

**INTERNATIONAL ORGANISATION FOR STANDARDISATION  
ORGANISATION INTERNATIONALE DE NORMALISATION  
ISO/IEC JTC1/SC29/WG11  
CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC1/SC29/WG11 N14751  
July 2014, Sapporo, Japan**

**Title**      **White paper on AAC Transport Formats**  
**Source**     **Audio Subgroup**

## **AAC Transport Formats**

### ***1 AAC Transport Protocols and File Formats***

As the MPEG-4 standard states, “In all of the MPEG-4 tools for audio coding, the coding standard ends at the point of constructing access units that contain the compressed data.” Since audio may be used alone or combined with video, and may be streamed or stored, several transport and storage formats have been developed.

Transport formats are usually specified so that a receiver may synchronize and decode a bitstream that is already being transmitted, while storage formats require reading from the beginning to understand a file. Of course, a transport stream can also be stored as a file, but it will likely take up more storage due to its synchronization or framing overhead, and may not support the metadata or indexing features present in storage formats.

Another classification is whether the format is specified solely for audio or may include video or other content types. Audio-only formats have less processing and transmission overhead and are popular for streaming music to mobile phones, for example.

Historically, the AAC formats ADIF and ADTS were defined in MPEG-2 part 7 as the original transport and storage formats for AAC. MPEG-4 introduced two additional formats LATM and LOAS, which are defined not only for traditional AAC, but also for newer variants such as AAC-LD and AAC-ELD v2.

Streaming of multimedia content, combining AAC with MPEG-4 part 2 or AVC video, is typically done using RFC 3640 protocol over IP networks, and using MPEG-2 Program Elementary Streams in MPEG-2 Transport Streams over DVB networks. A dichotomy of standards has developed because of the differences in architecture of one-way broadcast cable and satellite networks and the two-way unicast packet transport of IP networks. This is made more complex by the possibility of transporting an IP stream over an MPEG-2 Transport Stream, or of sending a MPEG-2 Transport Stream over IP. There are also standards for home media networks that stream audio content over HTTP.

In most application areas, the MPEG-4 File Format or the closely related 3GPP file format are used for storage. For example, iPod and iTunes content is based on the MP4 file format.

Transport formats are used in broadcasting, Internet radio, teleconferencing, and live streaming applications where receivers or player software can synchronize to a common bitstream broadcast to all users.

Although the web media industry may loosely refer to it as streaming, most on-demand new media content is not streamed but is progressively downloaded in a storage format. Here progressive merely means the storage format has been constrained to require all of the pointers or metadata to be at the beginning of the file, so that decoding and playback may begin while part of the file is still downloading. Later implementations, often called HTTP streaming, use the HTTP Get Range request to download parts of a file. This allows random access or “trick play” of the file’s content. This has recently been extended to “adaptive streaming” (see *3 Adaptive Streaming Protocols*)

where downloading is requested in small segments of a file to allow adaption to changes in the available network bandwidth as the content is viewed.

It should be understood that transport formats cause the receiving device's playback rate to be synchronized to the time base of the transmitter or server, while in downloading or HTTP streaming, the content is requested from a server by the receiver, and the playback rate is asynchronous to the server or original source.

## 2 Comparison Table of AAC Transport Formats

The following table shows the important features and defining standards of each AAC transport or storage format:

Protocol/ Format	Type	Defined by	Features	Typical Use
ADTS (Audio Data Transport Stream)	Transport Audio Only	MPEG-2 ISO/IEC 13818-7 [1] (also in MPEG-4)	Self-contained bitstream with Sync information, originating from MPEG-2 AAC.	Transport for MPEG-2 AAC Sometimes used for mobile handsets and devices, since parsing MP4FF has been quite demanding memory-wise in the past. Some mobile phones that are able to play AAC-LC or HE-AAC support only ADTS.
ADIF (Audio Data Interchange Format)	Storage Audio Only	MPEG-2 ISO/IEC 13818-7 (also in MPEG-4)	One header at the beginning of the file containing decoder-specific information followed by subsequent audio-access-units without further sync-information. No information about the position or length of access units available.	Lowest-overhead format for AAC access units. Since no random access and no multiplexing of other MPEG-4 data are possible, it is rarely used.
LATM (Low-overhead MPEG-4 Audio Transport Multiplex)	Transport Audio Only	MPEG-4 ISO/IEC 14496-3 [2]	Self-contained bitstream. Allows the use of the Error-Resilient Syntax from MPEG-4.	Used in 3GPP as transport for HE-AAC v2 (without the LOAS Sync-Layer). Random access is not possible (in a stored bitstream).
LATM/LOAS (Low Overhead Audio Stream)	Transport Audio Only	MPEG-4 ISO/IEC 14496-3	LATM with Sync information, which allows random access or skipping. May include MPEG-4 Error Protection and Resilience. Self-contained bitstream.	'Bitstream' format for AAC-Low Delay (since LD is only available in the Error Resilient variant) and xHE-AAC. Also used when AAC is carried in a MPEG-2 transport stream.
RFC 3016	Transport Audio/Video	IETF	Carries MPEG-4 Audio LATM Packets and MPEG-4 Video Packets in RTP (Real Time Protocol) streams. The RTP streams can be audio or video.	3GPP streaming, also used for video conferencing
RFC 3640 (RTP Payload Format for Transport of MPEG-4 Elementary Streams)	Transport Audio/Video	IETF	Carries MPEG-4 elementary streams including MPEG-4 Audio as raw Access Units in a RTP (Real Time Protocol) stream.	ISMA 2.0 streaming and in high-quality videoconferencing e.g. TIP. N/ACIP

Protocol/ Format	Type	Defined by	Features	Typical Use
MPEG-2 Transport Stream	Transport Audio/Video	ISO/IEC 13818-1	PES packets containing either ADTS, LOAS, or MPEG-4 Sync Layer stream.	DVB [ETSI TS 101 154] [3] uses the LOAS variant
MP4FF (MP4 File Format)	Storage Audio/Video	MPEG-4 ISO/IEC 14496-12 [4],14 [5]	MPEG-4 File Format – Storage of Audio and Video streams	Format for storage and as a storage container for IP-streaming from files. Used for iTunes and many Flash and YouTube files. Allows true random access.
3GPP File Format	Storage Audio/Video	ETSI TS 126 244	Very similar to MPEG-4 File Format, but includes support for non-MPEG codecs such as H.263 and GSM-AMR	
Application Standards:				
ShoutCast	Transport Audio Only	AOL [6]	HTTP Streaming of ADTS.	Internet Radio
Flash Video File Format	Storage Audio/Video	Adobe [7]	Some versions of the older FLV format may contain AAC bitstreams, but newer players also play MP4 files with a FLV or (preferred) F4V suffix	YouTube and other media websites

Table 1. AAC Transport and Storage Formats

### 3 Adaptive Streaming Protocols

Many transmission channels, such as mobile data networks, offer time-varying bandwidth or throughput. Four methods have evolved to send audio streams over such channels.

The simplest method is to use a bitrate that is supported under all expected network conditions. This might be determined by measuring the network bandwidth during the initialization of the stream, or a bitrate value that is known to work through operational experience may be used.

If the audio signal is being encoded as it is being streamed, another method is to adjust the bitrate of the encoder to match an estimate of the bitrate available during a future time interval. This time interval might range from a single audio frame to several seconds, depending on the network and the estimation technique employed.

For content that is stored as encoded files or bitstreams on a server, two techniques of varying the bitrate have been developed.

The most common method is stream switching, which uses several stored versions of a file, each encoded at a different bitrate. An estimate of the available bandwidth is taken and the version with the highest bitrate not exceeding the estimate is sent. Typically, the bandwidth estimates and streams are updated every few seconds.

With stream switching, the streams and decoder operation must be coordinated to insure that transitions between streams do not result in noticeable transients.

Several proprietary protocols for stream switching have been developed, such as Apple's HTTP Live Streaming [8], Adobe's Dynamic Streaming [9], and Microsoft's Smooth Streaming [10]. A similar protocol is being standardized by the MPEG DASH project [11]. All of these protocols are similar in operation, with the primary difference being that HTTP Live Streaming encapsulates the audio data in a MPEG-2 Transport Stream, while the others use the MPEG-4 File Format. The DASH specification will support both options.

Another technique is available with codecs that include a fine-grained scalability feature, such as HD-AAC. Because the enhancement layer of HD-AAC is encoded in a special bit-sliced manner, a server can very efficiently remove parts of the bitstream to vary the bitrate continuously as a file is being served.

#### 4 AAC Tools, Audio Object Types, Profiles, and Levels

The MPEG-4 Standard includes several versions of AAC, such as AAC-LC, HE-AAC and xHE-AAC, the low-delay versions AAC-LD and AAC-ELD v2, and other codecs such as the speech codec CELP. The standard defines a hierarchy of Tools, Audio Object Types, and Profiles to specify these codecs.

While MPEG-4 offers several Profiles to specify useful bundles of these codecs, the primary means of specifying a codec is by the Audio Object Type number of the tools it employs. Table 2 shows the most popular Audio Object Types for MPEG-4 audio.

AOT ID	Audio Object Type
AAC-LC	2
SBR	5
PS	29
AAC-LD	23
AAC-ELD	39
MPEG Surround	30
SLS (HD-AAC)	37
SLS (no core)	38
USAC (xHE-AAC)	42
Low Delay MPEG Surround (LD-mps)	44

Table 2. Popular MPEG-4 Audio Object Types

Some companies have used trade names for their implementations of AAC, as shown in Table 3. All the names refer to the same MPEG codec and are interoperable. HE-AAC is internally the combination of the SBR (spectral band replication) and AAC-LC Audio Object Types, while HE-AAC v2 is the combination of both these and the PS (parametric stereo) Audio Object Type.

	AAC-LC	HE-AAC	HE-AAC v2	xHE-AAC
Fraunhofer	AAC-LC	HE-AAC	HE-AAC v2	xHE-AAC
MPEG	AAC LC	HE AAC	HE AAC v2	Extended HE AAC
3GPP		aacPlus	Enhanced aacPlus	
Coding Technologies		aacPlus	aacPlus v2	
Dolby			Dolby Pulse	
MPEG Audio Object Type	2 (AAC-LC)	2 (AAC-LC), 5 (SBR)	2 (AAC-LC), 5 (SBR), 29 (PS)	2 (AAC-LC), 5 (SBR), 29 (PS), 42 (USAC)
MPEG Profile	AAC	High Efficiency AAC	High Efficiency AAC v2	Extended High Efficiency AAC

Table 3. Equivalent Trade Names and Profiles for popular MPEG Audio Codecs

Many profiles include level definitions to specify the maximum computational performance of decoders. The level definitions for HE-AAC v2 are shown in Table 4. Level definitions may have complex restrictions, as noted by the 24/48 entry in the table for level 3 and 4. The reader is advised

to refer to the standard [2] for the details.

Level	Maximum number of channels	Maximum AAC Sample Rate, no SBR	Maximum AAC Sample Rate, with SBR
1	n/a	n/a	n/a
2	2	48	24
3	2	48	24/48
4	5	48	24/48
5	5	96	48
6	7	48	24/48
7	7	96	48

Table 4. Level Definitions for the HE-AAC v2 Profile

### 5 The AudioSpecificConfig() Structure

In the transport or storage of MPEG-4 audio, the audio object types that must be decoded, as well as the fundamental audio parameters such as sampling rate, frame length [12], and the audio channel arrangement, are usually specified in the AudioSpecificConfig() data structure. The ASC allows understanding these parameters without parsing the AAC bitstream and is useful during codec negotiation such as for SIP or SDP setup.

MPEG-2 does not specify the ASC, so the formats ADIF and ADTS have a fixed frame length of 1024 samples, and in practice only the audio object type 2.

The ASC is not defined as a fixed-field structure, but by a pseudo-code description. The ASC consists of two parts, a generic one holding information common to most of the MPEG-4 Audio audio object types, and a second part with information specific to the audio object type such as the frame length. For AAC-LC, HE-AAC and AAC-LD, the second part is termed GASpecificConfig() in the standard (for detailed syntax see ISO/IEC 14496-3:2009, Table 4.1 “Syntax of GASpecificConfig()”), while for AAC-ELD the second part is named ELDSpecificConfig() (see ISO/IEC 14496-3:2009, Table 4.180 “Syntax of ELDSpecificConfig ()”). For xHE-AAC (USAC) this second part is called UsacConfig() (ISO/IEC 23003-3:2012 [13], Table 4).

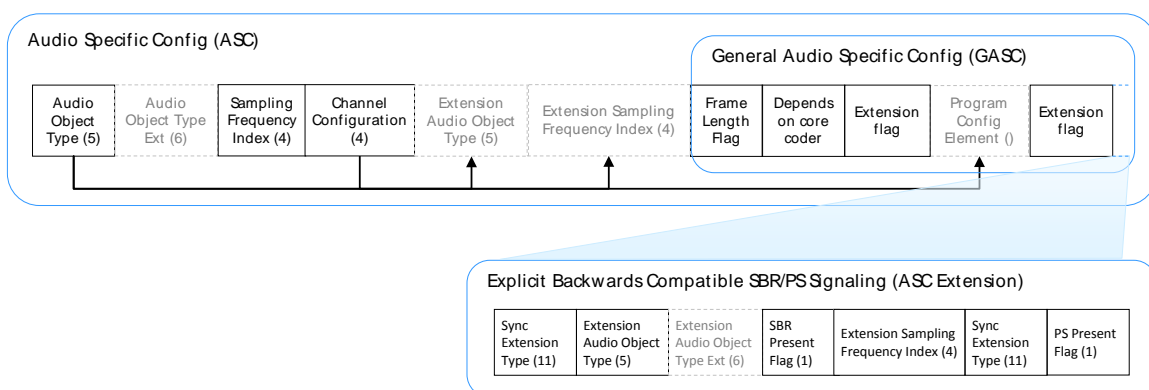


Figure 1. Audio Specific Config Bitstream for HE-AAC v2

If the value of the channel configuration field is zero, the GASpecificConfig() or GASC contains a Program Config Element (PCE) structure [14]. Its primary function is to specify the number and arrangement of audio channels in the bitstream. A list of Element Instance Tags (see 6 Raw Data Blocks or Access Units (the Payload or Audio Bitstream)) is given starting from the center front

channel, if any, and proceeding outwards to the remaining front channels, then the side channels, then the rear channels, and finally any rear center channel, in the order they are arranged in the listener's playback configuration.

### **6 Raw Data Blocks or Access Units (the Payload or Audio Bitstream)**

Each of the MPEG-4 transport or storage formats eventually consists of raw data blocks (MPEG-2 nomenclature) or access units (MPEG-4) that contain the actual bitstream produced by the audio encoder for an audio frame. The bitstream is divided into portions representing different audio channels in a flexible manner. For example, the data for a single audio channel is carried in the Single Channel Element (SCE). For stereo bitstreams, a Channel Pair Element (CPE) allows data for two channels to be combined so that joint coding information can be conveyed. 5.1 multichannel bitstreams usually contain an SCE for the center channel, two CPEs for the front and rear stereo pairs, and an LFE element for the LFE channel. Since the length of the data for an audio frame may be calculated from the number of audio samples in the frame length, the data is parsed without further synchronization information.

<b>Channel configuration</b>	<b>included channel elements</b>	<b>channel to speaker mapping</b>
0	-	defined in AOT related SpecificConfig (not available in ER bitstream syntax)
1	single_channel_element()	center front speaker
2	channel_pair_element()	left, right front speakers
3	single_channel_element() channel_pair_element()	center front speaker, left, right front speakers
4	single_channel_element() channel_pair_element() single_channel_element()	center front speaker, left, right center front speakers, rear surround speakers
5	single_channel_element() channel_pair_element() channel_pair_element()	center front speaker, left, right front speakers, left surround, right surround speakers
6	single_channel_element() channel_pair_element() channel_pair_element() lfe_channel_element()	center front speaker, left, right front speakers, left surround, right surround speakers, low frequency effects speaker
7	single_channel_element() channel_pair_element() channel_pair_element() channel_pair_element() lfe_channel_element()	center front speaker left, right front center speakers, left, right front speakers, left surround, right surround speakers, low frequency effects speaker
8 – 10	-	reserved

Channel configuration	included channel elements	channel to speaker mapping
11	single_channel_element(), channel_pair_element(), channel_pair_element(), single_channel_element(), lfe_element()	center front speaker, left, right front speakers, left surround, right surround speakers, rear center speaker, low frequency enhancement speaker
12	single_channel_element(), channel_pair_element(), channel_pair_element(), channel_pair_element(), lfe_element()	center front speaker left, right front speakers, left surround, right surround speakers, rear surround left, right speakers, low frequency enhancement speaker
14	single_channel_element(), channel_pair_element(), channel_pair_element(), lfe_element(), channel_pair_element()	center front speaker, left, right front speakers, left surround, right surround speakers, low frequency enhancement speaker, left, right front vertical height speakers
15	-	reserved

Table 5. channelConfiguration for AAC bitstream syntax

### 6.1 HE-AAC payload

An AAC access units may also contain a PCE element [15] (see 5 *The AudioSpecificConfig()* Structure), Fill elements (FIL) to pad the bitstream for maintaining an instantaneous bitrate, and Data Stream Elements (DSE) for user data and for DVB-specific metadata. The FIL element is discarded by an AAC-LC decoder and is also used to contain the SBR data used by an HE-AAC decoder, the metadata used for loudness normalization and dynamic range control, and the hidden spatial bitstream for MPEG Surround coding. A specialized element, the Coupling Channel Element (CCE) is not used in practice.

The access unit for a mono bitstream consists of:

- The ID code for SCE
- An instance number (0)
- The AAC compressed audio data for the frame
- The ID for end of a frame – TERM

A detailed technical description of the AAC payload syntax can be found in ISO/IEC 14496-3:2009, section “4.5.2.1 Top level payloads for the audio object types AAC main, AAC SSR, AAC LC and AAC LTP”.

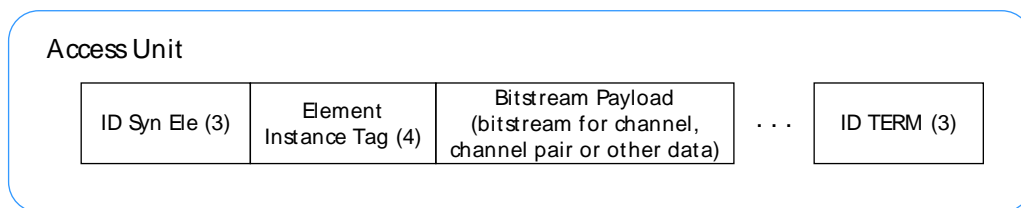


Figure 2. Access Unit Syntax

ID Syn Ele Encoding	Abbreviation	Syntactic Element
0	SCE	Single Channel Element
1	CPE	Channel Pair Element
2	CCE	Coupling Chanel Element
3	LFE	LFE Channel Element
4	DSE	Data Stream Element
5	PCE	Program Config Element
6	FIL	Fill Element
7	TERM	

Table 6. Access Unit Elements [16]

## 6.2 AAC-ELD payload

AAC-LD and AAC-ELD are using the error resilient bitstream syntax, which does not contain `id_syn_ele` elements or element instance tags but just channel elements. A detailed technical description of the ER AAC payload syntax can be found in ISO/IEC 14496-3:2009, section “4.5.2.4 Payloads for the audio object types ER AAC LC, ER AAC LTP, ER AAC LD, ER AAC ELD and ER AAC scalable”

## 7 Implicit versus Explicit Signaling

Extensions to the original AAC standard, such as HE-AAC, are designed to be transmitted in a compatible way, so that earlier decoders will ignore the extra bitstream elements. When a HE-AAC stream or file is played on a decoder that supports only an earlier version of the standard, the top octave of the signal, reproduced by the SBR technique in HE-AAC, is lost since SBR is not supported. When a HE-AAC v2 stream or file is played on an HE-AAC or AAC-LC decoder, the signal is reproduced as a monaural one since the stereo image is reproduced by the Parametric Stereo technique in HE-AAC v2. This is shown in the table below:

	AAC-LC Decoder	HE-AAC Decoder	HE-AAC v2 Decoder
<b>AAC-LC File or Stream</b>	Full Bandwidth	Full Bandwidth	Full Bandwidth
<b>HE-AAC File or Stream</b>	Reduced bandwidth	Full Bandwidth	Full Bandwidth
<b>HE-AAC v2 File or Stream</b>	Mono, reduced bandwidth [17]	Mono, Full Bandwidth	Full Bandwidth

Table 7. Compatibility of HE-AAC with Earlier Version Decoders.

There are three ways HE-AAC can be signaled in a transport or storage stream. One is to do nothing, and rely on newer decoders to detect the extra SBR information in the AAC elementary bitstream. This is termed *Implicit Signaling* and is the only method possible for transport formats such as ADTS that do not include an `AudioSpecificConfig()` data structure.

For formats that include an ASC structure, two other options are available. One is to begin the ASC with a descriptor for AAC-LC, which will be interpreted by a legacy decoder as a valid ASC and stop parsing of the ASC structure. An HE-AAC capable decoder will continue to parse the ASC and see a descriptor for the SBR data. This is termed *Explicit Backwards-Compatible Signaling* and is recommended for use with the MPEG-4 File Format. It can only be used with formats where the



size of the ASC is known, such as in the MPEG-4 File Format, thus it is not usable with LATM/LOAS.

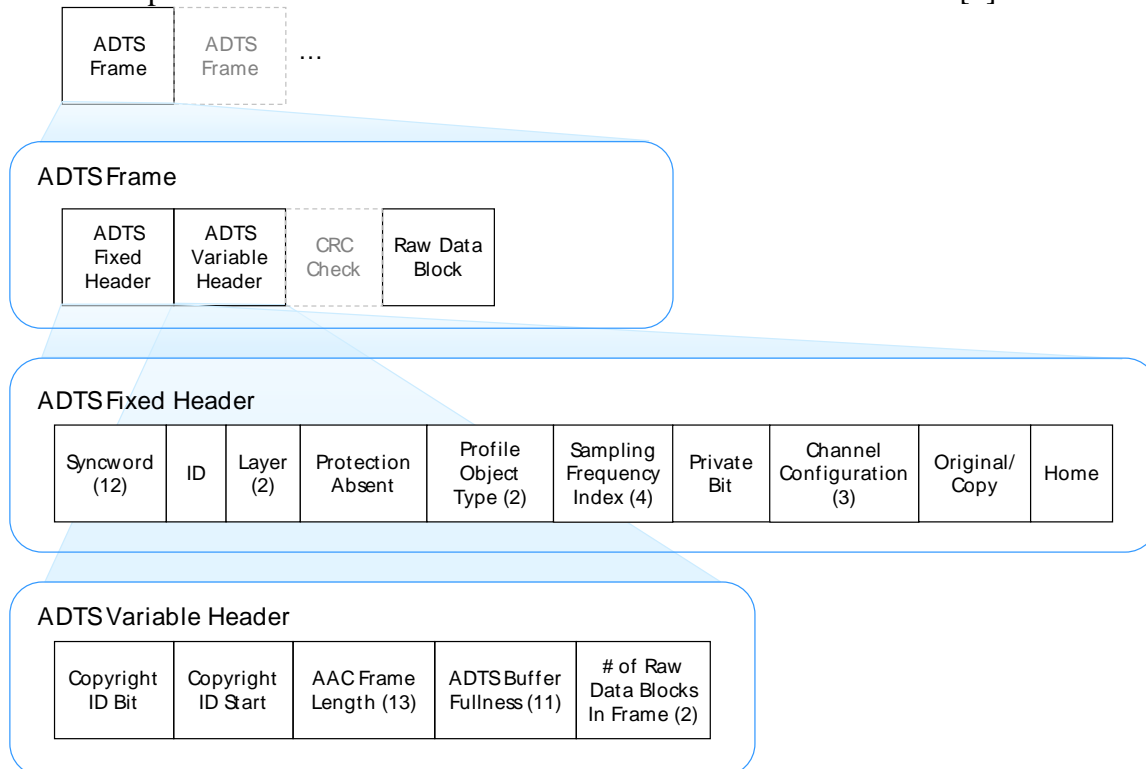
For LATM/LOAS, the third option is recommended: specifying HE-AAC as the first descriptor in the ASC, and following with a descriptor for the AAC-LC core. Since the first descriptor seen is for HE-AAC, a legacy decoder will see an unknown audio object type and not parse further or play the stream. This is termed *Hierarchical Signaling*.

Signaling Mode	Method	Recommended for	Advantages	Disadvantages
<b>Implicit</b>	No signaling is done, HE-AAC decoders look for SBR information in the bitstream	ADTS	Legacy Compatibility – legacy decoder ignores hidden SBR information	Playback may require re-initialization of the decoder once the SBR information is detected, due to the need to double the output sample rate (or change from mono to stereo if PS is detected)
<b>Explicit Backwards-Compatible</b>	AudioSpecificConfig() element signals AAC-LC, but also contains SBR information	MPEG-4 File Format	Legacy Compatibility – legacy decoder will parse AAC-LC part of ASC only.	Does not work with LATM/LOAS formats. [18] Can only be used with formats where the size of the ASC is known – such as MP4FF
<b>Hierarchical</b>	AudioSpecificConfig() element signals SBR audio object type and also indicates AAC-LC core	LATM, LOAS	Half-bandwidth or mono audio is never played by a legacy decoder	No legacy compatibility – legacy decoder will stop parsing ASC when it sees SBR audio object type.

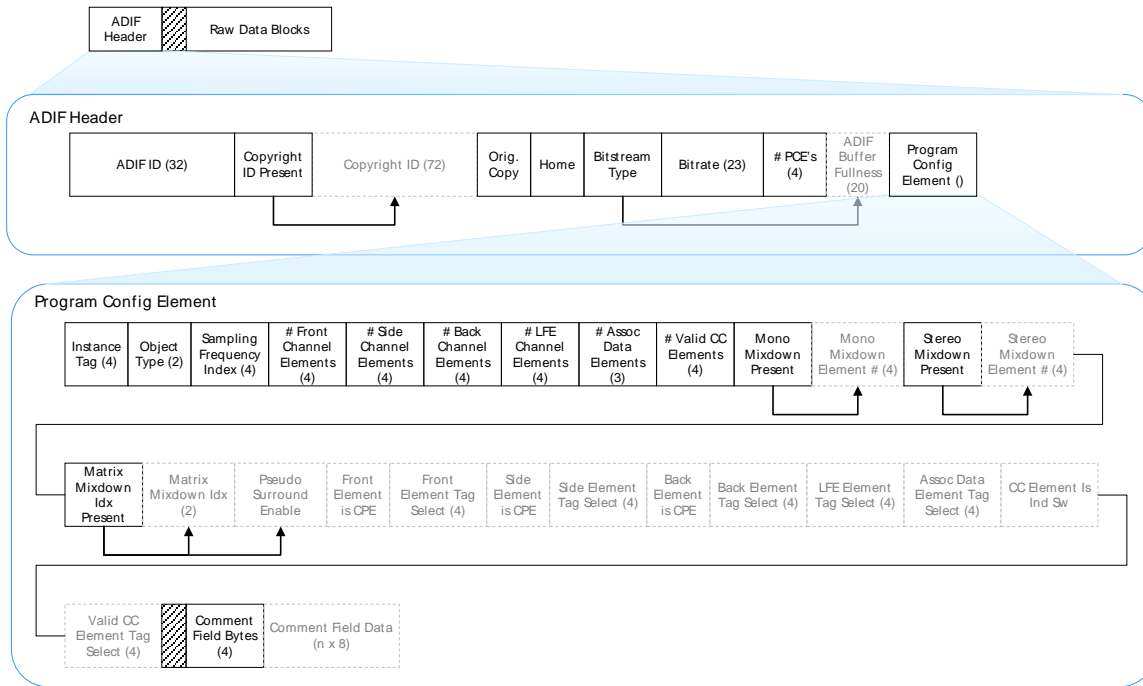
Table 8. HE-AAC Signaling Methods.

## 8 Bitstream Diagrams of Common Formats

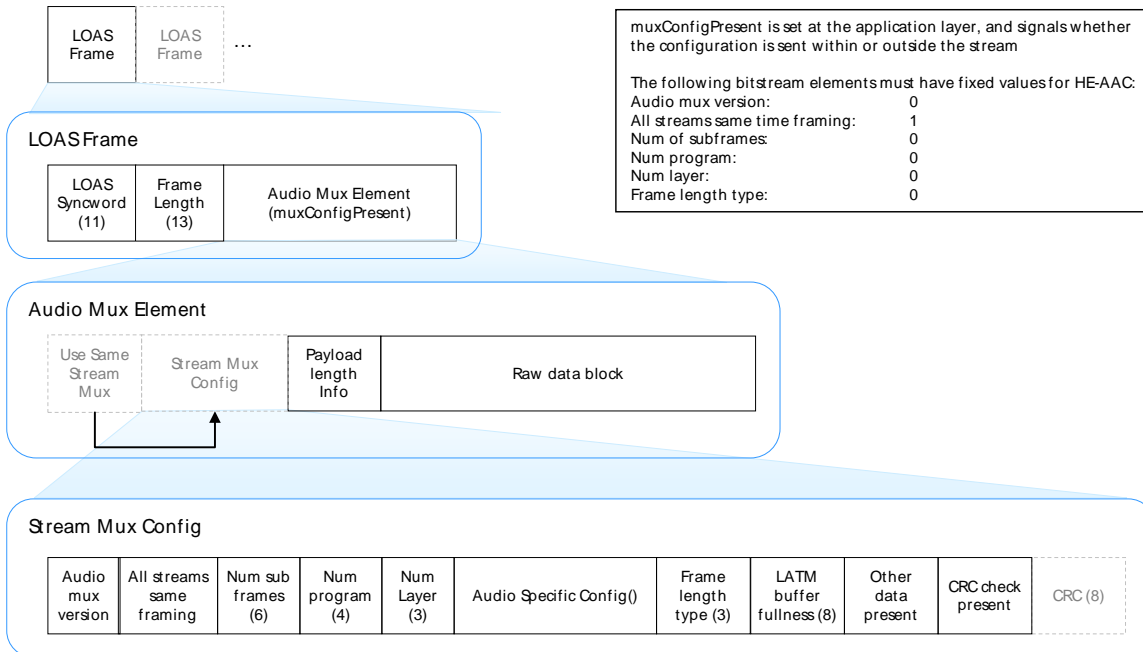
As an aid to understanding the structure of the ADTS, ADIF, and LATM/LOAS formats, diagrams showing their basic structure are presented below. This simplified presentation does not include all details or special cases of the formats as defined in the MPEG-4 standard. [2]




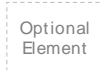
**Figure 3. ADTS Bitstream Diagram**



**Figure 4. ADIF Bitstream Diagram**



**Figure 5. LATM/LOAS Bitstream Diagram**

-  Byte Alignment
- (4) Number of bits(1 if not given)
-  Optional Element

Based on ISO/IEC 14496-3. Some syntax summarized or simplified. Refer to ISO/IEC Standards for official and complete specification

**Figure 6. Bitstream Diagram Legend**

## References

- [1] ISO/IEC 13818-7 Information technology -- Generic coding of moving pictures and associated audio information - Part 7: Advanced Audio Coding (AAC)
- [2] ISO/IEC 14496-3 Information technology - Coding of audio-visual objects – Part 3: Audio.
- [3] ETSI TS 101 154, Digital Video Broadcasting (DVB), Specification for the use of Video and Audio Coding in Broadcasting Applications based on the MPEG-2 Transport Stream, European Telecommunications Standards Institute.
- [4] ISO/IEC 14496-12 Information Technology - Coding of audio-visual objects - Part 12: ISO Base Media File Format.
- [5] ISO/IEC 14496-14 Information Technology - Coding of audio-visual objects - Part 14: MP4 File Format.
- [6] SHOUTcast 2 (Ultravox 2.1) Protocol Details, Winamp Developer Network, AOL, Inc. [http://wiki.winamp.com/wiki/SHOUTcast\\_Developer](http://wiki.winamp.com/wiki/SHOUTcast_Developer)
- [7] Adobe Flash Video File Format Specification, Version 10.1, Adobe Systems, Inc.
- [8] HTTP Live Streaming, draft-pantos-http-live-streaming-07, Internet Engineering Task Force, R. Pantos, Ed., W. May, Apple, Inc., September 30, 2011. <http://tools.ietf.org/html/draft-pantos-http-live-streaming-07>
- [9] [http://help.adobe.com/en\\_US/HTTPStreaming/1.0/Using/WS9463dbe8dbe45c4c-1ae425bf126054c4d3f-7fff.html](http://help.adobe.com/en_US/HTTPStreaming/1.0/Using/WS9463dbe8dbe45c4c-1ae425bf126054c4d3f-7fff.html)
- [10] IIS Smooth Streaming Technical Overview, Microsoft Corporation. <http://learn.iis.net/page.aspx/626/smooth-streaming-technical-overview/>
- [11] ISO/IEC 23009-1, Information technology - Dynamic adaptive streaming over HTTP (DASH) - Part 1: Media presentation description and segment formats.
- [12] The usual frame length for AAC-LC is 1024 samples, but a 960 sample version is used for radio broadcasting, and 480 or 512 sample versions are used for the low-delay codecs AAC-LD and AAC-ELD.
- [13] ISO/IEC 23003-3 Information technology — MPEG audio technologies — Part 3: Unified speech and audio coding
- [14] The Channel Configuration field provides an alternate means of specifying common channel configurations, such as mono, stereo, or 5.1 channels.
- [15] Primarily a legacy from MPEG-2 AAC, not often used in MPEG-4
- [16] Note that some elements are not valid for ER (error resilient) versions of AAC, such as AAC-LD.
- [17] The bandwidth in this condition is dependent on the SBR crossover frequency used – which may be between 4.5 and 12 KHz for 48 KHz sampling.
- [18] Explicit backwards compatible signaling is also possible with LATM/LOAS formats, if audioMuxVersion is set to 1 in the LATM multiplex so that the length field for the AudioSpecificConfig is present.