBSP payload but does not include the length field itself.

NALUnit contains a single NAL unit. The syntax of a NAL unit is defined in ISO/IEC 14496-10 and includes both the one byte NAL header and the variable length encapsulated byte stream payload.

5.4.3.3 Track reference

A track reference of type 'avcp' in the video elementary stream track reference table, referencing the parameter set stream, is used to connect from the video elementary stream to the parameter set elementary stream.

5.4.4 Parameter sets

This subclause applies when a separate parameter set stream is not used.

Each AVC sample entry, which contains the AVC video stream decoder specific information, includes a group of SPSs and PPSs. This group of parameter sets functions much like a codebook. Each parameter set has an identifier, and each slice references the parameter set it was coded against using the parameter set's identifier.

When the sample entry name is 'avc1' or 'avc2', the following applies:

- In the file format each configuration of parameter sets is represented separately. A parameter set cannot be updated without causing a different sample entry to be used. For example, suppose that samples 1 to 4 use PPSs identified as 1, 2, 3 and a single SPS identified as 1. At sample 5 a new value of PPS 2 is required but PPSs 1 and 3 remain unaltered and are used until sample 10. In this case, the sample entry for samples 1 through 4 is the same and contains the initial values of PPSs 1, 2, 3 and SPS 1. At sample 5 the sample entry must change to a second sample entry, which contains the updated value for PPS 2 *as well as* the original values of PPSs 1 and 3 and SPS 1. This second sample entry is used for samples 5 through 10.
- Systems wishing to send SPS or PPS updates will need to compare the two configurations to find the differences in order to send the appropriate parameter set updates.

NOTE 1 It is recommended that when several parameter sets are used and parameter set updating is desired, a separate parameter set elementary stream be used.

NOTE 2 Decoders conforming to this specification are required to support both parameter sets stored in separate elementary streams as well as parameter sets stored in the AVC sample entries, unless restricted by another specification using this one.

When the sample entry name is 'avc3' or 'avc4', parameter sets may be present in both sample entry and as part of samples, and an update of a parameter set by a parameter set of the same type that is stored as part of a sample is possible.

5.4.5 Layering and sub-sequences

5.4.5.1 Introduction

Streams may be constructed so that the referential dependencies between samples allow only subsets of the samples to be sent to the decoder. This mechanism is called *thinning* a stream. Thinning discards entire sets of samples using knowledge of what other sets of pictures this set of pictures depends on and what picture sets in turn depend on it.

The referential dependencies between samples in a stream are structured into *layers* and *sub-sequences*. Samples in higher layers can only depend on samples in lower layers. Layers are numbered, and the samples are organized such that a sample in layer N has no dependencies on samples in layers greater than N.

Sub-sequences are as defined in the Annex D of ISO/IEC 14496-10. Dependency relations between sub-sequences represent the dependency structure of a stream. Each sub-sequence belongs to one and only one layer. A sample shall reside in one layer and in one sub-sequence only.

Layering and sub-sequence information is represented in the file format to allow systems reading the files to understand the ways in which stream thinning may be achieved without having to examine the dependency structure of every sample.

Layer and sub-sequences are represented in the AVC file format as Sample Group. An AVC file shall contain zero or one instance of a SampleToGroupBox (per track) with a `grouping_type` equal to 'avll'. This SampleToGroupBox instance represents the assignment of samples in a track to layers. An accompanying instance of the SampleGroupDescriptionBox with the same grouping type shall, if it exists, contain AVCLayerEntry sample group entries describing the layers. Similarly, an AVC file shall contain zero or one instance of a SampleToGroupBox (per track) with a `grouping_type` equal to 'avss'. This SampleToGroupBox instance represents the assignment of samples in a track to sub-sequences. An accompanying instance of the SampleGroupDescriptionBox with the same grouping type shall, if it exists, contain AVCSubSequenceEntry sample group entries describing the sub-sequences.

5.4.5.2 Sub-sequence description entry

5.4.5.2.1 Definition

Group Types: 'avss'
Container: Sample Group Description Box ('sgpd')
Mandatory: No
Quantity: Zero or more.

A sub-sequence description entry is a sample group entry that describes a sub-sequence. A sub-sequence is a set of samples in a track belonging to the same layer. A sub-sequence depends on another sub-sequence if and only if there exists a sample in the sub-sequence that is directly referentially dependent on some sample in the other sub-sequence. All samples in a sub-sequence shall directly depend only on (i.e., refer to) other samples within the same sub-sequence or samples in the sub-

sequences on which is it dependent. A sub-sequence can depend on zero or more sub-sequences in the lower layers. A sub-sequence shall not depend on any other sub-sequence in the same or higher layer.

At most one partition of an AVC stream into layers shall exist in the AVC file format; that is, there is either zero or one instances of the sample group boxes (SampleToGroupBox, SampleGroupDescriptionBox) per track concerning the grouping of samples into layers and sub-sequences.

5.4.5.2.2 Syntax

```
aligned(8) class DependencyInfo
{
    unsigned int(8)    subSeqDirectionFlag;
    unsigned int(8)    layerNumber;
    unsigned int(16)   subSequenceIdentifier;
}

class AVCSubSequenceEntry () extends VisualSampleGroupEntry ('avss')
{
    unsigned int(16) subSequenceIdentifier;
    unsigned int(8)  layerNumber;
    unsigned int(1)  durationFlag;
    unsigned int(1)  avgRateFlag;
    unsigned int(6)  reserved = 0;
    if (durationFlag)
        unsigned int(32) duration;
    if (avgRateFlag)
    {
        unsigned int(8)  accurateStatisticsFlag;
        unsigned int(16) avgBitRate;
        unsigned int(16) avgFrameRate;
    }
    unsigned int(8) numReferences;
    DependencyInfo dependency[numReferences];
}
}
```

5.4.5.2.3 Semantics

subSeqDirectionFlag, layerNumber and subSequenceIdentifier within the DependencyInfo class identify a sub-sequence that is used as a reference for this sub-sequence. Only direct, not indirect, referential dependencies shall be identified. The identified sub-sequence has sub-sequence identifier equal to subSequenceIdentifier and resides in the layer having the layer number equal to layerNumber. If subSeqDirectionFlag is 0, the sub-sequence used as a reference for this sub-sequence is the closest sub-sequence among all the candidate sub-sequences whose first picture precedes the first picture of this sub-sequence in decoding order and which resides in the indicated layer and has the indicated sub-sequence identifier; 'closest' means that among all the candidate sub-sequences the first picture of the referenced sub-sequence is the closest to the first picture of this sub-sequence in decoding order. If subSeqDirectionFlag is equal to 1, the sub-sequence used as a reference for this sub-sequence is the closest sub-sequence among all the candidate sub-sequences whose first picture succeeds the first picture of this sub-sequence in decoding order and which resides in the indicated layer and has the indicated sub-sequence identifier; 'closest' has the same meaning as above.

subSequenceIdentifier gives the identifier for the sub-sequence.

layerNumber gives the layer number to which the sub-sequence belongs.

`durationFlag` equal to 0 indicates that the duration of the target sub-sequence is not specified. Otherwise, a value of 1 indicates that the `duration` field indicates the duration of this sub-sequence.

`avgRateFlag` equal to 0 indicates that the average bit rate and the average frame rate of the target sub-sequence are unspecified. Otherwise, a value of 1 indicates that the average rate characteristics are described by the `accurateStatisticsFlag`, `avgBitRate`, and `avgFrameRate` fields.

`duration` indicates the duration of the target sub-sequence in clock ticks of a 90-kHz clock.

`accurateStatisticsFlag` indicates how reliable the values of `avgBitRate` and `avgFrameRate` are. `accurateStatisticsFlag` equal to 1 indicates that `avgBitRate` and `avgFrameRate` are rounded from statistically correct values. `accurateStatisticsFlag` equal to 0 indicates that `avgBitRate` and `avgFrameRate` are estimates and may deviate somewhat from the correct values.

`avgBitRate` gives the average bit rate in (1000 bits)/second of this sub-sequence. All NAL units of this sub-sequence are taken into account in the calculation. In the following, B is the number of bits in all NAL units in the sub-sequence. t_1 is the decoding timestamp of the first picture of the sub-sequence (in decoding order), and t_2 is the decoding timestamp of the last picture of the sub-sequence (in decoding order). Then, the `avgBitRate` is calculated as follows provided that $t_1 \neq t_2$: $\text{avgBitRate} = \text{round}(B \div ((t_2 - t_1) * 1000))$. If $t_1 = t_2$, `avgBitRate` shall be 0.

`avgFrameRate` gives the average frame rate in units of frames/(256 seconds) of this sub-sequence. All NAL units of this sub-sequence are taken into account in the calculation. The average frame rate is calculated according to the presentation timestamp of the frame. In the following, C is the number of frames in the sub-sequence. t_1 is the presentation timestamp of the first picture of the sub-sequence (in decoding order), and t_2 is the presentation timestamp (in seconds) of the last picture of the sub-sequence (in decoding order). Then, the `avgFrameRate` is calculated as follows provided that $t_1 \neq t_2$: $\text{avgFrameRate} = \text{round}(C * 256 \div (t_2 - t_1))$. If $t_1 = t_2$, `avgFrameRate` shall be 0. Value zero indicates an unspecified frame rate.

`numReferences` gives the number of sub-sequences directly referenced in this sub-sequence. `dependency` is an array of `DependencyInfo` structures giving the identifying referenced sub-sequences.

5.4.5.3 Layer description entry

5.4.5.3.1 Definition

Group Types: 'avll'
Container: Sample Group Description Box ('sgpd')
Mandatory: No
Quantity: Zero or more.

A layer sample group entry defines the layer information for all samples in a layer. Layers are numbered with non-negative integers. Layers are ordered hierarchically based on their dependency on each other: A layer having a larger layer number is a higher layer than a layer having a smaller layer number. The layers are ordered hierarchically based on their dependency on each other so that a layer does not depend on any higher layer and may depend on lower layers. The lowest layer is numbered as zero and other layers are given consecutive numbers. In other words, layer 0 is independently decodable, pictures in layer 1 may be predicted from layer 0, pictures in layer 2 may be predicted from layers 0 and 1, etc.

5.4.5.3.2 Syntax

```

class AVCLayerEntry() extends VisualSampleGroupEntry ('av11')
{
    unsigned int(8)    layerNumber;
    unsigned int(8)    accurateStatisticsFlag;
    unsigned int(16)   avgBitRate;
    unsigned int(16)   avgFrameRate;
}

```

5.4.5.3.3 Semantics

`layerNumber` gives the number of this layer with the base layer being numbered as zero and all enhancement layers being numbered as one or higher with consecutive numbers.

`accurateStatisticsFlag` indicates how reliable the values of `avgBitRate` and `avgFrameRate` are. `accurateStatisticsFlag` equal to 1 indicates that `avgBitRate` and `avgFrameRate` are rounded from statistically correct values. `accurateStatisticsFlag` equal to 0 indicates that `avgBitRate` and `avgFrameRate` are estimates and may deviate somewhat from the correct values.

`avgBitRate` gives the average bit rate in units of 1000 bits per second. All NAL units in this and lower sub-sequence layers are taken into account in the calculation. The average bit rate is calculated according to the decoding timestamp. In the following, B is the number of bits in all NAL units in this and lower sub-sequence layers. t_1 is the decoding timestamp of the first picture in this and lower sub-sequence layers in the presentation order, and t_2 is the decoding timestamp of the latest picture in this and lower sub-sequence layers in the presentation order. Then, `avgBitRate` is calculated as follows provided that $t_1 \neq t_2$: $\text{avgBitRate} = \text{round}(B \div ((t_2 - t_1) * 1000))$. If $t_1 = t_2$, `avgBitRate` shall be 0. Value zero indicates an unspecified bit rate.

`avgFrameRate` gives the average frame rate in units of frames/(256 seconds). All NAL units in this and lower sub-sequence layers are taken into account in the calculation. In the following, C is the number of frames in this and lower sub-sequence layers. t_1 is the presentation timestamp of the first picture in this and lower sub-sequence layers in presentation order, and t_2 is the presentation timestamp of the latest picture in this and lower sub-sequence layers in the presentation order. Then, the `avgFrameRate` is calculated as follows provided that $t_1 \neq t_2$: $\text{avgFrameRate} = \text{round}(C * 256 \div (t_2 - t_1))$. If $t_1 = t_2$, `avgFrameRate` shall be 0. Value zero indicates an unspecified frame rate.

5.4.6 Alternate streams and switching pictures

5.4.6.1 Switching Pictures

In typical streaming scenarios, one of the key requirements is to scale the bit rate of the compressed data in response to changing network conditions. The simplest way to achieve this is to encode multiple streams with different bandwidths and quality settings for representative network conditions. The server can then switch amongst these pre-coded streams in response to network conditions. In earlier standards, switching between streams is only possible at I-pictures, because the pictures can only be switched when there are no dependencies on prior pictures for reconstruction.

AVC includes supports for SP-pictures and SI-pictures ("switching pictures") that allow switching from one stream to another while still supporting inter coding of switching pictures. The following figure shows how SP pictures are used to switch between two different bit streams.

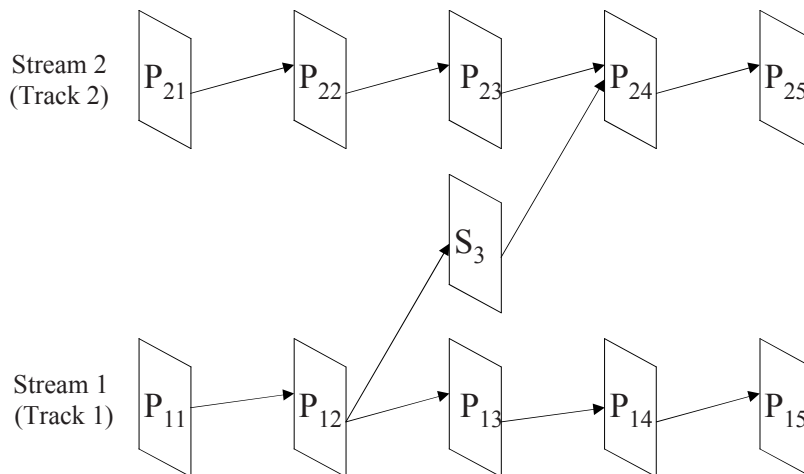


Figure 3 — Stream switching

In the file format, switching pictures are stored in *switching picture* tracks, which are tracks separate from the track that is being switched from and the track being switched to. Switching picture tracks can be identified by the existence of a specific required track reference in that track. A switching picture is an alternative to the sample in the destination track that has exactly the same decoding time. If all switching pictures are SI pictures, then no further information is needed.

If any of the pictures in the switching track are SP pictures, then two extra pieces of information may be needed. First, the source track that is being switched from must be identified by using a track reference (the source track may be the same track as the destination track). Second, the dependency of the switching picture on the samples in the source track may be needed, so that a switching picture is only used when the pictures on which it depends have been supplied to the decoder.

This dependency is represented by means of an optional extra sample table. There is one entry per sample in the switching track. Each entry records the relative sample numbers in the source track on which the switching picture depends. If this array is empty for a given sample, then that switching sample contains an SI picture. If the dependency box is not present, then only SI-frames shall be present in the track.

A switching sample may have multiple coded representations with different dependencies. For AVC video, the multiple representations of a switching sample are stored in different switching tracks (i.e. access unit). For example, one switch track might contain a SP-picture representation dependent on some earlier samples, used for stream switching, while another switch track may contain another representation as an SI-picture, used for random access.

5.4.6.2 Alternate group

The ISO Base Media File Format (but not the version one specification of the MPEG-4 file format, which is branded as 'mp41') supports the use of what is called *alternate tracks*. Each track can optionally specify an *alternate group* (in the track header box) that groups together alternate encodings of the same content. Thus, each alternate bit-stream can be stored as a separate track and related together as alternate tracks. All the tracks which form a group which may be switched between, but not the tracks containing the switching pictures, must be a member of an *alternate_group* with a non-zero group identifier.

An alternate group is not needed if there is only one primary track, with a switching track. This switching track may contain SI pictures, or SP pictures for trick modes or error resilience, which predict both from and to the same track.

5.4.6.3 Track references

The switching track must be linked to the track into which it switches (the destination track) by a track reference of type 'swto' in the switching picture track.

If the switching track contains SP pictures, the switching track must be linked to the track from which it switches (the source track) by a track reference of type 'swfr' in the switching picture track.

5.4.6.4 Sample dependency

5.4.6.4.1 Definition

Box Type: 'sdep'
 Container: Sample Table 'stbl'
 Mandatory: No
 Quantity: Zero or exactly one.

This subclause defines the dependencies of each switching sample on sample(s) in the source track. This table is only needed in a switching track that has a source ('swfr') track dependency.

This box contains the sample dependencies for each switching sample. The dependencies are stored in the table, one record for each sample. The size of the table, `sample_count` is taken from the `sample_count` in the Sample Size Box ('stsz') or Compact Sample Size Box ('stz2').

5.4.6.4.2 Syntax

```
aligned(8) class SampleDependencyBox
{
    extends FullBox('sdep', version = 0, 0) {
        for (i=0; i < sample_count; i++){
            unsigned int(16) dependency_count;
            for (k=0; k < dependency_count; k++) {
                signed int(16) relative_sample_number;
            }
        }
    }
}
```

5.4.6.4.3 Semantics

`dependency_count` is an integer that counts the number of samples in the source track on which this switching sample directly depends.

`relative_sample_number` is an integer that identifies a sample in the source track. The relative sample numbers are encoded as follows. If there is a sample in the source track with the same decoding time, it has a relative sample number of 0. Whether or not this sample exists, the sample in the source track which immediately precedes the decoding time of the switching sample has relative sample number -1, the sample before that -2, and so on. Similarly, the sample in the source track which immediately follows the decoding time of the switching sample has relative sample number +1, the sample after that +2, and so on.

5.4.7 Definition of a sub-sample for AVC

For the use of the sub-sample information box (8.7.7 of ISO/IEC 14496-12) in an AVC stream, a sub-sample is defined as one or more contiguous NAL units within a sample and having the same value of the following fields; `RefPicFlag`, `RedPicFlag` and `VclNalUnitFlag`. Each sub-sample includes both NAL unit(s) and their preceding NAL unit length field(s). The presence of this box is optional; however, if present in a track containing AVC data, it shall have the semantics defined here.

The subsample_priority field shall be set to a value in accordance with the specification of this field in ISO/IEC 14496-12.

The discardable field shall be set to 1 only if this sample can still be decoded if this sub-sample is discarded (e.g. the sub-sample consists of an SEI NAL unit, or a redundant coded picture).

[AC1] The codec_specific_parameters field of the Subsample Information box is defined for AVC as follows: **[AC1]**

```
unsigned int(1) RefPicFlag;  
unsigned int(1) RedPicFlag;  
unsigned int(1) VclNalUnitFlag;  
unsigned int(29) reserved = 0;
```

RefPicFlag equal to 0 indicates that all the NAL units in the sub-sample have nal_ref_idc equal to 0. RefPicFlag equal to 1 indicates that all the NAL units in the sub-sample have nal_ref_idc greater than 0.

RedPicFlag equal to 0 indicates that all the NAL units in the sub-sample have redundant_pic_cnt equal to 0. RedPicFlag equal to 1 indicates that all the NAL units in the sub-sample have redundant_pic_cnt greater than 0.

VclNalUnitFlag equal to 0 indicates that all NAL units in the sub-sample are non-VCL NAL units. Value 1 indicates that all NAL units in the sub-sample are VCL NAL units.

5.4.8 Post-decoder requirements scheme for signalling of SEI for AVC

5.4.8.1 General

In order to handle situations where the file author requires certain actions on the player or renderer, the ISO base media file format specifies the restricted-video mechanism where sample entries are hidden behind the generic sample entry 'resv'. The mechanism applies to AVC and for this case a file author can list occurring SEI message IDs [ISO/IEC 14496-10] and classify them into two categories: those that are deemed required by the file author for correct playback, and others. The occurrence of either type of SEI messages can be signalled in the SEI Information box.

The scheme for signalling of SEI for AVC is defined here.

The SchemeType 'aSEI' (AVC SEI) is used.

5.4.8.2 SEI Information box

5.4.8.2.1 Definition

Box Type: 'seii'
Container: Scheme Information box ('schi')
Mandatory: Yes (when the SchemeType is 'aSEI')
Quantity: One

The SEI Information box is contained in the Scheme Information box when the SchemeType is 'aSEI', and contains information about the SEI messages present in the AVC stream. Although the SEI messages are not required for decoding, the file author may require certain actions for rendering or other purposes. The box distinguishes between SEI which is required to be understood for correct playback and SEI which is not required for correct playback (but may enhance playback).

By inspecting the SEI Information box a player will know which SEI messages that occur in the bitstream and whether they are required for rendering.

When the coding system is AVC and the restriction is signaled with SEI messages, then the SEI messages listed here should be stored either in the bitstream or in the AVC Configuration Record. The SEI Information box does not contain the actual SEI messages, it only lists those that occur in the bitstream.

5.4.8.2.2 Syntax

```
aligned(8) class SeiInformationBox extends Box('sei') {
    unsigned int(16) numRequiredSEIs;
    for (i=0; i<numRequiredSEIs; i++) {
        unsigned int(16) requiredSEI_ID;
    }
    unsigned int(16) numNotRequiredSEIs;
    for (i=0; i<numNotRequiredSEIs; i++) {
        unsigned int(16) notrequiredSEI_ID;
    }
}
```

5.4.8.2.3 Semantics

`requiredSEI_ID` takes on the value “payloadType” of an SEI message present in the AVC stream that is deemed necessary by the file author for correct playback.

`notrequiredSEI_ID` takes on the value “payloadType” of an SEI message present in the AVC stream that is not deemed necessary by the file author for correct playback.

6 SVC elementary stream and sample definitions

6.1 Introduction

This clause specifies the storage format of SVC data. It extends the definitions of the storage format of AVC in clause 5.

The file format for storage of SVC content, as defined in this clause and Annexes Annex A to Annex D, uses the existing capabilities of the ISO base media file format and the plain AVC file format (i.e. the file format specified in clause 5). In addition, the following new extensions, among others, to support SVC-specific features are specified.

- Scalable or multiview grouping:
a structuring and grouping mechanism to indicate the association of NAL units with different types and hierarchy levels of scalability.
- Aggregator:
a structure to enable efficient scalable grouping of NAL units by changing irregular patterns of NAL units into regular patterns of aggregated data units.
- Extractor:
a structure to enable efficient extraction of NAL units from other tracks than the one containing the media data.
- Temporal metadata statements:
structures for storing time-aligned information of media samples.
- AVC compatibility:
a provision for storing an SVC bitstream in an AVC compatible manner, such that the AVC compatible base layer can be used by any plain AVC file format compliant reader.