INTERNATIONAL STANDARD

ISO/IEC 23000-19

Second edition
2020-03

Information technology — Multimedia application format (MPEG-A) —

Part 19: Common media application format (CMAF) for segmented media

Technologies de l'information — Format pour application multimédia (MPEG-A) —

Partie 19: Format CMAF (Common Media Application Format) pour médias segmentés

Reference number
ISO/IEC 23000-19:2020(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Fax: +41 22 749 09 47
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

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.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

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. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents) or the IEC list of patent declarations received (see http://patents.iec.ch).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see www.iso.org/iso/foreword.html.

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

This second edition cancels and replaces the first edition (ISO/IEC 23000-19:2018), which has been technically revised. It also incorporates the Amendments ISO/IEC 23000-19:2018/Amd.1:2018 and ISO/IEC 23000-19:2018/Amd.2:2019.

The main changes compared to the previous edition are as follows:

— addition of supplemental data brands;

— modification to the structural brand cmfc for compatibility with DASH segments;

— definition of a stricter brand '`cmf2`' for legacy devices;

— refinements and updates to HEVC media profiles for SDR and HDR;

— definition of the scalable HEVC media profile;

— definition of AAC multichannel media profiles;

— definition of MPEG-H 3D audio track format and CMAF media profile;

— definition of MPEG-D USAC track format and CMAF media profile;

— definition of IMSC1.1 media profile.

A list of all parts in the ISO/IEC 23000 series can be found on the ISO website.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

Common media application format (CMAF) combines and constrains several MPEG specifications to define a multimedia format that is optimized for delivery of a single adaptive multimedia presentation to a variety of devices, using a variety of adaptive streaming, broadcast, download and storage methods.

Several MPEG specifications have been adopted for much of the video delivered over the internet and other IP networks (cellular, cable, broadcast, etc.). Various organizations have taken MPEG's core coding, file format and system standards and combined them into their own specifications for their specific application. While these specifications are similar, their differences result in unnecessary duplication of engineering effort and duplication of identical content in slightly different formats, which results in increased storage and delivery costs.

CMAF provides a common media specification that application specifications, such as MPEG dynamic adaptive streaming over HTTP (DASH), can reference and a common media format that allows a single encoded multimedia presentation to be used by many applications.

Information technology — Multimedia application format (MPEG-A) —

Part 19: Common media application format (CMAF) for segmented media

1 Scope

This document specifies the CMAF multimedia format, which contains segmented media objects optimized for streaming delivery and decoding on end user devices in adaptive multimedia presentations.

CMAF specifies a track format derived from the ISO base media file format, then derives addressable media objects from CMAF tracks that can be used for storage and delivery.

CMAF specifies sets of tracks that share encoding and packaging constraints that enable the selection of multiple tracks to form a multimedia presentation and allow seamless switching of alternative encodings of the same content at different bit rates, frame rates, resolution, etc.

CMAF specifies a hypothetical application model that determines how tracks in a CMAF presentation are intended to be combined and synchronized to form a multimedia presentation. The model abstracts delivery to allow any delivery method. The hypothetical application model assumes a manifest and player, but CMAF does not specify a manifest, player, or delivery protocol, with the intent that any that support the hypothetical application model can be used.

CMAF specifies media profiles and brands that constrain media encoding and packaging of CMAF tracks to enable seamless adaptive switching of tracks and allow devices to identify compatible content by its brand.

CMAF specifies presentation profiles that conditionally require sets of CMAF tracks conforming to specified media profiles and allow content creators and devices to identify compatible multimedia presentations.

CMAF enables extensibility by specifying how new media profiles and presentation profiles can be specified and identified and includes guidelines for those specifications.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 14496-1, *Information technology — Coding of audio-visual objects — Part 1: Systems*

ISO/IEC 14496-3, *Information technology — Coding of audio-visual objects — Part 3: Audio*

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*

ISO/IEC 14496-14, *Information technology — Coding of audio-visual objects — Part 14: MP4 file format*

ISO/IEC 14496-15, *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*

ISO/IEC 14496-30, *Information technology — Coding of audio-visual objects — Part 30: Timed text and other visual overlays in ISO base media file format*

ISO/IEC 23001-7, *Information technology — MPEG systems technologies — Part 7: Common encryption in ISO base media file format files*

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

ISO/IEC 23009-1, *Information technology — Dynamic adaptive streaming over HTTP (DASH) — Part 1: Media presentation description and segment formats*

ISO/IEC 23008-3:2019,*Information technology — High efficiency coding and media delivery in heterogeneous environments — Part 3: 3D audio*

ISO/IEC 23091-3, *Information technology — Coding-independent code points — Part 3: Audio*

ISO/IEC 23003-4:2015, *MPEG audio technologies — Part 4: Dynamic range control*

ISO/IEC 23003-3:2012, *Information technology — MPEG audio technologies — Part 3: Unified speech and audio coding*

IETF RFC 5234, *Augmented BNF for Syntax Specifications: ABNF,* https://tools.ietf.org/html/rfc5234

IETF RFC 6381:2011, *The 'Codecs' and 'Profiles' Parameters for "Bucket" Media Types,* https://tools.ietf.org/html/rfc6381

ITU-R Recommendation BT.709, *Parameter values for the HDTV standards for production and international programme exchange*

ITU-R Recommendation BT.1886, *Reference electro-optical transfer function for flat panel displays used in HDTV studio production*

ITU-R Recommendation BT.2035, *A reference viewing environment for evaluation of HDTV program material or completed programmes*

ITU-T Recommendation X.667:2014, *Information technology — Open Systems Interconnection — Procedures for the operation of OSI Registration Authorities: Generation and registration of Universally Unique Identifiers (UUIDs) and their use as ASN.1 object identifier components,* https://www.itu.int/rec/T-REC-X.667

ANSI/CTA-608-E R-2014, *Line 21 Data Services,* http://www.techstreet.com/standards/cta-608-e-r2014?product_id=1815447

ANSI/CTA-708-E, *Digital Television (DTV) Closed Captioning*, http://www.techstreet.com/standards/cta-708-e?product_id=1860354

W3C IMSC1, *TTML Profiles for Internet Media Subtitles and Captions 1.0,* http://www.w3.org/TR/ttml-imsc1

W3C IMSC1.1, *TTML Profiles for Internet Media Subtitles and Captions 1.1,* http://www.w3.org/TR/ttml-imsc1.1

W3C, *TTML Media Type Definition and Profile Registry, W3C Working Group Note,* https://www.w3.org/TR/ttml-profile-registry

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— IEC Electropedia: available at http://www.electropedia.org/

— ISO Online browsing platform: available at http://www.iso.org/obp

3.1 Media objects

3.1.1
CMAF fragment
encoded ISO BMFF media segment conforming to CMAF constraints

3.1.2
CMAF header
sequence of CMAF constrained ISO BMFF boxes that do not reference any *media samples* (3.3.15), but are associated with a *CMAF track* (3.2.1) and necessary for the decoding of its *CMAF fragments* (3.1.1)

3.1.3
CMAF addressable media object
CMAF media object packaged for storage or delivery

Note 1 to entry: Examples include a *CMAF track file* (3.1.6) containing a *CMAF header* (3.1.2) and *CMAF fragments* (3.1.1), or a *CMAF segment* (3.1.5) containing one or more CMAF fragments, or a *CMAF chunk* (3.1.4) containing a partial sequence of the *media samples* (3.3.15) of a CMAF fragment.

3.1.4
CMAF chunk
CMAF media object that contains a consecutive subset of the *media samples* (3.3.15) of a *CMAF fragment* (3.1.1), where only the first CMAF chunk of a CMAF fragment is constrained to be an adaptive *switching* (3.3.9) point

3.1.5
CMAF segment
CMAF addressable media object (3.1.3) consisting of one or more consecutive *CMAF fragments* (3.1.1) from the same *CMAF track* (3.2.1)

Note 1 to entry: A "CMAF segment" is conformant to an "ISO BMFF segment" and a "DASH segment".

3.1.6
CMAF track file
one *CMAF track* (3.2.1) stored consecutively in a single ISO BMFF file with the earliest *CMAF fragment* (3.1.1) constrained to start at decode time zero

3.2 Logical structure

3.2.1
CMAF track
sequence of *CMAF fragments* (3.1.1) that are consecutive in presentation time, contain one media stream, conform to at least one structural CMAF brand, together with an associated *CMAF header* (3.1.2) that can initialize playback

3.2.2
CMAF switching set
set of one or more *CMAF tracks* (3.2.1), where each track is an alternative encoding of the same source content, and are constrained to enable seamless track *switching* (3.3.9)

3.2.3
aligned CMAF switching set
set of *CMAF switching sets* (3.2.2), the *CMAF tracks* (3.2.1) of which all contain alternative encodings of the same source content in time-aligned *CMAF fragments* (3.1.1), but all CMAF tracks do not conform to a single CMAF switching set

3.2.4
CMAF selection set
set of one or more *CMAF switching sets* (3.2.2), where each CMAF switching set encodes an alternative aspect of the same presentation over the same time period, only one of which is intended to be played at a time, e.g., an alternative language or codec

3.2.5
CMAF presentation
set of one or more *CMAF selection sets* (3.2.4) that can be simultaneously decoded to produce a multimedia user experience, potentially including synchronized audio, video, and subtitles

3.2.6
CMAF media profile
encoding constraint on a *CMAF track* (3.2.1) and its contained *media samples* (3.3.15) associated with a CMAF structural brand

3.2.7
CMAF presentation profile
requirement on the *CMAF media profiles* (3.2.6) contained in a *CMAF presentation* (3.2.5)

3.2.8
required media profile
CMAF media profile (3.2.6) conditionally required by a *CMAF presentation profile* (3.2.7)

3.2.9
manifest
document describing one or more *CMAF presentations* (3.2.5)

Note 1 to entry: Manifest formats are not specified in this document.

3.2.10
audio programme
complete collection of all audio programme components and, if present, a set of accompanying presets

3.2.11
audio programme component
smallest addressable unit of an audio programme

3.2.12
CMAF supplemental data
data that can be present in a *CMAF track* (3.2.1) and its contained *media samples* (3.3.15) conformant to a set of requirements identified by a brand

3.2.13
CMAF structural brand
four-character code used in brand-signaling boxes to indicate compliance to box-level constraints as opposed to media-level constraints

3.3 Application model

3.3.1
CMAF hypothetical application model
CMAF presentation (3.2.5) application model based on *late binding* (3.3.3) and synchronization of *CMAF tracks* (3.2.1) that partly determines the CMAF track encoding constraints necessary for an intended CMAF presentation

3.3.2
player
component of the *CMAF hypothetical application model* (3.3.1) responsible for interpreting a *manifest* (3.2.9), requesting resources, and rendering a *CMAF presentation* (3.2.5)

3.3.3
late binding
selection (3.3.8) and synchronization of separately stored *CMAF tracks* (3.2.1) by a *player* (3.3.2) resulting in a synchronized multimedia presentation

3.3.4
CMAF presentation timeline
timeline shared by all *CMAF tracks* (3.2.1) in a *CMAF presentation* (3.2.5), starting at CMAF presentation time zero, which is coincident with the earliest *media samples* (3.3.15) intended for presentation

3.3.5
presentation time offset
earliest presentation time of each *CMAF track* (3.2.1) at the start of a *CMAF presentation* (3.2.5)

Note 1 to entry: Presentation time offset is an encoded property of tracks in a presentation, but it can also refer to that value stored in a *manifest* (3.2.9).

3.3.6
CMAF fragment duration
sum of the *media sample* (3.3.15) durations documented in the `TrackFragmentRunBox` of all `MovieFragmentHeaderBoxes` in the *CMAF fragment* (3.1.1)

3.3.7
CMAF presentation duration
sum of the *CMAF fragment durations* (3.3.6) of the longest *CMAF track* (3.2.1) in a *CMAF presentation* (3.2.5), starting from its earliest presentation time on the *CMAF presentation timeline* (3.3.4)

3.3.8
selection
choice of a *CMAF track* (3.2.1) from alternatives in a selection set (e.g., selecting an audio track by language), possibly by user action or stored user preference

3.3.9
switching
changing to a different *CMAF track* (3.2.1) during presentation, including adaptively switching between *CMAF fragments* (3.1.1) in a *CMAF switching set* (3.2.2)

3.3.10
seamless switching
switching (3.3.9) between *CMAF tracks* (3.2.1) without interrupting presentation of the media content, i.e., decoding *media samples* (3.3.15), at the same time and quality as though their containing CMAF track was decoded without switching

3.3.11
CMAF switching set constraints
CMAF media profile (3.2.6) constraints that enable seamless *switching* (3.3.9) between *CMAF tracks* (3.2.1) in a *CMAF switching set* (3.2.2) conforming to that media profile

3.3.12
single initialization CMAF switching set constraints
additional *CMAF switching set constraints* (3.3.11) so *CMAF fragments* (3.1.1) do not depend on a different *CMAF header* (3.1.2) when *switching* (3.3.9)

3.3.13
resource identifier
externally specified identifier that identifies a *CMAF addressable media object* (3.1.3)

Note 1 to entry: An example is a URI or other object identifier specified by a delivery protocol and *manifest* (3.2.9).

3.3.14
stream access point
media sample (3.3.15) random access property

Note 1 to entry: This is numbered as in ISO/IEC 14496-12:2015, Annex I.

3.3.15
media sample
media data in a *CMAF fragment* (3.1.1) associated with a single decode start time and duration

Note 1 to entry: The term "sample" is often used in the context of video to refer to the spatial samples of an image and in the context of audio to refer to PCM waveform samples. In this document, each type of sample is identified by a defined term. A media sample defined by ISO BMFF is always identified by the term "media sample". The word "sample" is frequently used in ISO BMFF to refer to objects and parameters such as a "sample entry", "sample size", etc., and those terms are used without modification in this document.

3.3.16
audio PCM sample
digital sample quantizing the amplitude of an audio waveform at regular and frequent intervals, e.g., 48 kHz

3.3.17
video spatial sample
quantized values representing the colour and brightness of an area of an image corresponding to a two-dimensional spatial tessellation of the image

3.3.18
subsampling
video encoding using a smaller number of *video spatial samples* (3.3.17) than the source video, that number being an integer submultiple that can be scaled to the source video size based on video stream parameters without position shift or picture aspect ratio distortion

4 Abbreviated terms

AAC	advanced audio coding
ABNF	augmented backus-naur form
ADIF	audio data interchange format
ADTS	audio data transport stream
AOT	audio object type
ASC	audio specific configuration
AU	access unit
AVC	advanced video coding
CCE	coupling channel element
CDN	content delivery network
CMAF	common media application format
CPE	channel pair element

CVS	coded video sequence [A sequence of media samples (coded video frames), starting with a SAP type 1 or 2, and including all media samples prior to the next SAP type 1 or 2 in decoding order.]
DASH	dynamic adaptive streaming over HTTP
DRC	dynamic range control
DRM	digital rights management
DSE	data stream element
DTV	digital television
EBU	European Broadcast Union
EOTF	electro-optical transfer function
GOP	group of pictures
HDMI	high-definition multimedia interface
HDR	high dynamic range
HDTV	high definition television
HEVC	high efficiency video coding
HLG	hybrid log-gamma
HRD	hypothetical reference decoder
IDR	instantaneous decoding refresh
IMDCT	inverse modified discrete cosine transform
IPF	immediate playout frames
ISO BMFF	ISO base media file format, defined in ISO/IEC 14496-12
KID	key identifier, defined in ISO/IEC 23001-7
LFE	low frequency enhancement
LKFS	loudness, K-weighted, relative to nominal full scale
MAE	MPEG-H audio metadata information
MHAS	MPEG-H audio stream
MIME	multipurpose internet mail extensions
MPD	media presentation description
MPEG	moving picture experts group
MSE	media source extension
NAL	network adaptation layer
NTSC	National Television System Committee

OETF	opto-electronic-transfer-funktion
PCM	pulse code modulation
PNG	portable network graphic
PPS	picture parameter set
QMF	quadrature mirror filter
RAP	random access point
RGB	red blue green
SAP	stream access point, defined in ISO/IEC 14496-12
SAR	sample aspect ratio
SBR	spectral band replication
SCE	single channel element
SEI	supplemental enhancement information
SHVC	scalable high efficiency video coding
SPS	sequence parameter set
TTML	timed text markup language
UHD	ultra high definition
URI	uniform resource identifier
URL	uniform resource locator
URN	uniform resource name
USAC	unified speech and audio coding
UTC	coordinated universal time
UUID	universally unique identifier
VOD	video-on-demand
VCL	video coding layer
VPS	video parameter set
VUI	video usability information
XML	eXtensible Mark-up Language

5 Document organization

First-time readers of this document are advised to start with Clause 6 for a description of the objects and terminology specified, the CMAF object model, and the hypothetical application model, which defines how these objects can be combined to form adaptive multimedia presentations.

The normative specifications in Clause 7 through Clause 12 are terse to facilitate development and testing and assume an understanding of Clause 6. Clause 7 specifies ISO Base Media File Format boxes and structures such as movie fragments and tracks that are used to construct all CMAF media objects. Clauses 8 through 11 contain details specific to encryption, audio, video, and subtitle tracks. Clause 12 specifies the combination of CMAF tracks and media profiles into CMAF presentations. It also recommends how to specify additional CMAF media profiles and presentation profiles, which can be specified by other documents and organizations.

CMAF presentation profiles and CMAF media profiles are specified in annexes to allow the addition of new profiles without changing the core document. Additional informative annexes have been added to provide explanations and recommendations on specific topics.

The following is a list of the main clauses of this document, with a brief description of each.

Clause 6 describes the segmented media encoding and playback model using the media objects defined by the CMAF.

Clause 7 describes the use of ISO base media file format for the common media application format brand.

Clause 8 describes how digital rights management information and encryption is applied to the common media application format.

Clause 9 describes the general video track format, constraints for NAL structured video tracks, and the AVC video track format.

Clause 10 describes the general audio track format and specifies two AAC audio CMAF media profiles.

Clause 11 describes the subtitle track format, CMAF media profiles for WebVTT and IMSC1 TTML subtitles, and signalling of CTA 608/708 captions embedded in video streams.

Clause 12 describes the general requirements for CMAF media profiles and CMAF presentation profiles.

Annex A describes several CMAF media profiles, their brands, and a CMAF presentation profile that conditionally requires some of those media profiles. A CMAF presentation shall conform to the provisions of Annex A.

Annex B describes packaging and codec constraints for some CMAF media profiles using the HEVC video codec. Systems claiming conformance to CMAF using HEVC shall conform to the provisions of Annex B.

Annex C describes framing and encoding CMAF switching sets using subsampling and scaling of video to provide seamless playback with adaptive bit rate and scaling.

Annex D describes examples of player track selection, synchronization, and adaptive switching of a CMAF presentation.

Annex E describes the use of event messages attached to media objects to deliver metadata.

Annex F describes maintaining presentation timing and delivery in the event of missing media samples and resources.

Annex G describes encoding recommendations for AAC audio CMAF tracks conforming to adaptive CMAF switching sets.

Annex H specifies the CMAF media profile for scalable HEVC (SHVC). Systems claiming conformance to CMAF using scalable HEVC shall conform to the provisions of Annex H.

Annex I specifies the CMAF media profile for multichannel AAC. Systems claiming conformance to CMAF using multichannel AAC shall conform to the provisions of Annex I.

Annex J specifies the CMAF media profile for MPEG-H audio. Systems claiming conformance to CMAF using MPEG-H shall conform to the provisions of Annex J.

Annex K specifies the CMAF media profile for MPEG-D USAC. Systems claiming conformance to CMAF using MPEG-D USAC shall conform to the provisions of Annex K.

Annex L specifies the CMAF media profile for IMSC 1.1. Systems claiming conformance to CMAF using IMSC 1.1 shall conform to the provisions of Annex L.

6 CMAF hypothetical application model, media object model and profiles

6.1 Overview of the hypothetical application model and media object model

CMAF defines a hypothetical application model so that encoding to that model results in consistent CMAF track encoding, representation in manifests, track selection, late binding, synchronization, decoding, and rendering of CMAF presentations.

Decoding requirements can be inferred from encoding constraints and the hypothetical application model but are not directly specified by CMAF. CMAF does not specify manifest formats or associated resource identification and transport. However, CMAF does specify CMAF addressable media objects derived from encoded CMAF fragments, which can be referenced as resources by a manifest. External specifications can define how a manifest describes a CMAF presentation, including identifying CMAF addressable media objects as resources and representing their logical relationships determined by the CMAF tracks, CMAF switching sets, CMAF selection sets, and CMAF presentations they are derived from.

Figure 1 illustrates the media objects that are specified by CMAF, starting with the encoded CMAF fragments that form CMAF tracks, then logical CMAF track sets determined by CMAF track encoding constraints, then derived CMAF addressable media objects that can package encoded CMAF fragments or their media samples for storage and delivery.

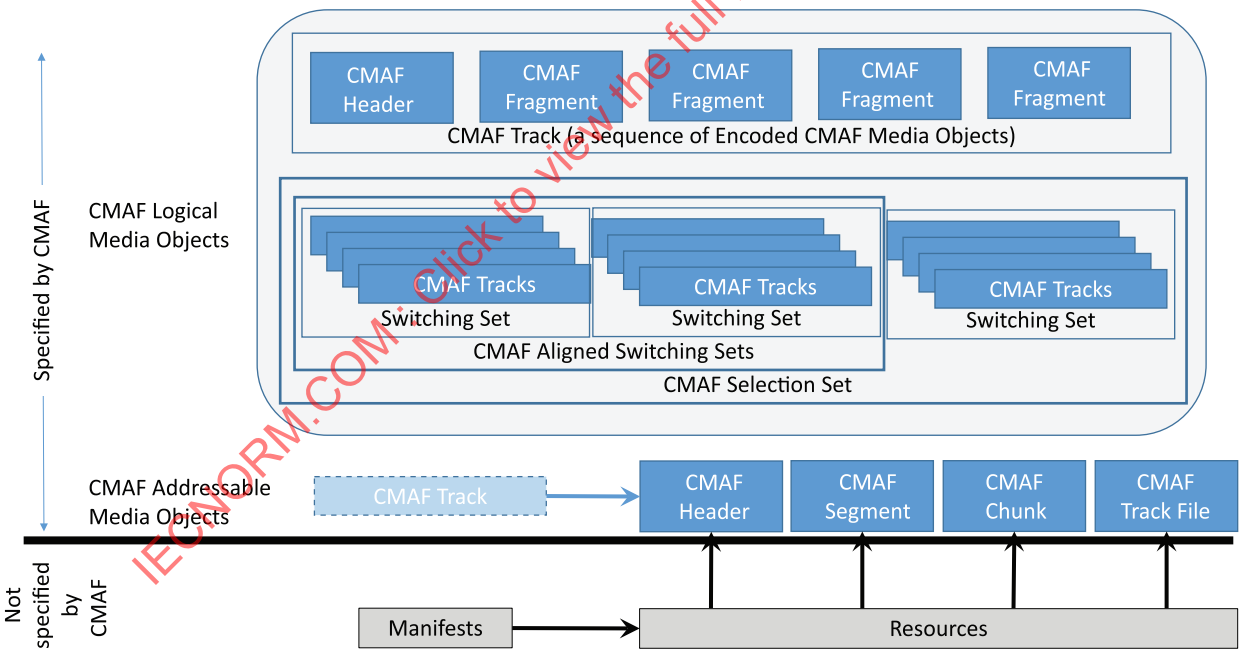


Figure 1 — Media objects specified in CMAF and presented by externally specified applications, such as adaptive streaming

Figure 1 illustrates the mapping between CMAF specified CMAF presentations, and externally specified manifests and resources. Multiple manifests may reference the same CMAF presentation and CMAF addressable media objects. Specification of manifests and resource delivery is outside the scope of this document.

To accurately represent a CMAF presentation, a manifest will describe CMAF track relationships determined by each track's source content and CMAF track encoding constraints, e.g., that CMAF tracks

belong to the same CMAF switching set, which belongs to a CMAF selection set. CMAF groups CMAF tracks based on their encoding constraints in logical media objects called CMAF selection sets and CMAF switching sets that also determine intended use in late binding, track selection, seamless switching, and synchronization. Additional CMAF track metadata such as CMAF media profile brands, "codecs" parameters, language fields, etc. can be included in manifests to enable adaptive track selection and playback, optimized for each user and device.

Manifests can reference CMAF addressable media objects by resource identifiers used by manifests and servers to select the identified CMAF addressable media objects for delivery and playback. Multiple CMAF addressable media object types are specified for different delivery use cases. Use cases include prerecorded content that is downloaded or streamed as files, and live and on demand adaptive streaming over the Internet. The size of the CMAF addressable media objects can be optimized for efficient download and CDN caching, or fast bit rate switching and low latency, depending on the application.

Figure 2 illustrates the relationship between CMAF and streaming specifications that can define a mapping between their manifest and resource formats, and CMAF presentations and the CMAF addressable media objects they include.

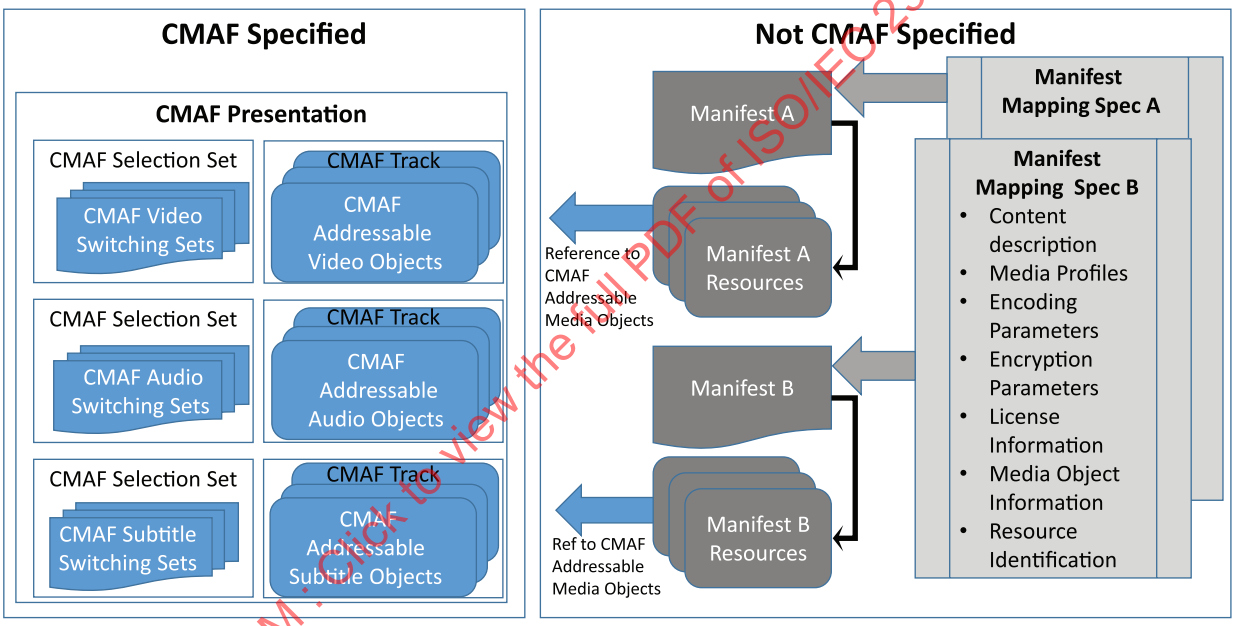


Figure 2 — CMAF hypothetical application model using externally defined manifests that describe the CMAF presentations and media objects

6.2 CMAF content processing model

The CMAF content processing model is shown in Figure 3. Each CMAF presentation is composed of one or more media content components from different source material, for example, audio components in different languages, video components with different views of the same subject, subtitles with different languages or functions, etc. A typical presentation includes audio, video, and subtitles. Presentations are also possible that only include audio components, or include multiple audio components, or include multiple video sources, e.g., side-by-side video, picture in picture, sign language overlay, etc.

Synchronized CMAF tracks can be created and encoded at different times and synchronized on playback if they share a common presentation timeline. Content packaged with the appropriate CMAF fragment decode and presentation times can form a multimedia presentation conforming to the hypothetical application model. Each CMAF track can be encoded, encrypted and packaged by an independent encoder if each encoder constrains timing, encoding and encryption parameters to the constraints of the intended CMAF presentation, CMAF selection set, and CMAF switching set. CMAF tracks in a CMAF switching set can be independently encoded by multiple encoders with the necessary CMAF fragment

time alignment by deriving CMAF fragment timing, spatial and temporal subsampling, etc., from the parameters of the shared media source they encode from.

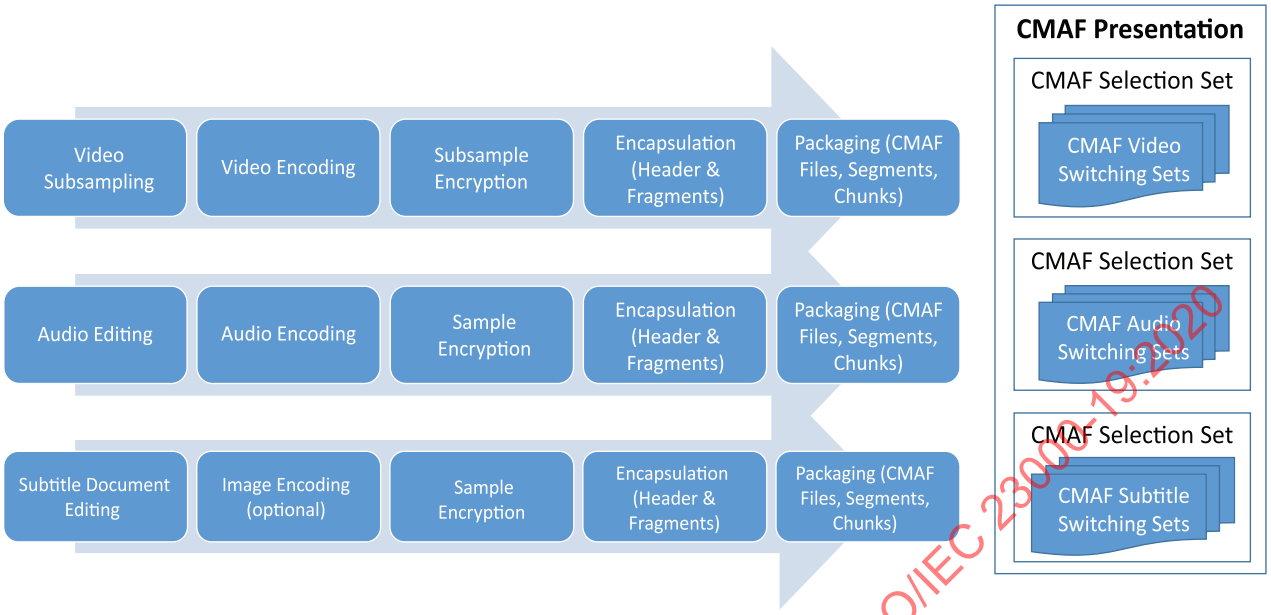


Figure 3 — CMAF presentation content generation

As illustrated in Figure 3, video media components are often preprocessed (e.g., subsampled) and encoded in multiple media streams, typically different in bit rate and encoded resolution. Media streams can be encrypted and encapsulated into CMAF fragments, which can then be packaged as CMAF addressable media objects appropriate for one or more delivery methods. CMAF addressable media objects can be made available as resources and identified as defined by other media application specifications.

CMAF addressable media objects are derived from CMAF fragments and CMAF chunks as specified in Clause 7, and inherit CMAF fragment constraints such as alignment of media sample sequences within a CMAF fragment, CMAF fragment time alignment within a CMAF switching set, CMAF track and CMAF media profile constraints, etc. This makes CMAF fragment encoding and decoding independent of the CMAF addressable media objects used to store and transport the CMAF fragments. But CMAF addressable media objects need not be constructed by packaging complete CMAF fragments as long as the resulting CMAF addressable media objects conform to CMAF fragment and CMAF track constraints.

For live streaming, a CMAF fragment in each CMAF track can be encoded simultaneously from each media content source component, and each CMAF fragment packaged and made available as soon as possible as a CMAF addressable media object, such as a CMAF segment or a CMAF chunk.

6.3 Late binding CMAF track synchronization

The feature of synchronizing separately stored CMAF tracks during playback is referred to as *late binding*. The late binding timing model and accurate recording of CMAF fragment presentation times during encoding are necessary to synchronize late bound CMAF tracks during delivery and presentation.

In the CMAF hypothetical application model, all CMAF tracks in a CMAF presentation share the same CMAF presentation timeline, which has a value of zero at the earliest media sample intended for presentation. All CMAF tracks in a CMAF presentation also share a common track decode timeline origin. A media sample of a CMAF track T1 that has a CMAF track presentation time of X is presented at the same time as the media sample of any other CMAF track T2 that also has the CMAF track presentation time X (the duration X having been divided by each track's timescale).

As defined in ISO BMFF, the presentation time of a media sample in a file is obtained from the decode time (derived from the `baseMediaDecodeTime` in the `TrackFragmentBaseMediaDecodeTimeBox` and the

duration of previous samples) and possibly from composition time offsets and edit lists that offset each track's presentation time on the file's movie presentation timeline, which starts at zero. Equivalent decode and presentation times in different tracks can have different integer values if the tracks have different timescales.

CMAF tracks are not contained in a single multitrack ISO BMFF file with a shared movie timeline, but the CMAF presentation timeline provides an equivalent shared timeline and start point for all CMAF tracks in a CMAF presentation. If a CMAF presentation contains only CMAF track files, the presentation time of the first media sample of each CMAF track file is 0. However, in CMAF presentations using CMAF tracks, the presentation time of the earliest presented media sample can be a non-zero value, but is limited to an equivalent presentation time across all CMAF tracks. All CMAF tracks in a CMAF presentation are required to share a common timeline origin, but they may use different timescales, such as 90 kHz for video and 48 kHz for audio, so an equivalent decode or presentation time can have a different integer value stored in `baseMediaDecodeTime`.

A shared CMAF presentation timeline permits synchronization of different CMAF tracks at the CMAF player by, for example, starting presentation with the earliest media sample of the earliest video CMAF fragment and simultaneously presenting selected audio and subtitle CMAF tracks at their presentation time equivalent to the presentation time of the earliest video media sample. If an audio or subtitle CMAF fragment contains media samples with earlier presentation times than the earliest video media sample, the leading media samples are not expected to be presented. If a CMAF presentation contains no video, then the starting CMAF presentation time can be the earliest audio media sample.

A CMAF track's presentation time at the start of the CMAF presentation can be specified in a manifest presentation time offset so long as the equivalent offset is specified for all CMAF tracks in a CMAF presentation. For instance, each period in a DASH manifest is considered a CMAF presentation. The requirement that CMAF presentations share a common presentation timeline origin means that all presentation time offsets in a DASH Period will be equivalent values.

6.4 Adaptive switching of CMAF tracks in CMAF switching sets

CMAF tracks are contained in CMAF switching sets. Only CMAF tracks encoded from the same media content component can belong to the same switching set, but CMAF tracks encoded from the same media source may also be grouped into different switching sets that each have different encoding constraints necessary for seamless adaptive switching.

CMAF tracks in a CMAF switching set have the following characteristics.

— The CMAF tracks are alternative encodings of a single media content component and media type, e.g., the same audio or video source.

— The CMAF tracks are perceptually equivalent, e.g., the same aspect ratio, colour space, duration, etc., typically resulting from encoding the same media source, i.e., master file or input stream.

— The CMAF tracks conform to a common CMAF media profile.

— The CMAF tracks are seamlessly switchable based on the general constraints on CMAF tracks and encoding constraints defined by each CMAF media profile in the CMAF switching set.

Track switching refers to the presentation of decoded media samples of one CMAF track up to a presentation time *t*, and presentation of decoded media samples of another CMAF track from time *t* onwards. For instance, this could be accomplished using two decoders, downloading overlapping portions of any two streams, decoding both streams, and switching on any decoded media sample. Track switching could be user initiated or programmatic.

Track switching within a CMAF switching set conforming to a CMAF media profile allows a player to download a sequence of non-overlapping CMAF fragments from different CMAF tracks and feed them to a single decoder for seamless playback. When a player performs CMAF track switching automatically during playback in response to available bandwidth, video quality, decoding capacity, etc., that is called "adaptive switching".

6.5 CMAF specified objects and profiles

6.5.1 Object derivation and interoperability code points

Common media application format specifies track formats and media formats for encoding multimedia and metadata derived from the fragmented ISO base media file format and various audio, video, and subtitle media formats.

— CMAF specifies encoded media objects in a CMAF track format that applies to all media types and determines the objects that are encoded and decoded.

— CMAF specifies logical sets of encoded media objects and their general encoding constraints.

— CMAF specifies profiles that determine the specific encoding constraints on media objects that conform to identified interoperability points called profiles.

— CMAF addressable media objects are derived by packaging CMAF headers, CMAF chunks, and CMAF fragments as ISO BMFF data structures available for storage and delivery.

CMAF addressable media objects inherit the properties of encoded media objects, their profiles, and logical media object and set constraints, and make the process of encoding and decoding of CMAF fragments independent of the delivery method and packaging. Different delivery specifications can map the same CMAF addressable media objects to their different manifests to enable interoperability. CMAF addressable media objects, such as CMAF chunks and CMAF segments, are logically derived from CMAF fragments, but it is up to implementations whether to physically derive them from CMAF fragments by post processing or encode and package them directly.

6.5.2 Encoded media objects

— CMAF specifies **CMAF headers** and **CMAF fragments** that form **CMAF tracks** that conform to **CMAF media profiles.** These are the encoded objects from which all other objects are derived.

6.5.3 Logical media object sets

— CMAF specifies **CMAF switching sets** that logically group CMAF tracks that are alternative encodings of the same content constrained to enable seamless switching. A CMAF switching set is constrained to simplify seamless switching between tracks by requiring time-aligned CMAF fragments, constrained encoding parameters, and shared encryption keys.

— CMAF specifies **aligned CMAF switching sets** that logically group multiple CMAF switching sets that are alternative encodings of the same content with time-aligned CMAF fragments. Some players can seamlessly switch between time-aligned CMAF fragments in different CMAF switching sets that are encoded with different characteristics, such as different codecs or encryption keys.

— CMAF specifies **CMAF selection sets** that logically group alternative CMAF switching sets based on their content and encoding constraints, e.g., alternative languages or codecs. At most, one CMAF track at a time is intended to be selected from each CMAF selection set in a CMAF presentation.

— CMAF specifies **CMAF presentations** that logically group CMAF tracks containing related synchronized content that can be selected, streamed, adaptively switched, decoded, decrypted, and synchronized on playback to render a multimedia presentation.

6.5.4 Addressable media objects

— CMAF specifies **CMAF addressable media objects** that are derived from CMAF tracks and fragments, and can be used for storage and delivery of a CMAF presentation. CMAF addressable media objects include **CMAF headers**, **CMAF track files**, **CMAF segments**, and **CMAF chunks**. CMAF fragment encoding and decoding are independent of the addressable media objects used to transport those CMAF fragments.

6.5.5 CMAF profiles, brand and identifiers

6.5.5.1 Overview

CMAF specifies the following brands and identifiers to support the creation and identification of interoperable content. They provide standardized conformance points for consistent content creation and validation. Players and devices can use these brands and identifiers to signal or recognize playback compatibility.

— CMAF specifies CMAF media profiles, CMAF supplemental data, and ISO BMFF brands for widely used audio, video, and subtitle formats. Each CMAF media profile specifies constraints on codecs, media samples, CMAF fragments, and CMAF tracks to identify encoder/decoder interoperability and optional functionality including random access and seamless adaptive switching specified as CMAF switching set constraints. CMAF also specifies guidelines for the specification of CMAF media profiles and matching brands to enable other specifications to define CMAF media profiles that conform to general CMAF requirements. See Table A.1, Table B.1, Table A.2, and Table A.3.

— CMAF specifies CMAF presentation profiles and identifiers that conditionally require inclusion of CMAF tracks conforming to specific CMAF media profiles. CMAF presentations conforming to the CMAF presentation profiles in A.1 are intended to be compatible with most media playback devices.

6.5.5.2 CMAF presentation profile identifiers

CMAF presentation profile identifiers identify CMAF presentations containing CMAF tracks conforming to CMAF media profiles that are conditionally required by the presentation profile. Players can rely on the availability of required CMAF tracks based on the presentation profile identifier. Other CMAF switching sets and media profiles can also be included in the CMAF presentation to enable additional features that might not be supported by all players conforming to the CMAF presentation profile. A CMAF presentation may conform to and signal multiple presentation profiles if all the required CMAF tracks for each presentation profile are available. Presentation profile conformance includes start alignment and synchronization of all CMAF tracks.

6.5.5.3 CMAF media profiles and ISO BMFF brands

CMAF file and media profile brands identify a CMAF track's file and media conformance to CMAF track and media profile constraints. Each media profile defines codec-specific properties in addition to those required by all CMAF tracks, such as the CMAF header sample entry, media sample format, CMAF fragment constraints, media sample presentation synchronization, scaling, mixing, layering, random access and encoding constraints such as codec profiles and levels. CMAF media profiles can specify constraints between multiple CMAF tracks in a CMAF switching set, and thereby imply player requirements necessary to seamlessly switch between CMAF fragments in those CMAF switching sets.

6.5.5.4 Signalling of supplemental data in CMAF tracks

CMAF defines brands, called "supplemental data brands" to identify the presence of additional information in a CMAF track that is not required by the CMAF media profile of that track. For example, NAL units, SEI messages, ISO BMFF Boxes can be added to an otherwise conformant CMAF track file, where this supplemental data does not make the CMAF track file non-conformant. Signalling its presence may be useful during production workflow, and to decoders for track selection and decoder initialization.

6.5.5.5 Single initialization CMAF switching set constraints identifier

Single initialization constraints specify that a CMAF switching set conforms to general CMAF switching set constraints and to additional CMAF media profile defined constraints that enable seamless switching and decoding without dependence on the CMAF header for each CMAF track. CMAF switching set single initialization constraints can be signalled in a manifest, such as the `bitstreamSwitching` attribute in a

DASH manifest. Players can take advantage of these constraints by processing a single CMAF header only once prior to sequencing CMAF fragments from the CMAF switching set.

6.5.5.6 ISO BMFF segment brands for CMAF media objects

Segment type brands are specified in 7.2 that can be stored in a `SegmentTypeBox` that start-delimits and identifies the type of CMAF media object that follows.

6.5.5.7 Specification of CMAF presentation and media profiles

The general requirements for CMAF media profiles and CMAF presentation profiles are specified in Clause 12.

This document contains the following CMAF profiles.

— CMAF presentation profiles are specified in A.1.

— CMAF media profiles and CMAF supplemental data are specified in A.2, A.3, A.4, and B.5.

Additional CMAF media profiles can be specified if they conform to the general CMAF track requirements, specify and register ISO BMFF brands, and define codec-specific CMAF track formats, codec constraints, and optional CMAF switching set constraints for seamless switching following the guidelines in 12.1.

Additional CMAF presentation profiles can be specified if they conform to the general CMAF requirements and specify a URL that identifies the conditionally required media profiles, as described in 12.2.

6.6 CMAF media object model

6.6.1 CMAF fragments

In the CMAF hypothetical application model, CMAF fragments are the media objects that are encoded and decoded.

Each CMAF media profile further constrains CMAF fragments and media sample data.

The CMAF media profiles in Annex A and their referenced track formats constrain CMAF fragments to be decodable independently of each other; for example, NAL structured video CMAF fragments contain one or more complete coded video sequences to make them randomly accessible and independently decodable. Audio CMAF fragments contain a sequence of audio access units, and subtitle CMAF fragments contain a single subtitle document in a media sample, constrained so that each CMAF fragment is randomly accessible and decodable. Common Encryption can also signal the necessary decryption parameters in each CMAF fragment to make it independently decryptable, in combination with an associated CMAF header and keys.

A CMAF fragment typically consists of one `MovieFragmentBox` and `MediaDataBox` pair, but can contain more than one of these pairs. When a fragment contains multiple pairs like this, each pair is called a CMAF chunk, and each CMAF chunk contains a consecutive subset of the CMAF fragment's media samples.

This is illustrated in Figure 4.

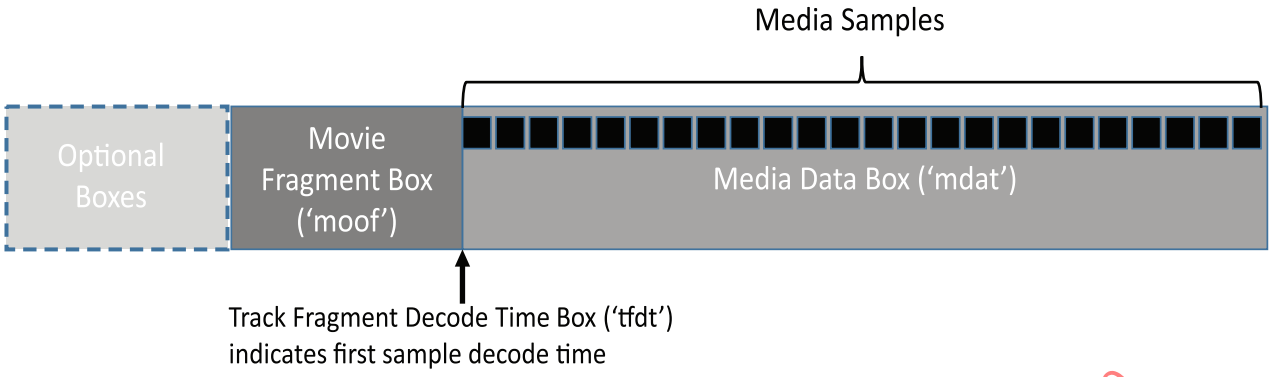


Figure 4 — Example CMAF fragment data structure containing a single CMAF chunk

6.6.2 CMAF tracks

A CMAF track is a continuous sequence of one or more CMAF fragments in presentation order conforming to a CMAF media profile and an associated CMAF header. The CMAF header contains a `MovieBox` sufficient to process and present all CMAF fragments in the CMAF track. A CMAF track can be produced by an encoder and ISO BMFF file packager, but it is made accessible in the form of CMAF addressable media objects that can be referenced as resources defined by an external media application specification.

This is illustrated in Figure 5.

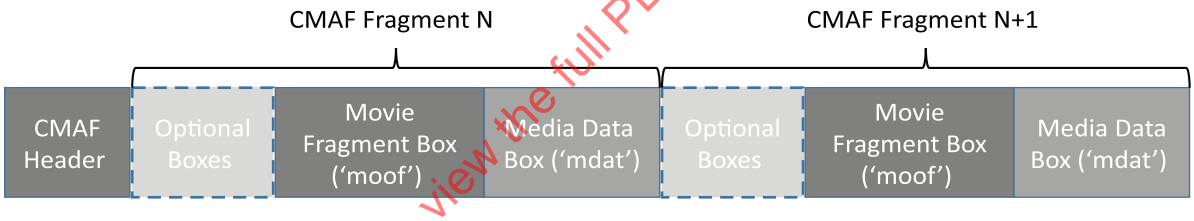


Figure 5 — CMAF track data structure

See Clauses 7 through 11 for additional details on the construction of CMAF fragments and CMAF tracks of different media types.

6.6.3 CMAF track files

A CMAF track file is a CMAF addressable media object consisting of a CMAF track stored in a single ISO BMFF file. A CMAF track file is a stored CMAF track with the following constraints.

— It starts with a CMAF header.

— The CMAF header is followed by a continuous sequence of one or more CMAF fragments stored in presentation order.

— The first CMAF fragment has a `baseMediaDecodeTime` of zero.

Additional boxes, such as `SegmentIndexBoxes`, can be present between the CMAF header and the first CMAF fragment.

This is illustrated in Figure 6.

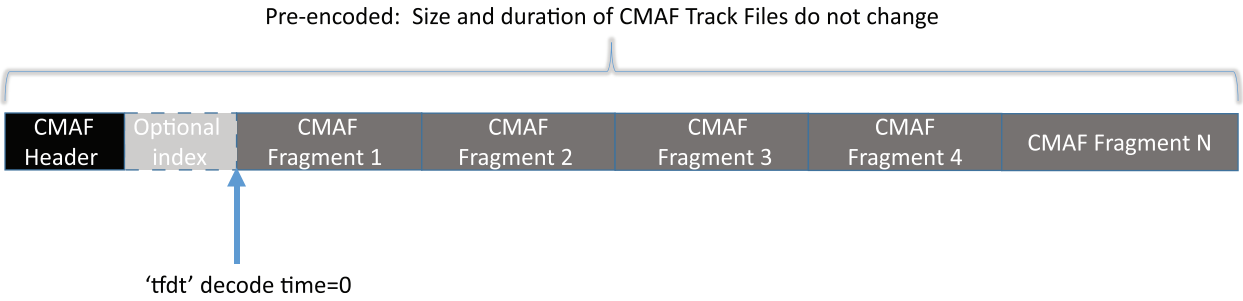


Figure 6 — CMAF track file, addressable media object

A sequence of one or more CMAF fragments can be requested from a CMAF track file resource, for example, using a resource identifier based on HTTP 1.1 consisting of the file URL and a byte range. CMAF resource identifiers can be listed in a manifest or determined by some other delivery format defined method, such as downloading a `SegmentIndexBox` to determine the byte ranges of CMAF fragments. The manifest and request method are out of scope of CMAF. A CMAF track file can also be downloaded or progressively downloaded.

6.6.4 CMAF segments

A CMAF segment is a CMAF addressable media object containing one or more consecutive CMAF fragments from a CMAF track. External application specifications can define how to reference a CMAF segment with a resource identifier that can be used by servers and manifests to reference and deliver each CMAF segment in a CMAF presentation.

This is illustrated in Figure 7.

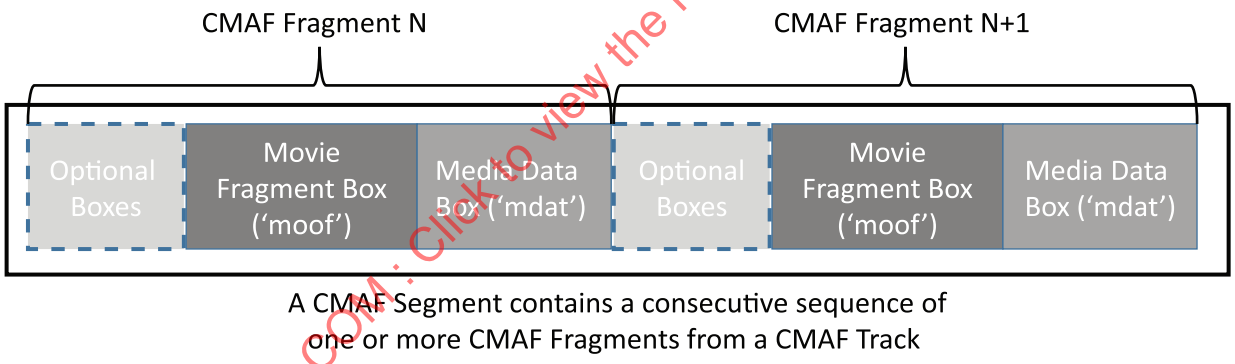


Figure 7 — Example CMAF segment, addressable media object

CMAF video fragment durations are typically 2 seconds to 6 seconds for coding efficiency, and CMAF segment durations are typically not greater than 10 seconds to 12 seconds to limit delivery latency in live streaming and limit bitrate rate adaptation response time.

Subtitle segment durations are usually close to video segment durations in live presentations to avoid increasing presentation delay. In a prerecorded CMAF presentation, a single subtitle CMAF segment can have a duration up to the duration of the CMAF track that contains it.

6.6.5 CMAF chunks

CMAF chunks are CMAF addressable media objects that contain a consecutive subset of the media samples in a CMAF fragment. CMAF chunks can be used by a delivery protocol to deliver media samples as soon as possible during live encoding and streaming, i.e., typically less than a second. The same media samples delivered in a CMAF segment containing one or more CMAF fragments would be delayed for the duration of the CMAF segment, i.e., typically several seconds. CMAF chunks enable the progressive

encoding, delivery, and decoding of each CMAF fragment. A `MovieFragmentBox` at the start of each CMAF chunk provides access to the contained media samples. Broadcast, unicast, and multicast protocols can deliver and identify CMAF chunks through various methods, including resource IDs that include both a CMAF segment number corresponding to a CMAF fragment and the number of the CMAF chunk within the fragment to allow both fragment and chunk identification.

Figure 8 illustrates that CMAF chunks inherit the media sample constraints of the CMAF fragment they are derived from. In this example of a video CMAF fragment containing a coded video sequence, the first media sample of the first CMAF chunk is SAP type 1 or 2, e.g., an IDR picture in an AVC CMAF fragment. The switching and splicing constraints are determined by the CMAF fragment. Boxes that lead the `MovieFragmentBox` in a CMAF fragment can also lead the `MovieFragmentBox` of the first CMAF chunk, although they are not shown in this example.

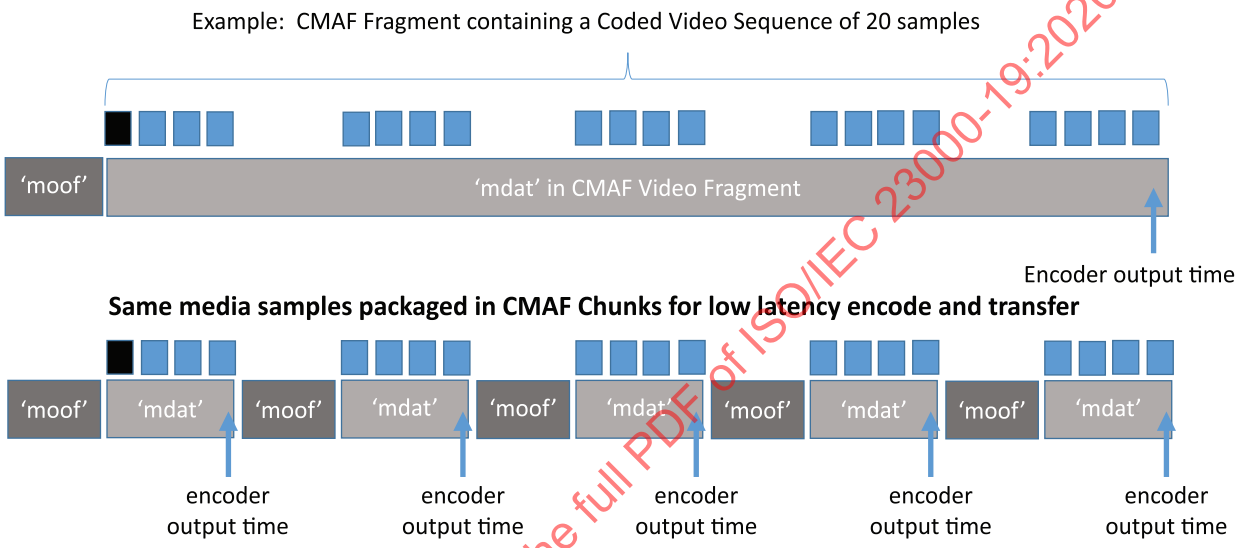


Figure 8 — CMAF chunk, addressable media objects

6.6.6 CMAF switching sets and adaptive switching

6.6.6.1 Overview

As illustrated in Figure 9, CMAF switching sets contain time-aligned CMAF fragments that start with stream access points (SAP type 1 or 2) and timestamps to simplify switching between tracks by sequencing CMAF fragments from different CMAF tracks during playback.

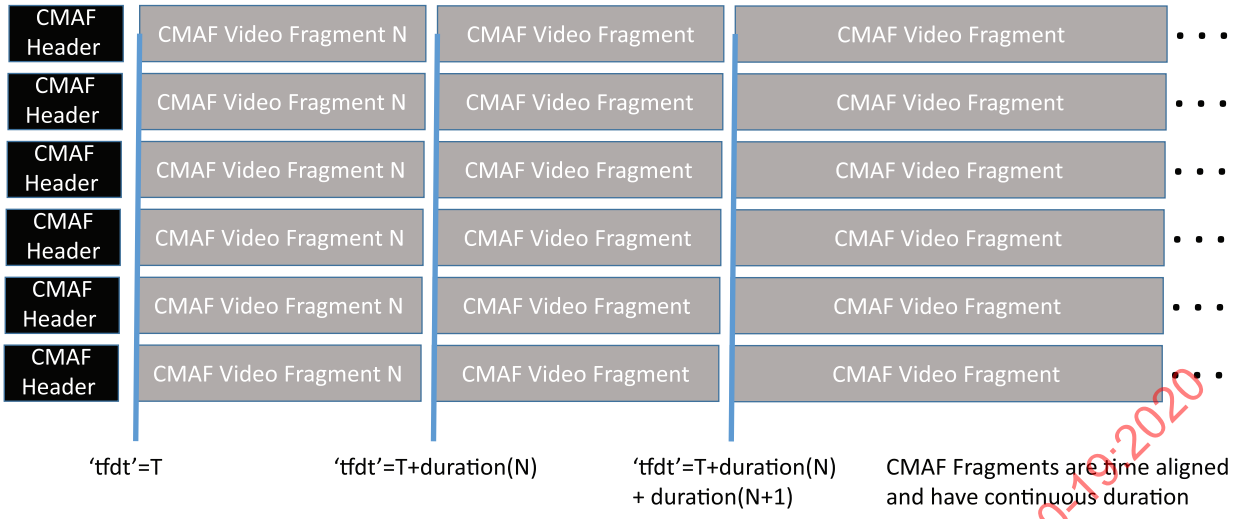


Figure 9 — A CMAF switching set

The manifest description of CMAF switching sets and the CMAF tracks they contain is out of scope of the CMAF specification, but manifests are expected to contain enough CMAF track and addressable media object information to enable automatic CMAF track selection and adaptive switching by players.

6.6.6.2 Decoding adaptively switched CMAF tracks

Each CMAF media profile can optionally specify one or more CMAF switching set constraints that enable players to seamlessly switch between alternative CMAF fragments in a CMAF switching set. CMAF switching set constraints enable player functions including decoder and decryptor initialization, video scaling, audio mixing, media sample presentation timing, etc. that are intended to make adaptive switching as perceptually seamless as possible.

The media profiles in the CMAF specification reference or define CMAF switching set constraints that enable a stream of CMAF headers and CMAF fragments to decode on decoders without downloading overlapping CMAF fragments or requiring the use of multiple decoders.

An adaptively delivered CMAF switching set results in a timed sequence of CMAF headers and fragments a player selects from multiple CMAF tracks contained in a CMAF switching set. The encoding constraints of the CMAF switching set and CMAF media profile determine how the CMAF switching set can be encoded and decoded. They are independent from CMAF addressable media object packaging applied during storage or transport.

The CMAF track format specifies two types of CMAF switching set processing, determined by specific constraints on NAL structured video track formats referenced by some CMAF media profiles. If a CMAF switching set conforms to general constraints, a CMAF track's header is assumed be processed before every switch to that CMAF track. Additional constraints are specified for a CMAF switching set conforming to single initialization constraints that do not depend on processing a new CMAF header to switch to other CMAF tracks within a CMAF switching set.

Figure 10 illustrates the player generated bitstreams resulting from the two types of CMAF switching set constraints and the decoding process each enables. Those two processes and bitstreams are described as "single initialization track switching" and "multiple initialization track switching".

One method of single initialization track switching relies on a "common" CMAF header containing all referenced sample entries and decoding parameters in the CMAF switching set. It is assumed that players download CMAF headers only once per CMAF track. Consequently, new video parameter sets and indexes referenced by portions of a CMAF track encoded after the CMAF header has been downloaded would not be available to a player. Another method relies on parameter sets delivered in-band when those parameter sets are not included in the initially downloaded CMAF header or differ between CMAF tracks.

Audio CMAF switching sets typically only require single initialization because parameters that are allowed to change are contained in the audio media samples and CMAF fragments.

Subtitle CMAF tracks are typically not adaptively switched, and therefore have a single CMAF track per CMAF switching set.

See 7.3.4 for details on CMAF switching sets.

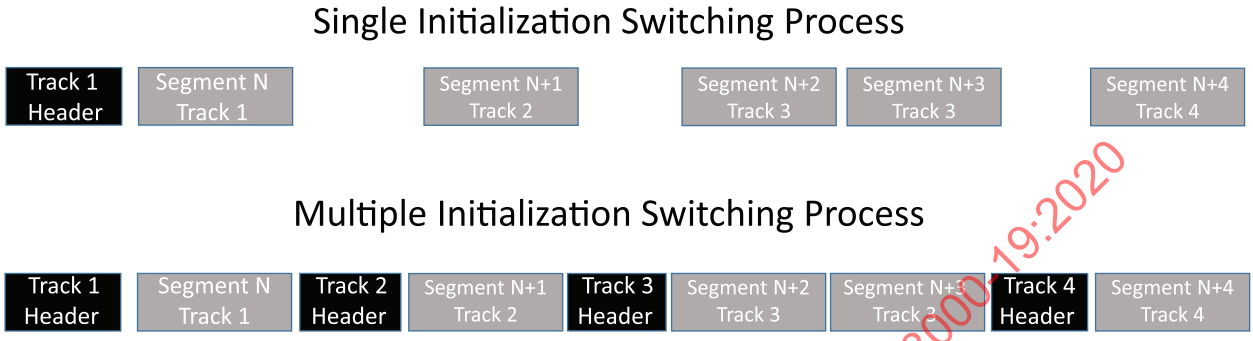


Figure 10 — Adaptively switched resource streams for different CMAF switching set constraints

6.6.6.3 Displaying adaptively switched video CMAF tracks

The normative behaviour of a video decoder processing a single track is well-specified in each video codec specification, such as AVC, but display processing is considered out of scope of those codec specifications. For instance, AVC does not specify the conversion by a display processor of YCbCr 4:2:0 video spatial samples and colour subsamples output by a decoder to 32-bit or 48-bit RGB pixels in a size, colour space, resolution, and refresh rate appropriate for a specific display.

Adaptive switching, enabled by CMAF media profile constraints, allows a player's decoder and display system to deterministically scale different video spatial sampling used to encode different CMAF fragments to the same display size, position, and refresh rate. The encoding of CMAF tracks in a CMAF switching set is constrained so that a sequence of CMAF fragments can be accurately rescaled to the same presentation height and width with precise registration and visual continuity of the video content, including its size, shape, location, colour, brightness, etc. The variations allowed between CMAF tracks conforming to CMAF switching set constraints determine the scaling behaviour needed in display processors, but display processing and other player requirements are not directly specified by CMAF.

6.6.6.4 Video parameter set storage and processing

The CMAF header of a CMAF video track needs to be processed on each CMAF track switch if decoding parameters and parameter indexes are not identical between CMAF tracks, and parameters that can change are not signalled in each CMAF fragment. This is often the case when each video CMAF track stores the video parameter set(s) for that CMAF track in its CMAF header sample table, e.g., the `'avc1'` sample and track format. In that case, a decoder and display processor need to index the correct encoded video spatial sample counts and cropping parameters from a sample entry containing the correct SPS, PPS, and VUI information matching the parameter index encoded in each video NAL slice header. If CMAF tracks in a CMAF switching set use different spatial subsampling, they will have different parameter sets and sample entries, and different CMAF tracks might use the same parameter set index value for a different parameter set. That makes it necessary to process the CMAF header and sample table of each CMAF track when switching to that CMAF track so that the correct decoding parameters in a sample entry are in effect.

Multiple initialization switching can be used to process all CMAF switching sets, but may not result in seamless presentation when a system processes the entire CMAF header rather than just the video parameters allowed to change in a CMAF switching set, such as height, width, and cropping parameters. Multiple initialization adaptive switching is necessary for a NAL structured video CMAF switching set unless it conforms to single initialization constraints specified by its CMAF media profile.

A CMAF header of a CMAF video track only needs to be processed once for a CMAF switching set conforming to single initialization constraints. There could be several ways to achieve this functionality that can be specified by each CMAF media profile.

Two methods of specifying single initialization constraints are the following.

a) A common CMAF header and sample entries for a CMAF switching set

   In this case, each CMAF header and its sample table contain all parameter sets and sample entries used to encode CMAF fragments in the CMAF switching set. All the parameter indexes are unique and correspond to the index values encoded in the video NAL slice headers. There would be a different sample entry for each different subsampling used in the CMAF switching set.

b) Generic CMAF headers and parameter sets stored in each CMAF fragment

   In this case, each CMAF header contains initialization information in the video configuration record and/or additional boxes as specified in Clause 9 and/or a manifest, but the SPS, PPS, and VUI information used for decoding is stored and indexed within each coded video sequence and CMAF fragment. The initialization information is sufficient for the CMAF media profile identified by brand in each CMAF header, and re-initialization is only necessary if switching to a higher CMAF media profile. For example, when a CMAF switching set contains some CMAF tracks that conform to both SD and HD profile, but others that only conform to HD profile, a decoder initialized to HD could switch all CMAF tracks without reinitializing, but a decoder initialized to SD would need to limit playback to SD media profile CMAF tracks, or reinitialize to HD, if possible.

For NAL structured video, single initialization encoding can be accomplished by encoding the Sequence Parameter Set and Picture Parameter Set NALs in the IDR access unit that begins that coded video sequence (e.g., using the `'avc3'` sample description and track format). Then the decoder and display processor can reference the parameter set necessary to decode and scale each CMAF fragment from the parameter set stored in the first media sample of each CMAF fragment.

Continuity between a sequence of CMAF presentations is not specified by CMAF (typically a manifest defines sequences of CMAF presentations). However, seamless playback of sequenced CMAF presentations is possible with CMAF presentations that conform to the same single initialization constraints and signal in-band parameters in each CMAF fragment.

6.6.7 CMAF selection sets and late binding

A CMAF selection set is a set of CMAF switching sets, where each CMAF switching set encodes an alternative aspect of the same presentation over the same presentation time, for example, different audio languages, video camera angles, video formats, or codecs.

CMAF players are expected to select one CMAF switching set from each CMAF selection set at the start of playback based on player compatibility and user preferences. Users or playback applications may switch between CMAF switching sets in a CMAF selection set during playback, but seamless presentation is not expected, either because the content differs (e.g., a different language or camera view) or because CMAF fragment time alignment and decoding are not constrained to decode seamlessly.

Aligned CMAF switching sets are encodings of the same source content with time-aligned CMAF fragments to enable compatible players to seamlessly switch between tracks in different but aligned CMAF switching sets that contain CMAF tracks that do not conform to all the constraints of a single CMAF switching set.

For example, in Figure 11, the SD, HD, and UHD10 CMAF media profile CMAF switching sets could be aligned CMAF switching sets, and some players could seamlessly switch between them. In one case, only the encryption key might be different between the SD, HD and UHD CMAF switching sets. In another case, codecs could differ. Each CMAF switching set individually conforms to normal constraints for its media profile, and each player can determine if it can process the differences between aligned CMAF switching sets seamlessly.

Late binding allows CMAF tracks to be encoded once and used in different combinations. See 6.3 for more information.

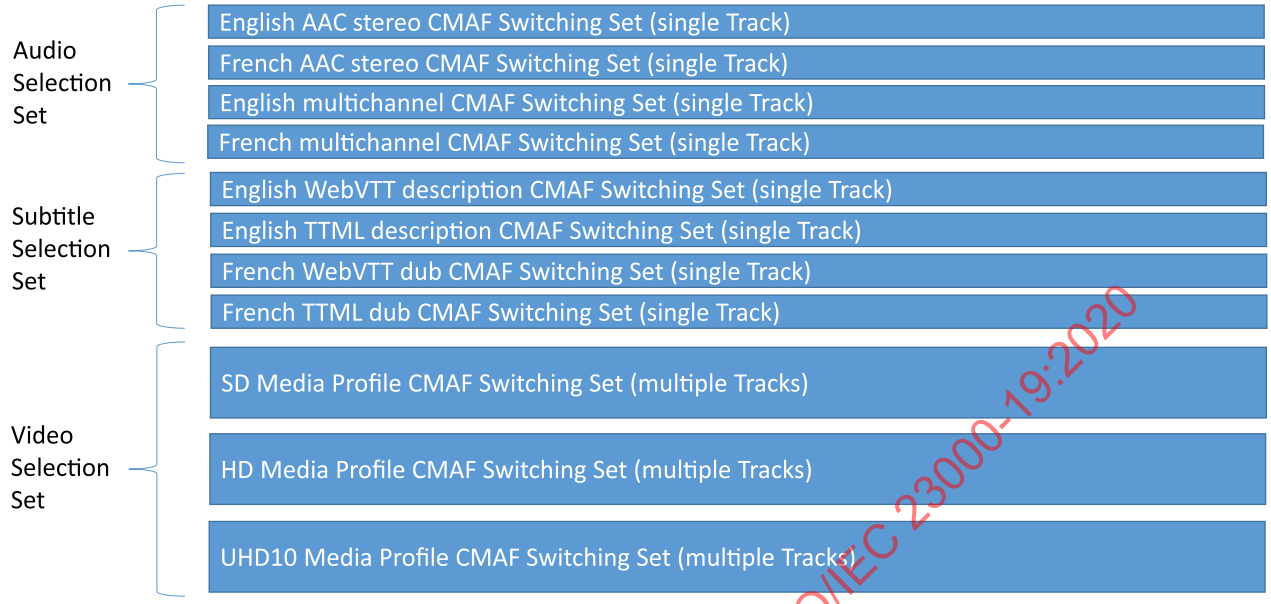


Figure 11 — Example of CMAF selection sets

The description of selection sets in a manifest is not defined by the CMAF specification; however, each manifest format is expected to provide sufficient information to enable players to select optimal CMAF switching sets from CMAF selection sets.

When there are multiple tracks in a selection set that differ by language, the language of each track needs to be identified to at least the precision needed to differentiate the tracks, using the language field and optionally the `ExtendedLanguageBox` as defined in ISO/IEC 14496-12.

When there is more than one track that is visually presented (e.g., video and subtitles), the visual layering of the tracks is indicated by the track header's layer field. CMAF recommends that subtitle CMAF tracks be positioned on layer –1 (ISO BMFF tracks with lower layer numbers overlay higher layer numbered tracks) and video CMAF tracks may be packaged with layer zero, or values relative to subtitles per the content creator's intended effect.

6.6.8 CMAF presentation timing model

There are multiple timelines involved in synchronizing a CMAF presentation. Each timeline has a timescale in units per second, increases at a continuous rate over time, and has an origin where the measure equals zero.

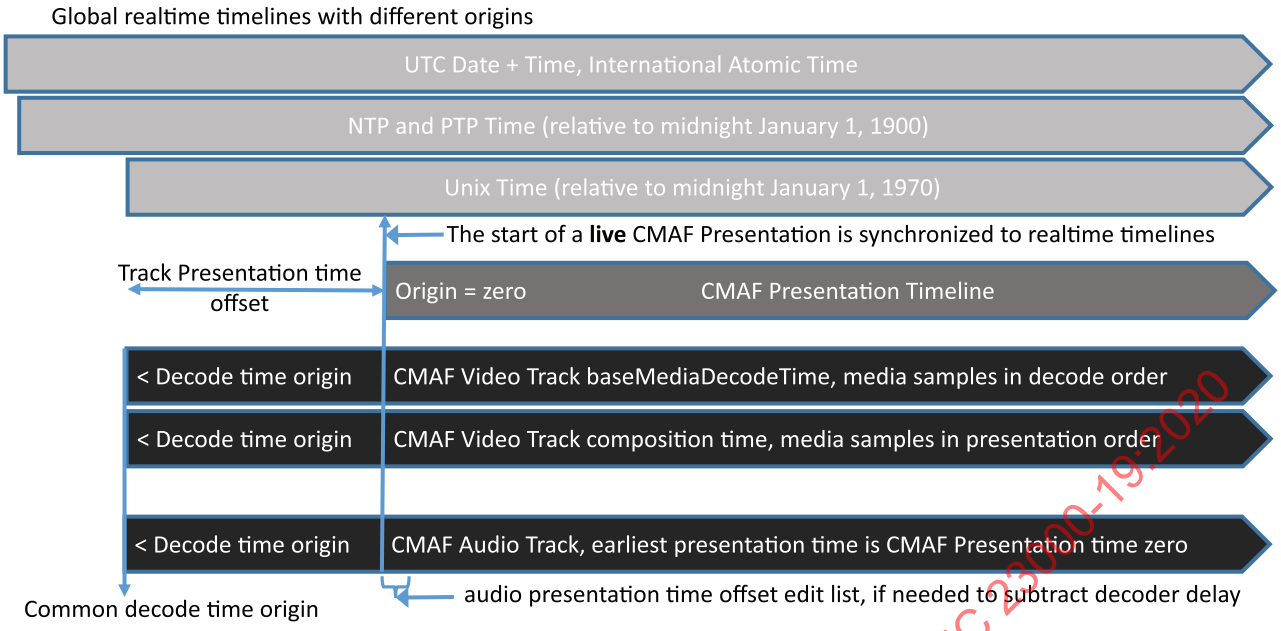


Figure 12 — CMAF timelines and synchronization model

All CMAF tracks in a CMAF presentation are specified in Clause 7 to measure `baseMediaDecodeTime` in the `TrackFragmentBaseMediaDecodeTimeBox` from the same timeline origin and adjust each media sample's presentation time relative to its decode time using composition offsets and offset edit lists, when necessary. The resulting presentation timelines and decode timelines all reference the same (coincident) timeline origin. The timelines and their relationships are illustrated in Figure 12.

The timelines are:

a) **Track decode time** — Is determined by the storage sequence and duration of each media sample, and for any media sample, equals the sum of all prior media sample durations added to the `baseMediaDecodeTime` of the first CMAF fragment.

b) **Track composition time** — Is determined by each sample's composition offset in the `TrackRunBox` relative to its decode time and is used to reorder video media samples to their presentation order.

c) **Track presentation time** — Is determined by applying any offset edit list in the CMAF header to each media sample's composition time, as shown for a CMAF audio track in Figure 12.

d) **CMAF presentation timeline** — Is defined in CMAF to start with a CMAF presentation time equal to zero, at the earliest video media sample presentation time, or the earliest audio media sample presentation time if there are no video CMAF tracks. The CMAF track presentation time at CMAF presentation time zero is its "presentation time offset", which can also be represented in some manifest formats, such as DASH.

e) **Wall clock time** — This timeline is important for live presentations, e.g., UTC time at the time of CMAF chunk availability. A wall clock timestamp may be stored in a `ProducerReferenceTimeBox` in a CMAF chunk or fragment, and that timestamp linked to a media sample decode time. Manifests can indicate the wall clock time that coincides with the start of a CMAF presentation so a player can determine when each CMAF fragment or chunk will be available for download.

CMAF tracks in a CMAF presentation are required to use the same decode timeline origin, similar to the shared movie timeline of multiple tracks in one ISO BMFF file. But CMAF tracks can start at a non-zero decode time (`baseMediaDecodeTime`) stored in the earliest video CMAF fragment. The CMAF presentation timeline starts at zero, coincident with the earliest media sample presentation time, and is equivalent to the ISO BMFF movie timeline.

As illustrated in Figure 13, CMAF tracks with audio CMAF fragments that overlap the earliest video media sample are intended to start presenting simultaneously with the earliest video media sample. Audio CMAF track presentation time is determined by the value of `baseMediaDecodeTime` recorded in each CMAF chunk or fragment, and any presentation time offset in an edit list, if an edit list is present in the CMAF track's CMAF header.

ISO BMFF defines files and requires that the decode time of each media sample is the sum of all prior media sample durations in that track in stored order. The first media sample in each track in a file therefore has a decode time of zero. Movie fragmentation is optional. In contrast, all media samples in CMAF tracks are stored in movie fragments, and the first CMAF fragment can have a non-zero `baseMediaDecodeTime` in the `TrackFragmentBaseMediaDecodeTimeBox`. The decode time of each media sample equals the sum of prior media sample durations in the CMAF track added to the `baseMediaDecodeTime` of the first CMAF fragment in the CMAF track. The decode time of each media sample also equals the sum of prior media sample durations in the CMAF fragment that contains it added to the CMAF fragment's `baseMediaDecodeTime`.

Manifests can specify a presentation time offset for each CMAF switching set to determine the CMAF track presentation time at the start of each CMAF presentation. The start of CMAF presentations is assumed to be the earliest video media sample when video is included in a CMAF presentation, otherwise the earliest audio media sample. To maintain audio, video, and subtitle synchronization encoded in the CMAF tracks, the presentation time offset of every CMAF switching set in a manifest will be equivalent, which means equal presentation time offsets, but possibly different integer values and timescales per CMAF switching set. Matching presentation time offsets plus CMAF presentation time in each CMAF track being presented maintains the encoded synchronization between audio, video, and text content on playback.

A manifest can select different start and end times with presentation time offsets and durations to present different timespans of the CMAF presentation timeline to play a portion of a CMAF presentation.

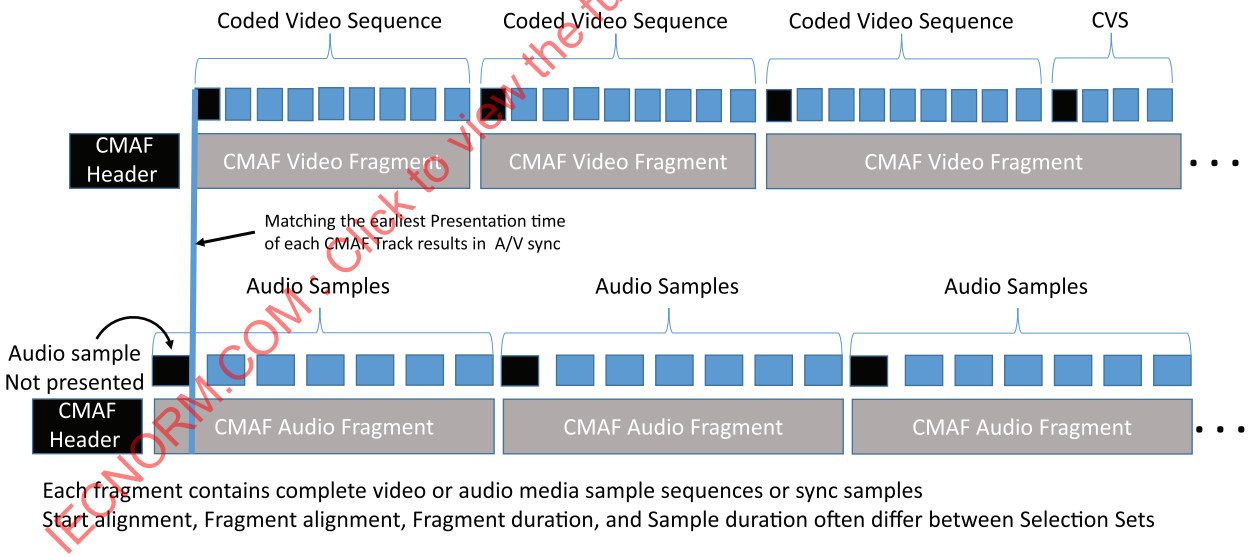


Figure 13 — Audio/video synchronization and start alignment

CMAF tracks containing video can use negative composition offsets where necessary to reorder media samples without adding composition delay, so that the earliest composition time and presentation time of each CMAF fragment equal its earliest decode time (with the optional exception of CMAF track files, mentioned above). The `baseMediaDecodeTime` in the `TrackFragmentBaseMediaDecodeTimeBox` is the earliest media sample presentation time of each CMAF fragment, and the `baseMediaDecodeTime` of the earliest video CMAF fragment is the presentation time offset of the CMAF track at CMAF presentation time zero.

CMAF tracks, illustrated in Figure 6, can also be encoded using positive composition offsets and packaged with an edit list in the CMAF header to remove composition delay.

During random access and "trick play" (fast forward, reverse, slow motion, jump, etc.), the `TrackFragmentBaseMediaDecodeTimeBox` and the `TrackRunBox` can be used to determine each media sample decode time relative to the CMAF fragment `baseMediaDecodeTime`, then any edit list present in the associated CMAF header can be used to calculate the presentation time offset relative to each media sample's composition time.

Decode time discontinuities between CMAF fragments can result from starting a new CMAF presentation in a playlist, for example, in sequenced programs or programs interspersed with prerecorded ads. A manifest can locate the new CMAF presentation on its manifest presentation timeline and adjust CMAF switching set presentation time offsets accordingly.

However, in cases where a decode time discontinuity results from one or more missing or damaged CMAF fragments in a continuous CMAF presentation, players can use the `baseMediaDecodeTime` of the next available CMAF fragment to resume presentation synchronized to the current presentation timeline. Error concealment could include repeating video frames and skipping over missing content in the case of on demand or buffered media, but all CMAF tracks have to skip the same duration to maintain A/V synchronization.

6.6.9 Manifest information

Although CMAF does not define the form or the content of the manifest, it does define its role. A manifest is a document that describes one or more CMAF presentations, e.g., an MPEG DASH Media Presentation Description (MPD). A manifest is responsible for describing the combination and synchronization of independently packaged CMAF tracks grouped in CMAF switching sets and selection sets to form a synchronized multimedia presentation.

A manifest provides the player with information to select, initialize, start align, and synchronize the CMAF track(s) to be played and identify CMAF media objects as resources to access and possibly download them. CMAF tracks and CMAF fragments contain sufficient information to enable decryption, decoding, and presentation scheduling. A manifest can also provide information on delivery protocol, network management, authorization, license acquisition, etc., in addition to resource identification and presentation description. The manifest can also signal that tracks conform to a CMAF media profile.

In cases where there are multiple CMAF tracks in a CMAF switching set or multiple CMAF presentations in a sequence, a manifest can indicate to a player that the CMAF switching set conforms to a CMAF media profile and CMAF switching set constraints, so a player can select and initialize compatible CMAF tracks and can switch seamlessly between those CMAF tracks, reinitializing only when necessary.

6.6.10 CMAF addressable media objects, resources, and resource identifiers

CMAF headers, CMAF chunks, and CMAF fragments can be packaged and referenced as CMAF addressable media objects for storage and delivery. Each CMAF addressable media object can be referenced as a resource as specified by an external specification, e.g., MPEG DASH (ISO/IEC 23009-1).

External specifications can define resource identifiers and the CMAF addressable media objects they identify so that servers, CDNs, and manifests can access CMAF addressable media objects and sequence them for decoding and presentation. A common example is a CMAF addressable media object packaged as the body of an HTTP response and identified as a resource by a unique segment URL. A CMAF track file can be segmented during delivery by adding a byte range to each HTTP request for the file's URL, using byte ranges matching one or more CMAF fragments or the CMAF header of the CMAF track file.

Use of the same URI per CMAF addressable media object is recommended to improve efficiency in content distribution networks and caches, even when different manifests or delivery protocols are used. The use of different URIs for the same CMAF addressable media object causes duplicate storage and delivery of the same CMAF media objects. For example, in cases where the same CMAF segments can be delivered over broadcast, multicast, and unicast, a CMAF player can request a CMAF segment with a single resource identifier and retrieve it from the player's cache regardless of which network path and encapsulation delivered it.

7 CMAF track format

7.1 Overview

The CMAF track format is derived from the ISO base media file format in this clause and structural brands are specified. At this point, the `'cmfc'` and the `'cmf2'` CMAF structural brands are defined. The `'cmf2'` brand further restricts the `'cmfc'` brand.

Several CMAF media objects are derived from the CMAF track format.

7.2 CMAF brands

Clause 7 defines the requirements and constraints that apply to all CMAF tracks.

A CMAF track should include all structural CMAF brands that it conforms to, indicating conformance with Clause 7 and Clause 8 (specifying file format and optional encryption, but not media encoding). The file brands also declare that boxes specified in common encryption (ISO/IEC 23001-7) and DASH (ISO/IEC 23009-1) may be present.

The common constraints of all structural file type brands are described in subclause 7.3.2.2.

Media profiles and their brands may specify additional box requirements and CMAF track constraints, such as the CMAF media profiles that reference the NAL structured video CMAF track format defined in Clause 9, incorporating boxes specified in ISO/IEC 14496-15, and AAC audio CMAF track format specified in Clause 10, incorporating boxes specified in ISO/IEC 14496-3.

CMAF also specifies addressable media objects in subclause 7.3.3 that are derived from CMAF tracks and have segment type brands that may be included in a `SegmentTypeBox` prepended to an addressable media object to identify its type (see Table 1).

Table 1 — CMAF brands

Brand	Location	Conformance requirements
`cmfc`	`FileTypeBox` and `SegmentTypeBox`	7.6
`cmf2`	`FileTypeBox` and `SegmentTypeBox`	7.7
`cmfs`	`SegmentTypeBox`	7.3.3.1
`cmfl`	`SegmentTypeBox`	7.3.3.2
`cmff`	`SegmentTypeBox`	7.3.2.3

An ISO brand, such as `'iso9'`, that indicates the ISO BMFF boxes actually present in each CMAF track should be listed in `compatible_brands` to improve interoperability with file readers. The `'isoX'` brand (where "X" represents a number or letter) specified in each new edition of ISO/IEC 14496-12 determines what box versions are allowed in CMAF tracks, unless specifically constrained by the common constraints of CMAF structural brands as specified in this clause.

If any of the structural CMAF brands is the `major_brand`, the `minor_version` shall be set to 0, and file names should use the file extensions in Table 2. Otherwise, file names should use the file extension and Internet Media Type specified to match the major brand, e.g., `*.mp4`, `*.3gp`, `*.uvv` and `*.uva`.

NOTE It is expected that file readers read possible future versions of CMAF that increment the `minor_version` number.

Table 2 — Common media application format file extensions

Track type	File extension	Internet media type (MIME type)
Video	`.cmfv`	`video/mp4`
Audio	`.cmfa`	`audio/mp4`
Text (subtitle)	`.cmft`	`application/mp4`

Each CMAF track has a CMAF header associated with it, although the CMAF header and CMAF fragments might not be stored as an ISO BMFF file, and might not be stored at all, i.e., only existing temporarily as streams.

7.3 CMAF media objects

7.3.1 CMAF boxes

CMAF tracks shall include the following ISO BMFF boxes with nesting, optionality and cardinality specified in Table 3 through Table 5, with box definitions incorporated by reference to the clauses listed in the "Specification" column that reference ISO/IEC 14496-12, ISO/IEC 23001-7, or ISO/IEC 23009-1.

Some boxes are additionally constrained by CMAF, as specified by clauses in this document referenced in the "Constraints" column of Table 3 through Table 5. The normative references are dated, e.g., the 2015 edition of ISO/IEC 14496-12, so the box versions and features defined in that edition are valid in a CMAF track unless additionally constrained by CMAF or by the `'isoX'` brand promised in the `FileTypeBox`.

CMAF addressable media objects are derived from the CMAF track, CMAF header, CMAF chunk, and CMAF fragment recorded media objects and specify additional box constraints in subclause 7.3.3.

Additional boxes and constraints are specified by CMAF track formats and CMAF media profiles for specific audio, video, and subtitle formats (see subclause 7.4.4).

Legend for Table 3, Table 4, and Table 5

The "Format Req." column indicates the number of boxes that are required to be present in a CMAF track, where:

- "*" means "zero or more" may be present;
- "+" means "one or more" shall be present;
- "0/1" indicates only one box may be present, and it is optional;
- "1" indicates one box shall be present;
- "CR" indicates that a box is conditionally required as specified in the CMAF track format or a specific media profile and shall be present under those conditions (see "Constraints" column).

Table 3 — CMAF header boxes

NL 0	NL 1	NL 2	NL 3	NL 4	NL 5	Format Req.	Specification	Constraints	Description
`ftyp`						1	ISO/IEC 14496-12	CMAF 7.2	File type and Compatibility
`moov`						1	ISO/IEC 14496-12		Container for functional metadata
	`mvhd`					1	ISO/IEC 14496-12	CMAF 7.5.1	Movie header
	`trak`					1	ISO/IEC 14496-12		Container for each track
		`tkhd`				1	ISO/IEC 14496-12	CMAF 7.5.4	Track header
		`edts`				CR	ISO/IEC 14496-12	CMAF 7.5.13	Edit box
			`elst`			1	ISO/IEC 14496-12	CMAF 7.5.13	Edit list box

Table 3 *(continued)*

NL 0	NL 1	NL 2	NL 3	NL 4	NL 5	Format Req.	Specification	Constraints	Description
		mdia				1	ISO/IEC 14496-12		Track media information
			mdhd			1	ISO/IEC 14496-12	CMAF 7.5.5	Media header
			hdlr			1	ISO/IEC 14496-12		Declares the media handler type
			elng			0/1	ISO/IEC 14496-12		Extended language tag
			minf			1	ISO/IEC 14496-12		Media information container
				vmhd		CR	ISO/IEC 14496-12	CMAF 7.5.6	Video media header
				smhd		CR	ISO/IEC 14496-12	CMAF 7.5.7	Sound media header
				sthd		CR	ISO/IEC 14496-12	CMAF 7.5.8	Subtitle media header
				dinf		1	ISO/IEC 14496-12		Data information box
					dref	1	ISO/IEC 14496-12	CMAF 7.5.9	Data reference box, declares source of media data in track
				stbl		1	ISO/IEC 14496-12	CMAF 7.5.12	Sample table box, container for the time/space map
					stsd	1	ISO/IEC 14496-12	CMAF 7.5.10	Sample descriptions (see Table 4 for additional details)
					stts	1	ISO/IEC 14496-12	CMAF 7.5.12	Decoding, time to sample
					stsc	1	ISO/IEC 14496-12	CMAF 7.5.12	Sample-to-chunk
					stsz/ stz2	1	ISO/IEC 14496-12	CMAF 7.5.12	Sample size box
					stco	1	ISO/IEC 14496-12	CMAF 7.5.12	Chunk offset
					sgpd	CR	ISO/IEC 14496-12	CMAF 7.5.18	Sample group description box
					stss	CR	ISO/IEC 14496-12	CMAF 7.5.12	Sync sample box
		udta				0/1	ISO/IEC 14496-12		User data box
			cprt			*	ISO/IEC 14496-12		Copyright box
			kind			*	ISO/IEC 14496-12	CMAF 7.5.3	Track kind box
	mvex					1	ISO/IEC 14496-12		Movie extends box
		mehd				0/1	ISO/IEC 14496-12	CMAF 7.5.1	Movie extends header
		trex				1	ISO/IEC 14496-12	CMAF 7.5.14	Track extends box
	pssh					*	ISO/IEC 23001-7	CMAF 7.4.3	Protection system specific header box

Table 4 — CMAF header protected sample entry boxes

NL 5	NL 6	NL 7	NL 8	Format Req.	Specification	Constraints	Description
stsd				1	ISO/IEC 14496-12		Sample description box
	sinf			CR	ISO/IEC 14496-12	CMAF 7.5.11	Protection scheme information box
		frma		1	ISO/IEC 14496-12		Original format box
		schm		1	ISO/IEC 14496-12		Scheme type box
		schi		1	ISO/IEC 14496-12		Scheme information box
			tenc	1	ISO/IEC 23001-7	CMAF 7.4.1	Track encryption box
NOTE Table 4 is a continuation of Table 3 showing nesting levels 5 to 8 separately to reduce table width.							

Table 5 — CMAF chunk, CMAF fragment, CMAF segment, and CMAF track file boxes

NL 0	NL 1	NL 2	NL 3	NL 4	NL 5	Format Req.	Specification	Constraints	Description
styp						0/1	ISO/IEC 14496-12		Segment type
prft						0/1	ISO/IEC 14496-12		Producer reference time
emsg						*	ISO/IEC 23009-1	CMAF 7.4.5	Event message
moof						+	ISO/IEC 14496-12		Movie fragment
	mfhd					1	ISO/IEC 14496-12		Movie fragment header
	traf					1	ISO/IEC 14496-12		Track fragment
		tfhd				1	ISO/IEC 14496-12	CMAF 7.5.16	Track fragment header
		tfdt				1	ISO/IEC 14496-12		Track fragment base media decode time
		trun				1	ISO/IEC 14496-12	CMAF 7.5.17	Track fragment run box
		senc				0/1	ISO/IEC 23001-7	CMAF 8.2.2.1	Sample encryption box
		saio				CR	ISO/IEC 14496-12	CMAF 8.2.2.1	Sample auxiliary information offsets box
		saiz				CR	ISO/IEC 14496-12	CMAF 8.2.2.1	Sample auxiliary information sizes box
		sbgp				*	ISO/IEC 14496-12		Sample to group box
		sgpd				*	ISO/IEC 14496-12	CMAF 7.5.17	Sample group description box
		subs				CR	ISO/IEC 14496-12	CMAF 7.5.19	Sub-sample information box
mdat						+	ISO/IEC 14496-12	CMAF 7.5.18	Media data container for media samples

7.3.2 CMAF track media objects

7.3.2.1 CMAF header

A CMAF header as defined in 3.1.2 conforms to the following constraints.

a) A CMAF header shall contain the set of boxes in Table 3 and Table 4 with the conditions and optionality listed.

b) Each CMAF header shall form a valid CMAF track, as specified in subclause 7.3.2.2, when followed by a continuous sequence of associated CMAF fragments in decode order.

c) A CMAF header shall be conformant with ISO/IEC 14496-12 and the following additional constraints and requirements.

1) The CMAF header shall start with a `FileTypeBox`.

2) The CMAF header shall include exactly one `MovieBox`.

3) The `MovieBox` shall start with a `MovieHeaderBox`, as constrained in subclause 7.5.1.

4) The `MovieBox` shall contain exactly one track containing media data as specified in subclause 7.3.2.2.

NOTE Timed metadata tracks can be provided as separate CMAF tracks in a separate selection set.

5) The `MovieBox` shall contain a `MovieExtendsBox`, as defined in ISO/IEC 14496-12, to indicate that the file contains `MovieFragmentBox`es.

6) The `MovieExtendsBox` may contain a `MovieExtendsHeaderBox`, as defined in ISO/IEC 14496-12, and if so, shall provide the overall duration of the CMAF track. If the duration is unknown, this box shall be omitted.

d) If the CMAF header requires sample entries with a decoder configuration record, as specified by a CMAF media profile, with decoding and rendering parameter limits, such as codec profile, level, image height, width, etc., then the first sample entry shall contain parameters equal or greater than the maximum values contained in all CMAF fragments in the track to enable a single initialization of media decoding and rendering systems to render all CMAF fragments in the CMAF track.

7.3.2.2 CMAF track

Table 6 — CMAF track structure

NL 0	Format Req.	Specification	CMAF constraints	Description
CMAF Header	1		CMAF 7.3.2.1	CMAF header
CMAF Fragment	1+		CMAF 7.3.2.4	CMAF fragment

A CMAF track as defined in subclause 3.2.1 has the structure shown in Table 6 and conforms to the following constraints.

a) A CMAF track shall conform to at least one structural CMAF brand and contain the set of boxes in Table 3, Table 4, and Table 5, with the conditions and optionality listed.

b) The concatenation of a CMAF header and all CMAF fragments in the CMAF track in consecutive decode order shall be a valid fragmented ISO BMFF file, with the exception that the first CMAF fragment in a CMAF track may have a non-zero `baseMediaDecodeTime`.

c) Each CMAF fragment in a CMAF track shall have `baseMediaDecodeTime` equal to the sum of all prior CMAF fragment durations added to the first fragment's `baseMediaDecodeTime`. A CMAF fragment duration is the sum of the media sample durations, documented in the `TrackFragmentRunBox` in the `MovieFragmentHeaderBox`.

NOTE Valid CMAF tracks do not have media time discontinuities resulting from missing media samples or fragments. Gaps in decode time can result in audio-video synchronization errors. For recommendations on handling missing media samples and missing CMAF fragments, see Annex F.

d) Each CMAF track contains a single ISO BMFF track and `TrackBox`, as determined by CMAF header constraints specified in subclause 7.3.2.1.

Additional constraints specific for encryption, video, audio, and subtitle CMAF tracks are specified in Clauses 8, 9, 10 and 11, respectively, and are derived from the general CMAF track format specified in this clause.

7.3.2.3 CMAF chunk

CMAF chunks as defined in subclause 3.1.4 are the smallest CMAF media object that can be encoded, and they can be referenced as addressable media objects.

Table 7 — CMAF chunk structure

NL 0	Format Req.	Specification	CMAF constraints	Description
styp	0/1	ISO/IEC 14496-12	CMAF 7.2	Segment type signalling compatibility to CMAF chunk
prft	0/1	ISO/IEC 14496-12		Producer reference time
emsg	*	ISO/IEC 23009-1	CMAF 7.4.5	Event message
moof	1	ISO/IEC 14496-12	CMAF Table 5	Movie fragment box and the boxes it contains
mdat	1	ISO/IEC 14496-12	CMAF 7.5.19	Media data container for media samples

Each CMAF chunk conforms to the following constraints:

a) A CMAF chunk shall contain the boxes in Table 7 with the conditions and optionality indicated.

b) The `MovieFragmentBox` shall conform to the constraints of the structural CMAF brand constraints specified in Table 5, such as containing only one `TrackFragmentBox` that contains only one `TrackRunBox`.

NOTE 1 Since there is only one `TrackRunBox` per `MovieFragmentBox`, all media samples of a CMAF chunk are located in a single track run in one `MediaDataBox`.

c) The `MediaDataBox` shall contain all media samples referenced by the `TrackRunBox` and should immediately follow the `TrackRunBox` in byte order.

d) All media samples in a CMAF chunk shall be addressed by byte offsets in the `TrackRunBox` that are relative to the first byte of the `MovieFragmentBox`. See ISO/IEC 14496-12.

e) A CMAF chunk shall contain a consecutive subset of the media samples of a CMAF fragment.

f) A CMAF chunk shall contain a `TrackFragmentDecodeTimeBox` containing the `baseMediaDecodeTime` of the first media sample.

g) A sequence of CMAF chunks that spans the decode time and duration of a CMAF fragment containing those media samples shall contain all the media samples in the CMAF fragment, without duplicates, i.e., CMAF chunks in a CMAF track shall not overlap or have gaps in decode time.

h) The decode time and CMAF track presentation times of a media sample in a CMAF track shall not change whether the sample is contained in a CMAF chunk or a CMAF fragment.

NOTE 2 Because the CMAF fragment and the first CMAF chunk both start with the same media sample, they both have the same `baseMediaDecodeTime`.

i) A CMAF chunk may include a `SegmentTypeBox` preceding the `MovieFragmentBox` and may be referenced as a CMAF addressable media object. The `SegmentTypeBox` should include the CMAF

chunk brand 'cmfl' and may include any valid compatible_brands listed in the FileTypeBox of the CMAF track's CMAF header.

j) CMAF chunks may be referenced as CMAF addressable media objects, and optional file level boxes may be prepended when they are formatted as CMAF addressable media objects.

NOTE 3 Preceding boxes indicated in Table 7 can be prepended to CMAF chunks at time of encoding or prepended to CMAF chunks and CMAF segments at time of delivery. It is expected that players will process the boxes prepended to the start of each CMAF addressable media object, but they will ignore these boxes when located within a CMAF addressable media object, i.e., not prepended to the first CMAF chunk in a CMAF segment. The SegmentTypeBox is specified by ISO/IEC 14496-12 to be the first box of an ISO BMFF segment and can be ignored in other locations.

NOTE 4 A SegmentTypeBox prepended to CMAF chunks or fragments at the time of encoding can be useful as a media object delimiter and identifier. It allows indexing and packaging operations to rapidly locate and identify CMAF chunks and fragments within a CMAF track without deep inspection of elementary streams or other clues to identify CMAF fragment boundaries and preceding boxes intended to be part of each media object.

7.3.2.4 CMAF fragment

A CMAF fragment as defined in subclause 3.1.1 contains one or more CMAF chunks, therefore CMAF chunks may be present in CMAF addressable media objects, including CMAF segments and CMAF track files, and a CMAF chunk may be referenced as a CMAF addressable media object (see subclause 7.3.3.2).

Table 8 — CMAF fragment structure

NL 0	Format Req.	Specification	CMAF constraints	Description
styp	0/1	ISO/IEC 14496-12	CMAF 7.2	Segment type Signalling compatibility to CMAF fragment
prft	0/1	ISO/IEC 14496-12		Producer reference time
emsg	*	ISO/IEC 23009-1	CMAF 7.4.5	Event message
CMAF Chunk	1+	CMAF 7.3.2.3		CMAF chunk

Each CMAF fragment conforms to the following constraints:

a) A CMAF fragment shall contain the boxes in Table 8, with the conditions and optionality indicated.

b) Each CMAF fragment, in combination with its associated CMAF header and optional decryption key(s), shall contain sufficient metadata to enable the CMAF fragment to be decoded, decrypted, and displayed when it is independently accessed.

NOTE 1 For example, if sample groups and sample group descriptions are used to signal encryption key changes, then a SampleGroupDescriptionBox and SampleToGroupBox probably need to be present in the TrackFragmentBox to make the CMAF fragment randomly accessible and decryptable.

c) The first CMAF chunk in a CMAF fragment may be preceded by other boxes as long as the CMAF fragment remains conformant, including one SegmentTypeBox and/or one or more ProducerReferenceTimeBox(es) and/or DASHEventMessageBox(es). (See subclause 7.4.5 and Annex E for more information on Event Messages).

d) A CMAF fragment with a SegmentTypeBox preceding the first MovieFragmentBox should include the CMAF fragment brand 'cmff' and may include any compatible_brands listed in the FileTypeBox of the CMAF track's CMAF header.

e) A CMAF fragment can be referenced as a CMAF segment and may have a SegmentTypeBox prepended that includes the CMAF segment brand 'cmfs' to identify the start of the CMAF segment.

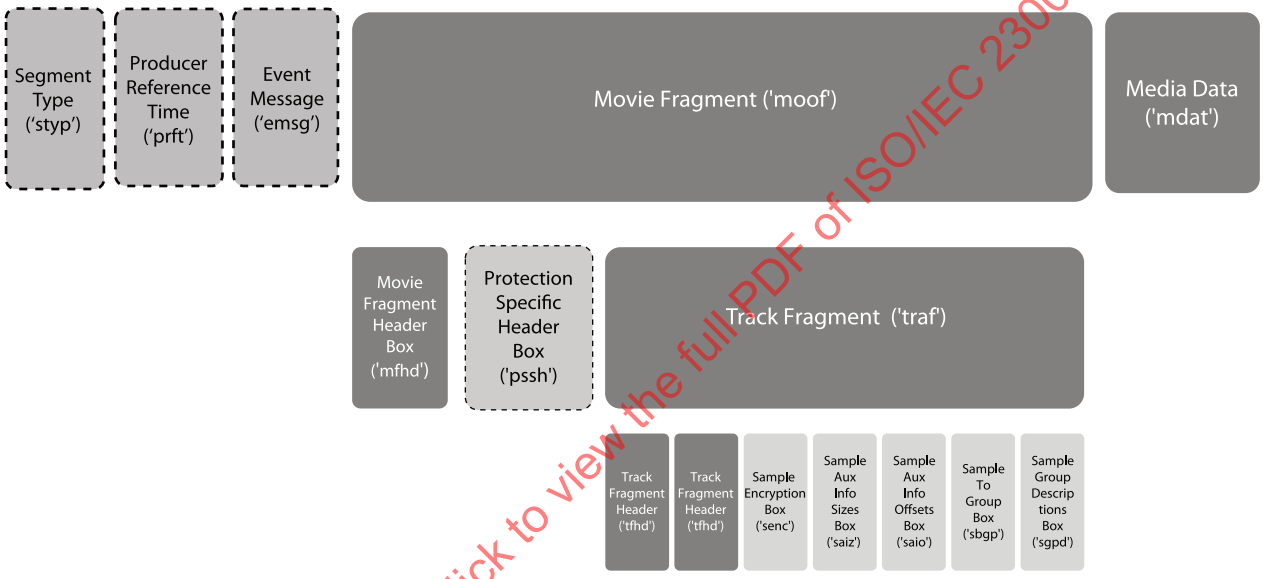
NOTE 2 It is possible that the first CMAF chunk in the first CMAF fragment in a CMAF segment would be preceded by a `SegmentTypeBox` that includes the `compatible_brands` `'cmfl'`, `'cmff'`, and `'cmfs'` to indicate that it is the start of a CMAF segment, fragment, and chunk. The start of a CMAF fragment that contains multiple CMAF chunks can be identified by including both the `'cmff'` and `'cmfl'` brands, whereas subsequent CMAF chunks in the CMAF fragment would only include the `'cmfl'` brand.

NOTE 3 As noted in subclause 7.3.2.3, only a `SegmentTypeBox` at the start of a CMAF addressable media object is expected to be processed by players, whereas a `SegmentTypeBox` located within a CMAF addressable media object is expected to be ignored by players.

f) Each CMAF fragment in a CMAF track should have a duration of at least one second, with the possible exception of the first and last CMAF fragments of the CMAF track.

NOTE 4 Additional constraints specific to video, audio, and subtitle media types are specified in Clauses 9, 10, Annex A and Annex B.

Figure 14 illustrates the box sequence (left to right) and containment (top to bottom) of a CMAF fragment.



NOTE 5 Lower boxes indicate containment in the box above. The sequence of boxes contained in the `TrackFragmentBox` (`'traf'`) is only recommended. Light shaded boxes are always optional. Light boxes in the bottom row are required or conditionally required when encryption is used. This example contains one CMAF chunk (`'moof'`, `'mdat'`) in the CMAF fragment.

Figure 14 — CMAF fragment box sequence and containment

7.3.3 CMAF addressable media objects

7.3.3.1 CMAF segment

Table 9 — CMAF segment structure

NL 0	Format Req.	Specification	CMAF con-straints	Description
styp	0/1	ISO/IEC 14496-12		Segment type Signalling compatibility to CMAF segment

Table 9 *(continued)*

NL 0	Format Req.	Specification	CMAF con-straints	Description
prft	0/1	ISO/IEC 14496-12		Producer reference time
emsg	*	ISO/IEC 23009-1	CMAF 7.4.5	Event message
CMAF Fragment	1+	CMAF 7.3.2.4		CMAF fragment

Each CMAF segment as defined in subclause 3.1.5 conforms to the following constraints.

a) A CMAF segment shall include the boxes in Table 9, with the conditions and optionality listed.

b) A CMAF segment shall contain one or more complete and consecutive CMAF fragments in decode order.

c) A CMAF segment may include a `ProducerReferenceTimeBox` and/or `DASHEventMessageBox`(es) preceding the first `MovieFragmentBox`.

d) A CMAF segment may include a `SegmentTypeBox` preceding all other boxes in the CMAF segment. The `SegmentTypeBox` should include the CMAF segment brand `'cmfs'` and may include any `compatible_brands` listed in the `FileTypeBox` of the CMAF track's CMAF header.

NOTE Preceding boxes can be prepended to CMAF chunks or CMAF fragments at time of encoding or prepended to CMAF segments at time of delivery. It is expected that players will process the boxes prepended to a CMAF addressable media object, but they will ignore these boxes when located within a CMAF addressable media object, i.e., not prepended to the first CMAF chunk in a CMAF segment. The `SegmentTypeBox` is specified by ISO/IEC 14496-12 to be the first box of an ISO BMFF segment and can be ignored in other locations.

7.3.3.2 CMAF chunk

A CMAF chunk is specified as a data structure within a CMAF fragment in subclause 7.3.2.3, but a CMAF chunk may also be referenced as a CMAF addressable media object, e.g., for low latency live streaming.

There is no direct restriction on the SAP type of the first media sample in a CMAF chunk, except by inheritance of CMAF fragment media sample constraints specified by a CMAF media profile. Because video CMAF media profiles typically constrain the initial media sample type to be SAP 1 or 2, the first CMAF chunk in a CMAF fragment inherits that media sample constraint.

Adaptive switching in players is normally limited to the earliest CMAF chunk in a CMAF fragment. Prepending a `SegmentTypeBox` to each CMAF chunk with appropriate brands can indicate the start of a CMAF chunk that is also the start of a CMAF fragment.

A `SegmentTypeBox`, `ProducerReferenceTimeBox`, and/or `DASHEventMessageBox(es)` may be prepended to a CMAF chunk at time of encoding or time of delivery. In a CMAF addressable media object, only when these boxes precede the first `MovieFragmentBox` are they expected to be processed by a player.

7.3.3.3 CMAF track file

A CMAF track as defined in subclause 3.1.6 is a logical collection of CMAF fragments and a CMAF header that might never be stored or stored in that form. In contrast, a CMAF track file is a single file that contains one CMAF track, stored in sequence, and is therefore addressable as a file. Since CMAF chunks and CMAF fragments can be preceded by a `SegmentTypeBox`, CMAF media objects may be identified by `SegmentTypeBox` within a CMAF track. CMAF media objects may be indexed in a `SegmentIndexBox` to enable addressing CMAF media objects in a CMAF track file by byte range.

Table 10 — CMAF track file structure

NL 0	Format Req.	Specification	CMAF constraints	Description
CMAF Header	1	CMAF 7.3.2.1	CMAF 7.2	CMAF header
sidx	0/1	ISO/IEC 14496-12		Segment index
CMAF Fragment	1+	CMAF 7.3.2.4	CMAF 7.3.2.4	CMAF fragments

Each CMAF track file conforms to the following constraints:

a) A CMAF track file shall include the boxes in Table 10, with the conditions and optionality listed.

b) Additional boxes, such as `SegmentIndexBox(es)`, may be present between the CMAF header and the first CMAF fragment.

c) If `SegmentIndexBoxes` exist, each subsegment referenced in the `SegmentIndexBox` shall be a single CMAF fragment contained in the CMAF track file.

d) A video CMAF track may use an offset edit list as specified in subclauses 7.5.13 and 10.2.6.

e) A video CMAF track may use v1 `TrackRunBoxes` using negative composition offsets to adjust the composition time of the earliest presented video media sample in each CMAF fragment to its `baseMediaDecodeTime` and the earliest video media sample in the CMAF track file to zero, without using an offset edit list.

7.3.4 CMAF switching sets

7.3.4.1 General constraints on a CMAF switching set

These general CMAF switching set constraints, plus specific constraints defined by a CMAF media profile that the CMAF switching set conforms to, are intended to enable seamless adaptive switching by typical players.

Each CMAF switching set as defined in subclause 3.2.2 conforms to the following constraints.

a) A CMAF switching set shall contain one or more CMAF tracks, each of which is a synchronized encoding of the same source content.

b) A CMAF switching set shall contain CMAF tracks of only one media type, i.e., audio or video or subtitles.

c) All CMAF tracks in a CMAF switching set shall have the same duration (computed from the media sample durations, optionally stored in a `MovieExtendsHeaderBox`).

d) All CMAF tracks in a CMAF switching set shall contain the same number of CMAF fragments.

e) For any CMAF fragment in one CMAF track in a CMAF switching set, there shall be a CMAF fragment with the same decode time in all other CMAF tracks.

f) All CMAF tracks in a CMAF switching set shall have the same value of `baseMediaDecodeTime` in the first CMAF fragment's `TrackFragmentBaseMediaDecodeTimeBox`, measured from the same timeline origin.

g) The presentation time of the earliest media sample of the earliest CMAF fragment in each CMAF track shall be equal.

h) Alternative CMAF fragments in a CMAF switching set with the same `baseMediaDecodeTime` shall contain alternative encodings of the same source media samples.

NOTE 1 These constraints do not prohibit different frame rates in different CMAF tracks in a CMAF switching set. However, these constraints restrict the combinable frame rates to those that result in identical CMAF fragment durations for alternative CMAF fragments.

i) All CMAF tracks in a CMAF switching set shall conform to one CMAF media profile. However, CMAF tracks included in a CMAF switching set may contain different CMAF media profile brands that all conform to one CMAF media profile. See subclause 12.1 for additional CMAF switching set constraints specified by CMAF media profiles.

NOTE 2 CMAF media profiles specify the maximum limits for encoding parameters, so subset or lower media profiles that can be decoded by a higher media profile decoder conform to the higher media profile.

j) CMAF header parameters shall not differ between CMAF tracks, except as allowed in Table 11. CMAF media profiles may specify additional constraints, as noted in the table.

k) Additional CMAF header boxes specified in Table 3 and Table 4 that are not listed in Table 11 may contain different boxes and information.

Table 11 — General constraints on CMAF header boxes in CMAF switching sets

Box	General CMAF header constraints in a CMAF switching set
`ftyp`	*Shall be identical except for media profile brands* [see i) in 7.3.4.1]
`mvhd`	*Shall be identical except for* `creation_time` and `modification_time`
`tkhd`	*Shall be identical except for* `width`, `height`, `creation_time`, and `modification_time`. See NOTE 1.
`trex`	*Identical*
`elst`	*Shall be identical except for video CMAF track files with a different composition offset*
`mdhd`	*Shall be identical except for* `creation_time` and `modification_time`
`mehd`	*Identical*
`meta`	*May contain different boxes and data*
`udta`	*May contain different boxes and data*
`cprt`	*Identical*
`kind`	*Identical*
`elng`	*Identical*
`hdlr`	*Identical*
`vmhd`	*Identical*
`smhd`	*Identical*
`sthd`	*Identical*
`dref`	*Identical*
`stsd`	*Sample entries shall have the same* `codingname` *(four-character code) and conform to other CMAF track format and media profile specified constraints.* See NOTE 2.
`sgpd`	*May differ*
`pssh`	*Identical*
`sinf`	*Identical*
`schi`	*Identical*
`schm`	*Identical*
`frma`	*Identical*
`tenc`	*Shall be identical except for initialization vector values*

NOTE 1 Track width and height can differ, but picture aspect ratio is the same for all CMAF tracks.

NOTE 2 Sample entry constraints for CMAF switching sets are defined by each CMAF media profile, its specified or referenced ISO BMFF track format, and its CMAF switching set constraints. A sample description can contain multiple sample entries.

There are two functional types of CMAF switching set initialization constraints:

1) Multiple initialization — CMAF switching sets conforming to the general constraints above may rely on decoding parameters stored in each CMAF track's header to decode a CMAF fragment from

that CMAF track. This implies that a player needs to access and process the CMAF header for each CMAF track it switches to.

2) Single initialization — If CMAF headers are constrained to be equivalent within a CMAF switching set, or CMAF tracks only change parameters that are stored in each CMAF fragment, then a player only needs to access one CMAF header and each CMAF fragment to find all the necessary decoding and rendering parameters. A player only needs to initialize the CMAF switching set once, then it can continue decoding CMAF fragments from any CMAF track in the CMAF switching set.

The two types of CMAF switching set constraints and processing are illustrated in Figure 10.

7.3.4.2 CMAF switching set single initialization constraints

Encoding of CMAF tracks in a CMAF switching set can be further constrained to allow one-time initialization of the CMAF switching set using a single CMAF header (see Figure 10). Each CMAF media profile or the CMAF track format it references can specify single initialization CMAF switching set constraints.

In general, single initialization constraints can be specified by:

— a “common CMAF header” in all CMAF tracks containing all decoding, decryption, and rendering parameters required by any CMAF fragment in the CMAF switching set;

— “inband parameter sets” stored and referenced in each CMAF fragment, where CMAF headers are only used to initialize decoding and display with the appropriate codec, profile, level, resolution, etc. This is typically the case for audio and can be specified for inband video parameter media sample formats such as `'avc3'` sample entry;

— a combination of both CMAF header and CMAF fragment stored parameters, but with the constraint that the combination of any CMAF header in the CMAF switching set and CMAF fragment will contain the necessary decoding parameters.

NOTE Enabling changes in inband parameters over time is useful for live streaming, where content from different encodings can be sequenced in a CMAF presentation, such as programs and ads in a live “channel”, or when a live video contribution feed is received containing inband parameters that change and are not known in advance when a CMAF header is packaged and delivered to a player.

7.3.4.3 CMAF switching set single initialization constraints identifier

The identifier `"urn:mpeg:cmaf:siss"` may be used to signal CMAF switching set conformance to single initialization constraints.

7.3.4.4 Aligned CMAF switching set constraints

Aligned CMAF switching sets are defined in subclause 3.2.3 to be a “set of CMAF switching sets, the CMAF tracks of which all contain alternative encodings of the same source content in time-aligned CMAF fragments, but all CMAF tracks do not conform to a single CMAF switching set”.

Aligned CMAF switching sets conform to the following constraints:

a) An aligned CMAF switching set shall contain two or more CMAF switching sets.

b) Aligned CMAF switching sets shall contain CMAF switching sets that are encoded from the same media stream.

c) Aligned CMAF switching sets shall contain CMAF switching sets of equal duration.

d) Aligned CMAF switching sets shall contain the same number of CMAF fragments in every CMAF track.

e) Aligned CMAF switching sets shall contain CMAF fragments in every CMAF track with matching `baseMediaDecodeTime` and duration, for each CMAF fragment in a CMAF track.

NOTE Some CMAF players can seamlessly switch between CMAF tracks in different CMAF switching sets that are aligned, such as CMAF switching sets with different encryption keys and their associated DRM license(s).

7.3.5 CMAF selection sets

CMAF selection sets as defined in subclause 3.2.4 conform to the following constraints.

a) A CMAF selection set shall contain one or more CMAF switching sets.

b) All CMAF switching sets within a CMAF selection set shall be of the same media type, e.g., audio, video, or subtitles.

c) All switching sets within a CMAF selection set shall be of the same duration, within a tolerance of the longest CMAF fragment duration of any CMAF track in the selection set.

d) Different CMAF switching sets that encode the same content with different codecs or video formats may be contained within a CMAF selection set to enable a player to select the most compatible codec or video format and the CMAF switching set that contains it.

e) A CMAF selection set may contain multiple aligned CMAF switching sets, which encode the same content in different CMAF switching sets that have time-aligned CMAF fragments, but with different keys, codecs, etc.

f) One CMAF track should be presented from each selection set in a presentation, e.g., one audio, one video, and one subtitle selection set.

7.3.6 CMAF presentations

A CMAF presentation as defined in subclause 3.2.5 conforms to the following constraints.

a) All CMAF tracks in a CMAF presentation shall have the same timeline origin.

b) Different CMAF switching sets may use different CMAF track `timescale` values. In that case, `baseMediaDecodeTime` integer values will be different even if the corresponding ISO BMFF samples would have the same presentation time.

c) All CMAF tracks in a CMAF presentation containing a video switching set shall be start aligned with CMAF presentation time zero equal to the earliest video media sample presentation start time in the earliest CMAF fragment (see subclause 6.6.8).

d) All CMAF tracks in a CMAF presentation that does not contain video shall be start aligned with the CMAF presentation time zero equal to the earliest audio media sample presentation start time in the earliest CMAF fragment.

e) Media samples in audio and subtitle CMAF tracks with earlier presentation times than CMAF presentation time zero should not be presented.

f) Media samples in audio and subtitle CMAF tracks whose durations overlap the earliest video media sample presentation time should be partially presented, starting at the earliest video media sample presentation time.

g) The duration of a CMAF presentation shall be the duration of its longest CMAF track.

h) CMAF tracks in a CMAF presentation shall equal the CMAF presentation duration, within a tolerance of the longest video CMAF fragment duration.

NOTE The CMAF hypothetical application model synchronizes media samples by presentation times relative to a common timeline origin. CMAF track presentation time is determined by each CMAF fragment's stored `baseMediaDecodeTime` and any composition or edit list offsets present. Synchronization during late binding does not rely on externally stored presentation time offsets. This means that audio, video and subtitle media samples that are intended to be played simultaneously in a CMAF presentation need to have overlapping presentation time ranges, where the presentation time range starts at the media sample presentation time (subclause 6.2) and ends at the media sample presentation time plus the media sample duration.

7.4 Additional boxes, not defined in the ISO Base Media File Format

7.4.1 Track Encryption Box (`'tenc'`)

The `TrackEncryptionBox` (`'tenc'`) specified in ISO/IEC 23001-7 indicates that media samples in the track might be encrypted using a specific scheme of Common Encryption, identified by a four character code, such as `'cenc'` or `'cbcs'`. The `TrackEncryptionBox` contains parameters and default values that apply to an entire track. See Clause 8 for additional information.

7.4.2 Sample Encryption Box (`'senc'`)

When Common Encryption sample auxiliary information is used:

— it shall be present in each CMAF fragment;

— the `SampleEncryptionBox` should be used to store sample auxiliary information.

See Clause 8 for more information.

7.4.3 Protection System Specific Header Box (`'pssh'`)

The `ProtectionSystemSpecificHeaderBox` is specified in the Common Encryption specification (ISO/IEC 23001-7) for the storage of content protection system information such as license acquisition information and DRM licenses. This box was primarily designed to store license information in download files. Either version 1 or version zero may be used, as constrained in Clause 8.

CMAF streaming applications should signal license acquisition information in the manifest and should not duplicate the information in this box in CMAF headers.

CMAF fragments may contain DRM information in this box intended for delivery to all players during playback.

See Clause 8 for more information.

7.4.4 Media profile specific boxes

Audio and video track formats typically derive sample entries and define decoder configuration boxes that can be required by one or more CMAF media profiles. For example, the `AVCConfigurationBox` is defined in ISO/IEC 14496-15 and required by the NAL structured video CMAF track format specified in Clause 9 and referenced by the CMAF video media profiles specified in Annex A. Other CMAF media profiles may specify boxes that are required by a conformant CMAF track that also conforms to the CMAF media profile brand.

7.4.5 Event Message Box (`'emsg'`)

The enhanced `DASHEventMessageBox` (`'emsg'`) described in this clause is version 1 of the `DASHEventMessageBox`, specified in ISO/IEC 23009-1.

Version 1 of this box adds the field `presentation_time`, which makes event message timing independent of box location in the CMAF track. Version 1 should be used for event messages in CMAF fragments and addressable media objects.

`presentation_time` provides the presentation time of the event measured on the CMAF track's presentation timeline, in the timescale declared in its `MovieHeaderBox`.

`message_data` is the body of the event message. The syntax and semantics of this field are defined by the owner of the scheme identified in the `scheme_id_uri` field. Message schemes may be defined for specific applications and users, or standardized for global use, such as SCTE-35 advertisement and program segmentation markers.

NOTE 1 `DASHEventMessageBoxes` can be included in CMAF fragments to indicate ad insertion points, etc. in the media stream, then other `DASHEventMessageBoxes` added to CMAF segments or CMAF chunks at time of delivery, e.g., to trigger manifest updates.

A `DASHEventMessageBox` in a CMAF track shall contain the value in its `timescale` field equal to the value of the timescale field in the `MediaHeaderBox` of the CMAF track that contains it.

If version 0 is used, then DASH defines the timing of an Event Message related to the earliest media sample presentation time of a DASH segment using the field `presentation_time_delta`, which "provides the media presentation time delta of the media presentation time of the event and the earliest presentation time in this segment." For CMAF fragments, the `presentation_time_delta` shall equal the media presentation time of the event minus the earliest presentation time of the following CMAF fragment.

NOTE 2 The earliest decode time in a CMAF track file is zero (defragmented or not), and if an edit list is present, the earliest presentation time is the earliest media sample composition time adjusted by the edit list offset.

See Annex E for more information on the use of event messages.

7.5 Constraints on ISO Base Media File Format boxes

7.5.1 Movie Header Box (`'mvhd'`)

In the `MovieHeaderBox`, the value of the duration field should be set to zero to indicate that the `MovieBox` contains no media samples and therefore has no duration.

NOTE The `duration` field in the `MediaHeaderBox` (`'mdhd'`) applies to the `TrackBox` (`'trak'`), which contains no media samples in a CMAF track. The duration of a CMAF track can optionally be stored in the `fragment_duration` field of the `MovieExtendsHeaderBox` (`'mehd'`), which is equal to the sum of all CMAF fragment durations in the CMAF track. If the duration is unknown, this box is omitted.

The fields `rate`, `volume`, and `matrix` shall be set to their default values.

7.5.2 Metadata Boxes

Metadata, carried in either `UserDataBox` or `MetaBoxes`, may be present. When present, they shall not occur at file level, i.e., they can only be contained in another box, as permitted by ISO/IEC 14496-12.

7.5.3 Kind Box (`'kind'`)

The `KindBox` (`'kind'`) may be used to store the role of a CMAF track. The `KindBox` is stored in the `UserDataBox` of the `TrackBox`, as documented in the ISO Base media file format (ISO/IEC 14496-12).

Any track can be labelled with role information describing the intended purpose of the track. This information can be captured at the time of encoding and later copied to a manifest describing the CMAF tracks in a selection set so that a user or an automatic algorithm can make an appropriate selection.

The `KindBox` can contain one or more tags from a variety of places, including:

— the DASH specification ISO/IEC 23009-1, as identified by the `schemeURI` `"urn:mpeg:dash:role:2011"` (without the quotation marks);

— the W3C HTML5 specification of track `'kind'`, as identified by the `schemeURI` `"about:html-kind"`.

Where multiple schemes define the same concepts, the DASH role scheme should be used.

7.5.4 Track Header Box (`'tkhd'`)

CMAF `TrackHeaderBoxes` shall conform to ISO/IEC 14496-12 with the following additional constraints.

— The field duration shall be set to a value of zero ('0'), indicating no media samples are referenced from the `TrackBox` (`'trak'`).

— The field matrix shall be set to their default values as defined in ISO/IEC 14496-12, except to indicate video orientation (i.e., portrait or landscape orientation relative to the captured scene). See subclause 9.2.3.

— The following fields shall be set to default values as defined in ISO/IEC 14496-12, unless specified otherwise in this document.

  — The layer field should equal 0 or greater for normally presented video tracks.

  — The layer field should equal –1 for subtitle tracks so they are normally presented over the video.

  — The `width` and `height` fields for a non-visual track (e.g., audio) shall be 0.

  — As defined in ISO/IEC 14496-12, the `width` and `height` fields for a CMAF video track shall specify the track's normalized presentation size as fixed-point 16.16 values expressed in square pixels after decoder cropping, and in the case of video encoded with a non-square video spatial sample shape, after horizontal scaling has been applied. See subclause 9.2.3 for normalized `width` and `height` calculation.

  — Subtitle tracks may set `width` and `height` to an intended layout size, in which case the text layout engine or graphics engine can scale the width and height to match the video display aperture (player implementation dependent).

  — As defined in ISO/IEC 14496-30, subtitle tracks encoded as text may use relative position coordinates and font sizes so that the text layout engine can adjust glyph and layout size to match the final video display aperture without relying on image scaling. For such tracks, the value of zero width and height should be used to indicate that the data can be rendered at any size, and the layout size may be determined by matching the size of the video display aperture.

  — For scalable text and subtitle tracks, the `flag track_size_is_aspect_ratio` may also be used.

7.5.5 Media Header Box (`'mdhd'`)

The CMAF `MediaHeaderBoxes` shall conform to ISO/IEC 14496-12 with the following additional constraints.

— The value of the `duration` field should be set to a value of zero ('0') (see subclause 7.5.1).

— Where possible, the value of the `timescale` field should be chosen such that when the frame rate is constant, the value of the media sample duration may also be constant.

— All tracks that are language-specific should identify the language as precisely as possible (e.g., a text track whose language can be written in different scripts should identify which script is used). When the language is not relevant or not known, the 'und' (undetermined) language tag should be used.

7.5.6 Video Media Header Box (`'vmhd'`)

The `VideoMediaHeaderBox` shall conform to ISO/IEC 14496-12 with the constraints specified in subclause 9.2.2.

7.5.7 Sound Media Header Box (`'smhd'`)

The `SoundMediaHeaderBox` shall conform to ISO/IEC 14496-12 and the following constraint.

The field `balance` shall equal 0 (centre).

7.5.8 Subtitle Media Header Box (`'sthd'`)

The `SubtitleMediaHeaderBox` shall conform to ISO/IEC 14496-12 with constraints specified in ISO/IEC 14496-30.

NOTE Subtitle media uses the `'subt'` `handler_type` in the `HandlerBox` in the `MediaHeaderBox`.

7.5.9 Data Reference Box (`'dref'`)

`DataReferenceBox`es in a CMAF track shall conform to ISO/IEC 14496-12 with the following additional constraints.

— The `DataReferenceBox` shall contain a single entry with the `entry_flags` field set to `0x000001` (which means that the media data is in the same file as the `MovieBox` containing this data reference).

7.5.10 Sample Description Box (`'stsd'`)

The `SampleDescriptionBox` in a CMAF track shall conform to version 0 as defined in ISO/IEC 14496-12 with the following additional constraints.

— Sample entries for encrypted tracks (those containing any encrypted media sample data) shall encapsulate the existing sample entry with the appropriate four-character-code listed in ISO/IEC 14496-12 and include a `ProtectionSchemeInfoBox` (`'sinf'`) that conforms to ISO/IEC 14496-12 and subclause 7.5.11.

— Constraints on visual sample entries are specified in subclause 9.2.4.

— Constraints on audio sample entries are specified in subclause 10.2.5.

— Constraints on subtitle sample entries are specified in ISO/IEC 14496-30.

NOTE A CMAF `SampleDescriptionBox` can contain multiple sample entries.

7.5.11 Protection Scheme Information Box (`'sinf'`)

CMAF shall use common encryption (ISO/IEC 23001-7) for CMAF tracks containing one or more encrypted CMAF fragments and use Scheme Signalling as defined in ISO/IEC 23001-7. An encrypted CMAF track shall include at least one `ProtectionSchemeInfoBox` (`'sinf'`) containing a `TrackEncryptionBox` (`'tenc'`) identifying a scheme specified in ISO/IEC 23001-7.

7.5.12 Track contained media sample information boxes

All boxes in the `SampleTableBox` have a sample count of 0 because CMAF does not reference media samples from the `TrackBox`. The mandatory boxes of ISO/IEC 14496-12 are mandatory, even if they document no samples.

The following boxes therefore shall have an `entry_count` of zero:

— `TimeToSampleBox` (`'stts'`);

— `SampleToChunkBox` (`'stsc'`);

— `ChunkOffsetBox` (`'stco'`);

— `SampleSizeBox` or `CompactSampleSizeBox` (`'stsz'` or `'stz2'`);

— SyncSampleBox ('stss'), if present.

NOTE 1 The presence of an empty SyncSampleBox in a CMAF header indicates that not all media samples in the CMAF track are sync samples.

NOTE 2 Media sample size, duration, and dependency information can be found in the TrackRunBox(es) in each CMAF fragment or CMAF chunk.

7.5.13 Edit List Box ('elst')

If the Edit List Box ('elst') is present, the following conditions apply:

— The EditBox shall contain a single EditListBox.

— The value of entry_count field in the EditListBox shall be set to 1.

— The value of the media_rate_integer field shall be set to 1 and the value of the media_rate_fraction field shall be set to 0.

— The value of the segment_duration field shall be set to 0.

Such conditions define an offset edit or offset edit list.

NOTE Since there is no media in the MovieBox, the duration in the edit is zero. However, as noted in ISO/IEC 14496-12, movie fragments implicitly extend any edit represented by an edit list in the MovieBox. This means that any offset edit list in the CMAF header is applied to the composition times of all samples to determine each sample's presentation time.

According to ISO BMFF, samples with a sum of movie presentation time and duration less than zero are not meant to be presented.

Using offset edits to set the presentation time of the earliest presented sample of each CMAF track to an equivalent CMAF track presentation time enables synchronization of all CMAF tracks to a common CMAF presentation timeline. Additional constraints specific to certain media types are in the following subclauses:

— audio edit lists in subclause 10.2.6;

— video edit lists in subclause 9.2.7.

7.5.14 Track Extends Box ('trex')

A TrackExtendsBox shall be present in a CMAF track since it is a fragmented file as defined in ISO/IEC 14496-12.

7.5.15 Movie Fragment Header Box ('mfhd')

A MovieFragmentHeaderBox in a CMAF track shall conform to ISO/IEC 14496-12.

NOTE The sequence_number integer value is not required to be unique within a CMAF track nor to increase with decode time.

7.5.16 Track Fragment Header Box ('tfhd')

A TrackFragmentHeaderBox in a CMAF track shall conform to ISO/IEC 14496-12 with the following additional constraints.

— The track_ID field shall contain the same value as the track_ID in the associated CMAF header.

— The base-data-offset-present flag (in the tf_flags field) shall be set to zero.

— The default-base-is-moof flag (in the tf_flags field) shall be set to one.

— Every TrackFragmentBox shall contain a TrackFragmentBaseMediaDecodeTimeBox, as defined in ISO/IEC 14496-12, to provide the decode time of the first media sample in the track fragment.

NOTE The baseMediaDecodeTime of the first available CMAF fragment in a CMAF track can be non-zero.

7.5.17 Track Run Box ('trun')

A TrackRunBox in a CMAF track shall conform to ISO/IEC 14496-12 with the following additional constraints.

— The version field shall be set to either '0' or '1',

— When the version field is set to '1', the sample_composition_time_offset of the first presented media sample in a CMAF fragment shall be such that its composition time is equal to the first media sample decode time (baseMediaDecodeTime).

— The data-offset-present flag (in the tf_flags field) shall be set to true in order to indicate that the data_offset field is present and contains the byte offset from the start of this fragment's MovieFragmentBox to the start of the first media sample in the following MediaDataBox.

NOTE This is called movie fragment relative addressing in ISO/IEC 14496-12.

— Within a CMAF track, any TrackRunBox that describes any non-sync media samples shall identify sample dependency with the CMAF chunk and CMAF fragment using a combination of the sample_flags and first_sample_flags fields and default values in the TrackFragmentHeaderBox:

  — sample_is_non_sync_sample shall be 0 for SAP type 1 or 2, and 1 otherwise;

  — an empty SyncSampleBox shall be present in the track.

  NOTE ISO/IEC 14496-12 specifies that absence of the SyncSampleBox indicates that all media samples are sync samples in the track, which allows a reader to know that all subsequent CMAF fragments will also consist of sync samples. If a SyncSampleBox is present, then dependency flags in each CMAF fragment indicate which media samples are sync samples, since the header contains no media samples and the SyncSampleBox therefore lists no media samples.

7.5.18 Sample Group Description Box ('sgpd')

As required by ISO/IEC 14496-12, when sample group descriptions can change within a CMAF track, a SampleGroupDescriptionBox shall be stored in each CMAF fragment that references that sample group description. If sample group information is the same for all CMAF fragments in a CMAF switching set, it may be stored in a SampleGroupDescriptionBox in the CMAF header SampleTableBox ('stbl'). Either version 1 or 2 may be used (version 0 is deprecated by ISO BMFF).

EXAMPLE When common encryption is used and KID values can change per CMAF fragment, a SampleToGroupBox stored in each TrackFragmentBox will reference a SampleGroupDescriptionBox containing the KID for that sample group, and both are stored in the TrackFragmentBox in order to support random access.

Pre-roll sample groups can be used for some audio as described in subclause 10.3.3.1.

7.5.19 Media Data Box ('mdat')

Each CMAF fragment shall contain one or more MediaDataBox(es) ('mdat') containing media samples. The MediaDataBox conforms to the definition in ISO/IEC 14496-12.

Each MediaDataBox in a CMAF chunk shall be immediately preceded by the MovieFragmentBox that references the media samples it contains.

7.5.20 Sub-sample Information Box (`'subs'`)

Each CMAF fragment in a TTML image subtitle track of CMAF media profile `'im1i'`, as specified in subclause 11.3.4, shall contain a `SubSampleInformationBox` in the `TrackFragmentBox` that indexes any images contained in the subtitle media sample. The field `entry_count` shall equal 1.

When one or more images are present in the subtitle media sample, the value of `subsample_count` shall equal 1 for the first image sub-sample, and the `subsample_count` of the TTML document shall equal 0.

When no images are present in the subtitle media sample, the value of `subsample_count` shall equal 0.

Version 0 of this box should be used, unless `subsample_size` exceeds a 16-bit byte address size (65 kibibytes), in which case version 1 should be used.

The fields `subsample_priority`, `discardable`, and `codec_specific_parameters` are undefined.

7.6 The Structural CMAF Brand `'cmfc'`

A CMAF track conforming to the CMAF structural brand `'cmfc'` shall conform to the CMAF track constraints defined in subclauses 7.1, 7.2, 7.3, 7.4, and 7.5.

7.7 The structural CMAF Brand `'cmf2'`

7.7.1 General

A CMAF track conforming to the CMAF structural brand `'cmf2'` shall conform to constraints of the CMAF structural brand `'cmfc'` and all remaining constraints in subclause 7.7.

7.7.2 Edit List Box (`'elst'`)

For video CMAF Tracks, the `EditBox` and in particular the `EditListBox` shall not be present.

For video CMAF Track files as well as any other media types, the `EditListBox` may be present following the constraints in subclause 7.5.13.

7.7.3 Track Run Box (`'trun'`)

A `TrackRunBox` in a CMAF track shall conform to the constraints in subclause 7.5.17 with the following additional constraints:

— For video CMAF Tracks not contained in Track Files, Version 1 shall be used.

— `default_sample_flags`, `sample_flags` and `first_sample_flags` shall be set in the `TrackFragmentHeaderBox` and/or `TrackRunBox` to provide sample dependency information within each CMAF chunk and CMAF fragment.

— Default values or per sample values of sample duration and sample size shall be stored in each CMAF chunk's `TrackRunBox` and/or `TrackFragmentHeaderBox`.

NOTE Default flags and sample parameters (duration, size, or sample description index) can be set and ignored in the `TrackExtendsBox`, as long as those values are also set in all CMAF chunks and CMAF fragments so each CMAF fragment is decodable without access to that track CMAF header.

8 Common encryption of CMAF tracks

8.1 Multiple DRM system support

Common encryption (ISO/IEC 23001-7) enables multiple DRM systems to provide a license for an encrypted CMAF track. The `default_KID` identifies the key and license required, and a registered

`SystemID` identifies a common encryption capable DRM system. License acquisition information can be provided to identify which DRM systems can provide licenses. License acquisition information usually includes the URL of an authorization and license server, DRM SystemID, DRM client identification, type of license requested, media key identifier, etc. and may be stored in a `ProtectionSystemSpecificHeaderBox`, a manifest, or an application in order to assist players in requesting a license to decrypt a CMAF track.

A single key and license may be sufficient to access all tracks in a presentation, or high value content may use different keys for audio and video tracks, since the audio path may be less secure. A content provider may also use different keys and possibly licenses for different qualities, such as SD, HD and UHD.

When streaming, it is recommended that any license acquisition information used to acquire CMAF decryption key(s) be signalled in a manifest or application, and not in the CMAF header. This enables a CMAF player application to parse license acquisition information one time, rather than each time the CMAF header is processed for adaptive switching purposes.

Manifest signalling makes it easier to:

— add or change license information without editing media files;

— offer different types of licenses for the same CMAF presentation, e.g., subscription, rental, ownership, SD, HD, UHD, etc.;

— support different distribution channels and license servers with the same media;

— request a license before the live media becomes available to authorize or purchase playback rights and download a license in advance.

8.2 Track encryption

8.2.1 General requirements

Encrypted media sample data in a CMAF track shall use an encryption scheme defined in ISO/IEC 23001-7. CMAF presentation profiles can constrain the allowable schemes in that CMAF presentation profile. See A.1.

Encrypted NAL structured video tracks shall use a media subsample encryption scheme specified in ISO/IEC 23001-7, which defines partial encryption of NAL units in a video media sample to allow access to unencrypted video NAL headers and slice headers in an encrypted NAL structured video elementary stream.

Encrypted non-video tracks shall use the schemes specified in ISO/IEC 23001-7, which define a full media sample encryption method for each scheme.

Conditional requirements are defined in ISO/IEC 23001-7 and subclause 8.2.2.1 in this document to include a `SampleAuxiliaryInformationOffsetsBox` and a `SampleAuxiliaryInformationSizesBox` in the `TrackFragmentBox`. As specified in ISO/IEC 14496-12, both boxes have a `default aux_info_type` of `'cenc'`.

The `SampleEncryptionBox` (`'senc'`) for storage of sample auxiliary information in encrypted CMAF tracks is documented in subclause 7.4.2 and its use specified in subclause 8.2.2.1.

The `TrackEncryptionBox` version 1 or zero for signalling CMAF track encryption parameters and their defaults is documented in subclause 7.4.1 and its use specified in subclause 8.2.2.2.

The following additional constraints apply to all encrypted CMAF tracks.

— All key identifier values shall uniquely identify one and only one key within their scope of use. To ensure this level of uniqueness, it is strongly recommended that the key identifier values be UUIDs generated according to X.667. A UUID shall be stored in the KID field as 16 bytes, sequenced as specified in X.667:2014, 6.2.

— A KID value may be represented in text form as a hyphenated hexadecimal string, as specified in ITU-T Recommendation X.667:2014, 6.4.

8.2.2 CMAF track constraints

8.2.2.1 Sample Encryption Box ('senc') and sample auxiliary information

Sample auxiliary information, such as per media sample initialization vectors and subsample byte ranges, should be stored in a SampleEncryptionBox ('senc') as described in subclause 7.4.2. Sample auxiliary information may be stored in any valid box within a CMAF chunk's or CMAF fragment's MovieFragmentBox, such as a 'uuid' box.

For encrypted CMAF fragments that contain sample auxiliary information, each TrackFragmentBox shall contain a SampleAuxiliaryInformationOffsetsBox with an aux_info_type value or default of 'cenc' as defined in ISO/IEC 23001-7 to provide access to sample auxiliary information data.

The SampleAuxiliaryInformationOffsetsBox shall conform to the following constraints.

a) The entry_count field of the SampleAuxiliaryInformationOffsetsBox shall equal 1. Therefore, data in the SampleEncryptionBox or other sample auxiliary information storage box shall be contiguous for all of the media samples in the movie fragment.

b) CMAF movie fragments use movie fragment relative addressing where no base data offset is provided in the track fragment header. Therefore, the offset field is calculated as the difference between the first byte of the containing MovieFragmentBox and the first byte of the first InitializationVector in the sample auxiliary information.

The SampleAuxiliaryInformationSizesBox may specify the size of the sample auxiliary data for each media sample with a type of 'cenc', as defined in ISO/IEC 23001-7, with the following constraints.

1) If subsample encryption is not used, the SampleAuxiliaryInformationSizesBox may be omitted.

   NOTE 1 For full media sample encryption with an IV per media sample, the size of the sample auxiliary information equals default_Per_Sample_IV_Size in the TrackEncryptionBox (see ISO/IEC 23001-7).

2) If subsample encryption is used and all the media samples have the same number of subsamples, then the size of the sample auxiliary information will be the same for all of the media samples. In that case, the default_sample_info_size of the SampleAuxiliaryInformationSizesBox ('saiz') may be used instead of storing a size per media sample.

3) If Per_Sample_IV_Size is also zero (because the scheme uses constant IVs and no subsamples), then there would be no sample auxiliary information, and the SampleEncryptionBox, SampleAuxiliaryInformationSizesBox, and SampleAuxiliaryInformationOffsetsBox should be omitted.

   NOTE 2 Sample auxiliary information is located by movie fragment byte offsets stored in the SampleAuxiliaryInformationOffsetsBox, and in some cases, by size information stored in the SampleAuxiliaryInformationSizesBox. Sample auxiliary information, such as per-sample initialization vectors and subsample byte ranges, is not intended to be read directly from the SampleEncryptionBox or other storage box. The SampleAuxiliaryInformationOffsetsBox should always be used to locate sample auxiliary information.

8.2.2.2 Track Encryption Box ('tenc')

A TrackEncryptionBox specified in subclause 7.4.1 shall be present in a CMAF header if any media samples in the track are encrypted.

8.2.2.3 Protection System Specific Header Box ('pssh')

Common encryption specifies version zero and version one ProtectionSystemSpecificHeaderBox. ProtectionSystemSpecificHeaderBoxes can be used to store licenses in downloaded files, signal in CMAF headers whether license downloads may be needed, and deliver licenses, keys, and usage information in

CMAF fragments. ProtectionSystemSpecificHeaderBoxes contain a SystemID that uniquely identifies the protection system intended to use the information. Information contained in the data[] array is considered opaque to players, file parsers, and other DRM systems and might be encrypted by the protection system.

Common encryption (ISO/IEC 23001-7) also specifies XML elements to contain license acquisition information for use in manifests. For CMAF presentations, manifest signalling is recommended.

NOTE A player can download all licenses that will be needed for playback as soon as it parses a manifest and before downloading any CMAF headers. It is particularly useful to acquire licenses in advance of a large live streaming event that would result in a large number of synchronized license requests when triggered by the simultaneous arrival of the first CMAF headers or CMAF fragments. Purchases, etc. can be completed prior to media availability.

The following constraints apply to ProtectionSystemSpecificHeaderBox.

— ProtectionSystemSpecificHeaderBoxes should not be present in CMAF headers, except as allowed below, and may be ignored if present.

— A common ProtectionSystemSpecificHeaderBox and additional ProtectionSystemSpecificHeaderBoxes in CMAF header may be used with presentation-specific playback applications that can read the KID from the common 'pssh' and implement application-specific license management. A common ProtectionSystemSpecificHeaderBox is defined by W3C Encrypted Media Extension specification as a version one ProtectionSystemSpecificHeaderBox with a W3C designated SystemID, and the track's default_KID listed in the version one ProtectionSystemSpecificHeaderBox KID array, but it contains no data in the data[] array.

— Version one ProtectionSystemSpecificHeaderBoxes may be present in CMAF fragments for delivering protection system information, such as encrypted keys in licenses, for use by the DRM system with a SystemID matching the ProtectionSystemSpecificHeaderBox SystemID.

— Multiple ProtectionSystemSpecificHeaderBoxes with different SystemIDs may be present to enable different protection systems on different devices.

  For example, version one ProtectionSystemSpecificHeaderBoxes in CMAF fragments can be used to deliver the same encrypted licenses to all players by a streaming or broadcast service, but only users who have purchased and downloaded an entitlement license bound to their particular player can decrypt the licenses in the CMAF fragments and decrypt media samples with the keys in the inband licenses. Since one license decrypts the other, they are said to be "chained", and chained licenses in combination with key changes can periodically verify that each player has a valid entitlement license to decrypt the inband licenses and continue the presentation (this process is often called "key rotation").

8.2.3 Encryption constraints

8.2.3.1 General

— For a given KID, initialization vectors and counter values shall follow the guidelines outlined in ISO/IEC 23001-7, including the requirement that the combination of values only be used on one cipher block.

— Initialization vectors for use with the 'cenc' scheme shall conform to ISO/IEC 23001-7 and shall be limited to 8-bytes to avoid block counter value overlap.

— Each KID and default_KID shall never reference more than one key value for all CMAF tracks in a CMAF presentation.

  NOTE It is possible for two different KIDs to reference the same key value.

— All CMAF tracks in a CMAF switching set shall use the same default_KID and key value. Any additional keys described by sample groups may be stored in version 1 ProtectionSystemSpeci

`ficHeaderBoxes` in CMAF fragments, identifying the contained KID(s), protected by DRM specific methods.

— It is recommended that any textual representation of KID and SystemID (e.g., in manifests) uses the hexadecimal string representation specified in ITU-T Recommendation X.667:2014, 6.4, derived from the 16 byte binary representation specified in ITU-T Recommendation X.667:2014, 6.2 and equivalent byte arrays specified in ISO/IEC 23001-7 boxes.

The following additional constraints shall be applied to the encryption of NAL structured video tracks.

— Slice headers, NAL type headers, NAL size headers, and all non-video NALs shall be unencrypted using subsample encryption, as required by ISO/IEC 23001-7.

This constraint is recommended in CENC, but required in CMAF.

— Video VCL data shall be protected by subsample encryption, with `bytesOfProtectedData` spanning all complete 16-byte blocks in the VCL slice data.

— `'cenc'` scheme `bytesOfProtectedData` shall be a multiple of 16 bytes.

This constraint is recommended in CENC, but required in CMAF.

— `'cbcs'` scheme `bytesOfProtectedData` shall start on the first complete byte of video data following the slice header and end on the end of the last complete 16-byte block of slice data in the video NAL unit.

These are constraints that are recommended in CENC, but required in CMAF.

8.2.3.2 Clear samples within an encrypted CMAF track

In an encrypted track, the `isProtected` flag in the `TrackEncryptionBox` shall be set to 1, indicating that all media samples are protected by default. Sample groups may indicate unprotected media samples, as specified in ISO/IEC 23001-7.

Each CMAF fragment shall be constrained to media samples that are all protected or all unprotected, not a mix.

8.2.4 CMAF presentation encryption

It is recommended that encryption keys not be shared between audio and video CMAF switching sets in a protected CMAF presentation. Audio decoding and decryption systems may have lower key security than video decoding systems so should not use and expose a key also used for video.

Premium UHD/HDR content may require hardware security for keys, a secure video path, secure output interfaces, etc. only available on some devices and DRM systems. Lower resolution video content with different keys could enable playback on devices that are less secure.

Standard definition, high definition, and/or ultra-high definition content in one CMAF presentation can be sold separately or accessed on different devices with different DRM licenses, keys, and DRM playback requirements. Separate CMAF switching sets for SD, HD and UHD may be functionally combined and possibly seamlessly switched using aligned CMAF switching sets, as specified in subclause 7.3.4.4.

9 Video CMAF tracks

9.1 Overview

CMAF video tracks are derived from the general CMAF track format specified in Clauses 7 and 8.

This clause specifies general constraints on all video CMAF tracks and video CMAF switching sets, then derives more specific constraints for NAL structured video from that. Then, AVC CMAF track format is derived from NAL structured video CMAF tracks.

Video media profiles can be derived from the general video track format, NAL structured video track format, or AVC video track format, as appropriate.

This clause specifies:

— general constraints on a video CMAF track, in addition to conforming to a structural CMAF brand;

— general constraints on multiple CMAF tracks conforming to a CMAF switching set intended to enable seamless adaptive switching of video;

— NAL structured video constraints for CMAF switching sets;

— single initialization constraints for NAL structured video CMAF switching sets;

— constraints between aligned CMAF switching sets containing NAL structured video;

— AVC video CMAF track constraints, used to derive the AVC video CMAF media profiles defined in A.2.

The CMAF track format for NAL structured video is derived from ISO/IEC 14496-15 and requires boxes defined in that specification.

Some CMAF video media profiles using the AVC codec are specified in Annex A. They define widely used video formats, such as SD, HD, and UHD, by specifying codec parameters, such as the maximum codec profile and level that can be encoded, and video parameters such as maximum resolution, frame rate, bit depth, transfer function, colour space, colour subsampling, etc.

Video CMAF tracks that carry a brand indicating parameter limits that also conform to a higher CMAF media profile (i.e., a lower codec profile and level) are considered to also conform to that higher CMAF media profile. For example, a CMAF track with an SD (standard definition) AVC brand also conforms to a CMAF HD (high definition) AVC media profile. In general, a media profile is considered contained in another media profile if it can always be decoded and rendered by all decoders conforming to the containing media profile.

Other CMAF media profiles for video codecs may be defined in other specifications, as long as they conform to CMAF track constraints specified in Clauses 7 and 8 and recommendations in Clause 12. Video media profiles may optionally specify CMAF track constraints for CMAF switching sets to enable seamless adaptive switching and may optionally specify single initialization CMAF switching set constraints.

An additional CMAF media profile for scalable HEVC (SHVC) is defined in Annex H.

For additional information, see Clause 12 and Annex A.

9.2 General video CMAF track format

9.2.1 General video CMAF track structure and constraints

Video CMAF tracks shall conform to at least one structural CMAF brand as specified in Clauses 7 and 8 and general CMAF video track constraints specified in subclause 9.2.

9.2.2 Video Media Header (`'vmhd'`)

`VideoMediaHeaderBoxes` in a video CMAF track shall conform to ISO/IEC 14496-12 with the following additional constraints.

— The following fields shall be set to their default values as defined in ISO/IEC 14496-12:

  — `version=0`;

  — `graphicsmode=0`;

  — `opcolor={0, 0, 0}`.

9.2.3 Track Header Box (`'tkhd'`)

For video tracks, the fields of the `TrackHeaderBox` shall be set to the values constrained below and specified in ISO/IEC 14496-12.

— `flags` = 0x000007

— The values of `width` and `height` (specified in ISO/IEC 14496-12 as the decoded and cropped image size in video spatial samples measured on a uniformly sampled square grid) in a CMAF `TrackHeaderBox` shall be normalized to width and height of the encoded video, as defined below:

  — The normalized presentation `height` shall be the number of vertical video spatial samples after codec cropping parameters are applied.

    NOTE 1 The `height` field of the visual sample entry is the number of encoded vertical video spatial samples after cropping.

  — The normalized presentation `width` shall be the number of horizontal video spatial samples after codec cropping parameters are applied, then multiplied by the video spatial sample aspect ratio. The video spatial sample aspect ratio is defined by the `PixelAspectRatioBox`, if present in the sample entry, or by a codec specific method.

    NOTE 2 The width field in the visual sample entry is the number of encoded horizontal video spatial samples after cropping, not the value of the track header `width` field.

  — The `CleanApertureBox` should not be present (since the cropped aperture is defined to be "clean").

— The value of the matrix field signals the video orientation. Non-identity matrices shall be rotations in multiples of 90 degrees.

  — When video is not rotated, matrix shall be {0x00010000, 0, 0, 0, 0x00010000, 0, 0, 0, 0x40000000}.

  — When video should be rotated 90 degrees clockwise for display, matrix should be {0, 0x00010000, 0, 0xFFFF0000, 0, 0, height<<16, 0, 0x40000000}.

  — When video should be rotated 180 degrees for display, matrix should be {0xFFFF0000, 0, 0, 0, 0xFFFF0000, 0, width<<16, height<<16, 0x40000000}.

  — When video should be rotated 90 degrees counter-clockwise for display, matrix should be {0, 0xFFFF0000, 0, 0x00010000, 0, 0, 0, width<<16, 0x40000000}.

    NOTE 3 A player is expected to adapt the picture aspect ratio (track header `width`/`height`) of the decoded and cropped video image to the size and shape of the current display. The player can frame the video aperture with letterbox bars, pillarbox bars, crop to fit, size to a window, etc. A player needs to scale, and possibly rotate, all CMAF fragments from a CMAF switching set to the selected video aperture to maintain seamless playback without size, shape, or location errors.

9.2.4 Sample Description Box (`'stsd'`)

The `SampleDescriptionBox` in a video track shall contain at least one visual sample entry which shall include:

— `width` and `height` field values of the first sample entry equal to the largest cropped horizontal and vertical video spatial sample counts of any image in the track;

— a decoder configuration record that:

  — should signal the lowest profile, level, height and width values that are equal or greater than all CMAF fragments in the CMAF track.

NOTE Although it is valid to signal a higher profile and level than is necessary to decode the CMAF fragments in the CMAF track, that could unnecessarily exclude decoders capable of decoding the CMAF fragments, but not able to initialize the unnecessarily high profile and level signalled in the CMAF header.

9.2.5 Video CMAF fragment presentation time

CMAF video tracks shall either

a) contain version 1 `TrackRunBoxes` in CMAF video fragments with composition offsets (negative composition offsets where necessary) to adjust the earliest media sample presentation time in each CMAF fragment to equal the earliest media sample decode time, which is equal to the `baseMediaDecodeTime` field of the `TrackFragmentBaseMediaDecodeTimeBox`, or

b) contain an offset edit list in the associated CMAF header to subtract the composition delay added by positive composition offsets, if and only if the CMAF track is contained in a CMAF track file containing version 0 `TrackRunBoxes`.

A video track shall not use both signed composition offsets and an `EditListBox`.

9.2.6 Video media sample dependencies

— Within a video CMAF track, any `TrackRunBox` that describes any non-sync pictures shall identify picture dependencies using a combination of the `sample_flags` and `first_sample_flags` fields, or default flags in the corresponding `TrackFragmentHeaderBox`:

  — `sample_is_non_sync_sample` shall be 0 for SAP type 1 or 2, and 1 otherwise;

  — `sample_depends_on` should be 1 or 2 (the value 2 identifies I pictures);

  — `sample_is_depended_on` should be 2 for disposable pictures.

9.2.7 Video edit lists

When unsigned composition offsets are used in version zero `TrackRunBoxes` in a CMAF track file, the earliest video media sample will normally have a composition time that is greater than zero (whenever video media samples are reordered).

The earliest video media sample in a CMAF track file is defined to have presentation time zero, so an offset edit list, as specified in subclause 7.5.13, shall be used when the earliest media sample's composition time is not zero. In this case, the value of the `media_time` field shall be set to the composition time of the earliest presented video media sample in the CMAF track file.

9.2.8 General video CMAF fragment random access constraints

a) The first media sample shall be Stream Access Point (SAP) type 1 or 2, as defined in ISO/IEC 14496-12 and a video CMAF media profile that specifies these SAP types for a specific video codec and track format.

   NOTE SAP type 1 or 2 means that video media samples in a coded video sequence can be decoded in stored order from the start of the CMAF fragment and presented with the correct presentation timing and ordering.

b) Each video CMAF fragment shall contain sufficient metadata to be decrypted and correctly displayed, possibly in combination with its CMAF header and separately delivered decryption keys.

9.2.9 Additional random access pictures within CMAF video fragments

When coded video sequences have durations longer than 2 seconds, pictures of SAP type 3 should be encoded every 2 seconds or less to provide additional random access video media samples.

For longer coded video sequences and resulting CMAF fragment durations, additional type 3 SAPs ("open GOP" independently decodable pictures) enable independent picture decoding for fast forward, fast reverse, and resumption of normal playback, while improving visual uniformity and lowering bit rate relative to groups of pictures with type 1 o 2 SAPs limited to the equivalent random access duration.

9.2.10 Image framing and encoding constraints

a) CMAF video tracks shall only encode video spatial samples intended for presentation. Image padding used to adjust the picture aspect ratio, such as letterbox and pillarbox bars, should not be encoded.

b) The video spatial samples intended for presentation should be upper left justified and only one row and/or one column of partially filled coding blocks encoded if the image height or width is not a multiple of the coding block size.

c) Decoder cropping parameters shall be encoded in the video stream to remove video spatial samples in a partially filled coding block that are not intended for presentation.

d) Each video CMAF fragment in a CMAF track or CMAF switching set may be encoded with different vertical and horizontal video spatial sample counts (and different corresponding sample entries or inband parameters), so devices are expected to scale each CMAF fragment to the same display aperture and maintain the correct picture aspect ratio, size, and position of each CMAF fragment during playback of CMAF switching sets.

NOTE Each player and display system is expected to frame the decoded and cropped video spatial samples to a player determined video display aperture using methods such as scaling, stretching, cropping, padding with letterbox or pillarbox bars, framing in a window, etc.

9.2.11 General video CMAF switching set constraints

9.2.11.1 Video CMAF switching set constraints

a) CMAF tracks in a CMAF switching set shall be encoded with the same display aspect ratio, although sample aspect ratio and size may differ. All video CMAF fragments within a CMAF switching set are intended to be displayed with the same height, width, and position.

b) Video tracks in a CMAF switching set shall conform to the general CMAF switching set constraints in subclause 7.3.4 that require each CMAF switching set to be encoded from the same source content and with the same visual characteristics such as transfer function, colour volume, colour primaries, colour subsampling, dynamic range, brightness, bit depth, and presentation timing, which would typically be the result of encoding from a single video master.

c) Video tracks in a CMAF switching set shall conform to initialization constraints specified by the CMAF media profile they conform to, so a player and decoder can determine when CMAF headers need to be processed during adaptive switching.

d) Switching between CMAF tracks at the start of CMAF fragments in a video CMAF switching set should result in continuous presentation and appearance when the CMAF fragments are scaled to the same device determined display aperture.

9.2.11.2 Aligned video CMAF switching set constraints

Aligned CMAF switching sets can enable players to switch between CMAF tracks in different CMAF switching sets. In aligned CMAF switching sets, CMAF fragments are time aligned so that track switching can be seamless at CMAF fragment boundaries, depending on the capabilities of the player and the differences between the aligned video CMAF switching sets.

Aligned video CMAF switching sets shall conform to subclause 7.3.4.4.

Video media samples with the same presentation time in aligned video CMAF switching sets shall contain perceptually equivalent images that only differ by their encoding or encryption, i.e., codec, resolution, bit rate, frame rate, default key ID, etc.

A common example of aligned video CMAF switching sets is one CMAF switching set containing CMAF f conforming to the CMAF HHD8 media profile, and another CMAF switching set containing CMAF tracks conforming to the CMAF UHD8 media profile, both encoded from the same source, using the same codec, with time aligned CMAF fragments. Because of different content protection requirements for high definition and ultra-high definition content, in this example, the high definition and ultra-high definition CMAF tracks are encrypted with different keys and digital rights management system playback rules.

An HHD8 capable player could select and play the HHD8 CMAF switching set normally. A UHD8 capable player could select and seamlessly switch both the UHD8 and HHD8 CMAF tracks in aligned CMAF switching sets, applying the different keys and licenses as needed.

9.2.11.3 Multiple initialization switching of video tracks

A CMAF header from each track in the switching set need only be downloaded once (it is assumed they do not change unless a new CMAF presentation is started, e.g., a DASH period).

CMAF switching sets conforming to general constraints or single initialization constraints can be adaptively switched using multiple initialization, meaning that the CMAF header of a CMAF track is processed before decoding the first CMAF fragment from that track on every track switch. The dynamic reinitialization behaviour of parsers, decoders, and display systems is rarely specified and may not be seamless.

It is sometimes possible for players that have full control over both switching and decoding to evaluate changes in CMAF header parameters when switching and leave most parameters unchanged to minimize presentation disruption. In playback environments, such as web browsers using an HTML5 media source extension (MSE) buffer, appending a new CMAF header can cause playback interruption when the media pipeline is fully reinitialized when each CMAF header is appended.

9.2.11.4 Single initialization switching of video tracks

CMAF switching sets containing video CMAF tracks that conform to single initialization constraints can be decoded and displayed by sequencing CMAF fragments from the same CMAF switching set after initializing once with a CMAF header.

Parameters that are allowed to change within a CMAF track or between CMAF tracks in a CMAF switching set can be contained in each CMAF fragment so that full reinitialization with a CMAF header is avoided on track switches. Parameters allowed to change are typically vertical and horizontal video spatial sample count and cropping, so changes in decoder configuration, display buffer sizes, etc. that might interrupt presentation can be avoided.

In a single initialization switching set with hierarchically nested CMAF media profiles, a CMAF header is sufficient to decode any CMAF tracks with equal or lower CMAF media profiles and resolutions than the initialized CMAF media profile and CMAF track.

Single initialization constraints for NAL structured video are specified in subclause 9.3.7.

9.3 NAL structured video CMAF tracks

9.3.1 Overview

The NAL structured video CMAF track format shall conform to the general video CMAF track format specified in subclause 9.2, the NAL structured video track format specified in ISO/IEC 14496-15, with the constraints specified in subclause 9.3.

9.3.2 CMAF track format constraints for NAL structured video

9.3.2.1 Track Header Box ('tkhd')

ISO/IEC 14496-12 specifies that the values of width and height are set to the decoded image size on a uniformly sampled square grid.

For NAL structured video CMAF tracks, the values are additionally constrained as follows.

The values of width and height shall be normalized to the width and height of the NAL structured video, as defined below.

— The normalized presentation `height` shall be the number of vertical video spatial samples in the sequence parameter set NAL VUI after cropping parameters are applied.

NOTE 1 The height field of the visual sample entry is also the number of encoded vertical video spatial samples after cropping.

— The normalized presentation `width` shall be the number of horizontal video spatial samples after sequence parameter set NAL VUI cropping parameters are applied, then multiplied by the video spatial sample aspect ratio. The sample aspect ratio is specified by `aspect_ratio_idc` (and if applicable by the `sar_width`/`sar_height` ratio) in SPS VUI parameters and the `PixelAspectRatioBox` if present in the sample entry.

NOTE 2 The width field of the visual sample entry is the number of encoded horizontal video spatial samples after cropping.

— The `CleanApertureBox` should not be present, since the cropped aperture is defined to be active image ("clean") in a video CMAF track.

9.3.2.2 Sample Description Box ('stsd')

The `SampleDescriptionBox` in a video track shall contain one or more visual sample entries conforming to ISO/IEC 14496-12 and ISO/IEC 14496-15.

The first `VisualSampleEntry` in the `SampleDescriptionBox`:

— shall include `width` and `height` field values that equal or exceed the largest cropped horizontal and vertical video spatial sample counts in any sequence parameter set referenced by a video slice in the NAL structured video track;

— should contain no more than one NAL parameter set of each type, e.g., for AVC video, one SPS NAL with VUI and one PPS NAL in the `AVCDecoderConfigurationRecord`;

— and contain a decoder configuration record that:

  — should signal the lowest codec profile, level, height and width values that are equal to or greater than the values required for all the CMAF fragments in the CMAF track. See general constraints in subclause 9.2.4;

  — should set `LengthSizeMinusOne` field to the value `"3"` (to indicate 4 bytes) to address large VCL NALs and simplify conversion of elementary streams between MPEG-2 TS bytestreams with startcodes and ISO/IEC 14496-15 with NAL length headers;

  NOTE 1 The size of the NAL header length field defined in video tracks conforming to ISO/IEC 14496-15 is stored in the field `LengthSizeMinusOne` in the corresponding decoder configuration record, e.g., for AVC video in the `AVCDecoderConfigurationRecord`.

  — shall contain one or more `ColorInformationBoxes` with sub-type `'nclx'` and a `PixelAspectRatioBox 'pasp'`, as documented in ISO/IEC 14496-12, if the first sample entry contains no SPS NAL with VUI in the decoder configuration record.

NOTE 2 A decoder configuration record without a parameter set is valid for a sample description such as `'avc3'` or `'hev1'` that can store the parameter set NALs necessary for decoding and display in each CMAF fragment.

Any subsequent sample entries in the `SampleDescriptionBox`:

— should contain one NAL video parameter set including VUI in each visual sample entry's decoder configuration box, e.g., for AVC, the `'avcC'` box;

— may be unreferenced by media samples in this CMAF track or other CMAF tracks in a CMAF switching set;

— may be constrained by dependency on other CMAF tracks and CMAF headers in a CMAF switching set. See subclause 9.3.6.

NOTE 3 CMAF headers can contain `VisualSampleEntries` intended for use by other CMAF tracks in a CMAF switching set or for the purpose of initializing decoding and display, but not referenced by any slice header in a media sample.

9.3.3 NAL structured video access units contained in media samples

Each media sample shall contain one NAL structured video access unit for one presentation time and duration, as defined in ISO/IEC 14496-15 and ISO/IEC 14496-12.

NOTE As specified in ISO/IEC 14496-15, timing information provided within a video elementary stream is ignored. Instead, media sample timing in the `TrackRunBox` determines picture presentation time and duration.

Access units shall conform to a sample description specified in ISO/IEC 14496-15 and signalled in the sample entry `codingname` field.

Each access unit may start with an access unit delimiter NAL.

Each access unit shall be stored as a media sample in a `MediaDataBox` in a CMAF chunk and/or CMAF fragment.

Access units conforming to sample descriptions with inband parameters, such as `'avc3'` and `'hev1'`, may retain filler data (NAL units or SEI messages) and SEI messages that might change hypothetical reference decoder bitstream conformance if removed.

The filler data and SEI messages may be retained if HRD conformance is desired.

9.3.4 NAL structured video coding sequences corresponding to CMAF fragments

Each CMAF fragment shall contain one or more complete coded video sequences, as specified by the video codec. Consequently, the first media sample in all NAL video CMAF fragments is SAP type 1 or 2, as specified in ISO/IEC 14496-12.

NAL structured video sample descriptions that allow inline parameter sets (e.g., `'avc3'` and `'hev1'`) shall contain all SPS and PPS NALs referenced by a coded video sequence in the first access unit of that coded video sequence, immediately following the access unit delimiter NAL, if present; followed by SEI NALs, if present; followed by the VCL NAL(s) of the first access unit.

If sample entries also exist in the CMAF tracks' CMAF header using the same sample entry and parameter set indexes as an inband parameter set, then they shall contain the same SPS and PPS NALs in their decoder configuration record as the inband parameter set.

9.3.5 Elementary stream constraints

9.3.5.1 Video colour and dynamic range mastering

Video streams conforming to CMAF shall be conformant to a CMAF media profile and that profile brand should be included in the CMAF header. CMAF media profiles can constrain video transfer characteristics, colour parameters, and grading. See Clause A.2.

For all video media profiles, unless signalled otherwise by a mechanism defined in that video media profile, default transfer characteristics and grading shall be assumed. Default grading is defined as a viewing environment that complies with BT.2035 for presentation on a display which uses the electro-optic transfer function specified in BT.1886 with a peak luminance of 100 cd/m^2.

Video may be graded for presentation on a display which uses an electro-optic transfer function not specified in BT.1886 and/or with a peak luminance greater than 100 cd/m^2, in which case the grading profile should be signalled by VUI and/or SEI messages contained in the sample entry decoder configuration box, or other mechanism defined in the video media profile.

9.3.5.2 Caption data in SEI messages

CTA 608/708 caption data (CTA-608-E, CTA-708-E) may be stored in video SEI messages defined by the associated video codec specification (see subclause 11.4). Video embedded captions are not considered a CMAF subtitle track, and it is expected that many CMAF compatible players will ignore these messages in video CMAF tracks. The presence of CTA 608/708 caption data in such SEI messages is CMAF supplemental data and should be indicated by the addition of the `'ccea'` brand in the CMAF header, as specified in Clause A.4.

For NAL structured video, CTA 608/708 caption data (CTA-608-E, CTA-708-E) may be stored in SEI messages in coded video sequences in CMAF fragments described as user data registered by Rec. ITU-T Recommendation T.35, with SEI `payloadType = 4` and the registered identifier in the field `user_data_registered_itu_t_t35`. See ISO/IEC 23008-2:2015, D.2.6.

Captions intended for CMAF applications should be carried in subtitle CMAF tracks to make delivery optional and selectable by each player, to allow several alternative languages, roles, etc. without duplicating the same video tracks with different SEI messages. Subtitle tracks also enable parsing and presentation by player applications when a device does not contain a built-in subtitle decoder.

9.3.6 General CMAF switching set constraints for NAL structured video

Each CMAF track in a CMAF switching set containing NAL structured video conforming to ISO/IEC 14496-15 (NAL structured video track format) shall conform to general CMAF switching set constraints in subclause 7.3.4 and general video CMAF switching set requirements in subclause 9.2.11.

9.3.7 Single initialization CMAF switching set constraints for NAL structured video tracks and media profiles

NAL structured video CMAF switching sets conforming to the general CMAF switching set constraints for NAL structured video in subclause 9.3.6 may also conform to single initialization constraints specified below.

Single initialization NAL structured video CMAF switching sets:

a) Shall conform to the general single initialization constraints in subclause 9.2.11.4.

b) Shall index and store video parameter sets (SPS and PPS NAL units) that are referenced by each video coding layer slice NAL for decoding and display, using parameter sets stored in:

   1) one or more sample entries in a CMAF header, in which case one CMAF header shall contain sample entries sufficient to decode and display every CMAF track in the CMAF switching set, or shall contain sample entries sufficient to decode and display every CMAF track with a CMAF

header containing the same CMAF media profile brand in its `FileTypeBox`. ISO/IEC 14496-15 requires `'avc1'` and `'hvc1'` sample descriptions to use sample entry storage;

2) the first (IDR) access unit of every coded video sequence that references the parameter set from a slice header, as specified in subclause 9.3.4. Parameter set indexes shall be valid within each coded video sequence, but need not be valid within other CMAF fragments or CMAF tracks. ISO/IEC 14496-15 allows `'avc3'` and `'hev1'` sample descriptions to use inband parameter set NAL storage, as additionally constrained in this clause;

3) a combination of sample entry parameter storage (1) and inband parameter storage (2) may be used for `'avc3'` and `'hev1'` sample description CMAF tracks. In that case, each sample entry and slice NAL parameter set index shall index the same parameter set if stored in both locations. In addition, any CMAF header shall provide all the sample entries referenced by any CMAF fragment in the CMAF switching set.

c) The first visual sample entry in each CMAF header shall be sufficient to initialize decoding, decryption, and display of all CMAF tracks in the CMAF switching set, or only CMAF tracks with matching CMAF media profile brands, if all CMAF headers contain CMAF media profile brands. Initialization information can be:

1) the first sample entry in the CMAF header shall contain a decoder configuration record containing a parameter set (e.g., SPS and PPS in `'avcC'`) conforming to c), or

2) may contain a sample entry without NALs that shall include one or more `ColorInformationBoxes` with sub-type `'nclx'` and a `PixelAspectRatioBox`, as specified in ISO/IEC 14496-12.

d) Any CMAF header shall be sufficient to initialize the CMAF switching set once and decode all CMAF fragments in the CMAF switching set, if the CMAF tracks in the CMAF switching set do not contain CMAF media profile brands.

e) Any CMAF header with a specific CMAF media profile brand shall be sufficient to initialize once and decode the other CMAF tracks with the same CMAF media profile brand, or CMAF media profile brands that conform to the brand initialized, if all CMAF tracks in the CMAF switching set contain CMAF media profile brands.

NOTE A manifest can reference a common CMAF header for every CMAF track in the CMAF switching set, or it can reference a different CMAF header for each CMAF media profile in the CMAF switching set when each CMAF header contains a CMAF media profile brand. In the second case, a player can initialize once and play a subset of the CMAF tracks that conform to the initialized media profile. If it initializes the highest signalled CMAF media profile, it can decode all CMAF fragments in the CMAF switching set.

9.4 AVC video CMAF tracks

9.4.1 Storage of AVC elementary streams

9.4.1.1 Conformance

AVC video tracks shall conform to the NAL structured video format specified in subclause 9.3, the AVC specification ISO/IEC 14496-10 and constraints specified below.

9.4.1.2 AVC Visual Sample Entry

The syntax and values for visual sample entry shall conform to `AVCSampleEntry` (`'avc1'`) or `AVCSampleEntry` (`'avc3'`) as defined in ISO/IEC 14496-15 and the general video sample entry requirements of subclause 9.3.2.2.

9.4.2 Constraints on AVC elementary streams

9.4.2.1 Picture type

All pictures shall be encoded as frames and shall not be encoded as fields.

9.4.2.2 Sequence parameter sets (SPS)

9.4.2.2.1 SPS field constraints

Sequence parameter set NAL units that occur in an AVC video CMAF track shall conform to ISO/IEC 14496-10 with the following additional constraints.

— The following fields have pre-determined values as follows:

  — `frame_mbs_only_flag` shall be set to 1;

  — `vui_parameters_present_flag` shall be set to 1;

  — `gaps_in_frame_num_value_allowed_flag` should be set to 0.

— The values of the following fields shall not change throughout a CMAF track.

  — `chroma_format_idc`

  — `bit_depth_luma_minus8`

  — `bit_depth_chroma_minus8`

— The maximum values of the following fields are specified by CMAF media profiles in A.2.

  — `profile_idc`

  — `level_idc`

  — `pic_width_in_mbs_minus1`

  — `pic_height_in_map_units_minus1`

— The following fields may change per CMAF fragment within a CMAF track. Changes in these parameters shall require a different sample entry, if the SPS NALs referenced are stored in sample entries.

  — `pic_width_in_mbs_minus1`

  — `pic_height_in_map_units_minus1`

  — `frame_crop_right_offset`

  — `frame_crop_bottom_offset`

  — `max_num_ref_frames`

9.4.2.2.2 Visual usability information (VUI) parameters

VUI parameters that occur within a CMAF AVC video track shall conform to ISO/IEC 14496-10 with the following additional constraints.

The following fields have pre-determined values as follows:

— `video_signal_type_present_flag` should be set to 1, and the value of `video_full_range_flag` when not present shall be assumed to be 0.

NOTE 1 This indicates normal black "setup", i.e., black level is 16 for 8-bit video.

— `aspect_ratio_info_present_flag` shall be set to 1. `aspect_ratio_idc` shall not be set to 0 (unknown).

NOTE 2 These parameters refer to the video spatial sample aspect ratio, not picture aspect ratio. Always setting the value distinguishes between content that omitted the value because it intended the default or just failed to set it properly.

— `overscan_info_present_flag`, if present, shall be set to 0.

NOTE 3 Exact scan encoding of the active image is used for reliable image framing by devices and precise adaptive scaling during adaptive switching.

— If `video_signal_type_present_flag` is set to 1, `colour_description_present_flag` should be set to 1.

NOTE 4 As defined in ISO/IEC 14496-10, if the `colour_description_present_flag` is set to 1, the `colour_primaries`, `transfer_characteristics` and `matrix_coefficients` fields are present.

— If `colour_description_present_flag` is set to 1, then `colour_primaries`, `transfer_characteristics` and `matrix_coefficients` shall be set to the values used in the AVC video CMAF track.

— If `colour_description_present_flag` is set to 0, the following values shall be assumed in AVC video CMAF tracks:

  — `colour_primaries` = 1;

  — `transfer_characteristics` = 1;

  — `matrix_coefficients` = 1.

— The presence and values of the following fields shall not change in a CMAF track.

  — `low_delay_hrd_flag`

  — `colour_primaries`

  — `transfer_characteristics`

  — `matrix_coefficients`.

9.4.2.3 Picture cropping

Any picture cropping needed due to partially filled coding blocks shall be indicated by the SPS cropping parameters `frame_crop_bottom_offset` or `frame_crop_right_offset`. See Annex C for examples.

NOTE The NAL video CMAF track format fixes top and left cropping values to zero (default when omitted), which means the image is upper left justified within the coded blocks, and there is at most one row and column of partially filled and padded blocks below and/or to the right of the image if the image size is not evenly divisible by the coded block size.

9.5 AVC video Internet Media Type parameters

9.5.1 AVC signalling of `"codecs"` parameters

The video codec profile and level of each AVC track and CMAF switching set should be signalled using parameters conforming to IETF RFC 6381 and ISO/IEC 14496-15.

10 Audio CMAF tracks

10.1 Overview

This clause specifies CMAF audio tracks derived from the CMAF track format, with additional constraints specific to CMAF audio tracks and audio CMAF media profiles.

Additional CMAF media profiles are defined in

— Annex I for multichannel AAC, and

— Annex J for MPEG-H 3D audio.

— Annex K for MPEG-D USAC.

See subclause 12.1 for guidelines on specification and registration of CMAF media profiles.

Clause 10 includes:

— constraints that apply to all audio CMAF tracks;

— constraints that apply to the audio track format and encoding of AAC audio using the system layer specified in ISO/IEC 14496-14;

— an "AAC Core" CMAF media profile that specifies constraints on AAC CMAF tracks to enable interoperability and random access of CMAF switching sets containing a single CMAF track encoded with one of three variants of AAC in stereo or mono;

— an "AAC adaptive" CMAF media profile that specifies additional constraints on CMAF AAC tracks, fragments, media samples, and metadata to allow seamless switching between alternative tracks and bit rates in an AAC adaptive CMAF switching set.

10.2 General audio CMAF track format

10.2.1 Derivation

Audio CMAF tracks shall conform to at least one CMAF structural brand and to the CMAF track format specified in Clause 7 and Clause 8 and the general audio CMAF track constraints specified in Clause 10.

The general CMAF audio track format applies to all audio CMAF tracks, regardless of audio codec or media profile.

10.2.2 Track Header Box (`'tkhd'`)

For audio tracks, the fields of the `TrackHeaderBox` shall be set to the values specified below. Other fields may be set per ISO/IEC 14496-12.

— `flags` = 0x000007

— `layer` = 0

— `volume` = 0x0100

— `matrix` = {0x00010000, 0, 0, 0, 0x00010000, 0, 0, 0, 0x40000000} // unity matrix

— `width` = 0

— `height` = 0

— `duration` = 0

10.2.3 Sound Media Header Box (`'smhd'`)

The syntax and values for the `SoundMediaHeaderBox` shall conform to ISO/IEC 14496-12 with constraints specified in subclause 7.5.7.

10.2.4 Sample Description Box (`'stsd'`)

As specified in ISO/IEC 14496-12, the `SampleDescriptionBox` contains the `AudioSampleEntry` box or `AudioSampleEntryV1` box.

10.2.5 AudioSampleEntry

For all audio media profiles in CMAF, the value of the samplesize parameter in the `AudioSampleEntry` box defined in ISO/IEC 14496-12 shall be set to 16.

Each `AudioSampleEntry` or `AudioSampleEntryV1` shall contain a "`(codingnamespecific)Box`" containing codec-specific information.

10.2.6 Audio offset edit list

An offset edit list in a CMAF audio track shall conform to constraints specified on `EditListBoxes` in subclause 7.5.13.

If audio media sample composition times differ from their intended presentation times in a CMAF presentation, then an offset edit list shall be included in that audio CMAF track to adjust to the intended CMAF track presentation time.

10.3 AAC audio CMAF tracks

10.3.1 Overview

AAC audio CMAF tracks containing AAC audio as defined in ISO/IEC 14496-3 shall conform to the general audio CMAF track constraints in subclause 10.2 and the AAC constraints in subclause 10.3. These constraints are used to derive the AAC audio media profiles specified in subclause 10.4 and subclause 10.5 and listed in Clause A.3.

Table 12 lists the AAC profiles allowed.

Table 12 — AAC profiles

AAC profile	codingname	SampleEntry Type
MPEG-4 AAC (AAC-LC)	`mp4a`	`MP4AudioSampleEntry`
MPEG-4 high efficiency AAC (HE-AAC)	`mp4a`	`MP4AudioSampleEntry`
MPEG-4 high efficiency AAC v2 (HE-AACv2)	`mp4a`	`MP4AudioSampleEntry`

The `SampleEntry` format in the `SampleDescriptionBox` is the same for each AAC audio profile.

10.3.2 "codecs" parameter signalling

The signalling of the MIME "`codecs`" parameter is per IETF RFC 6381, as shown in Table 13. The third value of the codecs parameter is the `audioObjectType`, as if explicit hierarchical signalling were used.

Table 13 — AAC MIME "codecs" parameter according to IETF RFC 6381

AAC profiles	MIME type	codecs parameter
MPEG-4 AAC (AAC-LC)	audio/mp4	`mp4a.40.2`
MPEG-4 high efficiency AAC (HE-AAC)	audio/mp4	`mp4a.40.5`

Table 13 *(continued)*

AAC profiles	MIME type	codecs parameter
MPEG-4 high efficiency AAC v2 (HE-AACv2)	audio/mp4	`mp4a.40.29`

NOTE HE-AAC is a superset of AAC-LC and HE-AACv2 is a superset of HE-AAC. It is assumed that an HE-AACv2 decoder is also capable of decoding HE-AAC or AAC-LC, and an AAC-LC decoder is capable of partially decoding HE-AAC and HE-AACv2 conforming to CMAF constraints (without reproduction of high frequencies coded with SBR).

10.3.3 Considerations for AAC audio encoding

10.3.3.1 Overview of AAC presentation timing

The AAC codec uses audio frames of a fixed length and a transform which applies over two frames. To obtain correct audio from a frame, both frames in the transform are needed, and hence the prior encoded frame and the current encoded frame need to be decoded to output the first frame. This is sometimes called “priming” and may be signalled using the `'roll'` sample group.

A full reconstruction of the first encoded audio frame is sometimes not possible since there is no previous access unit. To still achieve a full reconstruction, a common practice is to add silence to the beginning of the audio signal. A more detailed explanation of this approach can be found in ISO/IEC TR 14496-24.

In practice, an encoder might prepend an arbitrary amount of (invalid) audio waveform samples to the signal. This portion of the audio signal is sometimes called “encoder delay” and varies depending on the implementation.

10.3.3.2 Presentation delay compensation using an edit list

The most common approach to compensate for inserted extra audio is to add an offset edit list to the CMAF header.

In the case where padding has been added to the start of an audio stream, the `media_time` in the edit list is the length (in audio samples, as measured by the timescale of the track) of the inserted audio samples; 2112 is a common example for AAC.

If an edit list is used, a single `EditListBox` shall be recorded in the CMAF header, as specified in subclause 10.2.6.

10.3.3.3 Delay compensation before encapsulation

If the content has been generated according to Clause G.5, no `EditListBox` is present.

10.3.3.4 Delay compensation when using additional AAC coding tools

If the SBR and PS coding tools are present, they shall not be considered for the purpose of delay compensation.

10.3.3.5 Loudness and dynamic range control

The audio stream should contain DRC and loudness metadata according to ISO/IEC 14496-3. The audio encoder should set the program reference level to the loudness level of the audio stream.

The audio encoder should generate DRC metadata for light compression encoded in the `dyn_rng_ctl` and `dyn_rng_sgn` fields of `dynamic_range_info()` in the FIL element and DRC metadata for heavy compression in the `compression_value` field of `MPEG4_ancillary_data()` in the data stream element (DSE).

NOTE It is expected that the audio decoder will use the program reference level, if available, to achieve a desired target loudness, if applicable. It is expected that the audio decoder will apply the DRC metadata, if present, according to ISO/IEC 14496-3 including the DRC Presentation Mode value of the `drc_presentation_mode` fields.

10.3.4 AAC track constraints

10.3.4.1 Storage of AAC media samples

Storage of AAC elementary stream access units as media samples within a CMAF track shall conform to the CMAF audio track format specified in subclause 10.2.

The following additional constraints also apply.

— All audio media samples shall consist of one AAC audio access unit.

— All AAC access units in a CMAF track shall be encoded with one of AAC-LC, HE-AAC or HE-AACv2.

— The values given in `AudioSampleEntry`, `DecoderConfigDescriptor`, and `DecoderSpecificInfo` shall match the corresponding values in the AAC audio bitstream.

10.3.4.2 AAC audio sample entry

10.3.4.2.1 Field values

The syntax and values of the `AudioSampleEntry` shall conform to `MP4AudioSampleEntry` (`'mp4a'`) as defined in ISO/IEC 14496-14.

The sample entry and fields specified in subclause 10.3.4.2 shall not change within a CMAF track.

10.3.4.2.2 ESDBox

As defined in ISO/IEC 14496-14, the `MP4AudioSampleEntry` in the `SampleDescriptionBox` shall contain an `ESDBox`, which contains an `ES_Descriptor`.

10.3.4.2.3 ES_Descriptor

The syntax and values for `ES_Descriptor` shall conform to ISO/IEC 14496-1, and the fields of the `ES_Descriptor` shall be set to the following values.

— `ES_ID` = 0

— `streamDependenceFlag` = 0

— `URL_Flag` = 0

— `OCRstreamFlag` = 0

— `streamPriority` = 0

— `decConfigDescr` = `DecoderConfigDescriptor`

— `slConfigDescr` = `SLConfigDescriptor`, predefined type 2

Descriptors other than those specified in subclause 10.3.4.2.4 through subclause 10.3.4.2.6 shall not be used.

10.3.4.2.4 DecoderConfigDescriptor

The syntax and values for `DecoderConfigDescriptor` shall conform to ISO/IEC 14496-1, and the fields of this descriptor shall be constrained to the following values.

— `decoderSpecificInfo` shall be used and `ProfileLevelIndicationIndexDescriptor` shall NOT be used.

— `objectTypeIndication` = 0x40 (audio)

— `streamType` = 0x05 (audio stream)

— `upStream` = 0

— `decSpecificInfo` = `AudioSpecificConfig`

10.3.4.2.5 AudioSpecificConfig

The syntax and values for `AudioSpecificConfig` shall conform to ISO/IEC 14496-3.

The following fields of `AudioSpecificConfig` shall be set to AAC audio CMAF media profile specified values, for example, see subclause 10.4.

— `audioObjectType`

— `channelConfiguration`

— `extensionAudioObjectType`

— `GASpecificConfig`

10.3.4.2.6 GASpecificConfig

The syntax and values for `GASpecificConfig` shall conform to ISO/IEC 14496-3, and the fields of `GASpecificConfig` shall be set to the following values.

— `frameLengthFlag` = 0 (1024 lines IMDCT)

— `dependsOnCoreCoder` = 0

— `extensionFlag` = 0

10.3.5 AAC elementary stream constraints

10.3.5.1 General encoding constraints

AAC elementary streams shall conform to ISO/IEC 14496-3 and the constraints of an audio CMAF media profile it conforms to, for example, see Clause A.3.

— The elementary stream shall be a raw data stream, i.e., ADTS and ADIF headers shall not be present.

— CMAF fragments containing HE-AAC shall start with a type 1 SAP. Notably, the SBR configuration information shall be in the first packet.

— The transform length of the IMDCT for AAC shall be 1024 audio PCM samples for long blocks and 128 audio PCM samples for short blocks.

— The following parameters shall not change within the elementary stream:

  — audio object type;

  — sampling frequency;

  — channel configuration.

10.3.5.2 AAC elementary stream syntactic elements

10.3.5.2.1 Syntax and values of syntactic elements

The syntax and values for syntactic elements shall conform to ISO/IEC 14496-3.

The following element shall not be present in an MPEG-4 HE-AAC or HE-AACv2 elementary stream:

— coupling_channel_element (CCE).

If the program_config_element (PCE) element is present, then it shall only list a set of channels corresponding to one of the fixed channel configurations specific in ISO/IEC 14496-3, and the element shall not change for the duration of the track.

10.3.5.2.2 Arrangement of syntactic elements

The syntax and values for syntactic elements shall conform to ISO/IEC 14496-3.

Syntactic elements shall be arranged in the following order for the channel configurations below.

— <SCE>, <optional additional elements>, <TERM>... for HE-AACv2 and mono HE-AAC or AAC-LC.

— <CPE>, <optional additional elements>, <TERM>... for stereo HE-AAC or AAC-LC.

NOTE Angled brackets (<>) are used above to indicate separate syntactic elements, not stream syntax.

10.3.5.2.3 individual_channel_stream

The syntax and values for `individual_channel_stream` shall conform to ISO/IEC 14496-3. The following fields shall be set as defined:

— `gain_control_data_present` = 0.

10.3.5.2.4 Maximum bit rate

The maximum bit rate of AAC elementary streams shall be calculated in accordance with the AAC buffer requirements as defined in ISO/IEC 14496-3. Only the raw data stream shall be considered in determining the maximum bit rate (system-layer descriptors are excluded).

10.4 AAC core audio CMAF media profile

AAC core audio CMAF media profile shall conform to the AAC CMAF track format specified in 10.3.4, with the following constraints.

— Each AAC elementary stream shall be encoded using MPEG-4 AAC-LC, HE-AAC Level 2, or HE-AACv2 Level 2. Use of the MPEG-4 HE-AACv2 is recommended for 32 kbps or lower.

— When using HE-AAC and HE-AACv2 bitstreams, explicit backwards compatible signalling shall be used to indicate the use of the SBR and PS coding tools.

— AAC core CMAF tracks shall not exceed two audio channels.

— AAC core elementary streams shall not exceed 48 kHz sampling rate.

— The AAC core `FileTypeBox` brand shall be `'caac'` and should be used to indicate CMAF tracks that conform to this media profile.

See Clause A.3 for the table of AAC audio CMAF media profiles and their brands.

10.5 AAC adaptive switching audio CMAF media profile

10.5.1 General constraints

Producing audio content capable of seamless adaptive switching with codecs available in the AAC core audio CMAF media profile (AAC-LC, HE-AAC, HE-AACv2) requires constrained encoding at CMAF

fragment boundaries. Subclause 10.5 specifies constraints on AAC adaptive CMAF tracks and Annex G includes recommendations for their use in CMAF switching sets.

— The AAC adaptive audio CMAF media profile `FileTypeBox` brand shall be `'caaa'` and should be used to indicate CMAF tracks that conform to this CMAF media profile.

— CMAF tracks conforming to AAC adaptive CMAF media profile and brand shall conform to the AAC core CMAF media profile specified in subclause 10.4 and the constraints specified in subclause 10.5.

See Clause A.3 for the table of AAC audio CMAF media profiles and their brands.

10.5.2 CMAF fragment encoding constraints

To enable seamless switching between AAC CMAF tracks in a CMAF switching set, the AAC CMAF fragments comprising each CMAF track are encoded and packaged for random access and seamless splicing. The process of encoding the AUs (Access Units) requires periodic generation of stream access points (SAP) that are time aligned between CMAF tracks, and encapsulation requires CMAF fragment alignment to those SAPs. Tracks that belong to the same CMAF switching set can be generated from different encoder instances, but shall conform to the same CMAF switching set constraints.

All CMAF tracks in an AAC CMAF switching set shall have AAC SAPs at CMAF fragment boundaries (i.e., the first AU in each CMAF fragment), and corresponding CMAF fragments in alternative CMAF tracks shall be decode time aligned and presentation time aligned.

Video frames usually have well-defined SAP types, such as an IDR frame that always equals SAP type 1 or 2. For audio, frame SAP types are not usually specified. In MPEG-4 AAC, there is no special frame type defined for random access. Nevertheless, audio SAPs of type 1 can be generated by applying the constraints described in the following clauses and are referred to as an “SAP” frame, since there is only one type.

10.5.3 General considerations and requirements

The following values shall not change within a CMAF switching set:

— audio object type (AOT);

— channel configuration;

— sampling frequency.

The first point means that all the CMAF tracks within a CMAF switching set share the same AOT, i.e., they either use AAC-LC, HE-AAC or HE-AACv2.

The second and third constraints enable the output PCM to remain constant and not require, for example, a re-configuration of the audio device or sound card. A constant sampling frequency assures that AUs share the same temporal framing and, therefore, the AAC CMAF fragments in alternative CMAF tracks are decode time aligned, so no overlap or gap in presentation time results when switching.

All CMAF tracks in the same CMAF switching set shall use the same method of presentation delay compensation, either an offset edit list specified in subclause 10.3.3.2 or decode time compensation specified in subclause 10.3.3.3.

10.5.4 Constraints for AAC-LC

To guarantee seamless switching for AAC-LC encoded tracks, the window shape and window sequence shall be constrained as follows.

— To avoid artefacts from not cancelled time alignment components, the window shape and window sequence shall be synchronized across all tracks.

— At each CMAF fragment boundary (i.e., “right window half” of the last frame in a fragment and “left window half” of the first frame in a fragment), all tracks shall use the same window shape and a defined window sequence.

Using a short overlap for the window sequence (i.e., a long start or short window sequence) is recommended in the last frame of a CMAF fragment. This allows for the use of either a short or stop Window Sequence in the SAP frame (i.e., first frame in the next CMAF fragment). This concept is exemplified in Figure 15.

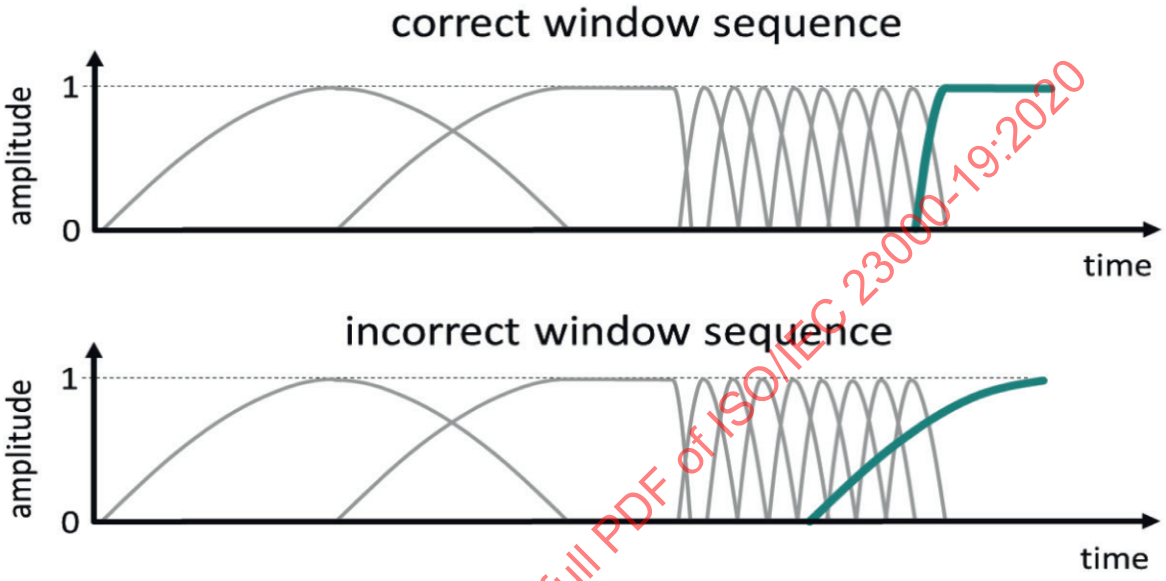


Figure 15 — Correct and incorrect windowing sequence at fragment boundaries for seamless switching of AAC-LC CMAF tracks

10.5.5 Constraints for HE-AAC

Since HE-AAC is based on AAC-LC, the restrictions described in subclause 10.5.4 apply also here. In addition, the following restrictions apply to the spectral band replication (SBR) tool.

10.5.5.1 Core bandwidth adjustment

The framing of the SBR decoder analysis is delayed by 6 QMF time slots (6 × 64 audio PCM samples) compared to the framing of the AAC core decoder. Furthermore, the QMF analysis adds another 320 audio PCM samples. This time shift between core and SBR framing might result in an energy gap when switching, for example, from a track encoded with low bit-rate to a track encoded with high bit-rate. This concept is depicted in Figure 16.

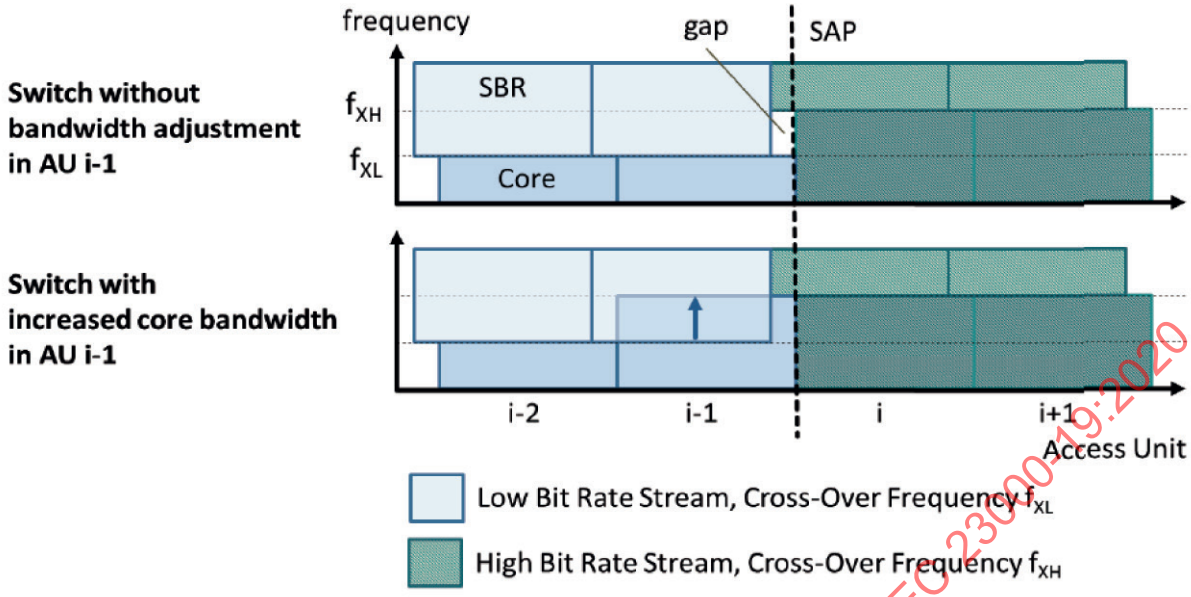


Figure 16 — Illustration of core bandwidth adjustment for HE-AAC switching points

To avoid this gap in the frequency range, it is recommended to have the AAC core bandwidth of the last frame of a fragment matching the highest AAC/SBR crossover frequency of the supported stream configuration. To properly encode the additional bandwidth, extra bits are necessary and the bit reservoir control should be adapted accordingly.

10.5.5.2 SBR header and time differential coding

In contrast to the AAC configuration that is completely signalled inside the audio specific configuration (ASC), the SBR decoder needs additional configuration parameters. These parameters are transmitted in the SBR header, which may not be contained in every access unit. The MPEG-4 standard recommends a transmission interval of 500 ms or whenever an instantaneous change of header parameters is required. To allow for seamless switching of HE-AAC encoded tracks, an SBR header shall be transmitted in each SAP frame. For frames containing the SBR header, the MPEG-4 audio conformance forbids the use of tools that rely on preceding frames. In other words, time differential coding of parameters contained in the SBR header is forbidden and, hence, this restriction guarantees that SAP frames can be completely decoded and processed.

10.5.5.3 SBR frame class

If the SBR frame class is “VARVAR” or “FIXVAR”, the SBR envelope for this particular frame could reach over the frame borders. To make sure all necessary information is self-contained in the SAP frame, so that it can be fully decoded, the SAP shall always start with a “FIX” border (“FIXVAR” or “FIXFIX”). Consequently, the last frame in a fragment (the frame before the SAP) shall also end with a “FIX” border (“FIXFIX” of “VARFIX”).

10.5.6 Constraints for HE-AACv2

Since HE-AACv2 is based on HE-AAC, the restrictions described in subclause 10.5.5 are valid also in this case. In addition, the following restriction applies to the parametric stereo (PS) tool.

Just like for SBR, the PS configuration parameters may not be transmitted with every access unit. To allow for seamless switching of HE-AACv2 encoded tracks, PS header shall be transmitted with each SAP frame. For frames containing a PS header, the MPEG-4 audio conformance again forbids the use of tools that rely on past information, i.e., time differential coding of parameters. These restrictions

assure that the SAP frame can be completely decoded and processed. Since HE-AACv2 conformance requires a PS header delivered together with each SBR header, this requirement is implicitly inherited from the HE-AAC requirements.

See Annex G for additional guidelines.

11 Subtitles and captions

11.1 Overview

CMAF defines the following formats for carrying subtitles and captions:

— WebVTT subtitle CMAF tracks;

— IMSC1 subtitle CMAF tracks;

— CTA-608/CTA-708 captions embedded in video CMAF tracks.

The term “subtitles” in this document is used to mean a CMAF track for the visual presentation of text or graphics that is synchronized with video and audio for purposes including “closed captions” for hearing impaired viewing. Subtitles are presented for various purposes such as dialogue language translation, localized titles, commentary, lyrics, and descriptive captions for hearing impaired viewers or presentation situations where audio is unavailable or inappropriate.

ANSI/CTA-708 (608) “closed captions” embedded in video elementary streams are also described, in part, to distinguish them from CMAF subtitle tracks. Closed captions are commonly used in North America to broadcast descriptive text and dialogue, primarily intended for hearing impaired viewers. See subclause 9.3.5.2. Embedded closed captions have limited flexibility because they cannot be independently created, selected, and delivered without duplicating every video track, and they require low level processing of video elementary streams to decode. Closed captions may be present in CMAF video tracks and can be ignored by players that only process subtitle CMAF tracks.

A CMAF presentation can offer the same subtitle content in more than one CMAF track and format to reach players with different decoding capabilities. Because each CMAF track can have different qualities, it is recommended that manifests indicate the preferred CMAF tracks in the selection set.

An additional CMAF media profile for IMSC 1.1 is defined in Annex L.

11.2 WebVTT

WebVTT is a subtitle format that is supported by several web browsers. It can also be used to render subtitles in other types of media presentation applications. In CMAF, a WebVTT document is encapsulated in a CMAF fragment contained in a CMAF track. WebVTT CMAF fragments may be sequentially downloaded, synchronized, and presented the same as audio and video CMAF fragments.

The encapsulation and format of WebVTT documents in ISO Base Media movie fragments and tracks shall conform to ISO/IEC 14496-30, to Clause 7, and to the `'cwvt'` media profile listed in Clause A.4.

As specified in ISO/IEC 14496-30, each WebVTT document is contained in a single ISO BMFF movie fragment. To conform to CMAF, WebVTT movie fragments shall also conform to the CMAF fragment constraints specified in Clause 7.

One CMAF fragment and WebVTT document may span the entire duration of a prerecorded presentation, or multiple CMAF fragments may sequentially span the duration of the presentation with CMAF fragment durations similar to audio and video CMAF fragments, which is necessary to approximate audio and video latency during live encoding and streaming.

WebVTT subtitle tracks are specified in ISO/IEC 14496-30 using a track `handler_type` of `'text'` with a codingname of `'wvtt'`. WebVTT subtitle tracks store a WebVTT document as a media sample containing zero or more WebVTT cues with presentation times contained within the presentation timespan of the

media sample and CMAF fragment that contains it. WebVTT subtitle CMAF fragments are synchronized during playback based on each CMAF fragment's presentation time and duration, and WebVTT cues are decoded and presented within the CMAF fragments presentation time interval based on their WebVTT timestamps.

By design in ISO/IEC 14496-30, no media sample can contain a WebVTT cue with a presentation time that is outside or crosses the media sample and CMAF fragment presentation time interval (i.e., starts before the CMAF fragment starts and/or ends after the CMAF fragment ends, that duration determined by the duration of the media sample contained in the CMAF fragment). ISO/IEC 14496-30 specifies how to split cues into multiple WebVTT documents and media samples to avoid this case.

NOTE Forced titles are supported in TTML/IMSC1, but not WebVTT. A workaround is to create two WebVTT tracks, one with only the forced titles and the other with both forced and regular subtitles. A player would then play the forced-only track by default, and switch to the forced + subtitles track when subtitles are selected.

11.3 IMSC text and image tracks

11.3.1 General

W3C IMSC1 and W3C IMSC1.1 are profiles of the timed text markup language (TTML) for subtitle and caption delivery. W3C IMSC1 defines a text-only profile and an image-only profile. W3C IMSC1.1 also defines a text-only profile and an image-only profile. W3C IMSC1.1 is a superset of W3C IMSC1, such that a valid W3C IMSC1 text or image profile document is a valid W3C IMSC1.1 text or image profile document, respectively. All IMSC tracks shall conform to the provisions of W3C IMSC1 or W3C IMSC1.1.

An IMSC1 text track is a CMAF track that conforms to the provisions of subclauses 11.3.2 and 11.3.3.

An IMSC1 image track is a CMAF track that conforms to the provisions of subclauses 11.3.2 and 11.3.4.

An IMSC1.1 text track is a CMAF track that conforms to the provisions of subclauses 11.3.2 and L.2.

An IMSC1.1 image track is a CMAF track that conforms to the provisions of subclauses 11.3.2 and L.3.

11.3.2 Common constraints

The CMAF track shall conform to ISO/IEC 14496-30.

The namespace field of the `XMLSubtitleSampleEntry` box shall contain one instance of the string `"http://www.w3.org/ns/ttml"`.

NOTE The `schema_location` field of the `XMLSubtitleSampleEntry` box is not expected to be set for the purpose of profile signalling. The CMAF media profile brand can be used to determine the profile.

The `XMLSubtitleSampleEntry` shall contain a `MIMEBox` as specified in ISO/IEC 14496-12 and its `content_type` field shall be constrained as follows.

— The type shall be "`application`".

— The subtype shall be "`ttml+xml`".

— The codecs parameter shall be present and set according to subclauses 11.3.3 and 11.3.4.

The media type of the CMAF track, as specified in IETF RFC 6381, shall conform to the following.

— The type shall be "`application`".

— The subtype shall be "`mp4`".

— The `codecs` parameter shall be present.

11.3.3 IMSC1 text track constraints

All subtitle media samples of the CMAF track shall conform to the text profile specified in W3C IMSC1.

The `codecs` parameter of the `content_type` field of the MIME box within the `XMLSubtitleSampleEntry` box (see subclause 11.3.2) shall contain the value `"im1t"`, which signals that an IMSC1 text processor is required, as specified in the W3C, TTML media type definition and profile registry.

NOTE 1 A document that conforms to EBU-TT-D, as specified in EBU TECH 3380, also generally conforms to IMSC1 text profile.

EXAMPLE 1 `"application/ttml+xml;codecs=im1t|etd1"` signals that the documents contained in an IMSC1 text track also conform to EBU TECH 3380.

The `codecs` parameter of the media type of the CMAF track shall contain the value `"stpp.ttml.im1t"`. Other TTML profiles the track conforms to may be additionally listed in the `codecs` parameter.

EXAMPLE 2 `"application/mp4;codecs=stpp.ttml.im1t"` is the media type of an IMSC1 text subtitle CMAF track.

EXAMPLE 3 `@mimeType="application/mp4"` and `@codecs="stpp.ttml.im1t"` signal an IMSC1 text subtitle CMAF track in a DASH manifest.

NOTE 2 Clause A.4 specifies a file brand that signals that a track is an IMSC1 text subtitle CMAF track.

11.3.4 IMSC1 image track constraints

All subtitle media samples of the CMAF track shall conform to the image profile specified in W3C IMSC1.

The codecs parameter of the `content_type` field of the MIME box within the `XMLSubtitleSampleEntry` box (see subclause 11.3.2) shall contain the value `"im1i"`, which signals that an IMSC1 image processor is required, as specified in the W3C, TTML media type definition and profile registry.

EXAMPLE 1 `"application/ttml+xml;codecs=im1i"` is a possible value for the `content_type` field of an IMSC1 image track.

EXAMPLE 2 `"application/mp4;codecs=stpp.ttml.im1i"` is the media type of an IMSC1 image track.

The `codecs` parameter of the media type of the CMAF track shall contain the value `"stpp.ttml.im1i"`.

EXAMPLE 3 `@mimeType="application/mp4"` and `@codecs="stpp.ttml.im1i"` signal an IMSC1 image track in a DASH manifest.

Each `smpte:backgroundImage` attribute shall be a URN as specified in ISO/IEC 14496-30 that may conform to the following bg-image-urn syntax expressed using IETF RFC 5234.

```
bg-image-urn = "urn:mpeg:14496-30:subs:" 1*DIGIT
```

NOTE 1 Since W3C IMSC1 requires images to conform to W3C PNG, the `auxiliary_mime_types` field of the `XMLSubtitleSampleEntry` box includes one instance of the media type `"image/png"` if any image is used.

NOTE 2 Annex A.4 specifies a file brand that signals that a track is an IMSC1 image subtitle CMAF track.

11.4 CTA-608 and CTA-708

ANSI/CTA-608 and ANSI/CTA-708 are formats developed for delivering closed captions for accessibility purposes for broadcast television in North America. Closed captions can be embedded in the video elementary stream's SEI messages.

NOTE Clause A.4 specifies a CMAF supplemental data brand that signals that a video track contains ANSI/CTA 608/708 captions in SEI messages and can be included in addition to the video CMAF media profile brand.

11.5 Metadata for subtitles

Text tracks should be labelled with their role. See subclause 7.5.3.

Text tracks should be tagged with their language. See subclause 7.5.5.

12 CMAF media profiles and CMAF presentation profiles

12.1 CMAF media profiles

12.1.1 General guidelines for specifying CMAF media profiles

12.1.1.1 CMAF media profile CMAF track format

All CMAF media profiles should specify a CMAF track format, either derived directly from a CMAF structural brand and format specified in Clause 7 or from a video or audio CMAF track format specified in Clause 9 or Clause 10. The media profile's CMAF track format should specify any requirements and constraints on CMAF headers and CMAF fragments.

12.1.1.2 Codecs, profiles and levels

All CMAF media profiles should specify codec specific features and operating points, such as profiles and levels, that are necessary to determine encoder/decoder interoperability. The maximum features and limits allowed in content should be specified. When related media profiles are specified, subset/superset relationships between those media profiles should be specified based on the ability of a decoder conforming to superset media profile being able to reliably decode content conforming to a subset media profile.

12.1.1.3 Media access unit mapping to media samples

All CMAF media profiles should specify one or more media access unit formats that are addressed as media samples in a CMAF track. In addition, all CMAF media profiles should specify the sample entry code point for the media samples in a CMAF track.

12.1.1.4 Media access unit sequence mapping to CMAF fragments

All CMAF media profiles should specify any limitation on storing sequences of media samples in a CMAF fragment in order to conform to the general random access and timing constraints of CMAF fragments. If all media samples are sync samples, there are no additional constraints. If sequential media sample dependencies are caused by codec prediction, media sample reordering transforms that overlap media samples, sample group and layer dependencies, etc., then media sample sequences should be specified that allow CMAF fragments to be randomly accessed and optionally, adaptively switched (if the media profile defines adaptive switching functionality and CMAF switching set constraints).

12.1.1.5 CMAF track constraints for CMAF switching sets

CMAF media profiles that support adaptive switching of multiple CMAF tracks in a CMAF switching set should specify the constraints between CMAF tracks in a CMAF switching set. CMAF media profiles should specify subset/superset relationships with compatible CMAF media profiles, e.g., a higher level of the same codec. A CMAF media profile should specify initialization constraints on CMAF switching sets and constraints related to the adaptive switching process enabled.

12.1.1.6 CMAF media profile internet media type

All CMAF media profiles should specify an internet media type and associated parameters, file extensions, etc. for this media profile in a CMAF track or more generally in an ISO BMFF file.

12.1.1.7 CMAF media profile brand

All CMAF media profiles should specify a brand and register that brand in accordance with the process specified in ISO/IEC 14496-12. The brand registry should reference a CMAF media profile specification conforming to subclause 12.1.1.

12.1.2 Guidelines for audio CMAF media profiles

12.1.2.1 General

Audio codecs often define features and other details that present options for implementers. In order to increase interoperability, CMAF media profiles usually limit some encoding options and specify an identifier for conforming CMAF tracks. While every codec offers its own set of choices, this subclause presents guidelines for what should be specified by audio CMAF media profiles in the form of testable assertions.

12.1.2.2 Audio track format

Audio CMAF media profiles should be derived from the general CMAF audio track format specified in subclause 10.2 or the AAC audio CMAF track format specified in subclause 10.2.

12.1.2.3 Loudness and dynamic range control

Audio CMAF media profiles should specify loudness and dynamic range control information.

12.1.2.4 Audio parameters

Audio CMAF media profiles should specify which parameters (such as loudness, dynamic range control information, channel configuration, etc.) are allowed and if they can change within a CMAF track or between CMAF tracks in a CMAF switching set.

12.1.2.5 Audio presentation time adjustment

Audio CMAF media profiles should specify any presentation time adjustment relative to decode time necessary on decoding, preferably without using media sample composition offsets in the `TrackRunBox`.

12.1.3 Guidelines for video CMAF media profiles

12.1.3.1 General

Video codecs often define features and performance limits, such as profiles, levels, video image characteristics, that present options for implementers. In order to increase interoperability, CMAF media profiles usually limit some encoding options and specify an identifier for conforming CMAF tracks. While every codec offers its own set of choices, this subclause presents guidelines for what should be specified by video CMAF media profiles in the form of testable assertions.

12.1.3.2 Video CMAF track format

Video CMAF media profiles should be derived from the general CMAF video track format specified in subclause 9.2 or NAL structured video CMAF track format specified in subclause 9.3.

12.1.3.3 Video image coding characteristics

Video CMAF media profiles should specify allowed options and constraints for image coding, such as colour primaries, transfer functions, colour and luminance matrix, colour spatial subsampling, bit depth, maximum height and width, frame rate limitations, etc.

12.1.3.4 Video rendering information

Video CMAF media profiles should specify any metadata in the video stream and CMAF track format that is necessary for accurate video rendering.

12.1.3.5 Video presentation time adjustment

Video CMAF media profiles should specify media sample composition offsets and any other means to align the presentation time of each coded video sequence to the earliest media sample decode time in that sequence, preferably using `TrackRunBox` composition offsets to reorder media samples without creating a presentation time delay.

12.1.3.6 Seamless adaptive video CMAF switching set constraints

Video CMAF media profiles should specify additional constraints on video CMAF tracks in a CMAF switching set that are necessary to enable seamless adaptive switching of CMAF fragments in decoders that conform to the video CMAF media profile. Only CMAF media profiles that support multiple CMAF tracks in a CMAF switching set should specify these CMAF switching set constraints. Seamless switching constraints should result in perceptually equivalent images when adaptive switching occurs.

12.1.3.7 Video decoder and display initialization

Video CMAF media profiles should specify CMAF track constraints and associated initialization process necessary to render a video CMAF track. If a video media profile also supports adaptive switching, it should specify CMAF switching set constraints and associated initialization process necessary to adaptively switch and seamlessly render a video CMAF switching set.

12.2 CMAF presentation profiles

12.2.1 General

Each CMAF presentation profile defines conditionally required media profiles for CMAF tracks that are required to be available in a CMAF presentation that conforms to the presentation profile.

Required media profiles are a minimum constraint. CMAF presentations containing the required CMAF media profiles may also include additional CMAF media profiles and remain conformant. A CMAF presentation may conform to more than one presentation profile if it contains CMAF tracks in all the conditionally required media profiles.

CMAF presentation profiles may be specified in specifications other than CMAF by entities other than MPEG. Externally specified CMAF presentation profiles should be identified by a URI in the namespace of the specifying entity. If a URL is defined, it should resolve to the specification document that defines the conditionally required media profiles of the presentation profile.

See Clause A.1 for CMAF presentation profiles defined in the CMAF specification.

12.2.2 CMAF profile conformance

CMAF defines several conformance points including CMAF fragments, CMAF tracks, CMAF addressable media objects, CMAF media profiles, CMAF switching sets, CMAF selection sets, and CMAF presentations.

CMAF objects and conformance points are hierarchical.

a) CMAF tracks are the conformance point for ISO BMFF box, track, and file format constraints associated with a CMAF structural ISO BMFF brand. CMAF headers, CMAF fragments, and CMAF chunks are objects within the CMAF track structure. CMAF track conformance provides basic compatibility between file packagers, players, parsers, and decryptors, independent of media encoders and decoders.

b) General CMAF media track formats for audio, video, and subtitles are conformance points that can be referenced by CMAF media profiles. The general CMAF track formats for audio, video, and subtitles and specific CMAF track formats for NAL structured video and AAC audio are specified in clauses 7 through 11, which are derived from ISO/IEC 14496-2, ISO/IEC 14496-3, ISO/IEC 14496-14, ISO/IEC 14496-15 and ISO/IEC 14496-30. Conformance to one of these media track formats can be specified by a CMAF media profile and that conformance indicated by the media profile brand. CMAF media track formats support CMAF adaptive multimedia presentations.

c) CMAF media profiles are conformance points for encoded media data stored in a specified CMAF track format. A CMAF media profile brand indicates encode/decode and adaptive switching conformance of media data, as well as a CMAF track format. CMAF media profile conformance enables encoding and decoding of media samples in CMAF fragments and CMAF tracks, independent of delivery.

d) CMAF addressable media objects are conformance points for CMAF media objects (headers, chunks, segments, and files) and their brands, which are derived from CMAF tracks, fragments, chunks, and headers, and optimized for external reference, storage, and delivery. They inherit the structure and media constraints of the CMAF track and media profile they contain, and constrain the structure of the CMAF addressable media object independent of the CMAF media. CMAF addressable media objects are a conformance point for streaming protocols and systems that is independent of the manifest format and delivery method used.

e) CMAF switching sets are a conformance point for a group of CMAF tracks that share content and encoding constraints defined by the CMAF track format and a CMAF media profile to enable seamless adaptive switching. Single initialization CMAF switching set constraints and an associated identifier are defined to indicate CMAF switching set functional conformance to single initialization constraints, the syntax of which is specified by a CMAF media profile. Conformance for content pertains to a set of CMAF tracks, and conformance for decoders pertains to decoding and seamlessly rendering sequences of consecutive CMAF fragments selected from any of the CMAF tracks.

f) CMAF selection sets are conformance points for a group of synchronized CMAF switching sets that share related content, timing and media type. For instance, alternative audio CMAF switching sets in a CMAF selection set can cover the same presentation timespan with synchronized alternative content or encoding. Since CMAF tracks contain a single media type (audio, video, or subtitles), multiple CMAF selection sets are necessary to compose a multimedia presentation using the CMAF hypothetical application model and late binding. Conformance indicates that all CMAF tracks are correctly grouped, synchronized, and have approximately the same duration to form a CMAF presentation.

g) CMAF presentation profiles are conformance points for a group of conditionally required CMAF tracks that conform to required media profiles and form a synchronized multimedia presentation. What content and timing are appropriate for a multimedia presentation is not specified by CMAF and can only be tested based on a presentation author's intended result. However, CMAF presentations can be tested against the CMAF hypothetical application model, which provides a common understanding between content creators and players as to the intended and expected result. A CMAF presentation profile constrains the media profiles, CMAF switching sets, track formats, etc. that it contains and is intended to identify basic content/device compatibility. A player that claims compatibility with a CMAF presentation profile is expected to decode all the required CMAF media profiles and implement the CMAF hypothetical application model to late bind CMAF tracks, synchronize them, decrypt them, render them, and perform seamless adaptive switching, if appropriate. If optional CMAF media profiles are included in a CMAF presentation, a player can select those instead, if that player supports the optional CMAF media profile.

A CMAF media profile that falls within the maximum encoding constraints of a higher CMAF media profile conforms to the higher CMAF media profile and does not need to indicate compatibility with a second CMAF media profile brand, since it is an intrinsic property of those media profiles.

EXAMPLE Standard definition video conforming to the CMAF SD media profile also conforms to the CMAF HD media profile because it will decode on an HD compatible system. It could be included in a CMAF switching set that conforms to the HD media profile. CMAF tracks and media profiles with different source formats, such as different colour spaces or transfer functions, do not conform to a single CMAF switching set or media profile because they are not considered the same source content and would not provide seamless adaptive switching in the same CMAF switching set and CMAF media profile.

A CMAF presentation shall conform to one or more CMAF presentation profiles by containing one or more CMAF tracks for each conditionally required media profile in the CMAF presentation profile. A CMAF presentation can conform to multiple CMAF presentation profiles by containing all the conditionally required CMAF tracks and CMAF media profiles. Additional CMAF media profiles in a CMAF presentation are optional by default when they are not conditionally required in a CMAF presentation profile that the CMAF presentation conforms to.

A presentation profile may specify a single encryption scheme to be used on every encrypted CMAF switching set that is conditionally required in the presentation.

Annex A
(normative)

CMAF presentation profiles, media profiles and supplemental data

A.1 CMAF presentation profiles

A.1.1 Overview

This clause defines some CMAF presentation profiles and their associated CMAF presentation profile identifiers. The CMAF presentation profiles below differ only by the common encryption schemes they require (ISO/IEC 23001-7).

NOTE In the current market, two encryption schemes, 'cenc' and 'cbcs', are widely deployed on popular platforms. Therefore, this document currently defines one presentation profile for each of these encryption schemes and another for unencrypted CMAF presentations. By defining multiple presentation profiles, a mechanism is provided to select a given encryption scheme by presentation profile within a given deployment. This identifies the potential incompatibilities between content and deployments conforming to similar but different CMAF presentation profiles.

A.1.2 CMFHD presentation profile

The CMFHD presentation profile is intended to provide basic interoperability of unprotected content on the widest range of Internet video devices in use today. Other CMAF media profiles that are not required in this presentation profile are considered optional.

Requirements of CMAF presentation profile CMFHD

— Presentation profile ID= "**urn:mpeg:cmaf:presentation_profile:cmfhd:2017**"

— If containing video, it shall include at least one CMAF switching set constrained to the `'cfhd'` media profile in Clause A.2.

— If containing audio, it shall include at least one audio CMAF switching set constrained to the `'caac'` media profile defined in Clause A.3.

— If containing subtitle tracks, it shall include at least one CMAF switching set for each language and role in the `'im1t'` media profile defined in Clause A.4.

— CMAF tracks containing optional CMAF media profiles in CMAF switching sets should be included in selection sets with CMAF switching sets containing required media profiles.

— All CMAF tracks shall not contain encrypted media samples or a `TrackEncryptionBox`.

A.1.3 The CMFHDc presentation profile

The CMFHDc presentation profile is intended to provide interoperability with a large portion of video devices in use today that support content protection using the `'cenc'` scheme of common encryption (see subclause 8.2.3). Other CMAF media profiles that are not required in this presentation profile are considered optional.

Requirements of CMAF presentation profile CMFHDc

— Presentation profile ID= "**urn:mpeg:cmaf:presentation_profile:cmfhdc:2017**"

— If containing video, it shall include at least one CMAF switching set constrained to the `'cfhd'` media profile in Clause A.2.

— If containing audio, it shall include at least one audio CMAF switching set constrained to the `'caac'` media profile defined in Clause A.3.

— If containing subtitle tracks, it shall include at least one CMAF switching set for each language and role in the `'im1t'` media profile defined in Clause A.4.

— CMAF tracks containing optional CMAF media profiles in CMAF switching sets should be included in selection sets with CMAF switching sets containing required media profiles.

— At least one CMAF switching set shall be encrypted. Any CMAF switching set that is encrypted shall be available in `'cenc'` common encryption scheme specified in Clause 8.

A.1.4 The CMFHDs presentation profile

The CMFHDs presentation profile is intended to provide interoperability with a large portion of video devices in use today that support content protection using the `'cbcs'` scheme of common encryption (see subclause 8.2.3). Other CMAF media profiles that are not required in this presentation profile are considered optional.

Requirements of CMAF presentation profile CMFHDs

— Presentation profile ID= "**urn:mpeg:cmaf:presentation_profile:cmfhds:2017**"

— If containing video, it shall include at least one CMAF switching set constrained to the `'cfhd'` media profile in Clause A.2.

— If containing audio, it shall include at least one audio CMAF switching set constrained to the `'caac'` media profile defined in Clause A.3.

— If containing subtitle tracks, it shall include at least one CMAF switching set for each language and role in the `'im1t'` media profile defined in Clause A.4.

— CMAF tracks containing optional CMAF media profiles in CMAF switching sets should be included in selection sets with CMAF switching sets containing required media profiles.

— At least one CMAF switching set shall be encrypted. Any CMAF switching set that is encrypted shall be available in `'cbcs'` common encryption scheme specified in Clause 8.

A.1.5 Protected CMAF presentations

Manifests that include equivalent CMAF presentations in CMFHDc and CMFHDs presentation profiles can declare both presentation profile IDs to indicate that all CMAF switching sets are available with both encryption schemes.

Presentations conforming to the CMFHD presentation profile are not protected by common encryption of media sample data, but do not preclude encryption during delivery by methods such as HTTPS or envelope encryption that are removed on arrival, prior to media processing of the unencrypted presentation.

A.2 AVC video CMAF media profiles and brands

A set of AVC video CMAF media profiles is defined together with an ISO BMFF brand and constraints in Table A.1.

For a CMAF track to comply with one of the media profiles in Table A.1, it

— shall conform to subclause 9.4,

— shall not exceed the profile or level listed in the table,

— shall conform to the `colour_primaries`, `transfer_characteristics` and `matrix_coefficients` values from the options listed in the table,

— shall not exceed the width, height or frame rate listed in the table, even if the AVC level would permit higher values, and

— should include the CMAF file brand in its CMAF header.

NOTE CMAF tracks conform to more than one media profile if they meet the requirements for multiple CMAF media profiles. For example, a video CMAF track could conform to both the SD and HD CMAF media profiles if its dimensions are within the constraints for the SD profile (the HD profile has higher dimension constraints and codec level) and the `colour_primaries`, `transfer_characteristics` and `matrix_coefficients` are values which are common to both profiles.

See Annex C for detailed encoding examples.

Table A.1 — AVC video CMAF media profiles

Media profile	Codec	Profile	Level	colour_ primaries field in VUI	transfer_ characteristics field in VUI	matrix_ coefficients field in VUI	Max frame height	Max frame width	Max frame rate	CMAF File Brand
SD	AVC	High	3.1	1 (BT.709) 5 (BT.601-7 PAL/ SECAM) 6 (BT.601-7 NTSC)	1 (BT.709 OETF) 6 (BT.601-7 OETF) (see NOTE)	1 (BT.709) 5 (BT.601-7 PAL/ SECAM) 6 (BT.601-7 NTSC)	576	864	60	`'cfsd'`
HD	AVC	High	4.0	1 (BT.709)	1 (BT.709 OETF)	1 (BT.709)	1080	1920	60	`'cfhd'`
HDHF	AVC	High	4.2	1 (BT.709)	1 (BT.709 OETF)	1 (BT.709)	1080	1920	60	`'chdf'`
NOTE Values 1 and 6 for `transfer_characteristics` are functionally equivalent, but the value which refers to the same specification as the `colour_primaries` is normally used.										

See Clause 9 and Annex C for general requirements and constraints on CMAF video tracks.

A.3 Audio CMAF media profiles and brands

The AAC audio CMAF media profiles in Table A.2 are specified in subclauses 10.4 and 10.5. Table A.2 summarizes key parameters and specifies the four-character code that can be used as an ISO BMFF brand to identify the CMAF media profile in CMAF tracks and CMAF media objects.

The AAC adaptive switching audio CMAF media profile (`'caaa'`) is a constrained subset of the AAC core audio CMAF media profile (`'caac'`), so `'caaa'` CMAF tracks always conform to the `'caac'` CMAF media profile.

Table A.2 — AAC audio CMAF media profiles

Media profile	Codec	Profiles	Levels	Number of channels	File brand
AAC core	AAC	AAC-LC, HE-AAC or HE-AACv2	2	Mono or stereo	`'caac'`
AAC adaptive	AAC	AAC-LC, HE-AAC or HE-AACv2 (see NOTE)	2	Mono or stereo	`'caaa'`
NOTE AAC adaptive CMAF media profile is AAC core constrained for adaptive switching.					

See Clause 10 for general constraints on audio CMAF tracks and subclause 12.1.2 for the specification of additional audio CMAF media profiles.

A.4 Subtitle CMAF media profiles and brands

CMAF specifies three subtitle CMAF track formats as shown in Table A.3 — WebVTT, IMSC1 text, and IMSC1 image — to provide interoperability between CMAF presentations and players. Subtitle media profiles shall conform to the CMAF subtitle track format specified in Clause 11.

Table A.3 — Subtitle and caption CMAF media profiles

Media profile	Format	Notes	File brand
WebVTT	Specified in 11.2	ISO/IEC 14496-30	'cwvt'
TTML IMSC1 text	Specified in 11.3.3	IMSC1 text profile	'im1t'
TTML IMSC1 image	Specified in 11.3.4	IMSC1 image profile	'im1i'
TTML IMSC1.1 text	Specified in L.2	IMSC1.1 text profile	'im2t'
TTML IMSC1.1 image	Specified in L.3	IMSC1.1 image profile	'im2i'

See Clause 11 for more details on subtitle CMAF track formats.

A.5 CMAF supplemental data

CMAF supplemental data is shown in Table A.4.

Table A.4 — CMAF supplemental data

Supplemental data	Format	Notes	Target CMAF tracks	File brand
CTA captions	CTA-608 and CTA-708 Specified in subclause 11.4	Caption data is embedded in SEI messages in video track; multiple closed caption streams may be present.	Video media profiles, including AVC and HEVC.	'ccea'
NOTE The 'ccea' brand can be included in addition to a video CMAF media profile brand to indicate the presence of captions embedded in the video elementary stream.				

Annex B
(normative)

HEVC video CMAF track format and CMAF media profiles

B.1 HEVC video CMAF tracks

HEVC tracks shall conform to subclause 9.3, as additionally constrained in this annex.

B.2 HEVC video track constraints

B.2.1 HEVC video CMAF switching set constraints

HEVC video CMAF switching set shall conform to constraints for NAL structured video CMAF switching sets specified in subclause 9.3.6 or subclause 9.3.7.

B.2.2 Sample Description Box (`'stsd'`)

The `Sample Description Box` (`'stsd'`) shall conform to subclause 9.2.4 and contain one or more sample entries.

B.2.3 Visual sample entry

The syntax and values of a visual sample entry shall conform to `HEVCSampleEntry` (`'hvc1'`) or `HEVCSampleEntry` (`'hev1'`) sample entries as defined in ISO/IEC 14496-15 and constrained as follows.

The `HEVCSampleEntry`:

— shall contain an `HEVCConfigurationBox` (`'hvcC'`) box containing an `HEVCDecoderConfigurationRecord`, as specified in ISO/IEC 14496-15;

— shall set sample entry fields consistent with the sequence parameter set and picture parameter set values in the video track as specified in ISO/IEC 14496-15 and constrained by B.3;

— should contain SEI messages containing colour and dynamic range mastering and rendering information if default values are not encoded, as required in subclause 9.3.5.1;

— shall contain a `ColourInformationBox` (`'colr'`) with `colour_type 'nclx'` and `PixelAspectRatioBox` (`'pasp'`) when required per subclause 9.3.5.1.

NOTE The `ColourInformationBox` with `colour_type 'nclx'` and `PixelAspectRatioBox` provide equivalent information to SPS VUI and are required when SPS VUI is not present in an `'hev1'` sample entry.

B.2.4 `HEVCDecoderConfigurationRecord` colour and dynamic range information

— As specified in subclause 9.3.5.1, video captured or colour graded with characteristics other than ITU-R Recommendation BT.709 defaults should include one or more SEI NALs with additional transfer characteristics or colour volume information stored in the `HEVCDecoderConfigurationRecord` to enable colour and dynamic range calibration during decoder and display initialization. This may include the following SEI messages specified in ISO/IEC 23008-2:2015, Annex D:

  — SEI payloadType 137, mastering_display_colour_volume;

  — SEI payloadType 144, content_light_level_info;

— SEI payloadType 147, alternative_transfer_characteristics.

— CMAF player display processors can use SEI colour grading, colour volume, and dynamic range messages in SEI NALs stored in the `HEVCDecoderConfigurationRecord` to calibrate rendering for display characteristics and viewing conditions.

— Alternative transfer characteristics in an SEI message can be used to optimize rendering of hybrid log gamma transfer function video on high dynamic range displays while using SPS signalled gamma rendering on standard range displays.

— A CMAF player can pass MaxFALL, MaxCLL, and other SEI message data to a display over a digital video interface, as specified in ANSI/CTA 861-G, to enable colour and transfer function optimization within a compatible UHD/HDR display.

— CMAF switching sets shall be constrained to include identical SEI NALs and SPS VUI colour mastering and dynamic range information in the first sample entry of every CMAF header in the CMAF switching set to provide consistent initialization and calibration.

B.2.5 Track Header Box (`'tkhd'`)

The requirements of subclause 9.2.3 apply.

NOTE Normalized `width` and `height` can be derived from the track's visual sample entry for `'hvc1'` video media samples (see subclause 9.3.2.2) or from a sequence parameter set NAL in each coded video sequence for `'hev1'` video media samples. See subclauses 9.3.3 and 9.3.4 for the storage and semantics of video sequence parameter sets.

B.3 Media sample and CMAF fragment constraints

B.3.1 Storage of HEVC elementary streams

HEVC video tracks shall comply with ISO/IEC 14496-15 and subclause 9.3.

B.3.2 Access units

Access units and media samples shall conform to subclause 9.3.

Access units shall conform to the requirements of a media sample of the indicated description (`'hvc1'` or `'hev1'`) as specified in ISO/IEC 14496-15.

CMAF fragments containing access units identified by the `'hev1'` sample description shall contain all SPS and PPS NALs referenced from a coded video sequence in the first access unit of that sequence, immediately following its first access unit delimiter NAL, if an access unit delimiter NAL is present.

Access units identified by the `'hev1'` sample description may retain filler data (in NAL units or SEI messages) and SEI messages that would change hypothetical reference decoder bitstream conformance if removed.

Access units of type `'hvc1'` shall reference a video parameter set in the sample entry of the CMAF header associated with the containing CMAF track.

B.3.3 Constraints on HEVC elementary streams

B.3.3.1 Overview

The following general constraints apply to all CMAF HEVC elementary streams, and their values are additionally constrained in Clause B.5 with constraints on tier, profile, level, resolution, video characteristics, and frame rates specified by HEVC video CMAF media profile in Table B.1.

B.3.3.2 Picture type

All pictures shall be encoded as coded frames and shall not be encoded as coded fields.

B.3.3.3 Video parameter sets (VPS)

Each HEVC video media sample in the CMAF track shall reference the VPS in the CMAF header sample entry. VPS shall not change within CMAF tracks or between CMAF tracks in a CMAF switching set. A CMAF HEVC track shall conform to ISO/IEC 23008-2 with the following additional constraints.

— The following fields shall have pre-determined values as follows.

  — `general_progressive_source_flag` shall be set to 1.

  — `general_frame_only_constraint_flag` shall be set to 1.

  — `general_interlaced_source_flag` shall be set to 0.

— The condition of the following fields shall not change throughout an HEVC elementary stream:

  — `general_profile_space`

  — `general_profile_idc`

  — `general_tier_flag`

  — `general_level_idc`

B.3.3.4 Sequence parameter sets (SPS)

B.3.3.4.1 SPS fields

Sequence parameter set NAL units that occur within a CMAF HEVC track shall conform to ISO/IEC 23008-2 with the following additional constraints.

— The following fields shall have pre-determined values as follows.

  — `vui_parameters_present_flag` shall be set to 1.

B.3.3.4.2 Visual usability information (VUI) parameters

VUI parameters that occur within a CMAF HEVC track shall conform to ISO/IEC 23008-2 with the following additional constraints.

— The following fields shall have pre-determined values as follows.

  — `aspect_ratio_info_present_flag` shall be set to 1.

  — `video_full_range_flag` shall be set to 0.

— The following fields have the following values.

  — `colour_description_present_flag` should be set to 1.

    NOTE As defined in ISO/IEC 23008-2, if the colour_description_present_flag is set to 1, the colour_primaries, transfer_characteristics and matrix_coefficients fields are present in the VUI.

  — If `colour_description_present_flag` is set to 1, then `colour_primaries`, `transfer_characteristics` and `matrix_coefficients` shall be set to one of the values permitted for the media profile (see Table B.1).

- If `colour_description_present_flag` is set to 0, this shall indicate the following values are to be assumed:
  - `colour_primaries` = 1;
  - `transfer_characteristics` = 1;
  - `matrix_coefficients` = 1.
- `overscan_info_present_flag` shall be set to 0, therefore overscan_appropriate shall not be present.

— The values of the following fields shall not change throughout a CMAF track and CMAF switching set.

- `low_delay_hrd_flag`
- `colour_description_present_flag`
- `colour_primaries`, when present
- `transfer_characteristics`, when present
- `matrix_coeffs`, when present

— The values of the following fields should not change throughout a CMAF track.

- `vui_time_scale`
- `vui_num_units_in_tick`

B.3.3.5 Frame rate in the elementary stream

The frame timing, including frame rate, is determined by the media sample presentation times and durations provided in the `TrackRunBox(es)` in each CMAF fragment.

B.4 Video codec parameters

B.4.1 HEVC signalling of "codecs" parameters

Presentation applications should signal video codec profile and levels of each HEVC track and CMAF switching set using parameters conforming to IETF RFC 6381 and ISO/IEC 14496-15.

B.4.2 Image cropping parameters

When necessary, picture cropping shall be indicated by setting SPS VUI cropping parameters `conf_win_bottom_offset` and/or `conf_win_right_offset` to remove video spatial samples not intended for display, and `conf_win_top_offset` and `conf_win_left_offset` set to zero.

B.5 HEVC video CMAF media profiles and brands

A set of HEVC CMAF video media profiles are defined, together with an ISO BMFF brand and constraints in Table B.1.

For a CMAF track to comply with one of the CMAF media profiles in Table B.1, it:

— shall conform to subclause 9.3 NAL structured video CMAF tracks,

— shall not exceed the tier, profile or level listed in the table

— shall conform to the `colour_primaries`, `transfer_characteristics` and `matrix_coefficients` values from the options listed in Table B.1 with the values defined in ISO/IEC 23008-2.

NOTE ISO/IEC 23091-2[12] may be consulted for additional details on the exact details of the values of these parameters.

— shall not exceed the width, height or frame rate listed in Table B.1, even if the HEVC level would permit higher values.

— shall not exceed the bit depth of the HEVC profile listed.

— should include the CMAF File Brand listed in its CMAF header

NOTE CMAF tracks conform to more than one CMAF media profile if they meet the requirements for multiple CMAF media profiles. For example, a CMAF track would conform to both the HHD8 and UHD8 media profiles if its dimensions were within the constraints for the HHD8 profile (the UHD8 profile has higher dimension constraints and codec level) and the `colour_primaries`, `transfer_characteristics` and `matrix_coefficients` are values which are common to both profiles. HHD8 also conforms to UHD10 because an HEVC Main10 decoder will decode the HEVC Main 8-bit content with Rec 709 video characteristics.

See Annex C for detailed encoding examples.

Table B.1 — HEVC video CMAF media profiles

Media Profile	Codec	Profile	Level	`colour_primaries` in VUI	`transfer_characteristics` in VUI	`matrix_coefficients` in VUI	Max Frame Height	Max Frame Width	Max Frame Rate	CMAF File Brand
HHD8	HEVC	Main MainTier 8-bit	4.1	1 (BT.709) [Note 4]	1 (BT.709 OETF) [Note 4]	1 (BT.709) [Note 4]	1080	1920	60	`'chhd'`
HHD10	HEVC	Main10 MainTier 10-bit	4.1	1 (BT.709) [Note 4]	1 (BT.709 OETF) [Note 4]	1 (BT.709) [Note 4]	1080	1920	60	`'chh1'`
UHD8	HEVC	Main MainTier 8-bit	5.0	1 (BT.709) [Note 4]	1 (BT.709 OETF) [Note 4]	1 (BT.709) [Note 4]	2160	3840	60	`'cud8'`
UHD10	HEVC	Main10 MainTier 10-bit	5.1	1 (BT.709) [Note 4] 9 (BT.2020) [Note 5]	1 (BT.709 OETF) [Note 4] 14 (BT.2020 OETF) [Note 1, 5]	1 (BT.709) [Note 4] 9 (BT.2020 Table 6 $Y'C'_BC'_R$) [Note 7]	2160	3840	60	`'cud1'`
HDR10	HEVC	Main10 MainTier 10-bit	5.1	9 (BT.2100) [Note 5]	16 (BT.2100 Table 4 PQ EOTF) [Note 3, 6]	9 (BT.2100 Table 6 $Y'C'_BC'_R$) [Note 7]	2160	3840	60	`'chd1'`
HLG10	HEVC	Main10 MainTier 10-bit	5.1	9 (BT-2100) [Note 5]	18 (BT.2100 Table 5 HLG OETF) [Note 8] 14 (BT.2020 OETF) [Note2]	9 (BT.2100 Table 6 $Y'C'_BC'_R$) [Note 7]	2160	3840	60	`'clg1'`

NOTE 1 Values 1 and 14 for the `transfer_characteristics` field are functionally equivalent. However, it is normal to use the value which refers to the same specification as the `colour_primaries` being used.

NOTE 2 When the HLG OETF function specified by BT.2100 is used, the `transfer_characteristics` syntax element in the VUI can be set to 14 (BT.2020) and the `alternative_transfer_characteristics` SEI message ISO/IEC 23008-2 included with the `preferred_transfer_characteristics` syntax element of that SEI message set to 18 (HLG) to enable both legacy and HDR-capable clients to display the content.

NOTE 3 HDR10 can optionally include dynamic range metadata such as ST-2086 [11], MaxFALL and MaxCLL.

NOTE 4 This value is equivalent to the definitions in ITU-R BT.709. For details refer to ISO/IEC 23008-2.

NOTE 5 This value is also equivalent to the definitions in ITU-R BT.2020 [9]. For details refer to ISO/IEC 23008-2.

NOTE 6 This value is also commonly also known as ITU-R BT.2100 [10] PQ EOTF. For details refer to ISO/IEC 23008-2.

NOTE 7 This value is commonly also known as ITU-R BT.2020 [9] non-constant luminance. For details refer to ISO/IEC 23008-2.

NOTE 8 This value is equivalent to the definition in ITU-R BT.2100 [10] HLG OETF. For details refer to ISO/IEC 23008-2.

Annex C
(informative)

Subsampling of NAL structured video tracks in CMAF switching sets

C.1 Spatial and temporal subsampling and scaling of video

Spatial and temporal subsampling are encoding methods commonly used to reduce bit rate during adaptive streaming while optimizing video quality.

— Spatial subsampling encodes a fraction of the number of spatial samples in the source video so that the source video is an exact integer multiple. For example, a subsample fraction of 50 % of the source's cropped vertical and horizontal spatial sample count, where the number of source samples is an even number, or a multiple of four when line pair cropping is required. See Clause C.2.

— Temporal subsampling encodes a fraction of the frames in the source video, e.g., 50 % of 24 Hz to result in a 12 Hz track with half the media samples, each of twice the duration, but the same CMAF fragment durations and start times. A CMAF switching set can include CMAF tracks at different frame rates, e.g., 15 Hz and 30 Hz (and perhaps 60 Hz and 120 Hz for UHD), but only if they are all exact multiples of the lowest frame rate.

In order for CMAF switching sets to be presented to viewers with minimal visible disturbance during adaptive bit rate switching, encoding and playback of all the alternative tracks, CMAF specifies the following constraints, which are summarized, but not defined here.

— Tracks in a CMAF switching set are alternative encodings of the same source content (same active video area, colour encoding, source spatial sampling, etc.).

— The number of video spatial subsamples encoded is an exact sub-multiple of the number of video spatial samples in the active area of the source video.

— The player determines the video display aperture, i.e., the number of vertical and horizontal pixels rendered. A player can select a display aperture based on the picture aspect ratio of the CMAF tracks in a CMAF switching set (`'tkhd' width/height`), the available display shape and size or output signal formats (such as HDMI EDIDs), as well as application and user preferences in framing the source aspect ratio within the application determined display area (full screen, windowed, letterboxed, portrait or landscape device orientation, etc.). Precise subsampling and rescaling are specified in this annex to avoid visible changes in image size, position, or shape between alternative CMAF tracks and fragments in a CMAF switching set. Also, different CMAF switching sets within a selection set need to maintain the same aspect ratio to prevent visible distortion when selecting, for example, different camera angles.

— Players are expected to scale the decoded and cropped samples in all tracks in a CMAF switching set to the same display aperture.

— Players are expected to display all tracks in a CMAF switching set at the same refresh rate. The player determines a display refresh rate for the CMAF switching set and maintains that by refreshing each decoded image multiple times if necessary. Frame rate changes are impractical for many video interfaces to smoothly handle, and they may lack operating points for low frame rates.

C.2 Spatial subsampling

Spatial subsampling can be a helpful tool for improving coding efficiency of a video elementary stream. It is achieved by reducing the resolution of the coded picture relative to the source picture, while adjusting the video spatial sample aspect ratio if necessary to compensate for any difference in the

ratio of horizontal and vertical subsampling. For example, by reducing the horizontal resolution of the coded picture by 50 % while increasing the sample aspect ratio from 1:1 to 2:1, the number of encoded video spatial samples is reduced by half. While this does not necessarily correspond to a 50 % decrease in the amount of encoded picture data, the decrease can nonetheless be significant.

The width and height in active image video spatial samples in a coded video sequence are specified by the combination of the following sequence parameter set fields in the video elementary stream or sample entry.

— AVC, ISO/IEC 14496-10:

  — `pic_width_in_mbs_minus1`, which defines the number of horizontal video spatial samples;

  — `pic_height_in_map_units_minus1`, which defines the number of vertical video spatial samples;

  — `aspect_ratio_idc`, which defines the aspect ratio of each video spatial sample;

  — `frame_crop_left_offset`, cropping parameter in SPS;

  — `frame_crop_right_offset`, cropping parameter in SPS;

  — `frame_crop_top_offset`, cropping parameter in SPS;

  — `frame_crop_bottom_offset`, cropping parameter in SPS.

— HEVC, ISO/IEC 23008-2:

  — `pic_width_in_luma_samples`, which defines the number of horizontal video spatial samples;

  — `pic_height_in_luma_samples`, which defines the number of vertical video spatial samples;

  — `aspect_ratio_idc`, which defines the aspect ratio of each video spatial sample;

  — `conf_win_left_offset`, cropping parameter in SPS;

  — `conf_win_right_offset`, cropping parameter in SPS;

  — `conf_win_top_offset`, cropping parameter in SPS;

  — `conf_win_bottom_offset` cropping parameter in SPS.

The presentation size in the video `TrackHeaderBox` is defined in terms of square pixels (i.e., 1:1 video spatial sample aspect ratio) in the `width` and `height` fields of the `TrackHeaderBox` of the video CMAF track. These values are used to determine the appropriate picture aspect ratio when displaying a track. The `width` and `height` in the sample entry by the cropped sample count, which can be converted to square pixels by multiplying the ratio indexed by `aspect_ratio_idc`.

All tracks in a CMAF switching set have the same picture aspect ratio, equal to that of the active image area of the source. A player may ignore the presentation size indicated in each track and scale all fragments to the player-determined presentation aperture for that CMAF switching set.

C.3 Subsample factor and sample aspect ratio

The original picture aspect ratio is conveyed in NAL-structured video by the video spatial sample aspect ratio of the source video as well as the cropped horizontal and vertical sample counts (SAR is the enumerated video spatial sample width to height ratio indexed by the SPS VUI parameter, `aspect_ratio_idc`). The picture aspect ratio is the cropped width divided by the cropped height times the SAR ratio.

Subsampling may change the encoded video spatial sample aspect ratio when it changes the encoded video spatial sample counts and has to be accurately encoded in the VUI information in sequence parameter sets in each fragment of every track. MPEG decoder models include an undefined display processor that uses SPS NAL and VUI information to convert decoded 4:2:0 numerical values to a scaled

image with the SAR, transfer function and colour space indicated in VUI. This annex further defines how multiple tracks in a CMAF switching set are scaled to a common display aperture with a fixed picture aspect ratio. The SAR and cropped video spatial sample counts are defined to equal the same picture aspect ratio as the source video.

The extent of subsampling applied to a track can be characterized by a *subsample factor* in each of the horizontal and vertical dimensions, defined as follows.

— The *horizontal subsample factor* is defined as the ratio of the number of columns of the *luma* video spatial sample array in the source frame after video spatial sample cropping has been applied, divided by the number of columns of the *luma* video spatial sample array after video spatial sample cropping has been applied in subsampled CMAF track. For example, a 1920 wide source image subsampled to 960 horizontal video spatial samples in a subsampled track would have a horizontal subsample factor of 0.5.

— The *vertical subsample factor* is defined as the ratio of the number of rows of the *luma* video spatial sample array in the source frame after video spatial sample cropping has been applied, divided by the number of rows of the *luma* video spatial sample array after video spatial sample cropping has been applied in a subsampled CMAF track. For example, 1088 vertical video spatial samples cropped to 1080 in the source, subsampled to 544 video spatial samples cropped to 540 in a subsampled track would have a vertical subsample factor of 0.5.

C.4 Examples of single dimension subsampling

If a 1920 × 1080 square pixel (SAR 1:1) source picture is horizontally subsampled and encoded at a resolution of 1440 × 1080 (SAR 4:3), which corresponds to a 1920 × 1080 square pixel (SAR 1:1) picture, then the horizontal subsample factor is 1440 ÷ 1920 = 0.75, while the vertical subsample factor is 1.0 since there is no change in the vertical dimension.

Similarly, if a 1280 × 720 (SAR 1:1) source picture is vertically subsampled and encoded at a resolution of 1280 × 540 (SAR 3:4), which corresponds to a 1280 × 720 (SAR 1:1) picture size, then the horizontal subsample factor is 1.0 since there is no change in the horizontal dimension, and the vertical subsample factor is 540 ÷ 720 = 0.75.

C.5 Example of mixed subsampling

If a 1280 × 1080 (SAR 3:2) source picture is vertically subsampled and encoded at a resolution of 1280 × 540 (SAR 3:4), corresponding to a 1920 × 1080 square pixel (SAR 1:1) picture size, then the horizontal subsample factor is 1280 ÷ 1920 = 2/3, and the vertical subsample factor is 540 ÷ 1080 = 0.5. To understand how this is an example of mixed subsampling, it is helpful to remember that the initial source picture resolution of 1280 × 1080 (SAR 3:2) can itself be thought of as having been horizontally subsampled from a higher resolution picture.

C.6 Cropping to active picture area

CMAF players typically control adaptation of the source image to whatever display environment is currently in use. Source content like movies, old TV, and videos from cellphones, etc. have a variety of picture aspect ratios, as do video devices that include phones, tablets, computers, TVs, projectors, and wall displays. In some cases, display aspect ratios will change dynamically when a device like a tablet is rotated from vertical to horizontal orientation, or a video is directed to a different display, and the video aperture may be switched from full screen to a window at any time. A player conforming to the CMAF application model is expected to adapt the source video size and shape to its display environment by methods such as scaling, padding, cropping and stretching, and apply the same adaptation to every CMAF fragment in a CMAF switching set, adjusted for subsampling. Image padding added during production to adapt images to a particular TV aspect ratio like 4:3 or 16:9 (i.e., letterbox bars or pillarbox bars) needs to be removed before encoding in order to allow devices to accurately frame the active picture area.

Since the subsampled picture area might not always fall exactly on the video spatial sample coding unit boundary employed by the video elementary stream, additional cropping parameters are used to further define the dimensions of the coded picture. It is a normative requirement of AVC and HEVC that decoders perform cropping as signalled in sequence parameter sets (SPS NALs).

— AVC (ISO/IEC 14496-10):

  — "Macroblocks" define the video spatial sample coding unit boundary (and are 16 × 16 blocks).

— HEVC (ISO/IEC 23008-2):

  — "Coding Tree Units" define the video spatial sample coding unit boundary (and are 64 × 64 blocks, 32 × 32 blocks, or 16 × 16 blocks). See Clause C.2.

C.7 Relationship of cropping and subsampling

When spatial subsampling is applied, additional cropping parameters are often needed to compensate for the mismatch between the coded picture size and the macroblock (ISO/IEC 14496-10)/coding tree unit (ISO/IEC 23008-2) boundaries. The specific relationship between these mechanisms is defined as follows.

— Each picture is decoded using the coding parameters, including horizontal and vertical sample counts and cropping fields, defined in the sequence parameter set corresponding to that picture's coded video sequence.

— The display aperture is determined by the CMAF player, and each fragment scaled to fill that aperture using the same method to maintain registration, e.g., common sides, common top and bottom, exact match, etc. For example, to output the video to an HDTV, the decoded image might need to be scaled to the display aperture width, then additional letterbox matting applied to match a valid HDMI television input format. A newer TV or projector might accept this picture aspect ratio directly without padding.

C.8 Example encoding and decoding process

The following example shows a typical movie picture aspect ratio that was padded and encoded with letterbox bars to fit a 16:9 TV display. The active image is extracted, subsampled, encoded, and partially filled macroblocks indicated by cropping parameters. AVC and HEVC parameters for the example are detailed in Figure C.1.

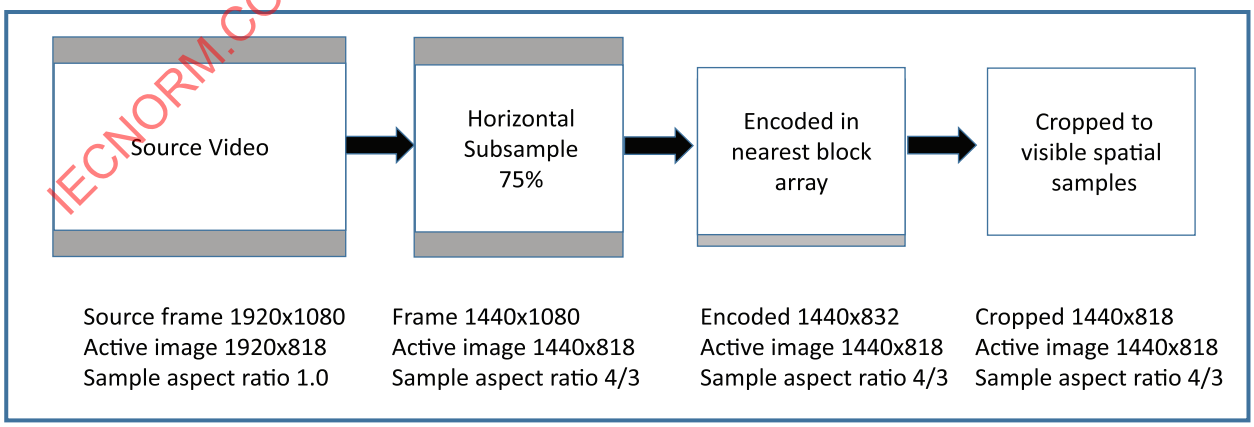


Figure C.1 — Example encode and decode process for letterboxed source content

Table C.1 shows the parameter values used.

Table C.1 — Subsample and cropping values for the example in Figure C.1

Object	Field	Value
Source picture format	width	1920
	height	1080
Subsample factor	horizontal	0.75
	vertical	1.0
`TrackHeaderBox`	width	1920
	height	818
ISO/IEC 14496-10 parameter values	chroma_format_idc	1 (4:2:0)
	aspect_ratio_idc	14 (4:3)
	pic_width_in_mbs_minus1	89
	pic_height_in_map_units_minus1	51
	frame_cropping_flag	1
	frame_crop_left_offset	0
	frame_crop_right_offset	0
	frame_crop_top_offset	0
	frame_crop_bottom_offset	7
ISO/IEC 23008-2 parameter values	chroma_format_idc	1 (4:2:0)
	MinCbSizeY	16
	log2_min_luma_coding_block_size_minus3	1
	aspect_ratio_idc	14 (4:3)
	pic_width_in_luma_samples	1440
	pic_height_in_luma_samples	832
	conformance_window_flag	1
	conf_win_left_offset	0
	conf_win_right_offset	0
	conf_win_top_offset	0
	conf_win_bottom_offset	7

NOTE 1 As `chroma_format_idc` is 1, `SubWidthC` and `SubHeightC` are set to 2 per ISO/IEC 14496-10 and ISO/IEC 23008-2. This results in a doubling of frame crop parameters (so `frame_crop_bottom_offset` and `conf_win_bottom_offset` both equate to 14 pixels in the above example).

NOTE 2 As ISO/IEC 23008-2 `MinCbSizeY` is 16 and `log2_min_luma_coding_block_size_minus3` is 1, the Coding Tree Unit size is 16 × 16 (matching the ISO/IEC 14496-10 macroblock size of 16 × 16).

The decoding and display process for this content is illustrated in Figure C.2.

In this example, the decoded picture dimensions are 1440 × 818, one line larger than the original active picture area. This is due to a limitation in the cropping parameters to crop only even pairs of lines. Cropping the source picture to 816 lines and 51 macroblocks might be more practical, but makes a less informative example.

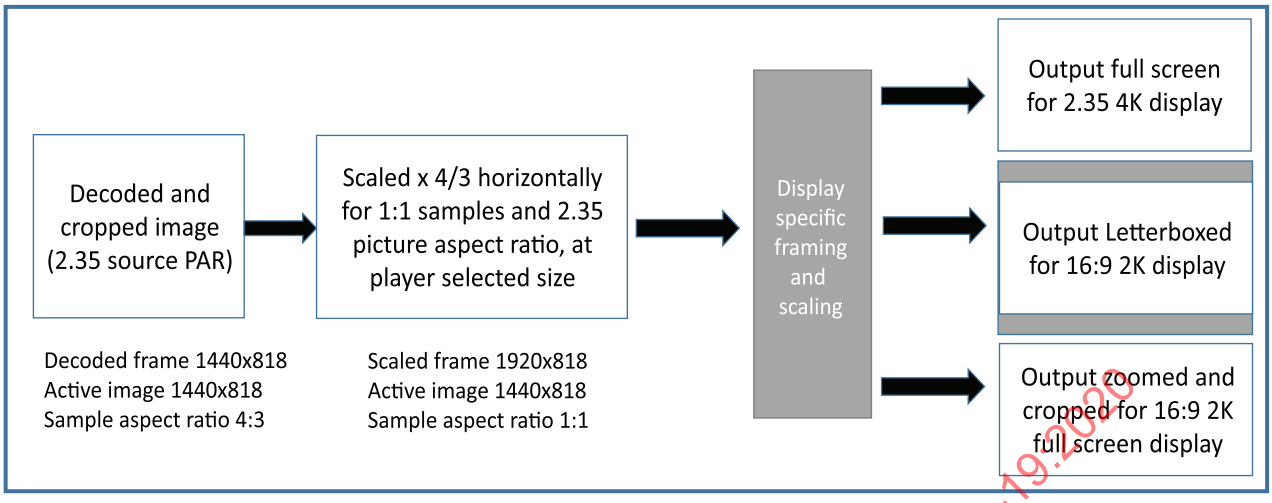


Figure C.2 — Example display process with and without letterbox bars added

Figure C.3 illustrates both subsampling and cropping applied to the horizontal dimension while encoding pillarboxed content.

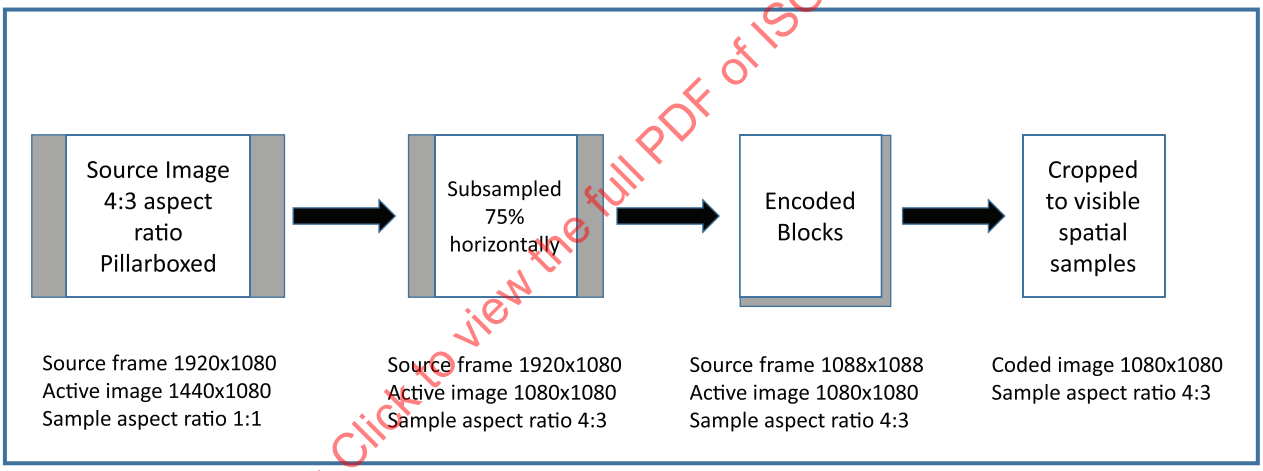


Figure C.3 — Example encode and decode process for pillarboxed source content

The original source picture content is first subsampled horizontally from a 1:1 sample aspect ratio at 1920 × 1080 to a video spatial sample aspect ratio of 4:3 at 1440 × 1080, then the 1080 × 1080 pixel active picture area of the subsampled image is encoded. However, the actual coded video spatial samples have a size of 1088 × 1088 samples due to the coding unit boundaries do not falling on even multiples of 16 video spatial samples in this example. Therefore, cropping parameters are provided in both horizontal and vertical dimensions. The parameters are shown in Table C.2.

Table C.2 — Subsample and cropping values for example in Figure C.3

Object	Field	Value
Source picture format	width	1920
	height	1080
Subsample factor	horizontal	0.75
	vertical	1.0
`TrackHeaderBox`	width	1440
	height	1080
ISO/IEC 14496-10 parameter values	chroma_format_idc	1 (4:2:0)
	aspect_ratio_idc	14 (4:3)
	pic_width_in_mbs_minus1	67
	pic_height_in_map_units_minus1	67
	frame_cropping_flag	1
	frame_crop_left_offset	0
	frame_crop_right_offset	4
	frame_crop_top_offset	0
	frame_crop_bottom_offset	4
ISO/IEC 23008-2 parameter values	chroma_format_idc	1 (4:2:0)
	MinCbSizeY	16
	log2_min_luma_coding_block_size_minus3	1
	aspect_ratio_idc	14 (4:3)
	pic_width_in_luma_samples	1088
	pic_height_in_luma_samples	1088
	conformance_window_flag	1
	conf_win_left_offset	0
	conf_win_right_offset	4
	conf_win_top_offset	0
	conf_win_bottom_offset	4

NOTE 1 As `chroma_format_idc` is 1, `SubWidthC` and `SubHeightC` are set to 2 per ISO/IEC 14496-10 and ISO/IEC 23008-2. This results in a doubling of frame crop parameters (so `frame_crop_bottom_offset` and `conf_win_bottom_offset` both equate to 14 pixels in the above example).

NOTE 2 As ISO/IEC 23008-2 `MinCbSizeY` is 16 and `log2_min_luma_coding_block_size_minus3` is 1, the Coding Tree Unit size is 16 × 16 (matching the ISO/IEC 14496-10 macroblock size of 16 × 16).

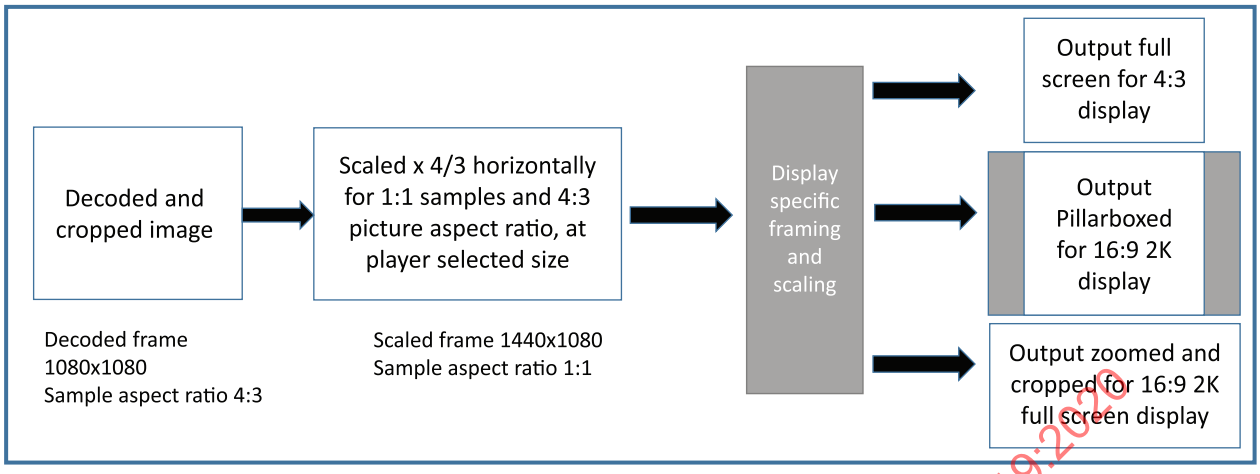


Figure C.4 — Example display process with and without pillarbox bars

The process for reconstructing the video for display is shown in Figure C.4.

As in the previous example, the decoded picture has to be scaled back up to the original 1:1 video spatial sample aspect ratio.

If this content is to be displayed on a 4:3 display, no further addition of pillarboxes is necessary to display the video image full screen. A user might elect to zoom the image to full width on a 16:9 display while cropping the top and bottom. Full frame 1.85 aspect ratio film often "protects" the 1.78 aspect ratio area (16:9) so that the top and bottom can be cropped when displayed full width on HDTVs without loss of significant content.

C.9 CMAF switching set video encoding considerations

C.9.1 General

Some of the variables that should be considered when designing a CMAF switching set for an application are:

a) valid CMAF framing, subsampling, and media profile encoding of multiple tracks from the same source video to avoid visible framing, position, and scaling errors on track switches;

b) the minimum and maximum bandwidth available for delivery;

c) the resolution of the source video;

d) the resolution and viewing conditions of target devices;

e) the compression efficiency of the codec and encoder;

f) the compression difficulty of each content source;

g) the number of CMAF tracks per CMAF switching set;

h) the bit rate difference ("gradient") between adjacent bit rate CMAF tracks;

i) the frequency of adaptive switching;

j) longer CVSs for better compression efficiency;

k) shorter CVSs and CMAF fragments for lower latency;

l) the size of the installed base for different codecs and media profiles;

m) structuring CMAF switching sets and aligned CMAF switching sets for content protection requirements.

Some general design choices and two different use cases can illustrate some of the tradeoffs based on prioritizing these parameters.

C.9.2 General configuration

The maximum bit rate that is optimal for a CMAF switching set will depend on the intended display resolution and video quality, as well as encoder and codec efficiency, and difficulty of the source material.

Network and device characteristics may determine the lowest bit rate and resolution of interest. Bit rates for small screens on variable cellular connections could be much lower than what would look acceptable on a connected UHD TV. Some codecs and CMAF media profiles would limit the installed base of capable decoders, unless the CMAF presentation also includes "fallback" CMAF media profiles such as HD AVC and AAC stereo, which are almost universally supported. Content protection and DRM tend to limit the number of compatible devices at UHD and HDR media profiles because they typically have higher device security requirements. Alternative CMAF switching sets encoded with different CMAF media profiles provide more reach at the cost of more encoding, storage, content management, caching, player intelligence, etc.

C.9.3 Use case 1 — High volume "video on demand" of long form content, e.g., movies

Prioritization of efficient CDN caching favours fewer tracks per CMAF switching set for better CDN utilization of finite edge cache storage and higher cache hit ratio with fewer alternative choices. Fewer tracks per CMAF switching set means a larger bit rate gradient between tracks for a given maximum and minimum bit rate. Somewhere around 40 % to 50 % gradient is a typical tradeoff ("gradient" is the percentage reduction in bit rate from each track to the next lower track). Video subsampling should reduce encoded resolution in rough proportion to the bit rate reduction to optimize video quality. The bit rate gradient should be consistent to have approximately the same quality difference on all track switches.

This large a gradient will probably result in noticeable quality differences between tracks, so a long duration player buffer (e.g., 2+ min) and low bit rate startup sequence to fill that buffer can be used with a slow adaptive switching algorithm to avoid frequent switching without risking buffer underrun (a long buffer can average out network throughput variations without underflowing the buffer.). Infrequent track switching makes it less critical to avoid processing a CMAF header on each track switch, which might result in frame drops, etc. on some systems.

An additional option is to package CMAF track files and store them near edge caches, then use byte range addressing to request portions of each track file for delivery (large CMAF segments). The byte ranges should not be cached because the responses typically duplicate content with slightly different start and end points so do not provide reuse and do not need to be cached because each byte range can be copied from one local track file without waiting for new segments from an origin server.

C.9.4 Use case 2 — Low latency live streaming

Prioritization of low latency live delivery can limit the duration of coded video sequences and CMAF fragments at the cost of coding efficiency or maintain efficient long duration coded video sequences, but reduce encoder removal and packaging delay by delivering short CMAF chunks whenever, for example, a half-second sequence of video media samples has been encoded and packaged in a half-second CMAF chunk. A half-second CMAF chunk can be delivered in less than a half second and can immediately start decoding (assuming the first CMAF chunk contains the first media samples of a CMAF fragment).

A player needs to start up with a time delay relative to the live edge and buffer size that is sufficient to avoid buffer underflow for the entire presentation (possibly with ad insertion, etc.), so a few seconds of buffer and delay from live may still be required.

To maximize bit rate for a given network, the rate adaptation algorithm needs to respond quickly to shrinking buffer fullness to avoid underflow and rebuffering, or make inefficient use of available bandwidth if it does not switch up when it can. Maintaining a short buffer results in frequent track switching, possibly as frequent as once per CMAF fragment duration, e.g., every 2 seconds to 4 seconds. If bit rate changes and video quality differences can be kept below a "least noticeable difference", then frequent switching can be done with consistent perceived video quality. A bit rate gradient of around 30 % can approximate a least noticeable difference for a typical AVC or HEVC encoder operating in a reasonable bit rate range relative to image size that avoids extreme artefacts due to insufficient bit rate for the encoded resolution.

A smaller bit rate gradient for a given minimum and maximum bit rate requires more CMAF tracks per CMAF switching set. Reliability for low latency delivery may require redundant origin servers and edge servers, possibly with multiple encoders and source streams to avoid latency and jitter on a single network node or path that would not be noticed under typical high latency delivery, e.g., several tens of seconds, but would cause streaming interruption when operating at low latency. Early requests either are to be avoided, or queued and handled, to avoid HTTP 400 errors such as "not found" or "not ready", that would flood the origin server.

Live encoding work flow often involves the splicing of independently encoded signal feeds, ads, and other recorded content and M2TS signal feeds that can change encoding parameters in any resulting CMAF fragment. CMAF switching sets with inband parameters and single initialization constraints can process CMAF fragments without needing to download and process a new CMAF header with new decoding parameters.

Encoding strategy, network distribution strategy, and player optimization need to be combined for optimum results.

Annex D
(informative)

Hypothetical player model

D.1 Overview

Player implementers should note that CMAF provides the following affordances.

— A CMAF segment is well-suited to network transfer because it is a compact, self-contained set of media samples that covers a short duration that can be downloaded relatively quickly.

— Each CMAF segment contains a media timestamp in the form of `baseMediaDecodeTime` in a `TrackFragmentBaseMediaDecodeTimeBox`, which allows individual CMAF segments and CMAF tracks that start with arbitrary timestamps to be synchronized to a presentation timeline using a manifest presentation time offset.

— The CMAF header, associated with a CMAF track, and the `MovieFragmentBox` and the media sample data referenced from it contain everything necessary to render the media samples at the correct place on the presentation timeline.

— CMAF tracks can be separately selected and delivered by a player, then synchronized at presentation time ("late bound"), thus allowing each player to customize different CMAF track combinations for each device, network, and user.

— The independently decodable picture at the start of each video CMAF fragment can be obtained with reduced download size for fast forward or fast reverse streaming, using a byte range request for a partial CMAF segment, or some other protocol.

— CMAF selection sets allow alternative content to be offered for playback without requiring the player to transfer content that it does not intend to play.

— CMAF switching sets allows alternative bit rate and resolution encodings of the same content to be seamlessly presented through a single video decoder.

— The use of Common Encryption supports access to the same encrypted content by multiple decryption key delivery systems.

— The `DASHEventMessageBox` allows CMAF addressable media objects to signal application-defined track events with low latency on CMAF addressable media objects already being downloaded during live presentations without requiring frequent manifest downloads.

D.2 Adaptive streaming playback examples

To play an adaptive streaming presentation, a player typically:

— parses the manifest and selects selection sets of media types it can present on that device (e.g., audio only, or audio/video/subtitles/picture in picture, etc.);

— compares CMAF switching set information, such as the CMAF media profile, to player, decoder, display, DRM, etc. capabilities to determine the compatible CMAF switching sets in those selection sets;

— selects the most preferred compatible CMAF switching set in each selection set, sometimes based on stored user preferences (language, accessibility, rating, stereo or multichannel audio, etc.);

— selects an initial CMAF track from each selected CMAF switching set, usually based on estimated network bandwidth, rapid start heuristics, display size, etc.;

— sequentially downloads the CMAF addressable media objects that CMAF track is packaged in;

— initializes, decodes, synchronizes, and presents the selected CMAF tracks and automatically switches the next requested CMAF track and bit rate if needed to maintain continuous playback within the limitations of network throughput.

D.3 Live streaming

Live CMAF presentation streaming is typically optimized for low latency by a player that only buffers a few seconds of each selected CMAF track to minimize presentation delay relative to the time content arrives at the encoder. A player can select a different bit rate for each CMAF segment if necessary to prevent buffer underflow in the player while maximizing media quality. In addition, some servers and players can use short CMAF chunks as CMAF addressable media objects to receive media samples from the encoder in a fraction of the time necessary for the encoder to complete a coded video sequence and make it available in a CMAF segment. Bit rate switching is still limited to CMAF fragment boundaries, i.e., the first CMAF chunk of a CMAF fragment. A player can reduce average bit rate in order to allow a short, low latency buffer while avoiding underflow during network and content variations, or increase the buffer duration and presentation latency to operate with more buffer underflow safety margin or higher bit rate.

Once a player selects the live presentation delay and buffer duration, the delay cannot be changed without halting playback and rebuffering, or decoding at a speed faster or slower than normal, which is usually not possible or acceptable. Measurements of network latency, jitter, throughput rate, throughput variation, CMAF segment duration variation, and server/client clock synchronization can help a player select an optimal presentation delay and next CMAF segment bit rate to request. Each measurement of network throughput made by a player is influenced by variable factors, such as local network congestion, e.g., other devices periodically downloading data or media segments at the same time, CMAF segments being cached on a local CDN node or not, variations in wireless network reception, etc. Some players can take advantage of network aware servers that can provide each player data to help those players optimize bit rate and buffer duration for reliable low latency playback.

Another factor in low latency streaming is the sequencing of multiple CMAF presentations in a manifest. Players are often required to insert short video advertisements in longer video presentations, like television shows. If a player uses a single buffer and decoder for all CMAF selection sets of one media type, then the presentation delay and buffer duration necessary for ads could be much larger than that needed for the live television show and harder to determine in advance. That is particularly true when ads are delivered from a different server and are selected during playback based on the viewer, device, location, etc. In this case, the player needs to select presentation delay and buffer duration sufficient for the worst-case ad delivery delay, and the CMAF fragments in the television show need to be time aligned to the duration of each advertisement in order to resume the live show on the start of a CMAF fragment without changing presentation delay. An alternative is to use separate buffers and decoders for the show and ads, in which case the presentation delays are decoupled. Prerecorded ads can be prefetched well in advance, and the return to the live show can happen on any frame by switching the view, since the show can be continuously buffering and decoding during ad playback.

To minimize visible changes between CMAF segments encoded at different bit rates, live CMAF switching sets may include more tracks with smaller bit rate differences, for instance, a decrease of 30 % to the next lower bit rate track. If changes in video quality between adjacent CMAF tracks in a CMAF switching set are kept below a “just noticeable difference”, then a player can use frequent bit rate switching and it will not be apparent to viewers.

In practice, a player processing a CMAF header on each bit rate switch is often not seamless because of the different behaviour and response times in Web browsers, decoders, etc. when processing a CMAF header and reinitializing some or all decoding and display parameters. CMAF switching sets with single initialization constraints provide more reliable seamless switching, so are particularly useful for low latency live streaming applications where frequent bit rate switching is used.